

Universidade do Minho

Escola de Engenharia

José Antônio da Cunha

**Um Sistema para Acompanhar a Aprendizagem do
Estudante em um Ambiente de ensino, Utilizando Data
Mining**

**Tese de Doutoramento
Engenharia Informática**

Trabalho efetuado sob a orientação do
Professor Doutor César Analide Rodrigues

Março de 2017

Agradecimentos

Gostaria de agradecer ao meu orientador o professor Doutor Cesar Analide Rodrigues, pela dedicação e disponibilidade a mim prestada no desenvolvimento desse trabalho.

Especialmente ao apoio, a compreensão, à confiança e a força de minha esposa “Francisca Eudocia de Medeiros Cunha”, minha companheira de muitas jornadas. E também não poderia deixar de lembrar do apoio de minha filha “Larissa de Medeiros Cunha”, que muito me incentivou na longa caminha dessa pesquisa.

E finalmente ao Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte – IFRN, pelo apoio a mim concedido a esta pesquisa.

Um Sistema para Acompanhar a Aprendizagem do Estudante em um Ambiente de ensino, Utilizando Data Mining

Resumo

O Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN), tem atuação nas áreas de ensino, de pesquisa e de extensão; contribui, de modo mais extensivo, para a formação humana e cidadã; e estimula o desenvolvimento socioeconômico, à medida que potencializa soluções científicas, técnicas e tecnológicas, com compromisso de estender benefícios à comunidade.

Essa ampla abrangência em todo território norte-rio-grandense contribui para posicionar tanto o IFRN como uma instituição de educação, ciência e tecnologia quanto os seus campis como elos de produção de conhecimento e de desenvolvimento social. Garante, assim, a manutenção da respeitabilidade junto às comunidades nas quais os campis se inserem e da credibilidade construída ao longo da história da Instituição [Projeto Político-Pedagógico do IFRN, 2012].

Nos últimos anos, têm-se observado, por parte dos educadores e gestores educacionais, um alto índice de repetência e evasão escolar, em alguns cursos e até mesmo, nota-se um certo desestímulo por parte dos alunos, em determinados cursos, ministrados nos diversos campus da instituição. É fato, por exemplo, que alguns cursos, as turmas de 3º e 4º períodos, chegam com média muita baixa de alunos, em torno de 20%, ou seja, essas turmas iniciaram com 40 alunos e, no 4º período estão com no máximo 8 a 10 alunos. Isso tem como consequência, vários professores, ministrando aulas para poucos alunos, diminuindo em muito, a relação aluno professor, um dos índices, que é utilizado pelo governo federal, para avaliar o Instituto. Diante desse quadro, surgiu a necessidade de fazer uma análise mais aprofundada de tais problemas. Ou seja, deseja-se analisar o **índice de reprovação**, o **índice de evasão**, o **índice de aprovação**, o **índice de conclusão** e também o **índice de matrículas canceladas**. Tendo como finalidade, tentar mapear quais são as causas que implicam, diretamente ou indiretamente nesses índices.

Para tanto, está se propondo nesse projeto a utilização dos recursos de aprendizagem de máquina, para se fazer uma análise na massa de dados do sistema acadêmico da instituição, a fim de tentar mapear os perfis da repetência

e da evasão escolar, duas causas que preocupam e muito todos os gestores da educação no Brasil, pois segundo dados publicados pelo **PNUD** (Programa das Nações Unidas para o Desenvolvimento), o Brasil, com uma taxa de evasão escolar de 24,3% (dados de 2013), tem a maior taxa de evasão escolar entre 100 países com maior **IDH** (Índice de Desenvolvimento Humano). Na América Latina, só Guatemala (35,2%) e Nicarágua (51,6%) têm taxa de evasão escolar superiores. Está sendo também proposto um sistema de **Data Warehouse (DW)**, onde será consolidado todos os dados das diversas fontes de dados, e tendo como finalidade maior, a geração de relatórios precisos para os gestores tomarem decisões de forma mais rápida e corretas. Além disso, está sendo proposto também um sistema de aconselhamento pedagógico, utilizando para isso, uma área da Inteligência Artificial (IA), conhecida como Sistema de Raciocínio Baseado em Casos (RBC) (em Inglês *Case-Based Reasoning – CBR*), para auxiliar a equipe pedagógica nas orientações aos alunos e seus responsáveis. E por fim, como nosso objetivo é a melhora dos índices de repetência e evasão escolar, está sendo proposto também um jogo na forma de Gamification, com o propósito de engajar cada vez mais os nossos alunos nos seus respectivos cursos e assim, diminuir esses índices.

Palavras-chave: Business Intellingence, Mineração de Dados, Descoberta de Conhecimento em Bases de Dados, Raciocínio Baseado em Casos, Aprendizagem de Máquina, Gamification.

Um Sistema para Acompanhar a Aprendizagem do Estudante em um Ambiente de ensino, Utilizando Data Mining

Abstract

Palavras-chave: Business Intellingence, Data Mining, Knowledge Discovery in Databases, Case-Based Reasonary, Machine Learning e Gamification.

SUMÁRIO

RESUMO	vi	
ABSTRACT	vii	
LISTA DE ABREVIATURAS e SIGLAS	xiv	
LISTA DE FIGURAS	xv	
LISTA DE TABELAS	xiii	
Parte I		
Capítulo 1		
1.	INTRODUÇÃO	18
1.2.	OBJETIVOS	24
1.2.1.	Objetivos Gerais	24
1.2.2.	Objetivos Específicos	
1.3.	MOTIVAÇÃO	25
1.4.	Metodologia	27
1.5.	ESTRUTURAÇÃO DA TESE	28
Capítulo 2		
2.	FUNDAMENTAÇÃO TEÓRICA	30
2.1.	BUSINESS INTELLIGENCE	30
2.1.1.	Arquitetura de Business Intelligence	30
2.2.	DATA WAREHOUSE	32
2.2.1.	Arquitetura de Data Warehouse	33
2.2.2.	Data Warehouse Empresarial	34
2.2.3.	Data Mart	34
2.2.4.	Virtual Data Warehouse	35
2.3.	PROCESSAMENTO ANÁLITICO ON-LINE	36
2.3.1.	Arquitetura OLAP	37
2.3.2.	Modelo Multidimensional	38
2.4.	Descoberta de Conhecimento em Base de dados	40
2.4.1	Caracterização do Processo de KDD	40
2.5.	Mineração de dados	44
2.5.1.	Definição	44
2.5.2.	Tarefas de Mineração de Dados	44

2.5.2.1.	Descoberta de Associação	45
2.5.2.2.	Classificação	46
2.5.2.3.	Agrupamento (Clustering)	49
2.5.3.	Métodos de Mineração de Dados	50
2.5.4.	Tecnologias de Suporte a Mineração de Dados	53
2.5.4.1.	Aprendizagem de Máquina	54
2.5.4.2.	Banco de Dados e Data Warehouse	55
2.5.4.3.	Estatística	56
2.5.4.4.	Visualização de Dados	56
2.6.	Sistema de Raciocínio Baseado em Casos	58
2.6.1.	Definição	58
2.6.2.	Ciclo do Sistema de Raciocínio Baseado em Casos	59
2.6.3.	Representação de Casos	60
2.6.4.	Indexação	61
2.6.5.	Recuperação	62
2.6.5.1.	Algoritmo da Vizinhança	62
2.6.5.2.	Algoritmo de Indução	63
2.6.6.	Adaptação	63
2.7.	Big Data	66
2.7.1.	Uso do Big Data	67
2.7.2.	Map-Reduce	68
2.7.2.1.	As Tarefas de Mapeamento	68
2.7.2.2.	Agrupamento e Agregação	69
2.7.2.3.	As Tarefas de Redução	70
2.7.2.4.	Detalhes de Execução de Map-Reduce	71
2.7.3.	Lidando com falhas nos nós	72
2.8.	Gamification	74
2.8.1.	Games	74
2.8.2.	Histórico do Gamification	75
2.8.3.	Definição	75
2.8.4.	Os Elementos do Game e do Gamification	76
2.9.	Bases Conceituais sobre a Evasão e a Retenção Escolar	81
2.9.1.	Evasão e Retenção Escolar	82

2.9.2.	Categorização das Causas da Evasão Escolar	83
2.9.3.	Indicadores de Evasão, Retenção e Conclusão	84
2.10.	Trabalhos Relacionados	86
Capítulo 3 - Desenvolvimento dos Módulos do Projeto		
3.	Business Intelligence – BI	86
3.1.	Visão do Problema	86
3.2.	Arquitetura do Business Intelligence	87
3.3.	O Processo do Business Intelligence	89
3.3.1.	Data Warehouse	90
3.3.2	Ciclo de Vida do Data Warehouse	90
3.3.2.1.	Estudo de Viabilidade	91
3.3.2.2.	Requisitos Funcionais e Não Funcionais	92
3.3.2.3.	Arquitetura de Fluxo de Dados	92
3.4.	Modelos de Dados Dimensionais (MDD)	94
3.5.	Modelos de Dados Multidimensionais (MDM)	103
3.5.1.	Projetando Agregações e Hierarquias	108
3.5.1.1.	Projetando Agregações para o Data Warehouse	109
3.5.1.2.	Projetando Hierarquias para o Data Warehouse	113
3.5.2.	Adicionando Medidas Calculadas ao Cubo	114
3.5.3.	Indicadores de Desempenho (KPI)	115
3.5.4.	Gerenciando Partições em Cubos OLAP	118
3.6.	Mineração de dados	119
3.6.1.	Pontuação de Interesse	122
3.6.2.	Entropia de Shannon	122
3.6.3.	Bayesiano com K2 a priori	123
3.6.4.	Bayesiano Dirichlet Equivalente com Uniforme a priori	123
3.6.5.	Personalizando o Algoritmo Microsoft Árvore de Decisão	124
3.6.6.	Aplicando o Algoritmo Microsoft Árvore de Decisão	126
3.6.7.	Aplicando o Algoritmo Microsoft Cluster	136
3.6.7.1.	Cluster EM	136
3.6.7.2.	Cluster K-means	137
3.6.7.3.	Personalizando o Algoritmo Microsoft Cluster	137
3.6.8.	Criando Consultas de Previsão	147

3.6.8.1.	Prevendo a Evasão e a Repetência Escolar	147
Parte II – Resultados		
Capítulo 4		
4.	Sistema de Raciocínio Baseado em Casos (RBC)	151
4.1.	Visão do Problema	152
4.2.	Metodologia	153
4.3.	Implementação do Sistema de RBC	157
4.3.1.	Casos	157
4.3.2.	Bases de Casos	160
4.3.3.	Indexação	160
4.3.4.	Recuperação	160
4.3.5.	Reutilização	161
4.4.	Desenvolvimento do Sistema RBC	161
4.4.1.	Arquitetura do Sistema RBC	164
4.4.2.	Interface com o Usuário	165
4.4.3.	Servidor de Aplicação	166
4.4.4.	Persistência de Dados	168
4.5.	Resultados	169
5.	Gamification	106
5.1.		106
5.2.		106
5.2.1.	Definição de Gestão do Conhecimento	106
5.2.2.	O que é Conhecimento	107
5.2.3.	Diferença entre Conhecimento, Informação e Dado	108
5.2.4.	Atividades da Gestão do Conhecimento	109
5.2.5.	Uma Metodologia para Gestão do Conhecimento	110
5.3.	Raciocínio Baseado em Casos	111
5.3.1.	Definição	112
5.3.2.	Representação de Casos	112
5.3.3.	Indexação	114
5.3.4.	Aquisição (Storage)	116
5.3.4.1.	Modelos de Memória Dinâmica	116

5.3.4.2.	Modelos de Categorias Exemplares	116
5.3.5.	Recuperação	117
5.3.5.1.	Algoritmo de Vizinhança	117
5.3.5.2.	Algoritmo de Indução	118
5.3.6.	Adaptação	118
5.4.	Conclusões	119
6.	REDES NEURAIS	121
6.1.	Introdução	121
6.2.	Definição	121
6.3.	O Neurônio	124
6.4.	Classificação e Propriedades	126
6.4.1.	Aprendizado RNA	127
6.4.2.	Tipos de Unidades	128
6.4.3.	Tipos de Arquiteturas de Conexões de Redes	130
6.5.	Tipos de Aplicações para Redes Neurais	132
6.6.	Vantagens das Redes Neurais	132
6.7.	Inconvenientes das Redes Neurais	133
6.8	Conclusões	134
7.	PROJETO	135
8.	DATA MINING APLICADA AO ESTUDO DE CASO	136
9.	CONCLUSÕES	137
10.	REFERENCIAL BIBLIOGRÁFICO	138

LISTA DE ABREVIATURAS

ADO	ActiveX Data Objects
ADOMD	ActiveX Data Object Multidimensional
AMO	Analysis Services Management Objects
BDE	Bayesiano Dirichlet Equivalente
BDEU	<i>Bayesiano Dirichlet Equivalente com Uniforme a priori</i>
BI	BUSINESS INTELLIGENCE
BISM	<i>Business Intelligence Semantic Model</i>
BLAP	Badges, Leaderboards, Achievements and Points
CBR	CASE-BASED REASONING
CEFETRN	Centro Federal de Educação Tecnológica do Rio Grande do Norte
CF	Constituição Federal
CODIR	Colégio de Diretrizes
CONSEPEX	Conselho de Ensino, Pesquisa e Extensão
CONSUP	Conselho Superior
CRISP-DM	<i>Cross-Industry Standard Process for Data Mining</i>
DAX	<i>Data Analysis Expressions</i>
DDS	<i>Dimensional Data Store</i>
DM	DATA MINING
DMX	Data Mining Extensions
DW	DATA WAREHOUSE
EIFRN	Escola Industrial Federal do Rio Grande do Norte
EIN	Escola Industrial de Natal
ETEP	Equipe Técnica Pedagógica
ETFRN	Escola Técnica Federal do Rio Grande do Norte
ER	ENTIDADE RELACIONAMENTO
ETL	EXTRACTION, TRANSFORMATION AND LOAD
IA	INTELIGÊNCIA ARTIFICIAL
IDH	Índice de Desenvolvimento Humano
IFRN	Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte
JSON	JavaScript Object Notation

KDD	KNOWLEDGE DISCOVERY IN DATABASES
KNN	<i>k-nearest neighbor</i>
KPI	<i>Key performance indicator</i>
LDB	Lei de Diretrizes e Base
MDX	<i>Multi-Dimensional eXpressions</i>
MEC	Ministérios de Educação e Cultura
MOLAP	MULTIDIMETIONAL OLAP
NDS	<i>Normalized Data Store</i>
ODS	<i>Operational Data Store</i>
OLAP	ONLINE ANALYTICAL PROCESSING
PNUD	Programa das Nações Unidas para o Desenvolvimento
PROEJA	Programa Nacional de Educação de Jovens e Adultos
RBC	RACIONARIO BASEADO EM CASOS
RIA	Rich Internet Application
ROLAP	RELACIONATIONAL OLAP
SDLC	System development life cycle
SGBD	SISTEMA GERENCIADOR DE BANCO DE DADOS
SGBDM	Sistemas de gerenciamento de banco de dados multidimensionais.
SISTEC	Sistema Nacional de Informações da Educação Profissional e Tecnológica
SQL	STRUCTURED QUERY LANGUAGE
SSOT	<i>Single Source of Truth</i>
WCF	Windows Communication Foundation
XMLA	XML for Analysis

LISTA DE FIGURAS

Figura 1	Uma arquitetura de data warehouse em camadas.	34
Figura 2	Mineração de Dados como um passo no processo de Descoberta de Conhecimento.	42
Figura 3	Ciclo do Raciocínio Baseado em Casos Fonte.	61
Figura 4	Esboço de Iteração de um Processo Map-Reduce.	73
Figura 5	Adaptada da formação Coursera em Gamification do professor Kevin Werbach.	77
Figura 6	As camadas conceituais do modelo BISM.	91
Figura 7	Arquitetura BISM adotada para o desenvolvimento do projeto.	92
Figura 8	O processo BI da solução.	93
Figura 9	Metodologia em Cascata.	94
Figura 10	Arquitetura Simples DDS combinado com o ETL dos dados das fontes de dados disponíveis.	96
Figura 11	Diagrama Banco de Dados do sistema acadêmico.	97
Figura 12	Data mart para as medidas de Evasão, cancelado e concluídos.	99
Figura 13	Modelo dimensional em função dos dados sociais dos alunos.	102
Figura 14	Data mart para o fato “FatoEvaEntradaSaida”.	103
Figura 15	Pacote ETL para extrair os dados das planilhas Excel para o banco de dados dimensional.	104
Figura 16	Data Mart para os dados relacionados ao boletim escolar.	104
Figura 17	Data mart para análise de repetência escolar agrupado por curso, disciplina e ano letivo.	105
Figura 18	Data Mart para dados relacionados a repetência escolar agrupados por professor.	105
Figura 19	Estrutura do Cubo para o fatoEvaCanCon.	106
Figura 20	Estrutura de cubo criada para os dados sociais dos alunos.	107
Figura 21	Estrutura de cubo criada para o fato “FatoEvaEntradaSaida”.	108

Figura 22	Estrutura de cubo para o fato “FatoAprovadosReprovadoProfAno”.	109
Figura 23	Estrutura de cubo para as medidas aprovados e reprovados por disciplina e curso.	109
Figura 24	Partição padrão do cubo Fato Evasão Entrada Saída.	112
Figura 25	Calculando agregações para o cubo Fato Evasão Entrada Saída.	113
Figura 26	Partição padrão com a agregação calculada pelo assistente.	113
Figura 27	O atributo “Situacao Matricula” com o valor 4 estimado para a contagem de linhas.	114
Figura 28	Agregações criadas para cada tabela de fato do cubo dados sociais.	115
Figura 29	Árvore de Decisão para o atributo previsível “Situacao=Reprovado”.	131
Figura 30	Rede de dependência dos atributos em relação ao atributo situação (Reprovado).	133
Figura 31	Rede de dependência dos atributos em relação ao atributo situação (Reprovado).	133
Figura 32	Gráfico de acurácia entre os métodos de divisão dos nós da árvore de decisão.	134
Figura 33	Árvore de Decisão usando o método de pontuação a Entropia.	135
Figura 34	Gráfico de correlação dos atributos de entrada com o atributo previsível “Situacao Ing” destacado na cor verde.	137
Figura 35	Gráfico dos cluster gerados para a situação “Reprovado” usando o método de clusterização o EM Evolutivo.	143
Figura 36	Cluster gerado com o método EM não Evolutivo.	143
Figura 37	Gráfico de clusters gerado pelo método K-means Evolutivo.	144
Figura 38	Gráfico de clusters gerado pelo método K-means Não Evolutivo.	145
Figura 39	Gráfico comparativo entre os métodos de cluster.	146
Figura 40	Perfis de Cluster algoritmo EM evolutivo.	147

Figura 41	Cluster EM Evolutivo para o atributo previsível “Situacao Ing=Evasão”.	148
Figura 42	Gráfico tipo histograma mostrando o quanto cada atributo influencia percentualmente na formação de cada cluster.	148
Figura 43	Arquitetura ADO.NET .	155
Figura 44	Arquitetura de Componentes do Analysis Services.	156
Figura 45	OLE DB for OLAP 9.0 Provider (MSOLAP.3).	156
Figura 46	Exemplo ADOMD.NET para se conectar ao Servidor do Analysis Services e executar uma consulta.	157
Figura 47	Método ADOMD.NET para processar um cubo no Servidor do Analysis Services.	158
Figura 48	Arquitetura usada para implementação do portal.	158
Figura 49	Tela principal do Portal de Análise de dados.	160
Figura 50	Dashboard Situação Escolar por Campus.	161
Figura 51	Dashboard Repetência por disciplina.	162
Figura 52	Dashboard Evasão por campus e ano.	163
Figura 53	Dashboard Evasão por curso e ano.	164
Figura 54	KPI Desempenho do Aluno.	165
Figura 55	KPI Desempenho do aluno em cada disciplina.	166
Figura 56	Janela Consultas sobre evasão escolar por campus.	167
Figura 57	Resulta após o utilizador clicar no quadradinho azul.	168
Figura 58	Reprovados por Curso e Disciplina/Ano.	169
Figura 59	Constas MDX Ad Hoc.	170
Figura 60	Consulta Ad Hoc com tabela dinâmica do Excel.	171
Figura 61	Um Dashboard exibindo a Evasão Escolar por campus do Power BI.	172
Figura 62	Dashboard de Índices Gerais de todos os campi do IFRN.	173
Figura 63	Dashboard Indicar de Ensino dos Curso no campus Natal-Central.	174
Figura 64	Indicador de Ensino do Curso 01434 do campus Natal Central.	175
Figura 65	Indicadores de Evasão das Licenciaturas do IFRN.	176

Figura 66	Probabilidade de Evasão em função do número de reprovações.	177
Figura 67	Probabilidade de Evasão em função do número de reprovações classificado por etnia.	178
Figura 68	Probabilidade de Evasão em função do número de reprovações classificado por tipo de escola de origem.	178
Figura 69	Probabilidade de Evasão em função do número de reprovações classificado por renda familiar.	179
Figura 70	Descrição de um caso ideal utilizando a notação JSON .	186
Figura 71	Fluxograma simplificado do algoritmo kNN.	190
Figura 72	Implementação do Cálculo de distância dos Casos na linguagem Javascript.	191
Figura 73	Modelo arquitetural de aplicação cliente-servidor.	193
Figura 74	Esquema de <i>Two-way data binding</i> na biblioteca AngularJS.	194
Figura 75	Mecanismo Event-Loop do NodeJS.	196
Figura 76	Comparação da representação de um mesmo documento usando JSON em texto pleno e na sua representação em binário, BSON .	197
Figura 77	Interface com o usuário do sistema. Passo 1 – Informações pessoais do aluno.	198
Figura 78	Passo 2 – Seleção de demandas do novo caso.	199
Figura 79	Relatório de Demandas mais atendidas.	201
Figura 80	Tela Principal do Jogo. Fonte: Autor (Visual Studio 2015).	211
Figura 81	Mapa de Jogos.	212
Figura 82	Tela do Jogo.	213
Figura 83	Gráfico mostrando a Evasão escolar no Campus Natal-central de 2000 a 2013.	217
Figura 84	Cancelamento de matrículas no Campus Natal-Central entre 2000 e 2013.	218
Figura 85	Árvore de Decisão mostrando a relação entre os atributos da base de dados Acadêmica.	220
Figura 86	Gráfico de cluster gerado pelo Analysis Service.	221
Figura 87	Características do cluster 1	222

Figura 88	Características do cluster 5	222
Figura 89	Representação dos clusters em forma de histograma.	223
Figura 90	Perfil do aluno com maior probabilidade de evasão escolar.	224
Figura 91	Composição de um Caso do Sistema de Aconselhamentos.	231
Figura 92	Relação dos atributos do perfil e variáveis da demanda.	232

LISTA DE TABELAS

Tabela 1	Lista de ofertas e indicadores de potencialidades ou fragilidades.	25
Tabela 2	Bases de Dados do Instituto Natal Central - RN	94
Tabela 3	Dimensões e tabela de fatos.	98
Tabela 4	Lista de dimensões e tabelas de fatos.	101
Tabela 5	Dimensões e fato para o “FatoEvaEntradaSaida”	102
Tabela 6	Hierarquias criadas para o cubo dados sociais.	116
Tabela 7	Formulário que documenta a execução do processo de KDD, no modelo CRISP.	121
Tabela 8	Parâmetros do algoritmo Microsoft Árvore de Decisão.	126
Tabela 9	Relação de atributos da tabela onde será aplicado o algoritmo Árvore de Decisão para análise.	128
Tabela 10	Especifica o método de cluster para o algoritmo a ser usado.	139
Tabela 11	Formulário que documenta a execução do processo de KDD.	141
Tabela 12	Tabela comparativa entre os clusters 6 e 9 da Figura 38.	146
Tabela 13	Valores dos atributos.	150
Tabela 14	Previsão de evasão de acordo com o perfil especificado na tabela.	151
Tabela 15	Previsão de evasão de acordo com o perfil especificado na tabela.	152
Tabela 16	Descrição de classes e seus respectivos pesos, em ordem de relevância.	182
Tabela 17	Distribuição de demandas nas respectivas classes e pesos, conforme análise da especialista psicopedagógico por ordem alfabética.	182
Tabela 18	Tempo para a ser disponibilizado para a solução de cada questão.	209
Tabela 19	Pontuação de jogo por nível.	209
Tabela 20	Quadro de medalhas.	209

Tabela 21	Índice de reprovação em Disciplinas no IFRN para o ano de 2010.	219
Tabela 22	Dados relativos à Educação no Relatório do PNUD.	225
Tabela 23	Taxa de Abandono Escolar Precoce na Europa.	226
Tabela 24	Alunos evadidos, por tipos de cursos, de ciclos de matrícula iniciados a partir de 2004 e encerrados até dezembro de 2011.	228

Capítulo 1

1. INTRODUÇÃO

Para uma melhor compreensão do perfil institucional, na perspectiva de consolidar a função social, os objetivos e os princípios orientadores do instituto, se faz necessário uma análise histórica, políticas e sociais do IFRN.

Criada pelo Decreto 7.566, de 23 de setembro de 1909, como Escola de Aprendizes Artífices, a instituição passou por diversas mudanças e recebeu várias denominações ao longo do tempo. Instalada, inicialmente, em janeiro de 1910, no prédio do antigo Hospital da Caridade, na Praça Coronel Lins Caldas, nº 678, Cidade Alta, onde hoje funciona a casa de estudantes de Natal, a Escola de Aprendizes Artífices oferecia cursos primários de desenho e oficinas de trabalhos manuais. Em 1914, o estabelecimento foi transferido para a Avenida Rio Branco, nº 743, ocupando, durante cinquenta e três anos, um edifício construído no início do século XX.

Mais tarde, ocorreu a mudança de denominação para Liceu Industrial de Natal, orientada pela reforma instituída pela Lei 378, de 13 de janeiro de 1937, do Ministério da Educação e Saúde, órgão a que a Instituição estava subordinada desde 1930. Na época, eram oferecidas oficinas de desenho, de sapataria, de funilaria, de marcenaria e de alfaiataria, inspiradas, segundo Meireles (2006, p.55), em “modelos exteriores ao Brasil, o que evidencia a influência de outros formatos culturais, educacionais, tecnológicos e produtivos na realidade brasileira do século XX”.

Designada como Escola Industrial de Natal (**EIN**), no ano de 1942, após a promulgação da Lei Orgânica do Ensino Industrial, a Instituição transformou as oficinas em cursos básicos de primeiro ciclo, organizados em quatro seções: Trabalhos de Metal, Indústria Mecânica, Eletrotécnica e Artes Industriais. Ademais, a Escola também estava autorizada a oferecer cursos de mestria para os professores atuantes nessas áreas.

Transformada em autarquia pela Lei Federal 3.552, de 16 de fevereiro de 1959, todas as Escolas Industriais do Brasil conseguiram autonomia

administrativa, didática e financeira, transformando-se em instituições federais destinadas a ministrar cursos técnicos de nível médio. Porém, somente em 1963, EIN implantou seus primeiros cursos técnicos de nível médio, com as ofertas de Mineração e de Estradas. O novo modelo, tinha equivalência ao ensino de 2º grau, o que permitia a continuidade de estudos no ensino superior para os egressos que assim o desejasse.

Em 1965, o Estabelecimento passou a nomear-se Escola Industrial Federal do Rio Grande do Norte (**EIFRN**). Nessa década, no dia 11 de março de 1967, ocorreu a inauguração da “nova” Escola Industrial nas recém-construídas instalações do prédio situado na Avenida Salgado Filho nº 1559, no bairro Tirol, atendendo a uma comunidade escolar de 233 servidores e cerca de 1.100 estudantes.

Na condição de Escola Técnica Federal do Rio Grande do Norte (**ETFRN**), mudança impetrada pela Portaria Ministerial 331, de 16 de junho de 1968, o Conselho de Representantes deliberou a extinção gradativa dos cursos industriais básicos, passando-se a ministrar somente o ensino profissional de nível técnico. Em consequência, foram criados, em 1969 e 1973, os cursos técnicos de nível médio em Eletrotécnica, em Mecânica, em Edificações, em Saneamento e em Geologia, sob a orientação da Lei 5.692/71, a qual definia a estrutura do ensino de 2º grau como ensino profissionalizante obrigatório. A partir de então, a ETFRN passou a dedicar-se, exclusivamente, ao ensino técnico profissionalizante de 2º grau.

No ano de 1994, iniciou-se outro processo de transição das escolas técnicas federais. No entanto, somente no dia 18 de janeiro de 1999, efetivou-se a mudança de ETFRN para Centro Federal de Educação Tecnológica do Rio Grande do Norte (**CEFETRN**). De acordo com os documentos oficiais, os centros de educação tecnológica foram implantados com a finalidade de formar e qualificar profissionais no âmbito da educação tecnológica, em diferentes níveis e modalidades de ensino, para os diversos setores da economia. Também objetivavam realizar pesquisas aplicadas e promover o desenvolvimento tecnológico de novos processos, produtos e serviços, em estreita articulação com os setores produtivos e com a sociedade.

Em 2006, o Governo Federal lançou um arrojado plano de expansão da rede federal. Com características de interiorização de educação profissional e

tecnológica para todo o País, e foram implantadas mais três unidades de ensino vinculadas ao CEFET-RN: na Zona Norte da cidade de Natal, na cidade de Ipanguaçu e Currais Novos. Ainda nesse mesmo ano, devido ao lançamento do Programa Nacional de Educação de Jovens e Adultos (**PROEJA**), o CEFET-RN começou a atuar na educação profissional técnica de nível médio, na modalidade de educação de jovens e adultos, oferecendo também, para educadores que atuam nessa modalidade, cursos de formação em nível de pós-graduação lato sensu.

Ao limiar de um século de existência, a Instituição adquiriu nova configuração, transformando-se em Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (**IFRN**), nos termos da Lei 11.892, de 29 de dezembro de 2008. De acordo com o estabelecido oficialmente, os institutos federais são instituições, pluricurriculares e multicampis, de educação superior, básica e profissional. São especializadas, na oferta de educação profissional e tecnológica nas diferentes modalidades de ensino, ancorando-se na conjuntura de conhecimentos técnicos e tecnológicos com as práticas pedagógicas. Esse processo, uma vez criada a rede federal de educação profissional, científica e tecnológica, constituiu-se em elemento de redefinição do sistema de ensino brasileiro. Reforçou, por outro lado, a autonomia administrativa, patrimonial, financeira, didático-pedagógica e disciplinar dessas instituições educativas, além de imprimir a equivalência às universidades federais, no que refere às disposições que regem a regulação, a avaliação e a supervisão das instituições e dos cursos de educação superior.

A expansão da rede federal de educação profissional e tecnológica está pautada na interiorização da educação profissional, com o compromisso de contribuir, significativamente, para o desenvolvimento socioeconômico do País. Nessa perspectiva, a criação dos institutos federais responde à necessidade da institucionalização definitiva da educação profissional e tecnológica como política pública permanente de Estado.

Esse processo de interiorização da educação profissional e tecnológica contribui para o combate às desigualdades estruturais de diversas ordens, proporcionando o desenvolvimento social por meio da formação humana integral dos sujeitos atendidos. Propicia, ainda, o desenvolvimento econômico, a partir da articulação das ofertas educacionais e das ações de pesquisas e de extensão.

Tal articulação vincula-se aos arranjos produtivos sociais e culturais, com possibilidades de permanência e de emancipação dos cidadãos assim como de desenvolvimento das diversas regiões do Rio Grande do Norte.

De organização pluricurricular, o IFRN oferece um ensino público laico, gratuito e de qualidade. Oferta, nesse sentido, cursos em sintonia com a função social que desempenha, visando a consolidação e o fortalecimento, dos arranjos produtivos, culturais e sociais locais. Apresenta, para tanto, um currículo organizado a partir de quatro eixos – ciência, trabalho, cultura e tecnologia – que atuam, de modo entrelaçado e inter complementar, como princípios norteadores da prática educativa. O Instituto desenvolve à pesquisa e a extensão, na perspectiva da produção, socialização e difusão de conhecimentos. Estimula a produção cultural e realiza processos pedagógicos que levem à geração de trabalho e renda. Em um contexto mais amplo, a Instituição visa contribuir para as transformações da sociedade, visto que esses processos educacionais são construídos nas relações sociais.

No que concerne à comunidade acadêmica, há os sujeitos sociais diretamente envolvidos com os processos pedagógicos e administrativos do IFRN. Essa comunidade é constituída por três seguimentos: estudantes, professores e técnicos-administrativos. Numa perspectiva mais abrangente, acrescenta-se, a esse coletivo, a comunidade local, composta tanto por pais dos estudantes e/ou responsáveis pelos estudantes, quanto representantes da sociedade civil.

Para efeito de regulação, avaliação e supervisão da Instituição e dos cursos de educação superior, equipara-se às universidades federais. Além de se submeter à legislação federal específica, rege-se pelos seguintes instrumentos normativos: estatuto; regimento geral; regimento interno dos campis e dos demais órgãos componentes da estrutura organizacional dos institutos federais; resolução do Conselho Superior (**CONSUP**); deliberações do Colégio de Diretrizes (**CODIR**) e do Conselho de Ensino, Pesquisa e Extensão (**CONSEPEX**); e atos da Reitoria.

A expansão do IFRN amplia, significativamente, a atuação nas áreas de ensino, de pesquisa e de extensão; contribui, de modo mais extensivo, para a formação humana e cidadã; e estimula o desenvolvimento socioeconômico, à

medida que potencializa soluções científicas, técnicas e tecnológicas, com compromisso de estender benefícios à comunidade.

Essa ampla abrangência em todo território norte-rio-grandense contribui para posicionar tanto o IFRN como uma instituição de educação, ciência e tecnologia quanto os seus campis como elos de produção de conhecimento e de desenvolvimento social. Garante, assim, a manutenção da respeitabilidade junto às comunidades nas quais os campis se inserem e da credibilidade construída ao longo da história da Instituição [Projeto Político-Pedagógico do IFRN, 2012].

Talvez em função da expansão, pois houve um aumento significativo da nossa massa de alunos, nos últimos anos, têm-se observado, por parte dos educadores e gestores educacionais, um alto índice de repetência e evasão escolar, em alguns cursos e até mesmo, nota-se um certo desestímulo por parte dos alunos, em determinados cursos, ministrados nos diversos campus da instituição. É fato, por exemplo, que alguns cursos, as turmas de 3º e 4º períodos, chegam com média muita baixa de alunos, em torno de 20%, ou seja, essas turmas iniciaram com 40 alunos e, no 4º período estão com no máximo 8 a 10 alunos. Isso tem como consequência, vários professores, ministrando aulas para poucos alunos, diminuindo em muito, a relação aluno professor, um dos índices, que é utilizado pelo governo federal, para avaliar o Instituto. Diante desse quadro, surgiu a necessidade de fazer uma análise mais aprofundada de tais problemas. Ou seja, deseja-se analisar o **índice de reprovação**, o **índice de evasão**, o **índice de aprovação**, o **índice de conclusão** e também o **índice de matrículas canceladas**. Tendo como finalidade, tentar mapear quais são as causas que implicam, diretamente ou indiretamente nesses índices.

Para tanto, está se propondo nesse projeto a utilização dos recursos de aprendizagem de máquina, para se fazer uma análise na massa de dados do sistema acadêmico da instituição, a fim de tentar mapear os perfis da repetência e da evasão escolar, duas causas que preocupam e muito todos os gestores da educação no Brasil, pois segundo dados publicados pelo **PNUD** (Programa das Nações Unidas para o Desenvolvimento), o Brasil, com uma taxa de evasão escolar de 24,3% (dados de 2013), tem a maior taxa de evasão escolar entre 100 países com maior **IDH** (Índice de Desenvolvimento Humano). Na América Latina, só Guatemala (35,2%) e Nicarágua (51,6%) têm taxa de evasão escolar superiores. Está sendo também proposto um sistema de **Data Warehouse (DW)**,

onde será consolidado todos os dados das diversas fontes de dados, e tendo como finalidade maior, a geração de relatórios precisos para os gestores tomarem decisões de forma mais rápida e corretas. Além disso, está sendo proposto também um sistema de aconselhamento pedagógico, utilizando para isso, uma área da Inteligência Artificial (IA), conhecida como Sistema de Raciocínio Baseado em Casos (RBC) (em Inglês *Case-Based Reasoning* – **CBR**), para auxiliar a equipe pedagógica nas orientações aos alunos e seus responsáveis. E por fim, como nosso objetivo é a melhora dos índices de repetência e evasão escolar, está sendo proposto também um jogo na forma de Gamification, com o propósito de engajar cada vez mais os nossos alunos nos seus respectivos cursos e assim, diminuir esses índices.

1.2. OBJETIVOS

1.2.1. Objetivos Gerais

Nesta pesquisa pretende-se desenvolver uma ferramenta para auxiliar o professor ou os gestores a diagnosticar problemas relacionados ao ensino aprendizagem, tais como: alto índices de evasão e repetências escolar e o mau desempenho dos alunos nas respectivas disciplinas e sugerir algumas ações, no intuito de diminuir os índices de repetência e evasão escolar no instituto federal do Rio Grande do Norte. Para isto, será desenvolvida uma solução tecnológica que permita aos gestores da educação no Instituto Federal de Educação, Ciências e Tecnologia do Rio Grande do Norte – **IFRN**, detectar através de relatórios e gráficos, a real situação dos alunos, em relação a evasão e a repetência escolar. Além disso, o professor poderá através desta ferramenta, acompanhar o desempenho dos alunos, diagnosticando aqueles alunos que estão em situação de risco e, dessa forma, orientar esses alunos em tempo hábil, na tentativa de recuperar o desempenho desses alunos, sugerindo aos mesmos, atividades engajadoras (baseadas em **Gamification**), tais como jogos, e também um sistema de aconselhamento ao aluno. Esta pesquisa tem também como foco, o uso de técnicas de Mineração de Dados, para descobrir relações entre os atributos da base de dados, que indique o perfil dos alunos evadidos ou retidos nos diversos cursos ministrados pelo instituto e também prever futuros índices de evasão escolar.

O principal objetivo é o “**Desenvolvimento de um Sistema de Auxílio à Gestão Acadêmica, na prevenção da repetência e da evasão escolar no IFRN**”.

1.2.2. Objetivos Específicos

- Criar um armazém de dados (Em inglês *Data Warehouse*), coletar, extrair, transformar e carregar os dados do sistema acadêmico para o armazém de dados;
- Aplicar as técnicas Mineração de Dados e/ou Aprendizagem de Máquina, para traçar o perfil dos alunos propícios a repetência e a evasão escolar no IFRN;
- Aplicar as técnicas de Mineração de Dados para fazer previsões futuras sobre a repetência e a evasão escolar no IFRN;
- Desenvolver o Sistema de Aconselhamento Pedagógico ao aluno (Raciocínio Baseado em Casos – em Inglês *Case-Based Racionary – CBR*);
- Desenvolver um Portal Acadêmico que será utilizado pelo usuário final.

1.3. MOTIVAÇÃO

O Instituto Federal de Educação, Ciências e Tecnologia do Rio Grande do Norte tem diferentes ofertas educacionais, com os seus múltiplos perfis de conhecimento, e as peculiaridades regionais requerem pensar a organização e o desenvolvimento de todas as suas ações educativas de modo globalizado, mantendo indicadores de qualidade social.

Lidar com essa pluralidade curricular implica considerar os desafios que lhe são inerentes. Um deles a se destacar, consiste em manter a qualidade do ensino coerente com as demandas sócio educacionais e as exigências legais.

Com o propósito de se ter uma visão geral do ensino no IFRN, avaliar e, posteriormente, construir indicadores para revisão e reestruturação dos cursos ofertados, foram aplicados instrumentos de escuta à comunidade acadêmica e realizados fóruns de debates e análise da situação atual.

A pesquisa de caráter diagnóstico, teve a participação de professores, membros da equipe técnico-pedagógica, estudantes e gestores. Como resultado, a pesquisa apontou alguns indicadores de potencialidades e de fragilidades nas ações didático-pedagógicas relativas as ofertas dos cursos. Foram muitos os indicadores levantados na pesquisa. No entanto, serão mostrados na Tabela 1.1 apenas os indicadores mais relevantes para esta pesquisa.

Tabela 1.1 Lista de ofertas e indicadores de potencialidades ou fragilidades.

Fonte: Equipe de sistematização do PPP (2011).

OFERTA	INDICADOR
EDUCACIONAL	
CURSOS TÉCNICOS INTEGRADOS REGULARES	<ul style="list-style-type: none">• Muita procura pelos cursos no processo seletivo.• Índice de reprovação em torno de 15%;• Necessidade de avaliação periódica dos cursos;
CURSOS TÉCNICOS INTEGRADOS EJA	<ul style="list-style-type: none">• Procura insuficiente pelas vagas nos processos seletivos;• Índice de reprovação aproximando de 20%;• Índice médio de conclusão dos cursos de 40%;• Alto índice de matrículas perdidas, cerca de 32%
CURSOS TÉCNICOS SUBSEQUENTES	<ul style="list-style-type: none">• Procura considerável pelos cursos no processo seletivo;• Índice de reprovação aproximado de 18%;• Índice médio de conclusão dos cursos de 45%;• Alto índice de matrículas perdidas, cerca de 34%.
CURSOS SUPERIORES DE TECNOLOGIA	<ul style="list-style-type: none">• Procura considerável pelos cursos no processo seletivo;• Índice de reprovação aproximado de 23%;• Índice médio de conclusão dos cursos em torno de 50%;• Alto índice de matrículas perdidas, cerca de 50%.
CURSOS DE LICENCIATURA	<ul style="list-style-type: none">• Procura considerável pelos cursos no processo seletivo;• Índice de reprovação entre 17% e 20%;

- Índice médio de conclusão dos cursos com possibilidade de até 42%;
- Alto índice de matrículas perdidas, cerca de 58%.

Os dados apresentados na Tabela 1.1 balizaram a motivação dessa pesquisa, provocando uma reflexão sobre os altos índices de reprovação e matrículas perdidas (ou alunos evadidos) nos cursos ofertados pelo IFRN, levando-se em conta, os compromissos assumidos na função social e da compreensão de educação como um direito universal, faz-se necessário encontrar caminhos para superar o quadro de reprovação, de repetência e de evasão escolar nos diferentes níveis e nas várias modalidades ofertadas.

A motivação deste trabalho é desenvolver uma ferramenta que permita que a comunidade acadêmica (professores e gestores e a equipe didático-pedagógica) do IFRN, possam acompanhar o processo de ensino aprendizagem e, dessa forma, desencadear ações que venham motivar e engajar os alunos a permanecerem nos seus referidos cursos. E assim, tentar melhorar os índices de repetência e evasão de nossos alunos.

1.4. METODOLOGIA

Esta sessão tem como finalidade descrever o modo como será desenvolvida esta pesquisa, em relação aos objetivos definidos. Primeiramente será feita uma pesquisa do referencial bibliográfico, que dará um embasamento teórico para o desenvolvimento do projeto. Para tanto, serão discutidos os seguintes conceitos Banco de Dados, Modelagem Multidimensional, **OLAP**, *Data Warehouse* e Descoberta de Conhecimento em Bases de Dados (Em inglês *Knowledge Discovery Data - KDD*), Aprendizagem de Máquinas, Raciocínio Baseado em Casos – em Inglês *Case-Based Racionary – CBR* e Gamificação.

Concluído o referencial teórico, na etapa seguinte, será feita uma análise na Base de Dados dos Sistema Acadêmico (modelo relacional normalizado), para que se possa fazer a seleção dos dados a serem manipulados pelos sistemas **OLAP** e de Mineração de dados. Os dados extraídos do sistema acadêmico, serão consolidados em uma nova base de dados esses dados, em

um modelo dimensional, com tabelas dimensões e tabelas fatos. Esta nova base de dados será justamente os **Data Marts (DM)** do sistema.

Posteriormente será criado o modelo multidimensionais do sistema, onde serão utilizados os recursos **OLAP** para a criação dos cubos que serão utilizados, tanto para a geração de relatórios administrativos, quanto para serem utilizados, pelas ferramentas de mineração de dados, no processo de **KDD**.

O passo seguinte é aplicar sobre as bases de dados disponíveis (o armazém de dados e a base dimensional), as técnicas de Aprendizagem de Máquina, dentro de um processo maior que é o **KDD**, para descoberta de novos padrões. Analisar esses novos padrões descobertos e decidir se os mesmos são úteis aos objetivos desejados pelos usuários. O principal objetivo nesse processo de **KDD** é traçar o perfil dos alunos em relação a repetência e a evasão escolar. Além disso, é nessa etapa do projeto que serão feitas as previsões futuras sobre a repetência e a evasão escolar no **IFRN**.

Na ordem do desenvolvimento será construído um sistema de aconselhamento pedagógico, usando a técnica “Raciocínio Baseado em Casos - **RBC**” (Em inglês *Case-Based Racionay – CBR*), para ser utilizado pela equipe pedagógica do **IFRN**, na orientação dos alunos e seus responsáveis. O objetivo desse módulo é facilitar o atendimento por parte da equipe pedagógica aos alunos, auxiliando a equipe pedagógica na identificação dos problemas apresentados pelos alunos e na solução dos mesmos. Quanto mais rápido, são solucionados os problemas dos alunos, maiores são as chances desses alunos se recuperarem nas disciplinas e, assim aumenta-se as possibilidades diminuir os índices de repetências desses alunos e, consequentemente, diminuir também o índice de evasão escolar.

Como meta engajadora do aluno em seus respectivos cursos, está sendo proposto nesse trabalho um sistema de gamificação. Este aplicativo em forma de jogo, implementará alguns assuntos ou disciplinas, que tenham sido identificados no processo de KDD, como aquele assunto ou aquela disciplina onde apresentam os maiores índices de repetência escolar.

1.4.1. Questão Ética e Gestão dos Dados

Os dados utilizados para teste e validação das estruturas de Mineração de Dados, bem como os dados utilizados para apresentação dos resultados, são dados reais extraídos do sistema acadêmico, ora em utilização pelo IFRN. Por essa razão os dados dos alunos não podem ser acessados por quaisquer outras pessoas que, não estejam envolvidas no projeto.

Apesar dos dados serem sigilosos, foi preferido utilizar os dados reais, para termos a garantia de que, os resultados obtidos representam as regras de negócio solicitadas pelos objetivos. Além do fato de ser implícito gerar tantos dados fictícios como aqueles encontrados no sistema acadêmico do IFRN.

1.5. ESTRUTURAÇÃO DA TESE

Esta pesquisa está estruturada em cinco capítulos. O primeiro capítulo trata de uma introdução, onde é dada uma visão ampla da pesquisa, os objetivos que se almeja alcançar, a motivação que deu incentivo ao desenvolvimento dessa pesquisa e a metodologia a ser aplicada no desenvolvimento da pesquisa.

No segundo capítulo é feito um estudo do estado da arte das tecnologias que serão utilizadas no desenvolvimento dessa investigação. Este capítulo é essencial, pois a pesquisa feita aqui, dará o embasamento teórico para desenvolvimento de todo projeto.

No terceiro capítulo corresponde ao desenvolvimento do portal, do sistema de aconselhamento pedagógico e do jogo.

No capítulo quarto serão apresentados os resultados dos módulos desenvolvidos no capítulo 3.

No capítulo cinco, serão apresentadas as conclusões e trabalhos futuros.

Capítulo 2

2. Fundamentação Teórica

2.1. Business Intelligence

Segundo Rob (2011), o termo *Business intelligence (BI)* é utilizado para descrever um conjunto amplo, coeso e integrado de ferramentas e processos utilizados para captar, coletar, integrar, armazenar e analisar dados para a geração e a apresentação de informações que deem suporte à tomada de decisões de negócio. Como o próprio nome diz, **BI** trata da criação de inteligência sobre o negócio. Portanto, o **BI**, é um modelo que permite à empresa transformar dado em informação, informação em conhecimento e conhecimento em sabedoria.

O **BI** não é, por si só, um produto, mas um modelo de conceitos, práticas, ferramentas e tecnologias (*data warehouse*, *data mart*, **OLAP** e/ou ferramentas de mineração de dados) que auxiliam uma empresa a compreender melhor seus recursos centrais, identificando oportunidades fundamentais para criar competitividade (Rob, 2011). Em geral, o **BI** envolve as seguintes etapas:

- Coleta e armazenamento de dados operacionais.
- Agregação de dados operacionais em dados de suporte a decisões.
- Análise de dados de suporte a decisões para gerar informações.
- Apresentação dessas informações ao usuário final para dar suporte a decisões de negócios.
- Tomada de decisões de negócio, o que, por sua vez, gera mais dados que são coletados, armazenados etc. (reiniciando o processo).
- Monitoramento para avaliar os resultados das decisões de negócio (Rob, 2011).

2.1.1 Arquitetura de Business Intelligence

Segundo Rob (2011), o **BI** utiliza-se de tecnologias e aplicações para o gerenciamento de todo o ciclo de vida dos dados, da aquisição ao armazenamento, transformação, integração, análise, monitoramento e apresentação. Não existe uma arquitetura única de **BI**, no entanto, há alguns tipos gerais de recursos, que são compartilhados por todas as implementações de **BI**.

Uma arquitetura de BI deve ser composta de dados, pessoas, processos, tecnologias e gerenciamento desses componentes.

Uma arquitetura de **BI**, é composta de componentes básicos que fazem parte de sua infraestrutura. Alguns desses componentes, possuem recursos adicionais. Porém, há quatro componentes básicos que todos os ambientes de **BI** devem fornecer, descritos as seguir (Rob, 2011):

- **Extração, transformação e carregamento (ETL) de dados:** esse componente é encarregado de coletar, filtrar, integrar e agregar dados operacionais a serem salvos em um armazém de dados otimizado para o suporte a decisões.
- **Armazenamento de dados (Data warehouse):** o armazém de dados é otimizado para o suporte a decisões e costuma ser representado por um *data warehouse* ou data mart. Ele contém dados de negócios extraídos de bancos de dados operacionais e de fontes externas. Esses dados são armazenados em estruturas otimizadas, com foco na velocidade de análise e consulta.
- **Processamento analítico online (OLAP):** esse componente executa as tarefas de recuperação, análise e mineração, utilizando os dados no armazém de dados e os modelos de análise de dados de negócio. Tal componente é utilizado pelo analista de dados para criar as consultas que acessam o banco de dados. Essa ferramenta orienta o usuário sobre quais dados selecionar e como construir um modelo de dados confiáveis.
- **Ferramentas de apresentação e visualização de dados:** esse componente é encarregado de apresentar os dados ao usuário final de várias formas. É utilizado pelo analista de dados para organizar e apesentar os dados. Essa ferramenta ajuda o usuário final a selecionar o formato de apresentação mais adequado, como relatório resumido, mapa ou gráfico.

2.2. DATA WAREHOUSE

Segundo Inmon (1994), o termo ***data warehouse*** é “um conjunto de dados integrado, orientado por assunto, variável no tempo e não volátil, que fornece suporte a tomada de decisões”.

- **Integrado.** O *data warehouse* é um banco de dados consolidado e centralizado, que integra dados proveniente de toda a organização e de várias fontes de dados, com diversos formatos.
- **Orientado por assunto.** Os dados do *data warehouse* são dispostos e otimizados de modo a fornecerem respostas a perguntas provenientes de diversas áreas funcionais da empresa. São organizados e resumidos por temas, contendo assuntos de interesse específico – produtos, clientes, departamentos, regiões, promoções, e assim por diante.
- **Variável no tempo.** Os dados são carregados periodicamente no *data warehouse*, e quando isso acontece, todas as agregações dependentes do tempo, são recalculadas. Por exemplo, se os dados de vendas da semana, são carregados no *data warehouse*, serão atualizadas todas as agregações dependentes dessa carga.
- **Não volátil.** Uma vez inserido um dado no *data warehouse*, ele nunca será removido. Uma vez que ele representa o histórico da empresa.

Resumindo, o *data warehouse* é um repositório de dados semanticamente consistente, que serve como uma implementação física de um modelo de dados de apoio a decisões. Ele armazena as informações que uma empresa necessita para tomar decisões (Han & Kamber, 2011). Normalmente é um banco de dados apenas de leitura, otimizado para processamento de análises e consultas. Em geral, os dados são extraídos de diversas fontes e, em seguida, transformados e integrados, antes de serem carregados no *data warehouse* (Inmon, 1994).

2.2.1 Arquitetura de Data Warehouse

Segundo Han & Kamber (2011), um ***data warehouse*** adota uma arquitetura em camadas, como ilustra a Figura 1.

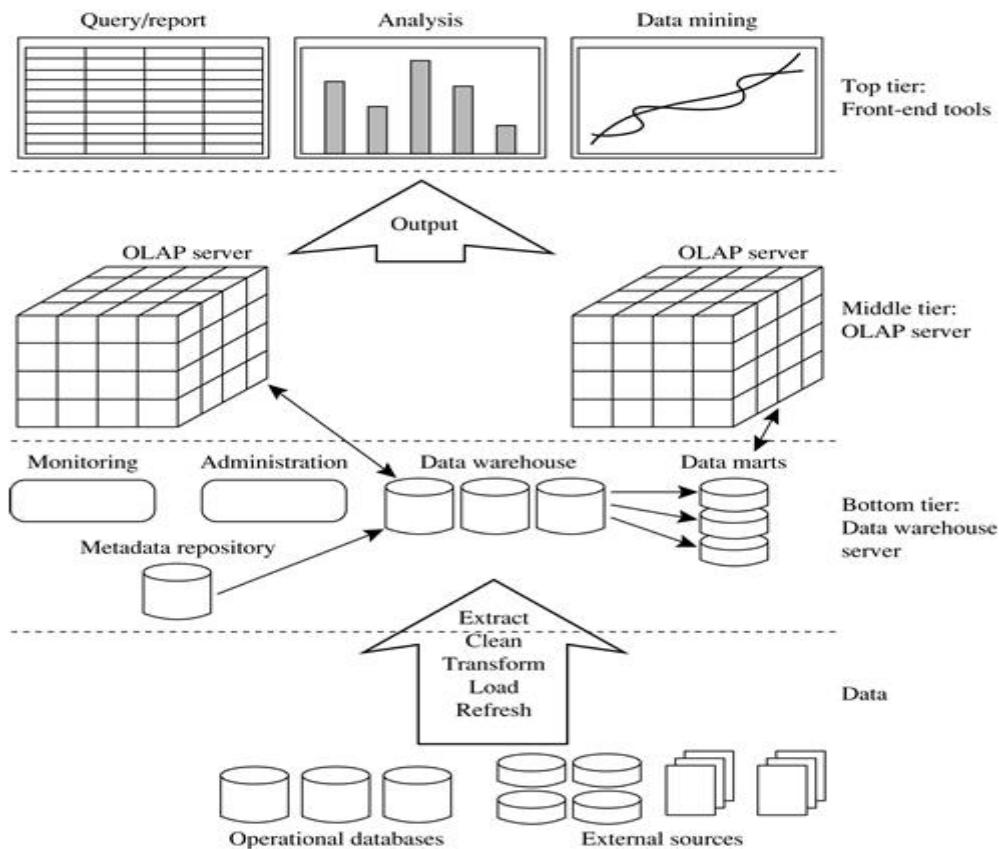


Figura 1. Uma arquitetura de data warehouse em camadas.

Fonte: (Han & Kamber, 2011).

1. **A camada inferior** é um servidor de *data warehouse*, que quase sempre, é um sistema de banco de dados relacional. Segundo Han & Kamber (2011), são usadas ferramentas de back-end e utilitários para extrair dados dessa camada e alimentar a camada superior.
2. **A camada intermediária**, segundo Han & Kamber (2011), é um servidor OLAP que geralmente é implementado usando (1) um modelo relacional **OLAP (ROLAP)** (fornece recursos de OLAP utilizando bancos de dados relacionais e ferramentas familiares de consulta relacional para armazenar dados multidimensionais; ou (2) um modelo multidimensional **OLAP (MOLAP)** (amplia os recursos de OLAP para sistemas de gerenciamento de banco de dados multidimensionais (**SGBDM**)).
3. **A terceira camada**, segundo Han & Kamber (2011), é o fronte-end do cliente, a qual contém as ferramentas de consulta, de relatório, de análise

e mineração de dados (por exemplo, análise de tendência, previsão, e assim por diante).

Segundo Han & Kamber, do ponto de vista da arquitetura, há três modelos de data warehouse: o Data Warehouse Empresarial, o Data Mart e Data Warehouse Virtual.

2.2.1. Data Warehouse Empresarial

Um data warehouse empresarial, segundo Han & Kamber (2011), coleta todas as informações sobre todos os assuntos da organização. Ele fornece informações de dados de toda empresa, geralmente, oriundas de um ou mais sistemas operacionais ou informações obtidas externamente. Ele contém dados detalhados e/ou sumarizados, e pode variar em tamanho de poucos gigabytes para centenas de gigabytes, terabytes ou superiores. A sua implementação pode ser em um mainframe, supercomputador, ou em uma plataforma de arquitetura paralela. Requer uma modelagem comercial extensiva e pode levar muitos anos para ser projetado e construído.

2.2.2. Data Mart

De acordo com Inmon (1994), embora o *data warehouse*, seja uma proposta muito atraente, que traga muitos benefícios, os gerentes podem relutar em adotar essa estratégia, pelo fato de que, a criação de um *data warehouse* exige tempo, dinheiro e considerável esforço gerencial. Estes fatos, fazem com que muitas empresas iniciem na criação de *data warehouse*, focando em conjuntos de dados gerenciais, orientados a atender pequenas áreas de negócio, dentro da empresa. Esses armazenamentos menores são chamados de ***data marts***. Um ***data mart*** é, portanto, segundo Inmon (1994), um pequeno subconjunto de um *data warehouse*, sobre um único assunto, que fornece suporte às decisões de um pequeno grupo de pessoas. No entanto, pode-se criar um *data mart* a partir de dados extraídos de um *data warehouse*, com a finalidade específica de dar suporte a um acesso mais rápido a determinado grupo ou função. Dessa forma, os *data marts* e o *data warehouse* podem coexistir em um ambiente de business *intelligence*.

2.2.3. Virtual Data Warehouse

De acordo com Han & Kamber (2011), um warehouse virtual é um conjunto de visões sobre bases de dados operacionais. Você pode materializar algumas visões operacionais, para obter um processamento de consultas eficientes. O warehouse virtual é o estado de visibilidade global de recursos, com base na aquisição e processamento de dados operacionais em tempo real. Informações disponíveis no armazém virtual tem o potencial de reduzir custos e melhorar o serviço ao cliente. A infraestrutura já está disponível para captura de dados em tempo real.

2.3. PROCESSAMENTO ANÁLITICO ON-LINE

De acordo com Rob (2011), a necessidade de suporte a decisões mais intensivo, levou à introdução de uma nova geração de ferramentas. Tais ferramentas, foram denominadas de **processamento analítico on-line (OLAP – Online Analytical Processing)**. Essa nova ferramenta cria um ambiente avançado de análise de dados que dá suporte à tomada de decisões, modelagem comercial e pesquisa operacional. Ainda segundo Rob (2011), esses sistemas comportam quatro características principais:

- **Utilizam técnicas de análise de dados multidimensionais:** De acordo com Rob (2011), a característica mais evidente das modernas ferramentas **OLAP**, é a capacidade de análise multidimensional, onde, os dados são processados e visualizados como parte de uma estrutura multidimensional. Essas técnicas de análise de dados multidimensionais, utilizam as seguintes funções: apresentação de dados em gráfico 3D, pivô, tabulações cruzadas, cubos dimensionais, e assim por diante. Além da capacidade de utilizar funções financeiras, estatísticas e de previsão.
- **Proporcionam suporte avançado a bancos de dados:** As ferramentas **OLAP**, para apresentar suporte eficiente a decisões, deve ter recursos avançados de acesso a dados. Tais recursos incluem (Rob, 2011):
 - Acesso a fontes de dados variadas, recursos de *drill down* e *roll up* e particionamento de bases de dados.
 - **Fornecem interface fácil de utilizar para o usuário final:** diversas ferramentas de geração de relatório, planilhas e de visualização de dados,

fornecem acesso aos cubos **OLAP**, facilitando, dessa forma a interação do usuário final, inclusive com interfaces gráficas fáceis de utilizar.

- **Dão suporte a arquitetura cliente/servidor:** Um ambiente cliente/servidor possibilita que um sistema OLAP seja dividido em vários componentes que definem sua arquitetura. Esses componentes podem, então, ser colocados no mesmo computador ou distribuídos entre diversas máquinas. Assim, segundo Rob (2011), o OLAP é projetado para atender a exigências de facilidades de utilização, ao mesmo tempo em que mantém a flexibilidade do sistema.

2.3.1 Arquitetura OLAP

Segundo Rob (2011), os sistemas OLAP são projetados para utilizar, tanto dados operacionais, quanto **data warehouse**. Para isso, existem várias arquiteturas de instalação de sistema OLAP, baseados nas regras de negócio de cada empresa. O fato é que, um sistema OLAP pode acessar ambos tipos de armazenamento de dados (operacional ou *data warehouse*) ou apenas um, dependendo da implementação que se deseje configurar. Em todo caso, a análise multidimensional de dados exige algum tipo de representação de dados multidimensionais, o que normalmente é fornecido pelo mecanismo OLAP.

De acordo com Rob (2011), há diversas formas de gerenciar e armazenar os dados em um sistema OLAP, como já foi visto anteriormente, o OLAP relacional (**ROLAP**) e o OLAP multidimensional (**MOLAP**):

- **O processamento analítico on-line relacional (ROLAP**, sigla em inglês para *Relational Online Analytical Processing*): Essa abordagem se estrutura a partir de tecnologias relacionais existentes e representa uma extensão natural para todas as empresas que já utilizem sistemas de gerenciamento de banco de dados relacionais. O ROLAP utiliza uma técnica especial de projeto que permite à tecnologia SGBDR dar suporte a representações de dados multidimensionais, conhecida como “Esquema estrela”.

- **O processamento analítico on-line multidimensional (MOLAP,** sigla em inglês para Multidimensional *Online Analytical Processing*): O MOLAP amplia os recursos de OLAP para sistemas de gerenciamento de banco de dados multidimensionais (**SGBDMs**). O pressuposto do MOLAP é que os bancos de dados multidimensionais são os mais adequados para gerenciar, armazenar e analisar dados multidimensionais. Nessa arquitetura os usuários visualizam os dados armazenados como um **cubo de dados**.

2.3.2. Modelo Multidimensional

A modelagem multidimensional é uma forma de Modelagem de Dados voltada para concepção e visualização de conjunto de medidas que descrevem aspectos comuns de um determinado assunto. É utilizada especialmente para sumarizar e reestruturar dados, apresentando-os em visões que suportem a análise dos dados envolvidos (Passos & Goldschimdt, 2005).

De acordo com Han & Kamber (2011), o *data warehouse* e as ferramentas **OLAP** são baseadas em um **modelo de dados multidimensional**. Nesses modelos, os dados são vistos na forma de um cubo de dados. Um modelo multidimensional possui três componentes básicos: Fatos (**facts tables**), Dimensões (**dimensions**) e Medidas (**measures**). Existem diversas formas de modelagem física de um data warehouse, incluindo esquema estrela (**star schema**), esquema floco de neves (**snowflake**) e constelação de fatos (**fact constellation**). Estes conceitos serão discutidos a seguir:

- **Esquema Estrela:** O esquema estrela, segundo Rob (2011), é uma técnica de modelagem de dados multidimensionais de suporte a decisões em um banco de dados relacional. O esquema estrela básico possui quatro componentes: **fatos, dimensões, atributos e hierarquias de atributos**.
 - **Fatos:** Um **fato** é uma coleção de itens de dados, composta de dados de medidas e de contexto. Representa um item, ou uma transação ou um evento associado ao tema da modelagem. São medidas numéricas (valores) que representam um aspecto ou atividade específica dos negócios. Os fatos são

armazenados em tabelas de fatos que constituem o centro do esquema estrela. A **tabela de fatos (fact table)** contém fatos vinculados por meio de suas dimensões (Kimball, 2002, Passos & Goldschimdt, 2005).

- **Dimensões:** Uma dimensão é um tipo de informação que participa da definição de um fato. As dimensões determinam o contexto do assunto. As **dimensões** são características de qualificação que fornecem perspectivas adicionais a um determinado fato. Essas dimensões são armazenadas em **tabelas de dimensões**.
- **Medidas:** Uma medida é um atributo ou variável numérica que representa um fato. Exemplos: valor da ação, número de evasões escolares, quantidade de produtos vendidos, valor total de venda, e assim por diante.
- **Atributos:** De acordo com Kimball (2002), cada tabela de dimensão contém atributos. Os atributos costumam ser utilizados para buscar, filtrar e classificar fatos. As dimensões fornecem características descritivas sobre os fatos por meio de seus atributos.
- **Hierarquias de Atributos:** De acordo com Kimball (2002) e Rob (2011), os atributos no interior de dimensões podem ser ordenados em hierarquias bem definidas. A hierarquia de atributos, fornecem uma organização vertical utilizada para duas finalidades principais: agregação e análise de dados por **drill down** e **roll up**.
- **Esquema Floco de Neves:** Segundo Rob (2011), para facilitar a navegação do usuário final, utiliza-se a técnica de normalização das tabelas dimensionais. Esse esquema normalizado é conhecido como esquema floco de neves.

Resumindo, projetar um *data warehouse* significa receber a oportunidade de ajudar a desenvolver um modelo integrado que capture os dados considerados essenciais para a organização, tanto da perspectiva do usuário

final, como da perspectiva dos negócios. Para tanto, um projeto de *data warehouse*, deve satisfazer:

- Critérios de integração e carregamento de dados.
- Recursos de análises de dados com desempenho aceitável de consulta.
- Necessidades de análises de dados do usuário final

Nessa seção foi apresentada uma visão geral sobre *Business Intelligence* (**BI**), onde deu para perceber que, o BI é um conjunto amplo, coeso e integrado de ferramentas e processos utilizados para captar, coletar, integrar, armazenar e analisar dados para a geração e a apresentação de informações que deem suporte à tomada de decisões para os gestores da empresa.

2.4. Descoberta de conhecimento em bases de dados

Os constantes avanços na área da tecnologia da informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Tecnologias como a Internet, sistemas gerenciadores de banco de dados, dispositivos de armazenamento de dados de maior capacidade e de menor custo e sistemas de informação em geral são alguns dos exemplos que, têm viabilizado a proliferação de inúmeras bases de dados de natureza comercial, administrativa, governamental e científica (Han & Kamber & Pei, 2011).

Pesquisas científicas, tais como missões espaciais da **NASA**, monitoramento temporal, em redes sociais, como Google, Facebook, e assim por diante, tem manipulado grandes massas de dados, que muitas vezes, têm alcançado proporções gigantescas, na ordem de zetabytes e ou petabytes de informações.

Diante desse cenário, a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portanto, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver estratégias de ação em cada contexto de aplicação (Han; Kamber & Pei, 2011).

Então, para atender a este novo contexto, surge segundo Han; Kamber & Pei (2011), uma nova área denominada a mineração de dados (em inglês **Data Mining** - DM), que vem despertando grande interesse junto às comunidades

científicas e industrial. No entanto, a mineração de dados é apenas uma fase em um processo maior chamado de Processo de descoberta de em bases de dados (em inglês **Knowledge Discovery in Databases – KDD**).

2.4.1. Caracterização do Processo de KDD

Basicamente, uma aplicação de **KDD** é composta por três tipos de componentes: o problema em que será aplicado o processo de **KDD**, os recursos disponíveis para a solução do problema e os resultados obtidos a partir da aplicação dos recursos disponíveis em busca da solução do problema (Passos & Goldschmidt, 2005).

- 1. O problema a ser submetido ao processo de KDD:** Este componente pode ser caracterizado por três elementos: conjunto de dados, o especialista do domínio da aplicação e objetivos da aplicação.
- 2. Os recursos disponíveis para solução do problema em questão -** Entre eles podem ser destacados: o especialista em KDD, as ferramentas de KDD e plataforma computacional disponível (Passos & Goldschmidt, 2005).
- 3. Os resultados obtidos a partir da aplicação dos recursos no problema** - Compreende, fundamentalmente, os modelos de conhecimento descobertos ao longo da aplicação de KDD e o histórico das ações realizadas (Passos & Goldschmidt, 2005).

Segundo Han et al. (2011); Witten & Frank (2005); Elmasri (2005), o processo de KDD é mostrado na Figura 2 e consiste da sequência iterativa dos seguintes passos: Seleção de dados, Limpeza dos dados, Enriquecimento, Transformação de dados, Data Mining, Avaliação dos padrões e representação do conhecimento.

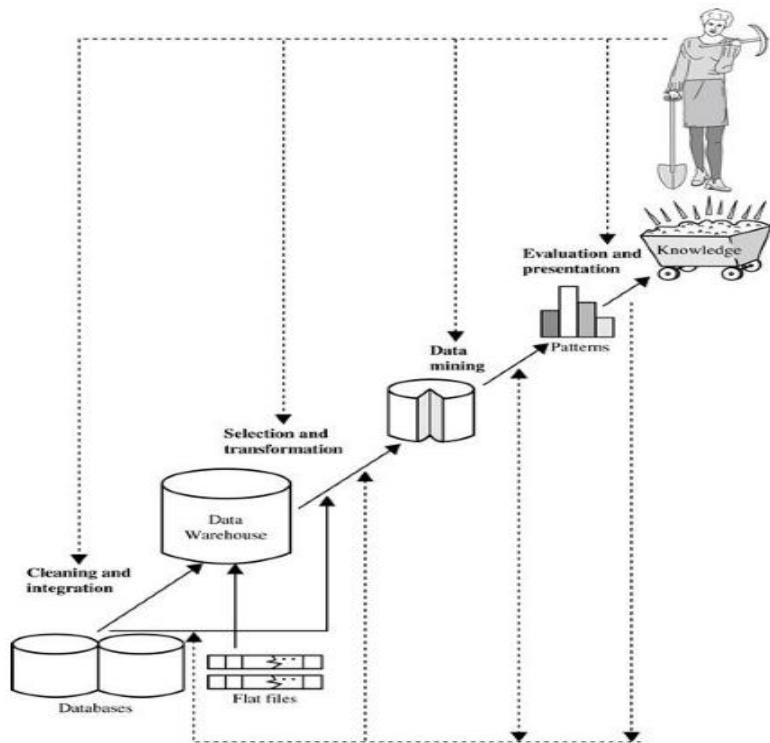


Figura 2 Mineração de Dados como um passo no processo de Descoberta de Conhecimento.

Fonte: Han; Kamber & Pei (2011).

- **Seleção dos dados:** Essa função, também denominada Redução de Dados, compreende, em essência, a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD. A seleção dos dados pode ter dois enfoques distintos: a escolha de atributos ou a escolha de registros que devem ser considerados no processo de KDD (Passos & Goldschmidt, 2005).
- **Limpeza dos Dados:** Abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos por eles representados. Informações ausentes ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD (Passos & Goldschmidt, 2005).
- **Transformação dos Dados:** Segundo Passos & Goldschmidt (2005), codificação de dados é a operação de pré-processamento responsável

pela forma como os dados serão representados durante o processo de KDD. A maneira como a informação é codificada tem forte influência sobre o tipo de conhecimento a ser encontrado. Em essência, a codificação pode ser: Numérica – Categórica, que transforma valores reais em categorias ou intervalos; Categórica – Numérica, que representa numericamente valores de atributos categóricos.

- **Enriquecimento dos dados:** De acordo com Passos & Goldschmidt (2005), a função de enriquecimento consiste em conseguir de alguma forma mais informações que possa ser empregada aos registros existentes, enriquecendo os dados, para que esses forneçam mais informações para o processo de descoberta de conhecimento. A seguir será comentada algumas das operações mais usualmente utilizadas no processo de enriquecimento das bases de dados.
- **Mineração de Dados:** De acordo com Han; Kamber & Pei (2011); Passos & Goldschmidt (2005), a execução da etapa de Mineração de Dados (Data Mining) compreende a aplicação de algoritmos sobre os dados procurando abstrair conhecimento. Estes algoritmos são fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento.
- **Avaliação dos padrões:** Segundo Liu & Hsu (1996), a obtenção do conhecimento não é o passo final do processo de Extração de Conhecimento de Bases de dados. O conhecimento extraído pode ser utilizado na resolução de problemas da vida real, seja por meio de um Sistema Inteligente ou de um ser humano como apoio a algum processo de tomada de decisão. Para isso é importante que algumas questões sejam respondidas aos usuários:
 - O conhecimento extraído representa o conhecimento do especialista?
 - De que maneira o conhecimento do especialista difere do conhecimento extraído?
 - Em que parte o conhecimento do especialista está correto?

Nessa seção foi mostrado de uma forma sucinta, o processo de descoberta de conhecimento em bases de dados e, na seção a seguir, será

detalhado o processo de mineração de dados. Como foi mostrado, o processo de mineração de dados é uma das fases do processo maior, o processo **KDD**.

2.5. Mineração de Dados

2.5.1. Definição

Existem, segundo Han; Kamber & Pei (2011), várias definições para Mineração de dados (em inglês **Data Mining – DM**), por ser um assunto verdadeiramente interdisciplinar, ela envolve um extenso campo de pesquisa, que associa técnicas e conceitos de diversas áreas como sistemas de banco de dados, sistemas baseados em conhecimento, inteligência artificial, aprendizado de máquina, aquisição do conhecimento, estatística, bancos de dados espaciais e visualização de dados. As várias tarefas desenvolvidas em Mineração de Dados têm como objetivos primário a predição e/ou a descrição. A predição usa atributos para predizer valores futuros de uma ou mais variáveis (atributos) de interesse. A descrição contempla o que foi descoberto nos dados de vista da interpretação humana.

Durante a etapa de Mineração de Dados é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de **KDD**. É por tanto, na Mineração de Dados, onde são definidas, as técnicas e os algoritmos a serem utilizados no problema em questão. A escolha da técnica depende, muitas vezes, do tipo de tarefa de KDD a ser realizada. São muitas as tarefas de Mineração de dados, dentre elas as tarefas de **Associação, Classificação, Árvores de Decisão e Agrupamento**.

De acordo com Han; Kamber & Pei (2011); Passos & Goldschmidt (2005), a execução da etapa de Mineração de Dados (Data Mining) compreende a aplicação de algoritmos sobre os dados procurando abstrair conhecimento. Estes algoritmos são fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento.

2.5.2. Tarefas de Mineração de Dados

As metas primárias que podem ser alcançadas através da Mineração de Dados, são as seguintes (Fayyad, et al, 1996):

- **Previsão** - Nesse caso busca-se um modelo de conhecimento que permita, a partir de um histórico de casos anteriores, prever os valores de determinados atributos em novas situações.
- **Descrição** - Nesse caso busca-se por um modelo que descreva, de forma compreensível pelo homem, o conhecimento existente em um conjunto de dados.

Ainda, segundo Fayyad, et al (1996), a mineração preditiva consiste na generalização de exemplos ou experiências passadas com respostas conhecidas ou regras de negócios estabelecidas por especialistas. Enquanto, a mineração descritiva, consiste na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe específica.

2.5.2.1. Descoberta de Associação

De uma forma geral, a tarefa clássica de busca por regras de associação, também denominada regras associativas, foi introduzida em (Agrawal et al., 1993). Intuitivamente essa tarefa consiste em encontrar conjunto de registros de itens que ocorram simultaneamente e de forma frequente em um banco de dados.

Uma **regra de associação** é uma implicação da forma $X \Rightarrow Y$, onde $X = \{x_1, x_2, \dots, x_n\}$ e $Y = \{y_1, y_2, \dots, y_m\}$ são conjunto de itens, com x_i e y_j sendo itens distintos para todo i e todo j . Essa associação estabelece que se um cliente comprar X , ele também estará propenso a comprar Y . Para que uma regra de associação seja de interesse para a mineração de dados, a regra precisa satisfazer algumas medidas. Duas medidas de interesse comuns fornecem suporte e confiança. Segue a seguir, as formulas para cálculo do fator de suporte e de confiança (Agrawal, et al., 1996; Elmasri, 2005).

$$F_s = \frac{|X \cup Y|}{N}, \text{ onde } N \text{ é o número total de tuplas.} \quad (\text{Equação 2.1})$$

$$F_c = \frac{|X \cup Y|}{X} \quad (\text{Equação 2.2})$$

O fator de **suporte** pode ser descrito como a probabilidade de uma transação qualquer satisfazer tanto X como Y, ao passo que o fator de **confiança** é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaça X. A tarefa de descobrir regras de associação consiste em extrair do banco de dados todas as regras com “Fs” e “Fc” maiores ou iguais a um “Fs” e “Fc” especificado pelo analista de dados (Agrawal, et al., 1996).

Uma associação é considerada frequente se, o número de vezes em que a união de conjuntos de itens ($X \cup Y$), ocorrer em relação ao número total de transações do banco de dados, for superior a uma frequência mínima (denominada suporte mínimo), que é estabelecida em cada aplicação. Busca-se por meio do suporte, identificar que associações surgem em uma quantidade expressiva a ponto de ser destacada das demais existentes (Agrawal, et al., 1996).

Ainda segundo Agrawal et al. (1996), uma associação é considerada válida se, o número de vezes em que $X \cup Y$ ocorrer, em relação ao número de vezes que X ocorrer, for superior a um valor denominado confiança mínima, que também é estabelecida em cada aplicação. A medida de confiança procura expressar a qualidade de uma regra, indicando o quanto a ocorrência do antecedente da regra pode assegurar a ocorrência do consequente desta regra.

Desta forma, a tarefa de Descoberta de Associações ou Descoberta de Regras de Associações, pode ser definida formalmente como a busca por **regras de associação frequentes e válidas** em um banco de dados, a partir da especificação dos parâmetros de suporte e confiança mínima (Agrawal et al., 1996).

Os valores dos parâmetros de suporte e confiança mínima devem ser especificados pelo especialista em KDD em conjunto com o especialista no domínio da aplicação (Agrawal et al., 1996).

Segundo Agrawal et al. (1996), existem diversos algoritmos desenvolvidos especificamente para aplicação na tarefa de descoberta de associações, dentre eles: Apriori, DHP (Direct Hashing and Pruning), Partition, DIC (Dynamic ItemSet Counting), Eclat, MaxEclat, Clique, MaxClique, etc. além dos mais, existem versões destes algoritmos para sistemas distribuídos.

2.5.2.2. Classificação

Essa tarefa pode ser compreendida como a busca por uma função que permita associar corretamente cada entrada X_i de um conjunto de dados de entrada, a um único rótulo Y_i , denominado classe. Uma vez identificada, essa função pode ser aplicada a novas entradas de forma a prever a classe em que tais dados se enquadram. Dados podem ser associados a classes ou a conceitos através de um processo de discriminação ou caracterização (Han; Kamber & Pei, 2011).

De acordo com Han; Kamber & Pei (2011), a tarefa de classificação é um processo de dois passos. No primeiro passo, constrói-se um modelo com base nos dados. No segundo passo, determina se a acurácia desse modelo é aceitável, se assim for, usa-se esse modelo para classificar novos dados.

Para formalizar a tarefa de classificação, consideremos um par ordenado da forma $(x, f(x))$, onde x é um vetor de entradas n-dimensional e $f(x)$ a saída de uma função f , desconhecida, aplicada a x . A tarefa de inferência indutiva consiste em, dada uma coleção de exemplos de f , obter uma função h que se aproxime de f . A função h é chamada de hipótese ou modelo de f (Passos, Goldschmidt, 2005).

Nos casos em que a imagem de f é formada por rótulos de classes, a tarefa de inferência indutiva é denominada classificação e toda hipótese h chamada de classificador. A identificação da função h consiste em um processo de busca no espaço de hipóteses H , pela função que mais se aproxime da função original f . Esse processo é denominado **aprendizado** (Russel e Norving, 1955, citado por Passos & Goldschmidt, 2005). Todo algoritmo que possa ser utilizado nesse processo é denominado de **algoritmo de aprendizado**. O conjunto de todas as hipóteses que podem ser obtidas a partir de um algoritmo de aprendizado L é representado por H_L . Cada hipótese pertencente ao H_L , é representado por h_L .

A acurácia de um classificador em um dado conjunto de teste é a percentagem do conjunto de tuplas de testes que são corretamente classificadas pelo classificador Han; Kamber & Pei (2011), ou seja, a acurácia da hipótese h em mapear corretamente cada vetor de entradas x em $f(x)$. O conjunto de pares

$(x, f(x))$ utilizado na identificação da função h é denominada conjunto de treinamento. Por outro lado, o conjunto de pares $(x, f(x))$ utilizados para avaliar a acurácia de h é denominado conjunto de testes. Dessa forma, o algoritmo L pode ser interpretado como uma função tal que:

$$L: T \rightarrow H_L \quad (\text{Equação 2.3})$$

Onde,

T é o espaço composto por todos os conjuntos de treinamento possíveis para L .

Segundo Utgolff (1996), cada algoritmo possui *bias* indutivo que direciona o processo de construção dos classificadores. O *bias* indutivo de um algoritmo, pode ser definido como o conjunto de fatores que, coletivamente influenciam na seleção de hipótese.

Em termos práticos, o *bias* de um algoritmo de aprendizado L afeta o processo de aprendizado de duas formas: restringem o tamanho de espaço de hipóteses H_L , e impõem uma ordem de preferência sobre as hipóteses em H_L (Bensusan, 1999, citado por Utgolff, 1996).

Conforme mencionado anteriormente, uma medida de desempenho de um classificador comumente utilizada é a acurácia ($\text{Acc}(h)$), também conhecida como precisão do classificador.

$$\text{Acc}(h) = 1 - \text{Err}(h) \quad (\text{Equação 2.4})$$

Onde,

$\text{Err}(h)$ é denominada taxa de erro ou taxa de classificação incorreta:

$$\text{Err}(h) = \frac{1}{n} \sum_{i=1}^n ||y_i \neq h(i)|| \quad (\text{Equação 2.5})$$

Onde,

O operador $||E||$ retorna 1 se a expressão E for verdadeira e 0, caso contrário;

n é o número de exemplos (registros da base de dados);

y_i é a classe real associada ao i -ésimo exemplo;

$h(i)$ é a classe indicada pelo classificador para o i -ésimo exemplo.

O modelo derivado pode ser apresentado de várias formas, tais como regras de classificação (***IF-THEN***), árvores de decisão, fórmulas matemáticas, ou redes neurais. Existem muitos outros métodos de construção de modelos de classificação, tais como classificação ***naïve Bayesian***, máquina de vetor de suporte (***support vector machines***), e classificação do vizinho mais próximo (***k-nearest neighbor***), Backpropagation, classificação usando Padrões frequentes, etc (Han; Kamber & Pei, 2011).

2.5.2.3. Agrupamento (Clustering)

Técnicas de agrupamento e classificação objetivam realizar uma separação ótima entre objetos de uma coleção, permitindo a descoberta de novos padrões, previamente desconhecidos. O resultado da segmentação, independentemente da ferramenta utilizada, pode ser interpretado eficientemente por um especialista na área de origem dos dados sob análise. A facilidade de visualização resultante do agrupamento, favorece a análise (Han; Kamber & Pei, 2011).

A tarefa de agrupamento, é usada para segmentar os dados de entrada em subconjuntos ou clusters, de tal forma que elementos de um cluster compartilhem um conjunto de propriedades comuns que os distingam dos elementos de outros clusters. O objetivo desta tarefa é maximizar similaridades intra-cluster e minimizar similaridades inter-cluster. Diferente da classificação que tem rótulos predefinidos, o agrupamento precisa automaticamente identificar os rótulos. Por essa razão, o agrupamento também é chamado de indução não supervisionada (Passos & Goldschmidt, 2005).

Segundo Han; Kamber & Pei (2011), em geral, a classe não é representada nos dados de treinamento, simplesmente porque elas não são conhecidas no começo. Entretanto, os agrupamentos (*clusters*) podem ser usados para gerar tais rótulos. Os objetos são agrupados baseados no princípio de máxima similaridade intra-classe e mínima similaridade inter-classe. Isto é, o agrupamento de objetos, é formado, de modo que os objetos dentro do agrupamento, tenham alta similaridade em comparação com um outro objeto, mas são muitos diferentes dos objetos em outro agrupamento. Cada

agrupamento formado pode ser visto como uma classe de objetos, da qual pode derivar regras.

Formalmente para o processo de agrupamento, supõe-se a existência de n pontos de dados x_1, x_2, \dots, x_n tais que cada ponto pertença a um espaço dimensional \mathbb{R}^d . A tarefa de agrupamento desses pontos de dados, separando-os em k clusters consiste em encontrar k pontos m_j em \mathbb{R}^d de tal forma que a expressão (Passos & Goldschmidt, 2005):

$$\frac{\sum_i \min_j d^2(x_i, m_j)}{N} \quad (\text{Equação 2.6})$$

Seja minimizada, onde $d^2(x_i, m_j)$ denota uma distância entre x_i e m_j . Os pontos m_j são denominados centroides ou médias dos clusters.

De forma resumida, o problema descrito acima consiste em encontrar k centroides de clusters de tal maneira que a distância entre cada ponto de dado e o centroide do cluster mais próximo seja minimizada (Passos & Goldschmidt, 2005).

2.5.3. Métodos de Mineração de Dados

De acordo com Passos & Goldschmidt (2005), cada método de Mineração de Dados requer diferentes necessidades de pré-processamento. Tais necessidades variam em função do aspecto extensional da base de dados em que o método será utilizado. Em decorrência da diversidade de métodos de pré-processamento de dados, são muitas as alternativas possíveis de combinações entre métodos. A escolha dentre estas alternativas pode influenciar na qualidade do resultado do processo de KDD (Morik, 2000; Engels, 1996; Engels et al., 1997).

Os métodos de Mineração de Dados, sendo um caso particular de um método KDD, podem ser considerados operadores definidos a partir de precondições e efeitos. Uma precondição de um método de KDD é um predicado que estabelece um requisito que deve ser cumprido antes da execução do método. Um efeito de um método de KDD também é um predicado que descreve uma situação gerada após a aplicação do método. Um plano de ação de KDD

válido é toda sequência de métodos de KDD onde as precondições para execução de cada um dos métodos da sequência, sejam devidamente atendidas Passos & Goldschmidt (2005).

Dentre os métodos de Mineração de dados, pode-se citar os baseados em redes neurais, os estatísticos, os métodos específicos Apriori, indução de árvore de decisão, lógica nebulosa, hierárquicos e os métodos baseados em densidade.

- **Métodos Baseados em Redes Neurais:** Segundo Han; Kember & Pei (2011); Passos & Goldschmidt (2005); Rezende (2003), diversos modelos de Redes Neurais podem ser utilizados na implementação de métodos de Mineração de Dados. Classificação, Regressão, Previsão de Séries Temporais e Agrupamento (Clusters) são exemplos de tarefas de Mineração de Dados que podem ser implementadas por métodos de Redes Neurais. Além do mais, alguns modelos de Redes Neurais podem ser aplicados em mais de um tipo de tarefa de Mineração de dados. Alguns dos algoritmos de aprendizado indicados na tarefa de Mineração de dados são:
 - **Back-Propagation:** O algoritmo Back-Propagation, também conhecido como algoritmo de retro-propagação do erro, é um algoritmo de aprendizado supervisionado, cuja aplicação é adequada a tarefa de Mineração de Dados tais como Classificação, Regressão ou Previsão. Esse algoritmo tem como objetivo minimizar a função de erro entre a saída gerada pela rede neural e a saída real desejada, utilizando o método do gradiente descendente.
 - **Kohonen:** O mapa Kohonen pertence à classe das Redes Neurais Auto organizáveis. Em uma Rede Neural Auto organizável o treinamento é não supervisionado, geralmente baseado em uma forma de competição entre os elementos processados. Entre as principais aplicações das Redes Auto organizáveis estão:
 - **Tarefas de Clusterização** – Tarefa na qual os dados de entrada devem ser agrupados em conjuntos que agregam padrões semelhantes.

- **Detecção de Regularidades** – Modelo em que o sistema deve extrair as características relevantes dos padrões de entrada.
- **Métodos Estatísticos:** De acordo com Passos & Goldschmidt (2005); Han; Kamber & Pei (2011), diversos algoritmos de Mineração de Dados são fundamentados em princípios e teorias da Estatística. A seguir serão apresentados alguns deles:
 - **Classificador Bayeasiano Ingênuo:** O classificador Bayeasiano Ingênuo baseia-se no Teorema de Bayes, estando relacionado ao cálculo de probabilidades condicionais. É aplicável, conforme o próprio nome sugere, em tarefas de classificação.
 - **K-Means:** O algoritmo k-means é um método popular da tarefa de agrupamento. Toma-se, randomicamente, k pontos de dados (dados numéricos) como sendo os centroides (elementos centrais) dos clusters. Em seguida, cada ponto (ou registro da base de dados) é atribuído ao cluster cuja distância deste ponto em relação ao centroide de cada cluster é a menor dentre todas as distâncias calculadas. Um novo centroide para cada cluster é computado pela média dos pontos do cluster, caracterizando a configuração dos clusters para a iteração seguinte. O processo termina quando os centroides dos clusters param de se modificar, ou após um número limitado de iterações que tenham sido especificados pelo usuário (Han, Kamber & Pei, 2011; Passos & Goldschmidt, 2005).
 - **K-Modes** – O algoritmo k-modes é uma variação do método k-means, só que utilizado para agrupamento de dados categóricos (nominais). Em geral, no lugar do cálculo da média, calcula-se a moda dos objetos, usando medidas de similaridades para tratar objetos categóricos, e usando métodos baseados em frequência para atualizar as modas dos clusters.
 - **K-Prototypes** – O método *k-prototypes* é a integração dos métodos *k-means* e *k-modes*. Esse método pode ser aplicado a bases de dados que contenham tanto atributos numéricos quanto atributos categóricos.

- **K-Medoids** - O algoritmo k-medoids baseia-se, primeiramente, em encontrar os medoids (objetos mais centralmente localizado em um cluster). Os objetos restantes são então agrupados com o medoid ao qual ele é mais similar. Há então uma troca iterativa, de um medoid por um não medoid, visando à melhoria do agrupamento. O método então, é realizado baseado no princípio de minimizar a soma das dissimilaridades entre cada objeto p e seu correspondente objeto representativo.
- **Método Específico – Apriori:** O *Apriori* é um algoritmo clássico de Mineração de Regras de Associação (Agrawal, 1993). Diversos algoritmos tais como **GSP**, **DHP**, **Partition**, **DIC**, etc., foram inspirados no funcionamento do *Apriori* e se baseiam no princípio da antimonotonicidade do suporte. Segundo este princípio, “Um k -itemset somente pode ser frequente se todos os seus $(k-1)$ -itemsets forem frequentes”. Assim sendo, a combinação de itemsets para gerar um novo itemset somente ocorre quando estes são frequentes.
- **Métodos Baseados em Indução de Árvores de Decisão:** Segundo Passos & Goldschmidt (2005), alguns dos principais métodos de Mineração de Dados são baseados na construção de árvores de decisão a partir da base de dados. Em geral a construção de uma árvore de decisão é realizada segundo alguma abordagem recursiva de particionamento da base de dados. Um exemplo clássico de método baseado na indução de árvores de decisão é o algoritmo C4.5.
 - **C4.5** – O C4.5, procura abstrair árvores de decisão a partir de uma abordagem recursiva de particionamento das bases de dados. Utiliza, para tanto, conceitos e medidas da Teoria da Informação.

Existem muitos outros métodos de Mineração de Dados, que serão aqui apenas citados, pois fogem ao escopo dessa pesquisa, tais como: Métodos Baseados em Lógica Nebulosa, Métodos Hierárquicos, Métodos Baseados em Densidade, dentre outros.

2.5.4. Tecnologias de Suporte a Mineração de Dados

O processo de **KDD** é realizado incorporando-se várias técnicas de diferentes áreas como Aprendizado de Máquina, Data Warehousing, Banco de Dados, Estatísticas, Visualização de Dados, dentre outras. A etapa de Extração de Conhecimento, por exemplo, utiliza muitos recursos da área de Aprendizado de Máquina, e as demais são consideradas como áreas de apoio ao processo de KDD (Han; Kamber & Pei, 2011).

2.5.4.1. Aprendizagem de Máquina

De acordo com Han; Kamber & Pei (2011), Aprendizagem de Máquina, investiga como os computadores pode aprender, com base em dados. Ou seja, a Aprendizagem de Máquina (do Inglês Machine Learning) é uma área da Inteligência Artificial (**IA**) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o processo de aprendizado [Rezende, 2003].

De acordo com Mitchell (1997), aprendizado de máquina, é uma subárea da **IA**, cujo objetivo é desenvolver métodos, técnicas e ferramentas para construir máquinas inteligentes, que se modificam para realizar cada vez melhor suas tarefas.

Para aprender, os sistemas, bem como os seres humanos, podem se valer de estratégias de aprendizado. A seguir, será apresentado, algumas dessas estratégicas clássicas de aprendizagem de máquina relacionado a mineração de dados (*data mining*) (Han; Kamber & Pei, 2011):

- **Aprendizado Supervisionado:** No aprendizado supervisionado, o objetivo é induzir conceitos a partir de exemplos que estão pré-classificados, ou seja, exemplos que estão rotulados com uma classe conhecida. Se as classes possuírem valores discretos, o problema é categorizado como classificação. Caso as classes possuam valores contínuos, o problema é categorizado como regressão [Han; Kamber & Pei, 2011].
- **Aprendizado Não-supervisionado:** É essencialmente, um sinônimo para agrupamento (*cluster*). No aprendizado não-supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou *clusters*. A tarefa do algoritmo é agrupar exemplos não rotulados, i.e.,

exemplos que não possuem o atributo classe especificado. Nesse caso, é possível utilizar algoritmos de aprendizado para descobrir padrões nos dados a partir de alguma caracterização de regularidade, sendo esses padrões denominados *clusters* (Decker & Focardi, 1995; McCallum, Nigan, & Ungar, 200). Tipicamente, pode-se usar agrupamento para descobrir classes dentro de dados (Han; Kamber & Pei, 2011).

- **Aprendizado semi-supervisionado:** o aprendizado semi-supervisionado, consiste em utilizar algoritmos que aprendem apartir de exemplos rotulados e não rotulados. Ou seja, o aprendizado semi-supervisionado, pode ser aplicável tanto em tarefas de classificação quanto em tarefas de agrupamento (Han; kamber & Pei, 2011).
- **Aprendizado ativo:** é uma abordagem de aprendizado de máquina, que permite que o usuário execute um papel ativo no processo de aprendizagem. Uma abordagem de aprendizado ativo, pode pedir ao usuário (i.e., o especialista do domínio) para rotular um exemplo, que pode ser a partir de um conjunto de exemplos não-rotulados, ou sintetizado pelo programa de aprendizagem. O objetivo é otimizar a qualidade do modelo, através de aquisição de conhecimento ativamente do usuário humano, a partir da quantidade exemplos a serem rotulados (Han; Kamber & Pei, 2011).

2.5.4.2. Banco de Dados e Data Warehousing

Um Banco de dados é uma coleção integrada de dados, organizada de tal forma a facilitar o armazenamento eficiente, assim como sua modificação e recuperação (DATE, 2003). Um Sistema Gerenciador de Banco de Dados (SGBD) é uma coleção de procedimentos e mecanismos para recuperação, armazenamento e manipulação de dados.

O *Data Warehousing*, se refere ao processo de coleta de dados e pré-processamento dos dados armazenados em um ou mais banco de dados operacionais, com o objetivo de servir de fonte para os Sistema de Suporte a Decisão. O resultado desse processo é a criação de um Depósito de Dados (tradução da forma em inglês *Data Warehouse*), uma coleção de dados integrados, consolidados e possivelmente estruturado no tempo (dados históricos) (GOLDSCHMIDT, 2015).

No entanto, para o processo de Mineração de Dados não é necessário ser implementado um **DW**. No entanto, se a empresa já possui um DW, o tempo

gasto na etapa de pré-processamento será reduzido drasticamente (Inmon, 1996).

2.5.4.3. Estatística

A estatística estuda as coleções, análise, interpretação ou explicação, e apresentação dos dados. A mineração de dados tem conexão própria com a estatísticas (Han; Kamber & Pei, 2011).

Segundo Han; Kamber & Pei (2011), um modelo estatístico, é um conjunto de funções matemáticas, que descrevem as ações dos objetos, em uma classe alvo, em termos de suas variáveis aleatórias e sua distribuição de probabilidades associadas. Ainda segundo os autores, os modelos estatísticos são muito usados para modelar dados e classes de dados. Por exemplo, pode-se utilizar os modelos estatísticos para caracterização e classificação de dados em tarefas de Mineração de Dados. Em outras palavras, tais modelos estatísticos podem ser o resultado de uma tarefa de Mineração de Dados. Alternativamente, tarefas de Mineração de Dados podem ser construídas em cima dos modelos estatísticos.

A Estatística, junto com a área de Aprendizado de Máquina, é considerada ancestral da área de KDD. Técnicas de reconhecimento de padrões e de análises exploratória de dados provenientes da estatística são muito utilizados em algoritmos de Mineração de Dados (GOLDSCHMIDT, 2015). Seleção de dados e amostragem, pré-processamento, transformação dos dados e avaliação de padrões extraídos são apenas alguns exemplos de métodos há muito tempo utilizados em estatística e que são aplicados durante o processo de KDD (GOLDSCHMIDT, 2015).

Segundo Goldschmidt (2015), o conhecimento que se infere a partir dos dados tem um componente estatístico fundamental, que é o grau de certeza com o qual se espera que este conhecimento descreva ou faça previsões sobre os dados. A estatística fornece uma linguagem para quantificação da incerteza resultante quando se tenta inferir padrões a partir de uma amostra de uma coleção de dados.

2.5.4.4. Visualização de Dados

As técnicas e ferramentas para Visualização de Dados são indispensáveis ao processo de Mineração de dados. Elas podem ser usadas durante a execução das etapas do processo de extração de conhecimento melhorando a compreensão dos resultados obtidos e a comunicação entre os usuários (Rezende et al., 1998).

As técnicas de Visualização de Dados estimulam naturalmente a percepção e a inteligência humana, aumentando a capacidade de entendimento e associação de novos padrões. Logo, a Visualização de Dados utiliza a percepção humana como um primeiro método para descobrir valores. Poderosas ferramentas de visualização que consigam gerar diversas formas de visualização (árvores, regras, gráficos 3D/2D, espectro) combinadas com técnicas de Mineração de Dados podem melhorar o processo de Mineração de dados (Fayyad, Grinstein, & Wierse, 2002).

Segundo Han, Kamber & Pei (2011), a visualização de dados, utiliza várias abordagens, incluindo técnicas orientada a pixel, técnicas de projeção geométrica, técnicas baseada em ícone, hierárquica e técnicas baseada em gráfico.

Nesta seção foi feita uma abordagem geral, mais sucinta, do processo de Mineração de dados, apresentando alguns de seus algoritmos e métodos utilizados na extração de conhecimento em grande massa de dados.

2.6. Sistema de Raciocínio Baseado em Casos

Os sistemas de raciocínio baseados em casos representam um modelo cognitivo de raciocínio. Sua técnica é utilizar experiências passadas, para encontrar soluções aos novos problemas.

No desejo de compreender como as pessoas conseguem recuperar informações e que elas, frequentemente resolvem problemas lembrando-se como solucionaram casos similares no passado, descreve-se o estímulo do desenvolvimento do RBC [Watson, 2003].

2.6.1. Definição

Raciocínio Baseado em Casos (**RBC**) (do Inglês *Case-Based Reasoning* – **CBR**), é um conjunto de técnicas para o funcionamento de um sistema

inteligente, tendo como centro de operações, uma base de conhecimento prévia, composta de experiências contextualizadas, que descrevem os problemas e suas respectivas soluções, chamada Base de Casos. Estas soluções podem ser sugeridas (reutilização) para serem aplicadas a quaisquer novos casos, ou problemas, que sejam apresentados, a partir da relação de similaridade (recuperação) do problema apresentado com os casos existentes, podendo inclusive, se servir destes novos casos (retenção) para ampliar a base de conhecimento (Watson 1997).

Fernandes (2005) descreve o Raciocínio Baseado em Casos como: Raciocínio Baseado em Casos (RBC) é uma ferramenta de raciocínio da Inteligência Artificial. A filosofia básica desta técnica é a de buscar a solução para uma situação atual através da comparação com uma experiência passada semelhante. O processo característico do **RBC** consiste em: identificar o problema atual, buscar a experiência mais semelhante na memória e aplicar o conhecimento dessa experiência passada no problema atual.

Riesbeck e Schank (1996), definem **CBR** como “Um sistema de CBR resolve problemas por adaptar soluções que foram utilizadas para resolver problemas anteriores”.

Segundo Wangenheim (2003, p. 10), os elementos básicos de um sistema RBC são:

- **Representação do conhecimento:** Em um sistema de RBC, o conhecimento é representado principalmente em forma de casos que descrevem experiências concretas. No entanto, se for necessário, também outros tipos de conhecimentos sobre o domínio de aplicação podem ser armazenados em um sistema de RBC (por exemplo, casos abstratos e generalizados, tipos de dados, modelos de objetos usados como informação).
- **Medida de similaridade:** Temos de ser capazes de encontrar um caso relevante para o problema atual na base de casos e responder à pergunta quando um caso relembrado for similar a um novo problema.
- **Adaptação:** Situações passadas representadas como casos dificilmente serão idênticas às do problema atual. Sistemas de RBC avançados têm mecanismos e conhecimento para adaptar os casos recuperados

completamente, para verificar se satisfazem às características da situação presente.

- **Aprendizado:** Para que um sistema se mantenha atualizado e evolua continuamente, sempre que ele resolver um problema com sucesso, deverá ser capaz de lembrar (armazenar) dessa situação no futuro como mais um novo caso.

2.6.2. Ciclo do Sistema Raciocínio Baseado em Casos

Alguns autores descrevem o processo que compõem um sistema de raciocínio baseado em casos em seis atividades recuperar, reusar, revisar, avaliar, manter e refinar. No entanto, WANGENHEIM (2003), descreve esse processo em quatro partes, as quais são:

- **Recuperação:** Recupera, na base de casos, o caso mais parecido com o novo problema. Identifica e pesquisa índices, calcula a similaridade entre o caso recuperado e o novo problema.
- **Reutilização:** Reutiliza a solução associada ao caso recuperado no contexto do novo problema, identificando as diferenças entre o caso recuperado e o problema, e identificando as 8 partes do caso recuperado que pode ser transferido ao novo contexto. Geralmente, a solução do caso recuperado é transferida ao novo problema diretamente como sua solução.
- **Revisão:** É necessário revisar a solução do caso recuperado, gerado pelo processo de reutilização, quando a solução não pode ser aplicada diretamente no novo problema. Esta etapa avalia as diferenças entre o problema recuperado da base de casos e o problema de entrada, e qual parte do caso recuperado pode ser transferida para o novo caso, adaptando, assim, a solução do caso recuperado à solução do novo caso.
- **Retenção:** É o processo de incorporar tudo que for útil no novo problema na base de casos. Isto envolve decidir que informações armazenar e de que forma armazenar, como indexar o caso para futuras recuperações e integrar o novo caso à base de casos.

A figura 3 mostra o ciclo do sistema raciocínio baseado em casos.

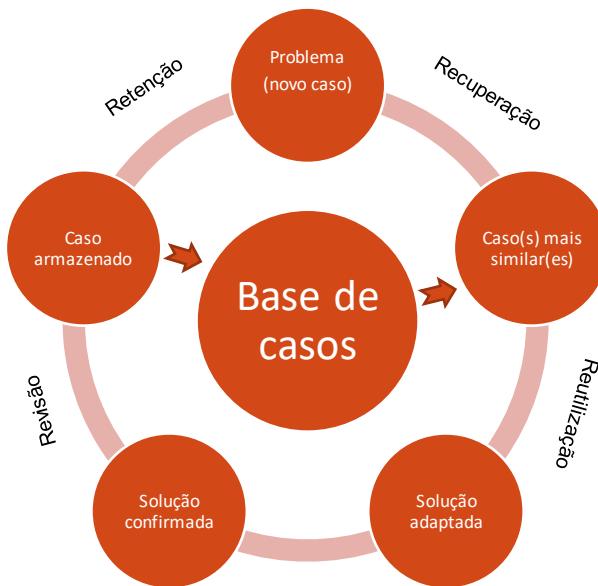


Figura 3 – Ciclo do Raciocínio Baseado em Casos Fonte.

Fonte: Adaptado de Wangenheim (2003)

O sistema efetua uma busca na base de casos, a partir de um novo problema, e recupera um conjunto de casos similares que possam atender à solução do problema em questão. Os casos recuperados são avaliados para posterior utilização e/ou armazenamento na base.

2.6.3. Representação de Casos

De acordo com Watson (2003), casos são registros de experiências que contém conhecimento, que pode ser ambos explícito e tácito. Por exemplo, ele pode ser casos de históricos de pacientes no sentido médico, detalhes de empréstimos bancários, ou descrição de situações de erros de equipamentos. Cada um desses registros de casos compreende:

- Uma descrição;
- O respectivo resultado ou solução.

Assim, um caso tipicamente compreende um par problema e solução. Uma coleção de casos, é chamado de uma base de casos, justamente como uma base de registros é chamado de banco de dados (Watson, 2003).

Não há um consenso por parte da comunidade de RBC, que informações exatamente, poderia ser um Caso. Embora, duas medidas pragmáticas poderiam ser tomadas para se decidir o que poderia ser representada em Casos: a funcionalidade da informação e a facilidade de aquisição da informação (Watson, 1999).

Segundo Watson (2003), as bases de casos dividem-se em duas grandes categorias:

- **Em bases de Casos homogêneas:** todos os casos compartilham os mesmos dados ou estrutura de registros; isto é, casos têm os mesmos atributos, mas variando os valores.
- **Em Bases de Casos Heterogêneas:** casos têm estrutura de registros variados; isto é, casos podem ter diferentes atributos e valores variados.

2.6.4. Indexação

Muitos sistemas de Banco de Dados utilizam-se de índices para agilizar a recuperação de dados. Um índice é computacionalmente, uma estrutura de dados que pode ser realizada em memória, tornando a localização da informação, muito rápida, sem ter que fazer a busca dos registros no disco. O RBC também faz uso de índice para agilizar a recuperação de Casos. Como diretrizes, os índices devem: [Watson, 1999].

- Ser preditivo.
- Indicar o propósito em que o Caso será usado.
- Ser abstrato o suficiente para permitir ampliar a base de casos e seu uso no futuro.
- Ser concreto o suficiente para ser reconhecido em futuras situações.

A escolha do índice, tanto pode ser manual com automatizada. Escolher um índice manualmente, envolve decidir o propósito do Caso em respeito ao objetivo do sistema e decidir em que circunstâncias o Caso vai ser útil (Watson, 1999).

Existem um número crescente de métodos de indexação automática na literatura, incluindo: MEDIATOR, CHEF e CYRUS, etc.

2.6.5. Recuperação

Dada uma descrição de um problema, um algoritmo de recuperação deveria encontrar os casos mais similares à situação atual, utilizando-se dos índices da memória de casos. Os algoritmos baseiam-se nos índices e na organização de memória para guiar a busca dos casos potencialmente úteis.

De acordo com Watson (2003), a recuperação de casos, está diretamente relacionado e dependente ao método de indexação usado. Em geral, duas técnicas são correntemente usadas pelas ferramentas de CBR comerciais: algoritmo de vizinhança (**Nearest-Neighbor**) e indutivo.

2.6.5.1. Algoritmo da Vizinhança

Esse método, segundo Watson (2003), baseia-se na comparação entre um novo caso e aqueles armazenados na base de casos, utilizando uma soma ponderada das suas características. Para isso é necessário atribuir um peso a cada uma das características que descrevem o caso e que serão utilizadas na recuperação.

Na prática, a similaridade (isto é, a proximidade) do caso destino para o caso fonte para cada atributo é determinado. Esta medida é multiplicada por um fator peso. Então a soma da similaridade de todos os atributos é calculada. Esta pode ser representada por uma equação relativamente simples

$$\text{Similarity}(T, S) = \sum_{i=1}^n f(T_i, S_i) \times w_i \quad (\text{Equação 2.6})$$

Onde,

T é o caso destino

S é o caso fonte

n é o número de atributos em cada caso

i é um atributo individual de 1 até n

f é a função de similaridade para atributo i nos casos T e S

w é o peso do atributo i

Algoritmos de similares a este são usados por muitas ferramentas RBC para realizar recuperação do caso mais similar. Similaridade são normalmente para cair dentro da faixa de 0 para 1 (onde 0 significa totalmente dissimilar e 1

exatamente similar) ou usando um percentual, onde 100% é totalmente similar (Watson, 1999; 2003).

2.6.5.2. Algoritmo de Indução

Indução é uma técnica desenvolvida por pesquisadores de Aprendizado de Máquinas para extrair regras ou construir de dados passados. Em sistema RBC, a base de casos é analisada por algoritmo de indução para produzir uma árvore de decisão que classifica (ou indexa) os casos. O algoritmo de indução foi amplamente usado pela ferramenta RBC chamada **ID3**.

2.6.6. Adaptação

A tarefa final do Sistema CBR é adaptar a solução associada a um caso recuperado para as necessidades do problema corrente. Quando uma situação é fornecida, o algoritmo de recuperação traz o melhor caso que ele encontrar para a memória. Normalmente, o caso selecionado não atende perfeitamente com a descrição do problema do usuário. Ou seja, existem diferenças entre o problema do usuário e o caso contido no banco de casos que devem ser levadas em conta. Então, o processo de adaptação procura por diferenças salientes entre as duas descrições e aplica regras de forma a compensá-las. Em geral, existem dois tipos de adaptação em CBR (Watson, 2003):

- **Adaptação Estrutural** - as regras de adaptação são aplicadas sobre a solução armazenada junto aos casos.
- **Adaptação Derivacional** – o algoritmo reusa os algoritmos, métodos ou regras que geraram a solução que consta no banco de casos para gerar uma nova solução para o problema corrente. Neste método, a sequência que construiu a solução original deve ser armazenada juntamente com o caso na memória de casos. O algoritmo de adaptação derivacional exige uma perfeita compreensão dos casos armazenados e da forma como as soluções foram geradas.

Segundo Watson (2003), várias técnicas têm sido usadas em sistema CBR. Incluindo as seguintes:

- **Adaptação nula** – ele simplesmente aplica a solução recuperada ao problema corrente sem modificação. Adaptação nula é útil para problemas envolvendo raciocínio complexo mais com solução simples.

Por exemplo, em um sistema para concessão de crédito, embora seja necessário coletar muitas informações do cliente, a solução final de conceder ou rejeitar o crédito é direta.

- **Ajuste por parâmetros** – é uma técnica de adaptação estrutural que compara parâmetros específicos entre o caso recuperado e o novo para modificar a solução armazenada na direção apropriada. Esta técnica foi usada no sistema RBC chamado JUDGE, que recomenda sentenças mais curtas para crimes menos violentos.
- **Reinstanciação** – instancia uma nova solução para um caso recuperado do banco de casos com novas características adequadas ao problema do usuário. Por exemplo, o sistema RBC CHEF, que a partir de uma receita existente, criar uma nova receita.
- **Substituição derivacional** – repete o método, ou parte do método que gerou uma solução armazenada em um caso similar de forma a obter a solução para o novo caso, substituindo os atributos distintos. Como no sistema BOGART que reaplica os planos de geração de projetos para novos problemas.
- **Repara guiado por modelos** – utiliza um modelo casual para adaptar as soluções armazenadas aos problemas do usuário. O sistema RBC, CELIA, utiliza-o para aprendizado e diagnóstico de problemas mecânicos de automóveis.

De acordo com Watson (2003), a adaptação é útil em muitas situações. Mais não significa que seja essencial. Muitos dos sistemas RBC comerciais não usam adaptação para tudo. Eles simplesmente reusam a solução sugerida, para o melhor caso correspondente (i.e., adaptação nula) ou eles deixam a adaptação para as pessoas.

Pode-se concluir que, **O Raciocínio Baseado em Casos** é um método em que problemas novos são resolvidos através de soluções adaptadas que foram usadas para resolver problemas mais antigos.

Um **Caso** é um pedaço contextualizado de conhecimento que representa uma experiência. Ao se analisar cada caso, se tem a descrição do problema e a solução armazenada. Caso já exista um problema semelhante já anteriormente armazenado no banco de dados, a solução será recuperada. Porém, se não existir um caso similar, a descrição desse novo problema, será enviado ao espaço de problemas, recuperando o caso com o problema mais similar possível, criando uma nova solução (Watson, 1997).

2.7. Big Data

Big Data é um termo que vem chamando a atenção pela acelerada escalada em que, volumes cada vez maiores de dados são criados pela sociedade. Fala-se comumente em petabytes de dados gerados a cada dia, e zetabytes começa a ser uma escala real e não mais futurista. A uma década atrás, terabytes de dados, era uma quantidade futurista, agora temos em nossos próprios computadores. Muito tem sido escrito sobre Big Data e como ele pode servir como base para a inovação, diferenciação e crescimento da análise de dados em grandes massas de dados (Kolb, 2013).

De acordo com Raj (2013), as tecnologias que sustentam o Big Data, podem ser analisadas sob duas óticas: as envolvidas com análise de dados, tendo Hadoop e Map-Reduce como as principais e as tecnologias de infraestrutura, que armazenam e processam os dados. Neste aspecto, destacam-se os bancos de dados NoSQL (Not Only SQL).

O termo Big Data está diretamente ligado a questões como volume, variedade, velocidade, complexidade e valor (Mayer-Schönberger, 2013):

- **Volume** – Organizações coletam dados de uma grande variedade de fontes, incluindo transações comerciais, redes sociais e informações de sensores ou dados transmitidos de máquina a máquina.
- **Variedade** - Os dados são gerados em todos os tipos de formatos - de dados estruturados, não estruturados (tais como documentos de texto não estruturados, e-mail, vídeo, áudio), dados de cotações da bolsa e transações financeiras.
- **Velocidade.** Os dados fluem em uma velocidade sem precedentes e devem ser tratados em tempo hábil. Etiquetas de RFID, sensores, celulares e contadores inteligentes estão impulsionando a necessidade de lidar com imensas quantidades de dados em tempo real, ou quase real.
- **Veracidade.** É necessário que haja processos que garantam o máximo possível a consistência dos dados.
- **Complexidade.** Os dados surgem de várias fontes, o que torna difícil estabelecer uma relação, corresponder, limpar e transformar dados entre diferentes sistemas. No entanto, para que seus dados não saiam

rapidamente de controle, é necessário ligar e correlacionar relações, hierarquias e as várias ligações de dados.

- **Valor:** uma solução de Big Data, se mostrará viável se o resultado trouxer benefícios significativos e que compensem o investimento.

2.7.1. Uso do Big Data

Os modelos relacionais, quando proposto por Edgar F. Codd, atenderam muito bem, a demanda era acessar dados estruturados, de acordo com (Elmasri & Navathe (2005), gerados pelos sistemas internos das corporações. Estes modelos não foram desenhados para tratar dados não estruturados e nem para volumes de dados na casa dos petabytes de dados.

Para tratar dados na escala de volume, variedade e velocidade do Big Data precisamos de outros modelos. Surgem os softwares de banco de dados NoSQL, desenhados para tratar imensos volumes de dados estruturados e não estruturados. Existem diversos modelos como sistemas colunares como o *Big Table* (De uso interno pelo Google), o modelo **Key/value** como *DynamoDB da Amazon*, o modelo “*document database*” baseado no conceito proposto pelo Lotus Notes da IBM e aplicado em softwares como MongoDB, e o modelo baseado em grafos como o *Neo4j*, e assim por diante (Kolb, 2013).

Aplicações modernas de mineração de dados, frequentemente chamada “*Big-Data Analytics*”, exigir-nos gerenciar grande quantidade de dados rapidamente e em muitas dessas aplicações, exige-se um amplo paralelismo (Kolb, 2013).

Para lidar com tais aplicações, novos tipos de software têm surgido. Estes sistemas de programação são projetados para obter o máximo do paralelismo. O novo tipo de software começa com uma nova forma de sistema de arquivos, chamada "Sistema de arquivos distribuídos", que contam com unidades muito maiores do que os blocos de disco dos sistemas operacionais convencionais. Além do mais, os sistemas distribuídos também fornecem replicação de dados ou redundância para proteger os dados, contra falhas frequentes de mídias, que ocorrem quando o dado é distribuído para milhões de nós de computadores (Kolb, 2013).

No topo destes sistemas de arquivos, diversos sistema de alto nível de programação foram desenvolvidos. No centro do novo software está o sistema de programação chamada ***Map-Reduce***. Implementações de ***Map-Reduce*** permite que os cálculos sob os dados em grande escala, sejam executados em clusters de computação de forma eficiente e tolerante a falhas de hardware (kolb, 2013).

2.7.2. Map-Reduce

Map-reduce não é um produto ou um software específico, mas sim uma tecnologia desenvolvida pelo Google para lidar com grande quantidade de dados (Kolb, 2013).

A ideia básica é que os dados que precisam ser processados, entram no sistema, e é cortado em pedaços chamados de *chunks*. Essas peças de software, são responsáveis em fazer esses cortes é chamado de “***Mapper***”. Os chunks, são então enviados para outras peças de software para fazer o processamento requerido sobre eles, e então eles são ainda enviados para outra peça de software chamado “***Reducers***” que combina o resultado final para a saída.

O importante que, a tecnologia de Map-Reduce é capaz de pegar uma grande quantidade de dados, que seria muito dispendioso rodar em apenas um servidor, e poder distribui-lo por vários servidores. Este é um novo paradigma de programação. Existem algumas ferramentas que implementam esse novo paradigma, entre elas cito, o **Hadoop** do *Apache Foundation* (Kolb, 2013).

2.7.2.1. A Tarefa de Mapeamento

De acordo com Kolb (2013), o arquivo de entrada para uma tarefa Mapeamento, consiste de elementos, que podem ser de qualquer tipo: uma tupla ou um documento, por exemplo. Um chunk é uma coleção de elementos, e nenhum elemento é armazenado em dois chunks. Tecnicamente, todas as entradas para as tarefas de mapeamento (*The Map Tasks*) e saídas para as tarefas Redução (*The Reduce Tasks*) são os pares na forma chave-valor (*key-value*), geradas por uma função hash. Essa forma de entradas e saídas são

motivadas pelo desejo de permitir a composição de vários processos Map-Reduce (Kolb, 2013).

A função de mapeamento (Map) recebe um elemento com seus argumentos e produz zero ou mais pares chave-valor. Os tipos de chaves e valores são arbitrários. Mais, as chaves não são "chaves" no sentido usual; elas não precisam ser únicas. Mais uma tarefa de mapeamento pode produzir vários pares chave-valor com a mesma chave, mesmo a partir do mesmo elemento (Kolb, 2013).

Exemplo 2.5.1: Suponha que se deseje contar o número de ocorrências para cada palavra em uma coleção de documentos. Neste exemplo, o arquivo de entrada é um repositório de documentos, e cada documento é um elemento. A função de mapeamento para este exemplo usa chaves que são do tipo String (a palavra) e valores que são inteiros. A tarefa de mapeamento lê um documento e quebra ele em uma sequência de palavras $w_1, w_2, w_3, \dots, w_n$. Ela então emite uma sequência de pares de chave-valor onde o valor é sempre 1. Isto é, a saída da tarefa de mapeamento para este documento é a sequência de pares chave-valor $(w_1, 1), (w_2, 1), \dots, (w_n, 1)$.

Note que uma simples tarefa de mapeamento irá processar muitos documentos - todos os documentos em um ou mais chunks. Assim, a saída produzida será mais do que a sequência para o documento sugerida acima. Note também que se uma palavra w aparece m vezes entre todos os documentos atribuídos a esse processo, então haverá m pares chave-valor $(w, 1)$ entre sua saída. Uma opção para resolver esse problema é usar agrupamento e agregação, que é combinar esses m pares em um simples par (w, m) , isso só é possível porque as tarefas Redução, aplica uma operação associativa e comutativa, para os valores.

2.7.2.2. Agrupamento e Agregação

O processo controlador mestre sabe quantas tarefas Reduce haverá, digamos r tarefas, pois o usuário normalmente informa ao sistema map-reduce quais são as r tarefas. Então o controlador mestre aplica uma função hash e produz uma tabela de chaves de números (códigos) de 0 até $r-1$. Cada chave produzida pela tarefa Map é um hash e seus pares chave-valor são colocados

em um arquivo local. Cada arquivo é destinado para uma das tarefas Reduce (Kolb, 2013).

Após todas as tarefas Map terem completadas com sucesso, o controlador mestre junta os arquivos de cada tarefa Map que são destinados para uma particular tarefa e alimenta o arquivo resultante com uma lista de pares chave-valor. Isto é, para chave k , a entrada para a tarefa Reduce que manipula a chave k é um par da forma $(k, [v_1, v_2, \dots, v_n])$, onde $(k, v_1), (k, v_2), \dots, (k, v_n)$ são todos pares chave-valor e k , vindo de todas as tarefas Map.

2.7.2.3. As Tarefas de Redução

Os argumentos da função Reduce é um par consistindo de uma chave e sua lista de valores associados. A saída da função Reduce é uma sequência de zero ou mais pares chave-valor. Esses pares chave-valor podem ser de tipo diferente daqueles enviados das tarefas map para as tarefas Reduce, mas normalmente elas são do mesmo tipo. Referimo-nos a aplicação da função Reduce que reduz para uma simples chave e seus valores associados de redutor (Kolb, 2013).

Uma tarefa reduce recebe uma ou mais chaves e sua lista de valores associados. Isto é, uma tarefa reduce executa um ou mais redutores. As saídas de todas as tarefas reduce são juntas em um simples arquivo. Redutores podem ser divididos em tarefas reduce menores e a função hash associa cada chave com um dos códigos da tabela hash (Kolb, 2013).

Exemplo 2.5.2: Vamos continuar com o exemplo conta palavras do Exemplo 4.1. A função Reduce simplesmente agrupa todos os valores. A saída de um redutor consiste da palavra e da soma. Isto é, a saída de todos as tarefas Reduce é uma sequência de pares (w, m) , onde w é uma palavra que aparece pelo menos uma vez entre todos os documentos e m é o total de ocorrências de w em todos os documentos.

2.7.2.4. Detalhes de Execução de Map-Reduce

A Figura 4 oferece um esboço de como processo, tarefas, e arquivos interagem. Aproveitando uma biblioteca fornecida por um sistema map-reduce tal como Hadoop, o programa do usuário bifurca o processo controlador mestre e alguns dos processos Worker para diferentes nós de computação. Normalmente, um Worker manipula suas tarefas Map (um Map worker) ou tarefas Reduce (um Reduce worker), mas não ambos (Kolb, 2013).

O mestre tem muitas responsabilidades. Uma é criar um certo número de tarefas Map e algumas tarefas Reduce, este número sendo selecionado pelo programa do usuário. Estas tarefas serão atribuídas para o processo Worker pelo Mestre. É razoável criar uma tarefa Map para cada chunk de arquivo de entrada, mas pode-se desejar criar poucas tarefas Reduce. A razão para limitar o número de tarefas Reduce é que é necessário para cada tarefa Map criar um arquivo intermediário para cada tarefa Reduce, e se existe muitas tarefas Reduce o número de arquivos intermediários aumenta bastante (Kolb, 2013).

O Mestre (Master) se matem informado do estado de cada tarefa Map e Reduce (ocioso, executando um particular worker, ou concluído). Um processo Worker relata para o Mestre quando ele termina uma tarefa, e uma nova tarefa é agendada pelo Mestre para esse processo Worker (Kolb, 2013).

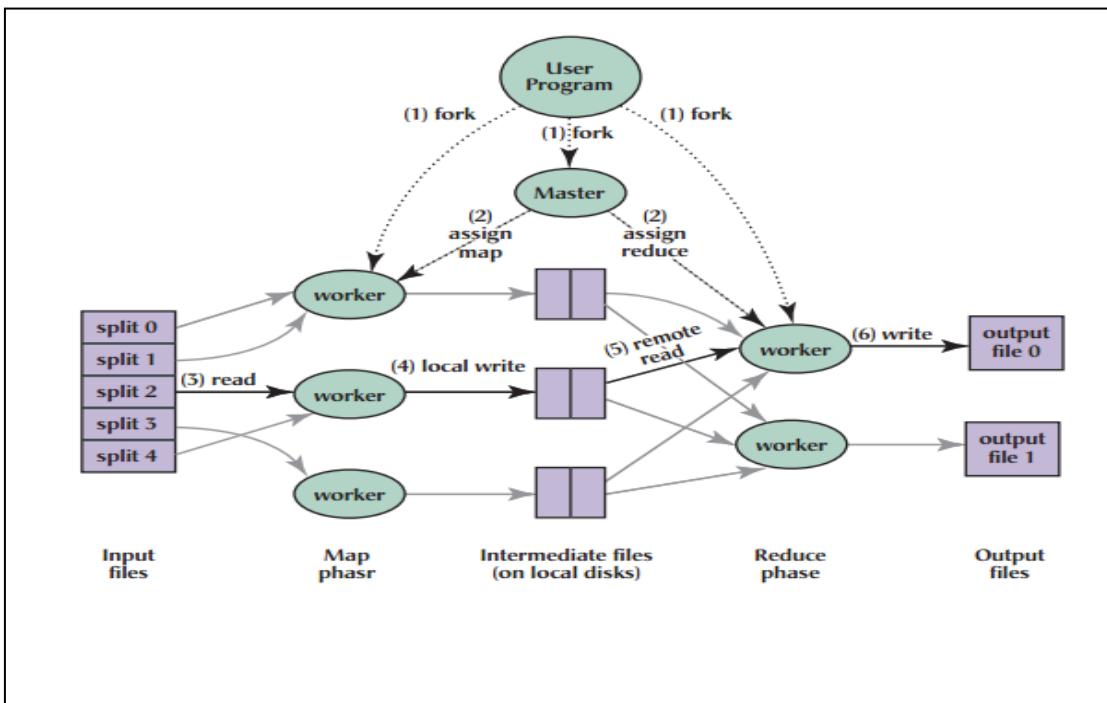


Figura 4. Esboço de Iteração de um Processo Map-Reduce.

Fonte: Dean & Ghemawat (2008)

2.7.3. Lidando com falhas nos nós

A pior coisa que pode acontecer é quando o nó de computação que está executando o Mestre falha. Neste caso, toda carga de entrada do map-reduce deve ser reiniciado. Mas somente este nó pode derrubar um processo inteiro; outras falhas iram ser gerenciadas pelo Mestre, e o trabalho do map-reduce irá completar eventualmente (Kolb, 2013).

Suponha que o nó de computação no qual reside o Map worker falha. Esta falha irá ser detectada pelo Mestre, porque ele periodicamente pings o processo Worker. Todas as tarefas Map que foram atribuídas para este Worker terão que ser refeitas, mesmo que tivessem concluído. A razão para refazer completamente as tarefas Map é que sua saída destinada as tarefas Reduce residem no nó de computação, e devido falha, está indisponível para as tarefas Reduce. O Mestre configura o estado de cada uma das tarefas Map para ociosa e reprograma-o para um Worker quando se tornar disponível. O Mestre também informa cada tarefa Reduce que a localização de suas entradas, que são as tarefas Map, mudaram (Kolb, 2013).

Lidar com uma falha para um Reduce Worker é simples. O Mestre, simplesmente, configura os estados de suas correntes tarefas Reduce em execução, para ociosa. Estas serão agendadas para outro reduce worker mais tarde (Kolb, 2013).

Concluindo, Map-Reduce é um modelo de programação, e framework introduzido pelo Google para suportar computações paralelas em grandes coleções de dados em clusters de computadores. Agora Map-Reduce é considerado um novo modelo computacional distribuído, inspirado pelas funções map e reduce usadas comumente em programação funcional. Map-Reduce é um “Data-Oriented” que processa dados em duas fases primárias: Map e Reduce. A filosofia por trás do Map-Reduce é: Diferentemente de data-stores centrais, como um banco de dados, você não pode assumir que todos os dados residem em um lugar central, portanto você não pode executar uma query e esperar obter os resultados em uma operação síncrona. Em vez disso, você precisa executar a query em cada fonte de dados simultaneamente. O processo de mapear a requisição do originador para o data source é chamado de ‘Map’, e o processo de agregação do resultado em um resultado consolidado é chamado de ‘Reduce’.

Hoje existem diversas implementações de Map-Reduce, como: Hadoop, Disco, Skynet, FileMap e Greenplum. Hadoop é a implementação mais famosa.

A tecnologia de big data não apenas suporta a habilidade de coletar grandes volumes de dados como também provê a habilidade de compreendê-los e tirar proveito de seu valor.

2.8. **Gamificação**

Para conceituar *Gamificação* é necessário antes de tudo a compreensão de sua origem. Podemos começar com o pensamento de que os games são uma superação voluntária de obstáculos desnecessários. Tomando como base essa simples reflexão, podemos avançar para o conceito apresentado pelo professor Kevin Werbach na sua formulação do “*Gamification*” oferecidas gratuitamente pela Coursera.

“O game é uma atividade ou ocupação voluntária exercida dentro de certos limites de tempo e espaço seguindo regras livremente consentidas, mas absolutamente obrigatórias, dotadas de um fim

em si mesmo e acompanhada de um sentimento de tensão, de alegria e da consciência de ser diferente da vida cotidiana.”.

O nosso objetivo é compreender os *games* como base conceitual para o *Gamification* aplicado à aprendizagem, então, faz-se necessário buscarmos uma definição que melhor fundamente nosso trabalho.

2.8.1. Games

A definição apresentada por Kapp (2012), nos oferece uma perspectiva melhor que servirá de base para a análise que vai nos conduzir ao *Gamification*.

“Um game é um sistema no qual jogadores se engajam em um desafio abstrato, definido por regras, interatividade e feedback; e que gera um resultado quantificável frequentemente e licitando uma reação emocional.”.

A definição correlaciona objetivos alcançáveis e mensuráveis a um sistema definido por regras. Estabelece a premissa da interatividade e a presença do feedback essencial para o acompanhamento da evolução da aprendizagem [Alves, 2014].

2.8.2. Histórico do *Gamification*

O Gamification surgiu há muito tempo quando, no ano de 1912, a marca americana Cracker Jack, de biscoitos e snacks, começou a introduzir brinquedos surpresa em suas embalagens. Em 1980 surge o primeiro jogo on-line o “MUD1”, projetado pelo pesquisador britânico Richard Bartle.

Em 2002, a categoria “*Serious games*” ganha proporções com o surgimento do “*Serious games movement*”. Mas é em 2003 que o termo *Gamification* surge no formato que conhecemos hoje. O termo é atribuído a Nick Pelling, programador de computador e inventor nascido na Inglaterra, na década de 60 [Alves, 2014].

Em 2007, a Bunchball lança uma moderna plataforma de Gamification que é a primeira a incorporar a mecânica de jogos com o uso de placar, pontos e distintivos para servir a propósitos de engajamento [Alves, 2014].

Segundo Alves (2014), foi no ano de 2010 que o Gamification se proliferou alcançando o mercado de massa na ocasião. No entanto, é no ano de 2011 que o conceito começa a amadurecer e surgem relatórios e estatísticas sobre o assunto que hoje, comprovadamente, agrega valor à categorias de negócios e aprendizagem diversificadas.

2.8.3. Definição

Gabe Zichermann define o Gamification como:

“Gamification consiste no processo de utilização de pensamento de jogos e mecânica de jogos para engajar audiências e resolver problemas”.

Já a autora Amy Jo Kim, define Gamification como:

“A utilização de técnicas de games para tornar atividades mais divertidas e engajadoras.”.

Dessa forma, das duas definições acima podemos extrair o mesmo princípio, ambas consideram que Gamification consiste no uso de elementos de jogos e técnicas de design de jogos em contextos diferentes de jogos [Alves, 2014].

3 Gabe Zichermann é autor de “Game-based marketing, Gamification by Design e The Gamification Revolution”.
4 Amy Jo Kim autora de “Community Building on the Web”.

Resumindo, Gamification não é a transformação de qualquer atividade em um game. Gamification é aprender a partir dos games, encontrar elementos dos games que podem melhorar uma experiência sem desprezar o mundo real. Encontrar o conceito central de uma experiência e torná-la mais divertida e engajadora [Alves, 2014].

2.8.4. Os elementos do game e do Gamification

Os elementos dos games são a caixa de ferramentas que é utilizada para criar a solução de aprendizagem gamificada. O professor Kevin Werbach, define os elementos de games como:

“Elementos são padrões regulares que podem ser combinados de diferentes maneiras para que você construa um jogo”.

Quando estamos pensando em um *Gamification* em busca de aprendizagem, estamos em busca de experiências que sejam engajadoras e que mantenham os jogadores focados em sua essência para aprenderem algo que impacte positivamente a sua performance [Alves, 2014].

Kevin Werbach produziu um modelo para definir os elementos e ele mesmo pontua que o modelo não cobre todos os elementos possíveis, mas mostra os mais comuns. A Figura 5 mostra o modelo de Kevin Werbach.

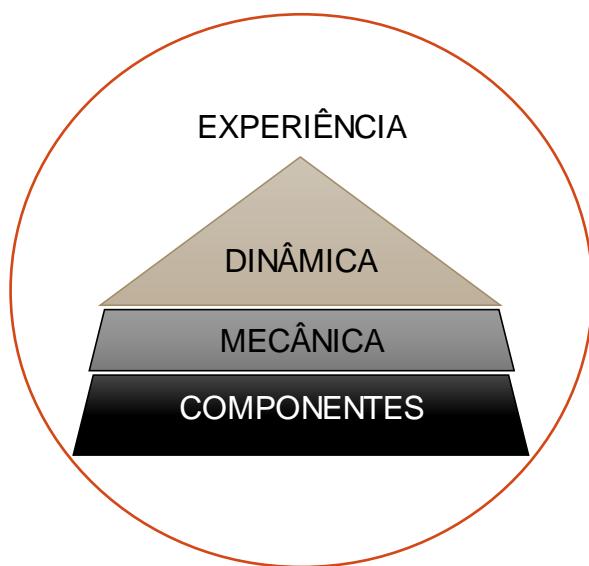


Figura 5. Adaptada da formação Coursera em Gamification do professor Kevin Werbach. Fonte: Autor

A Figura 2.5, representada por uma pirâmide, tem em sua parte inferior o que o autor chama de componentes, no meio se encontra a mecânica e no topo a dinâmica. O entorno da pirâmide representa a experiência que buscamos promover por meio de nosso sistema gamificado.

A dinâmica que está no topo da pirâmide, é constituída por elementos responsáveis por atribuir coerência e padrões regulares à experiência. Eles não são as regras, são a estrutura implícita e as regras podem estar em sua superfície [Alves, 2014]. Entre esses elementos estão:

- **Constricções:** responsáveis por restringir o alcance do objetivo pelo caminho mais óbvio e assim incentivar o pensamento criativo e

estratégico. São as constrições que criam no jogo escolhas que o jogador considera significativas.

- **Emoções:** um game pode provocar uma grande diversidade de emoções, desde alegria até tristeza. Com o Gamification não acontece o mesmo, pois de alguma forma estamos conectados à realidade, uma vez que, nosso objetivo é promover a aprendizagem, mas mesmo assim a emoção de alcançar um objetivo, ser motivado por feedback e recompensado pelo alcance de um resultado são essenciais.
- **Narrativa (*Storytelling*):** é a estrutura que de alguma forma une os elementos do sistema gamificado e faz com que haja um sentimento de coerência, um sentimento de todo. A narrativa pode ser explícita, e nesse caso é *storytelling*, mas diferente do contexto dos games não é necessário que haja uma história. O essencial é que a narrativa do sistema gamificado permita aos jogadores estabelecer uma correlação com o seu contexto, criando conexão e sentido para que o sistema não se torne um amontoado de elementos abstratos.
- **Progressão:** se refere ao oferecimento de mecanismos para que o jogador sinta que está progredindo de um ponto a outro, para que ele de alguma forma verifique que vale a pena progredir.
- **Relacionamento:** pessoas interagindo, amigos, colegas de time, oponentes, são os elementos da dinâmica social que são também essenciais para o ambiente do game.

No meio da pirâmide está a mecânica de games. Nesse nível estão os elementos que podem ser considerados “os verbos” pois são eles que promovem a ação, que movimenta as coisas adiante. Existe inúmeros mecanismos que podemos utilizar para movimentar um sistema gamificado e entre eles estão [Alves, 2014]:

- **Desafios:** são os objetivos que são propostos para os jogadores alcançarem durante o jogo. São eles que mobilizam o jogador a buscar o estado da vitória.
- **Sorte:** a possibilidade de envolver algum elemento no sistema gamificado que dê ao jogador a sensação que há alguma aleatoriedade ou sorte envolvida.

- **Cooperação e competição:** mesmos sendo opostas, ambas promovem no jogador o desejo de estar com outras pessoas engajadas em uma mesma atividade, seja competindo, ou construindo algo juntos em seus resultados, alcançando o estado de vitória.
- **Feedback:** o feedback é fundamental pois ele faz com que o jogador perceba que o objetivo proposto é alcançável e consiga acompanhar o seu progresso escolhendo estratégias diferentes quando aplicável.
- **Aquisição de recursos:** são elemento que o jogador deve adquirir ao longo do jogo para que consiga algo maior.
- **Recompensas:** são os benefícios que o jogador conquista e que podem ser representados por distintivos, vidas e direitos a jogar novamente.
- **Transações:** as mais comuns encontradas são as transações de compra, venda e troca. As transações podem ser utilizadas como mecanismos para a movimentação para uma fase seguinte de maior complexidade.
- **Turnos:** é a simples existência de jogadas alternadas entre um jogador e outro, presente até em games simples como o “jogo da velha”.
- **Estados de vitória:** indica que foi o vitorioso e pode ser representado de diversas formas.

Na base da pirâmide estão os componentes do jogo. São formas específicas de fazer o que a dinâmica e mecânica representam, complementando a analogia com um determinado idioma. E entre eles estão [Alves, 2014]:

- **Realizações:** diferente dos desafios, são os mecanismos de recompensar o jogador por cumprir um desafio.
- **Avatares:** mostram ao jogador alguma representação visual de seu personagem ou papel no sistema gamificado.
- **Badges:** são as representações visuais das realizações ou resultados alcançados.

- **Boss Fights:** consiste em um desafio grande como travar uma batalha muito difícil para que o jogador consiga passar de uma fase ou nível a outro.
- **Coleções:** significa coletar ou colecionar coisas ao longo do game.
- **Combate:** trata-se de uma luta que deve ser travada.
- **Desbloqueio de conteúdo:** significa que o jogador precisa fazer algo para que possa ganhar acesso a um conteúdo do sistema gamificado.
- **Doar:** o altruísmo ou as doações compõe um mecanismo que pode ser muito interessante e que faz com que o jogador deseje permanecer no game.
- **Placar ou “leaderbord”:** consiste no ranqueamento dos jogadores, permitindo que o jogador veja sua posição em relação a seus oponentes.
- **Níveis:** são graus diferentes de dificuldades que vão sendo apresentados ao jogador no decorrer do sistema gamificado, de forma que ele desenvolva suas habilidades enquanto avança de um nível ao outro.
- **Pontos:** dizem respeito ao score, à contagem de pontos acumulados no decorrer do game ou sistema gamificado.
- **Investigação ou exploração:** é o alcance de resultados implícitos no contexto do game ou sistema gamificado, que implica em buscar algo, fazer algo ou ainda explorar e investigar para alcançar um resultado.
- **Gráfico social:** consiste em fazer com que o game ou sistema gamificado seja uma extensão de seu círculo social a exemplo do Foursquare.
- **Bens virtuais:** são coisas virtuais pelas quais os jogadores estão dispostos a pagar com moeda virtual ou até real, como por exemplo um conjunto de cores diferentes para utilizar em desenhos durante o game ou sistema gamificado.

Resumindo no Gamification, a mecânica, a estética e o pensamento de games trabalham juntos para que o sistema gamificado funcione. Cabe ressaltar

a importância da narrativa, ou seja, do storytelling presente no sistema gamificado, pois sem uma história que crie significado para o jogador, a credibilidade do sistema fica prejudicada e a motivação para o engajamento no sistema deixa de existir porque perde a relevância [Alves, 2014].

2.9. Bases Conceituais sobre a Evasão e Retenção Escolar

A Constituição Federal (**CF**) de 1988, em seu art. 6º, define a educação como um direito social, ao lado de outros, como: saúde, trabalho, moradia, lazer, segurança, previdência social, proteção à maternidade e à infância, assistência aos desamparados. Como dever do Estado e da família, o direito à educação deve consolidar-se na promoção do pleno desenvolvimento da pessoa, no preparo para o exercício da cidadania e na qualificação para o trabalho [BRASIL, 1988, art. 205].

O direito à educação pode ser considerado como um dos alicerces da República Federativa do Brasil na medida em que é instrumento necessário

Á construção de uma sociedade livre, justa e solidária; a garantia do desenvolvimento nacional; à erradicação da pobreza e da marginalização, com a redução das desigualdades sociais e regionais; e à promoção do bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação [Garcia, 2004, s.p.].

Portanto, à educação, é parte da matriz que constitui em larga escala o respeito à dignidade humana. Esse preceito vem sendo expresso em inúmeros documentos, tratados, acordos nacionais e internacionais na **LDB** – Lei de Diretrizes e Bases [MEC, 2014].

Assim, à educação é considerada como [MEC, 2014]:

- **Direito social e dever do Estado**, para corresponder às aspirações da sociedade por um país democrático, justo e isonômico, traduzindo-se em ações que visem dar respostas públicas aos compromissos socialmente assumidos em cada uma das instituições, tanto fortalecendo o processo de inserção cidadã como contribuindo para o desenvolvimento pessoal e

profissional dos sujeitos e para o desenvolvimento local, regional e nacional do país;

- **Direito de cidadania**, para formar pessoas críticas, autônomas, emancipadas e competentes tecnicamente, ativas na dinâmica do convívio social e partícipes na definição dos projetos de desenvolvimento nos âmbitos público e privado, pessoal e coletivo;
- **Bem público**, na perspectiva da inclusão e valorização da educação profissional e tecnológica como política pública, comprometendo-se o Estado com a qualidade social; e
- **Questão de soberania conjunta Estado-cidadão**, para cumprir a função social e os compromissos firmados a expansão do direito e a universalização do acesso.

Portanto, o conceito de educação para cidadania impõe-se como requisito político e pedagógico para que as instituições cumpram sua função social. No entanto, não basta admitir a educação como direito fundamental. É essencial concretizar e prover as ações que permitem a garantia desse direito. Nesse sentido, tanto a CF, em seu art. 206, quanto a **LDB** (Lei de Diretrizes e base), em seu art. 3º, indicam os seguintes princípios, com relação direta com o sucesso escolar, para que o processo educacional ocorra de forma efetiva: a igualdade de condição para o acesso e permanência na escola, a garantia do padrão de qualidade, a valorização do profissional da educação escolar e a vinculação entre a educação escolar, o trabalho e as práticas sociais [MEC, 2014].

2.9.1. Evasão e a Retenção Escolar

Dentre as questões conflitantes, que envolve a relação entre educação, instituições de ensino e sociedade, a retenção e a evasão merecem destaque. Entre todas as modalidades de ensino, da educação básica à educação superior, esses problemas estão presentes [MEC, 2014].

Apesar das pesquisas relativas à evasão não identificar um conceito homogêneo, a partir de 1970, autores como Tinto (1995), professor da Syracuse University, passaram a abordar o modelo de integração do estudante, destacando que a decisão de evadir-se é tomada em função da falta de

integração com o ambiente acadêmico e social da instituição, sendo esta integração influenciada pelas características individuais, pelas expectativas para a carreira ou curso e, por último, pelas intenções/objetivos e compromissos assumidos antes do início do curso [MEC, 2014].

O modelo desenvolvido por Tinto (1995) sugere seis conjuntos de variáveis:

- Os atributos de pré-entrada, entendidos como habilidades do aluno, escolaridade anterior e antecedentes familiares;
- Os comprometimentos iniciais ou metas traçadas pelo próprio estudante;
- A integração acadêmica, tida como o vínculo entre o estudante e a estrutura da instituição de ensino;
- A integração social entre os grupos de estudantes e docentes como variável;
- Os comprometimentos subsequentes ou influencias das dimensões acadêmicas e sociais da integração no vínculo com a instituição e na intenção de alcançar o objetivo de conclusão de curso; e
- Os aspectos externos.

Tinto (1995) finalmente, descreve os resultados, constituídos pela decisão, persistência ou deserção do curso ou sistema, como variável. Após o embasamento teórico de Tinto (1995) pode-se pensar em explicações sociológicas e políticas no estudo da evasão. Considerando que a evasão escolar, entendida como interrupção no ciclo de estudos, deve ser vista como um fenômeno complexo e não um problema comum, uma vez que compromete o efetivo do direito à educação de qualidade para todos [MEC, 2014].

Segundo Dore (2011), no Brasil a evasão pode se referir à retenção e repetência do aluno na escola; à saída do aluno da instituição, do sistema de ensino, da escola e posterior retorno; ou à não conclusão de um determinado nível de ensino. Portanto, para a pesquisadora a evasão ou abandono escolar é um processo que tem natureza multiforme.

A escolha de abandonar ou permanecer na escola é fortemente condicionada por características individuais, por fatores sociais e familiares, por características do sistema escolar e pelo grau de atração que outras modalidades de socialização, fora do ambiente escolar, exercem sobre o estudante [DORE, 2013, p.5].

2.9.2. Categorização das causas da Evasão e da Retenção Escolar

A classificação proposta por Brasil (1996), para categorizar as causas da evasão e da retenção, em função de um plano estratégico e monitoramento desses problemas, mapeou os seguintes fatores ou categorias motivadores, da evasão e da retenção, adaptados às especificidades da contemporaneidade e das próprias instituições de ensino da Rede Federal:

- **Fatores individuais** – são problemas relacionados a aspectos peculiares às características do estudante, tais como: adaptação à vida acadêmica, capacidade de aprendizagem e habilidade de estudo, escolha precoce da profissão, qualidade de formação escolar anterior, questões financeiras do estudante ou da família, descoberta de novos interesses ou novo processo de seleção, dentre outros;
- **Fatores internos às instituições** – são problemas relacionados à infraestrutura, ao currículo, a gestão administrativa e didático-pedagógica da instituição, bem como outros fatores que desmotivam e conduzem o aluno a evadir do curso. Aqui se destacam os fatores: atualização, estrutura e flexibilidade curricular, formação do professor, infraestrutura física, material, tecnológica e de pessoal para ensino, motivação do professor, questões didático-pedagógica e relação escola-família;
- **Fatores externos** – relacionam-se às dificuldades financeiras do estudante permanecer no curso e às questões inerentes à futura profissão. Dentre estes fatores destacam-se: avanços tecnológicos, econômicos e sociais, oportunidade de trabalho para egressos do curso, políticas governamentais para a educação profissional e tecnológica e para a educação superior, reconhecimento social do curso e valorização da profissão.

Dessa forma, o conceito de evasão adotado aproxima-se dos conceitos propostos por Brasil (1996) e Dore (2013), sendo definido como a interrupção do aluno no ciclo do curso. Em tal situação, o estudante pode ter abandonado o curso, não ter realizado a renovação da matrícula ou formalizado o desligamento do curso. Por outro lado, a retenção consiste da não conclusão do curso no período previsto, fator concorrente para o aumento da propensão em relação à evasão escolar [MEC, 2014].

2.9.3. Indicadores de Evasão, Retenção e Conclusão

Os conceitos de evasão e de retenção servem de base para a construção de indicadores que relacionam esses conceitos ao número de estudantes ingressantes e matriculados nas instituições, fornecendo subsídios para identificação de necessidades de ações específicas [MEC, 2014].

O Sistema Nacional de Informações da Educação Profissional e Tecnológica (**SISTEC**), tem o registro efetivo da vida do estudante ou de um conjunto de estudantes (ciclo de matrículas) na instituição, desde seu ingresso até sua saída, e as mudanças que ocorrem durante esse período. Isso permite o acompanhamento dos indicadores de conclusão, evasão e retenção dentro de um mesmo ciclo.

Portanto, o SISTEC, a partir das situações de matrícula, estabelece conceitos de total de abandono, retenção e conclusão que serão utilizados no cálculo dos indicadores de evasão, retenção e conclusão. Nesse sentido, estabelecem-se os seguintes conceitos [MEC, 2014]:

- **Total de matrículas ativas:** número de matrículas que permanecem ativas com situação “em curso” ou “integralizado”.
- **Total de retenção:** número de matrículas que permanecem ativas com situação “em curso” ou “integralizado” após a data para o término do ciclo de matrícula do curso.
- **Total de saída sem êxito:** número de matrículas finalizadas com situação “transferido interno”, “transferido externo”, “desligado/desistente” ou “evadido”.
- **Total de saída com êxito:** número de matrículas finalizadas com situação “concluído”.

Para o cálculo das taxas de evasão, retenção e conclusão, pode ser realizado considerando a amostra escolhida como sendo os estudantes matriculados no período em análise ou sobre os estudantes matriculados em um ciclo, a partir dos dados de matrículas ativas ou finalizadas.

Nesta seção foi dada uma visão geral sobre os conceitos relacionados a evasão e a retenção escola, bem como os índices indicadores de evasão, retenção e concussão. Sem, no entanto, ter a pretensão de esgotar o assunto, uma vez que o mesmo envolve muitas varáveis em seu entorno.

Para a pesquisa realizada nessa seção, foi tomado como base o “Documento orientador para a superação da evasão e retenção na Rede Federal”, realizado pelo Ministério da Educação em 2014.

2.10. Trabalhos Relacionados

Na tentativa de encontrar trabalhos semelhantes, foi feito uma consulta em alguns repositórios de dados científicos, dentre eles o Repositorium da UMinho, o Repositório-Aberto da Universidade do Porto, o Repositório Scopus, Web + Knowledge, ISI Web of Cience, dentre outros.

Um dos trabalhos semelhantes encontrado, foi a tese de doutoramento de **Janete Santos** em 2015, cujo título é “A evasão na Universidade Federal do Recôncavo da Bahia (**UFRB**): um estudo inicial” (identificador do trabalho: <http://hdl.handle.net/1822/39988>). Este trabalho teve como objetivo, analisar os dados do levantamento quantitativo da evasão dos estudantes naquela instituição, analisando os aspectos influenciadores da evasão a partir do estudo de outros autores. No entanto, como já frisado, a pesquisa aborda uma análise quantitativa, com base em dados coletados no período de 2010 a 2013 na universidade UFRB.

Um outro trabalho encontrado no repositorium, com características semelhantes, foi a pesquisa de **Andreia Patrícia Gomes** de 2014 (registro: <http://hdl.handle.net/1822/35243>), que trata da implementação de uma plataforma de BI, aplicado a área da Educação. Seu objetivo é fornecer aos utilizadores, indicadores escolares, para que assim se possam tomar medidas que possam influenciar todo processo de transição escolar.

A pesquisa investigativa de Alice Maria Gonçalves de 2007, sobre a orientação de Dr. Paulo Cortez (registro: <http://hdl.handle.net/1822/7966>), onde foram aplicadas diversas técnicas de Data Mining, como redes neurais, máquinas de vetores de suporte e árvore de decisão, com vista à extração de conhecimento relevante na previsão do desempenho escolar, tem alguma semelhança a pesquisa proposta neste trabalho. A solução proposta por Alice pretendia prever o desempenho escolar dos alunos, através de um conjunto de variáveis de entrada. O objetivo do trabalho dela é ajudar a comunidade educativa a melhorar a gestão de recursos e a aplicação de estratégias, por formar a melhoria do desempenho dos alunos. Para o desenvolvimento dessa investigação, ela realizou análise de uma base de dados. Segundo ela, a base de dados foi criada no ano letivo de 2005/2006, com os resultados obtidos através da elaboração e aplicabilidade de um questionário em duas escolas

alentejanas (Escola Secundária Gabriel Pereira em Évora e Escola Secundária Mousinho da Silveira em Portalegre).

De fato, se pode dizer que, os trabalhos citados têm alguma semelhança com a nossa pesquisa, onde de um modo ou de outro, ambos estão relacionados ao melhoramento do desempenho escolar. O primeiro trabalho, investiga as causas da evasão escolar, no aspecto quantitativo. O segundo trabalho, trata da implementação de uma plataforma de BI, para dar suporte a gestão escolar e o terceiro aplica técnicas de Data Mining para extrair conhecimento em uma base de dados.

O diferencial destes trabalhos apresentados e do trabalho proposto por nós, é que o nosso encobrar os três, ou seja, é uma plataforma de **BI**, para dar suporte a gestão educativa do **IFRN**, tem aplicação de algoritmos de Data Mining para extrair conhecimentos da base de dados acadêmica, sobre os indicadores da repetência e evasão escolar, tudo isso com a finalidade também de melhorar o desempenho escolar de nossa instituição. Além do mais, foi implementado um sistema de Raciocínio Baseados em Casos, para suportar a equipe pedagógica nos problemas dos alunos, tendo como finalidade fim, o melhoramento do desempenho escolar dos alunos e, com isso diminuir os índices de repetência e evasão escolar nos IFRNs.

Capítulo 3 – Desenvolvimento do Projeto

3. Business Intelligence – BI

Nesse capítulo serão desenvolvidos os módulos do **BI**, onde será implementado um sistema de Data Warehouse (**DW**), cuja finalidade é consolidar as informações necessárias e consistentes para a geração de relatórios precisos para a tomada de decisões por parte dos gestores educativos do IFRN, o módulo de Mineração de dados ou *Data Mining*, cuja finalidade é a de descobrir padrões de comportamentos na base de dados consolidada, afim de mapear os perfis dos alunos em relação e repetência e a evasão escolar, o módulo do Raciocínio Baseado em Casos (RBC) e, o módulo do Gamification. Além de fazer previsões futuras sobre os índices de repetência e de evasão escolar, através de análise preditivas, utilizando as técnicas e algoritmos da aprendizagem de máquina.

3.1. Visão do problema

O Instituto Federal de Educação, Ciência e tecnologia do Rio Grande do Norte (**IFRN**), ministra diversos cursos em diferentes níveis de ensino: médio, técnico, técnico-subsequente, tecnológico, PROEJA e cursos de pós-graduação, latu e estrito sensu. Com um corpo docente amplo e uma grande massa de alunos, como já foi falado, nos parágrafos anteriores. A movimentação diária, e constantes, do sistema acadêmico, gera uma grande massa de dados e de diversos tipos, como por exemplo, dados demográficos (idade, sexo, situação financeira dos pais, nível de escolaridade dos pais, região de procedência, escola de origem, dentre outros), e dados comportamentais (conclusão, repetência, cancelamento de matrícula, evasão, transferência e assim por diante) e por fim, os dados sobre avaliações (Notas e faltas). Essa grande massa de dados está armazenada em um sistema de banco de dados relacional e, em algumas situações em planilhas eletrônicas, como no caso das pesquisas semestrais que são realizadas no IFRN, pela gestão e pela equipe pedagógica, entre alunos e servidores, com a finalidade de se mapear os problemas relacionados a infraestrutura e ensino aprendizagem.

Nos últimos anos, têm-se observado, por parte dos educadores e gestores educacionais, um alto índice de repetência e evasão escolar, em alguns cursos e até mesmo, nota-se um certo desestímulo por parte dos alunos, em determinados cursos, ministrados nos diversos campus da instituição. É fato, por exemplo, que alguns cursos, as turmas de 3º e 4º períodos, chegam com média muita baixa de alunos, em torno de 20%, ou seja, essas turmas iniciaram com 40 alunos e, no 4º período estão com no máximo 8 a 10 alunos. Isso tem como consequência, vários professores, ministrando aulas para poucos alunos, diminuindo em muito, a relação aluno professor, um dos índices, que é utilizado pelo governo federal, para avaliar o Instituto.

Diante desse quadro, surgiu a necessidade de fazer uma análise mais aprofundada de tais problemas. Ou seja, deseja-se analisar o **índice de reprovação**, o **índice de evasão**, o **índice de aprovação**, o **índice de conclusão** e também o **índice de matrículas canceladas**. Tendo como finalidade, tentar mapear quais são as causas que implicam, diretamente ou indiretamente nesses índices, ou seja, traçar o perfil para esses problemas.

Para tanto, inicialmente será desenvolvido um Armazém de dados (em inglês **Data Warehouse - DW**), onde será consolidado todos os dados das diversas fontes de dados disponíveis no IFRN.

3.2. Arquitetura de desenvolvimento do BI

Para o desenvolvimento do BI proposto, será adotada a arquitetura do *Business Intelligence Semantic Model (BISM)*, por a mesma ser bastante flexível para consumir dados, tanto de bases de dados relacionais, quanto de bases de dados tabulares, tais como: planilhas eletrônicas, arquivos OData, arquivos simples e também de serviços na nuvem. Além do mais, o BISM disponibiliza ferramentas de interface intuitiva para manipular dados, criação de relatórios, Dashboards e indicadores de desempenho (em Inglês **KPI**).

Conceitualmente o modelo pode ser apresentado em três camadas: modelo de dados, camada lógica e de consulta e a camada de acesso a dados. A Figura 6 ilustra esse modelo.



Figura 6. As camadas conceituais do modelo BISM

Fonte: adaptado de Harinath.

O modelo de dados (**Data model**), tanto pode ser um modelo multidimensional (com dimensões, medidas e cubos), quanto um modelo tabular (com tabelas, colunas e relações). Isso na perspectiva do desenvolvedor da aplicação.

Para a camada lógica de negócio e consultas (Business logic and queries), o modelo BISM oferece a linguagem *Multi-Dimensional eXpressions (MDX)* para o modelo multidimensional e a linguagem **DAX**, um acrônimo para *Data Analysis Expressions*, para o modelo tabular.

Para a camada de acesso a dados (Data Access), o BISM dá suporte a todas as funcionalidades para ambos modelos multidimensionais **MOLAP** e **ROLAP**. Por outro lado, para o modelo tabular, o BISM usa um mecanismo de armazenamento de colunas em cache, organizando os dados em colunas, e utiliza um mecanismo de acesso a dados muito rápido (**Vertipaq**) que por manipular os dados em memória, dispensa a necessidade de indexação ou agregação, pelo fato do mesmo simplesmente realizar um escaneamento na memória.

Portanto, de um lado temos a complexidade do modelo multidimensional, que suporta a construção de *Data warehousing*, com dimensões, factos e cubos, onde os dados podem ser analisados através do **OLAP** e técnicas de mineração de dados e são consumidos pelos utilizadores finais (analistas de negócio ou tomadores de decisão). E por outro lado, temos o modelo Tabular composto de tabelas e relacionamentos e, dessa forma, mais simples e mais intuitivo de trabalhar.

Então, para enfatizar ou justificar a escolha pela arquitetura **BISM**, para o desenvolvimento deste trabalho, considere o seguinte cenário.

Cenário: tenho a necessidade de acessar e consumir dados de bases de dados relacionais e multidimensionais, modelos tabulares, planilhas eletrônicas, e de diversos outros tipos de arquivos. Além dos mais, preciso apresentar para os usuários finais, os resultados obtidos através de relatórios, **Dashboards**, **KPI's**, **planilhas ou tabelas** e assim por diante. Então, diante dessa perspectiva, apresento na Figura 7 a arquitetura BISM que será utilizada neste trabalho.

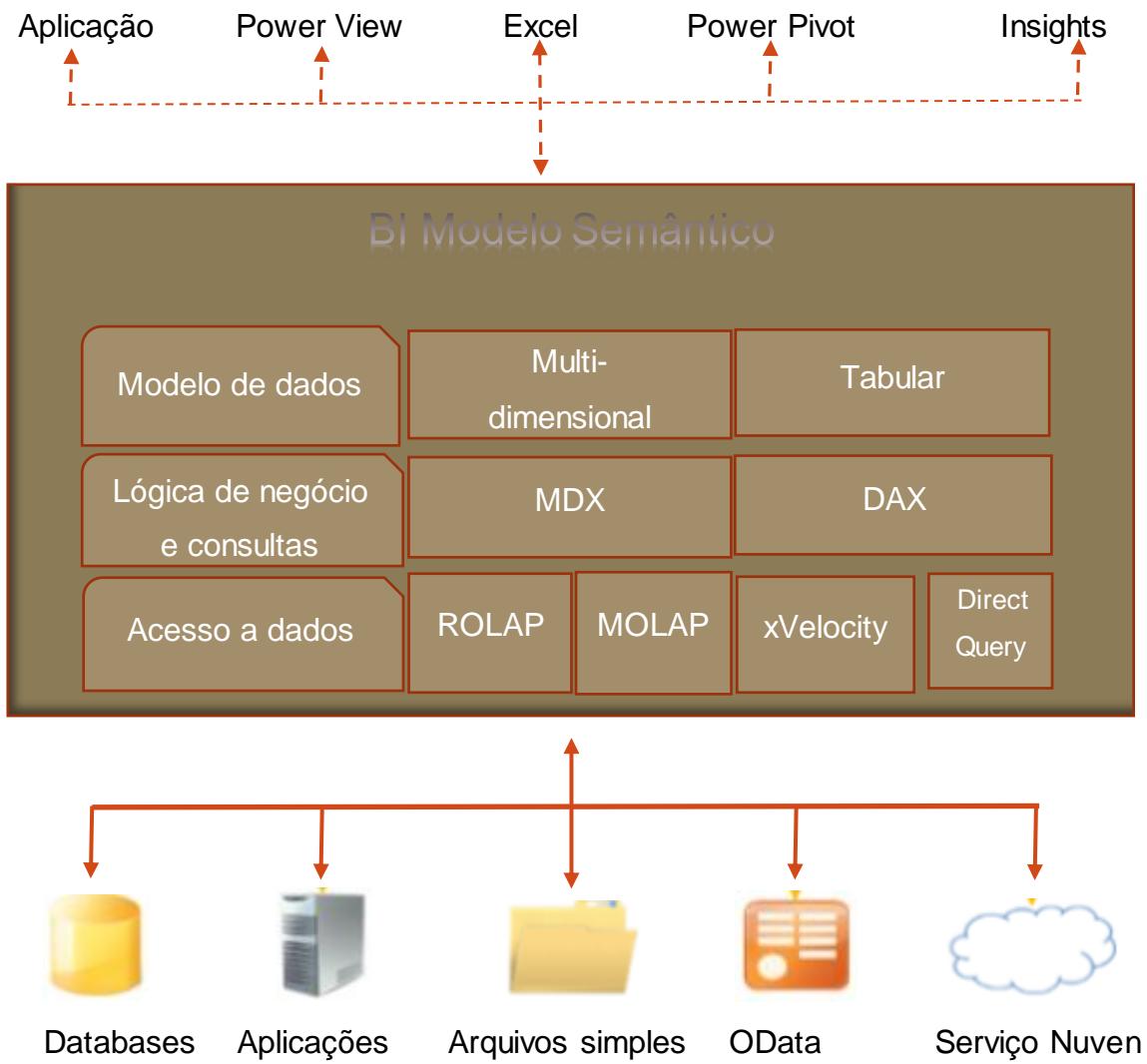


Figura 7. Arquitetura BISM adotada para o desenvolvimento do projeto.

Fonte: Adaptado de Harinath, 2012

A arquitetura **BISM** permite acessar diversas fontes de dados, desde bases de dados relacionais, arquivos simples ou mesmo dados na nuvem. Gerar modelos multidimensional ou tabular, manipular esses dados e disponibilizá-los aos utilizadores finais. Além do mais, esta arquitetura tem a flexibilidade, de

utilizar várias ferramentas na criação de relatório, tais como Excel, SQL Server Report Services, Power BI, PowerPivot, Power View, e assim por diante).

3.3. O Processo BI

O BI é um processo de converter dados em informações para que os envolvidos na tomada de decisões da empresa, possam fazê-las de forma rápida e precisa.

No contexto desse trabalho, foi desenvolvido primeiro o armazém de dados (**DW**). Então, foi realizado um estudo de viabilidade para a construção do mesmo. Em seguida, foi feita a extração, limpeza e carga dos dados das fontes de dados disponíveis para o armazém de dados. A Figura 8 mostra o processo da solução adotado.

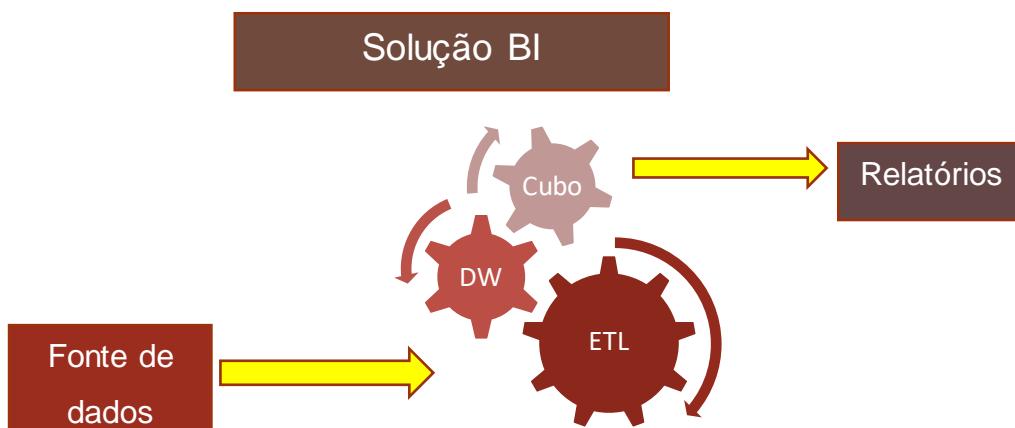


Figura 8. O processo BI da solução.

Fonte: Autor (adaptado de Savjani,2014)

3.3.1. Data Warehouse

Como sabemos, um DW pode ser implementado por assunto e, como o nosso sistema tem vários assuntos, então, cada assunto será implementado gerando um Data Marts, conseguintemente, o nosso DW é uma constelação de data marts.

3.3.2. Ciclo de vida do DW

Existem duas variantes principais para desenvolvimento de um sistema, onde os mais utilizados o modelo em cascata e o iterativo e incremental. Essas metodologias são também conhecidas com o ciclo de vida de desenvolvimento

do sistema, em inglês system development life cycle (SDLC) que, alguns autores chamam simplesmente de metodologia de desenvolvimento de sistema, em inglês system development methodology. Para esse projeto foi utilizado a metodologia em cascata, seguindo o diagrama da Figura 9 [Rainardi, 2008].

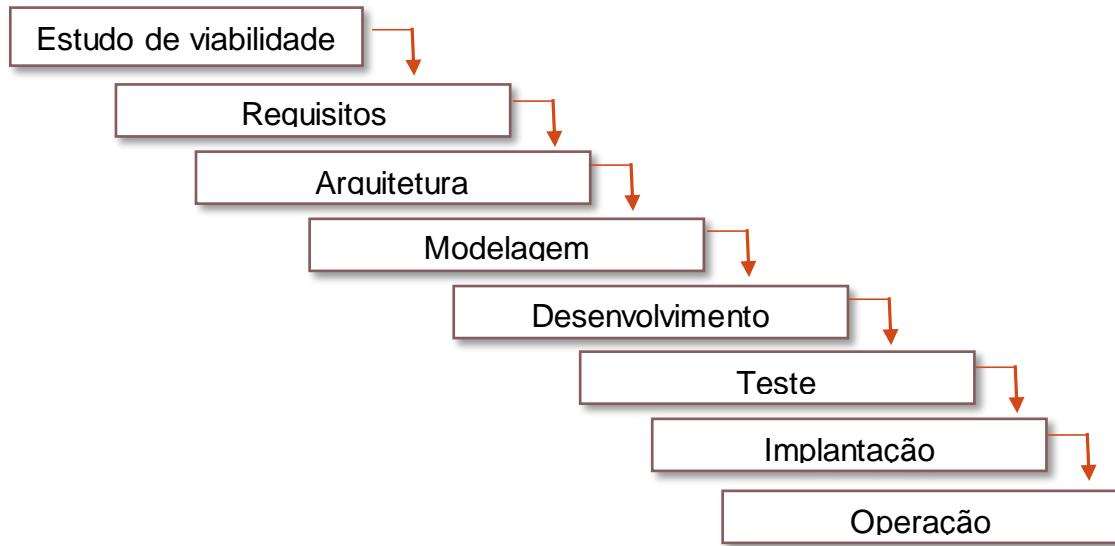


Figura 9. Metodologia em Cascata.

Fonte: Autor adaptado de Rainardi, 2008

3.3.2.1. Estudo de viabilidade

Foi feita uma análise em toda base de dados do sistema acadêmico e nas planilhas de dados que me foram disponibilizadas pela equipe pedagógica e outros setores do IFRN. Então, foram selecionados os dados que são de fundamental importância para esse sistema, descartando os dados irrelevantes. A Tabela 2, lista as fontes de dados consultadas.

Tabela 2. Bases de Dados do Instituto Natal Central – RN.

Fonte de dados	Modelo	Descrição
Controle acadêmico	Base de dados SQL Server	Vida acadêmica dos alunos, professores, cursos e matrizes curriculares.
Pedagogia	Planilhas eletrônicas	Dados de pesquisas aplicadas semestralmente, para avaliação do desempenho escolar dos alunos. Para avaliação do desempenho dos professores e pesquisas avaliativas

		do instituto, como por exemplo, infraestrutura.
Serviço social	Planilhas eletrônicas	Dados referentes a assistência social do instituto para com os alunos carentes.
Biblioteca	Base de dado MySql	Informações de acesso a biblioteca pelos alunos (requisições de livros).

Essas diversas bases de dados não estão interligadas, ou seja, elas não se comunicam entre si, dificultando se fazer uma análise estatística com esses dados. Portanto, a primeira tarefa a ser realizada, foi uma análise sucinta das bases de dados, para que fossem selecionados, apenas os dados relevantes ao sistema proposto. E em seguida foi feita a extração, transformação e carga dos dados selecionados, para uma base de dados consolidada. Justifica-se a criação dessa base consolidada, para se ter em uma mesma fonte de dados, todas as informações necessárias e consistentes, para se fazer análise de dados.

3.3.2.2. Requisitos Funcionais e Não funcionais

Foram levantadas, de forma sucinta, todas as características que cercam a educação no instituto federal. Essas informações foram coletadas de forma a elucidar os principais índices de desempenho que se queira analisar, de forma a facilitar a escolha do processo educativo, que se queira modelar. Um processo educativo, é por exemplo, analisar a repetência e a evasão escolar no IFRN. Além do mais foi definida a estrutura arquitetural de implantação do DW.

3.3.2.3. Arquitetura de Fluxo de Dados

Em um *Data Warehousing*, a arquitetura de fluxo de dados é a configuração dos armazéns de dados dentro do sistema DW, e como os dados fluem das fontes de dados, para os armazéns de dados e desses para os usuários finais do sistema. Isto inclui, como esses dados são controlados, monitorado, bem como os mecanismos para assegurar a qualidade desses dados. Existem diversas arquitetura para fluxo de dados em um DW, dentre elas pode-se citar: um simples *Dimensional Data Store (DDS)*, *Normalized Data Store (NDS) + DDS*, *Operational Data Store (ODS) + DDS*, e *Data Warehouse*

federado. Para este trabalho foi utilizar a arquitetura simples DDS, ilustrada pela Figura 10.

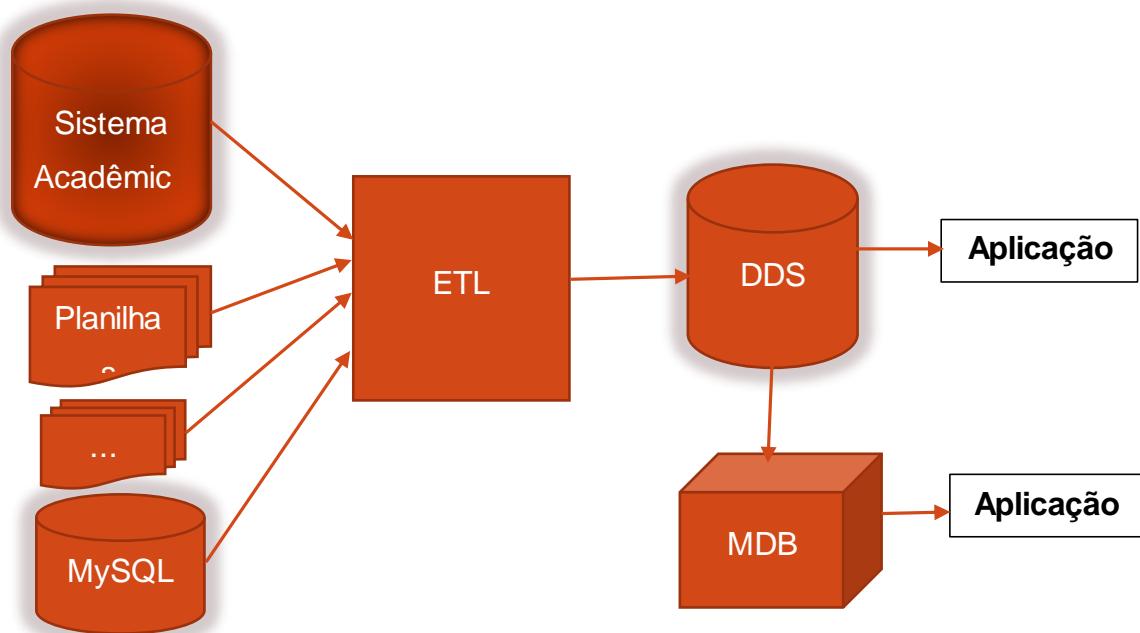


Figura 10. Arquitetura Simples DDS combinado com o ETL dos dados das fontes de dados disponíveis.

Fonte: Autor (adaptado de Rainardi, 2008).

A principal fonte de dados para esse trabalho é sem dúvida nenhuma, a base de dados do sistema acadêmico do IFRN, onde estão armazenados todos os dados referentes a vida acadêmica dos alunos, cursos, professores, dados sociais e assim por diante. A Figura 11 mostra o diagrama simplificado do sistema acadêmico.

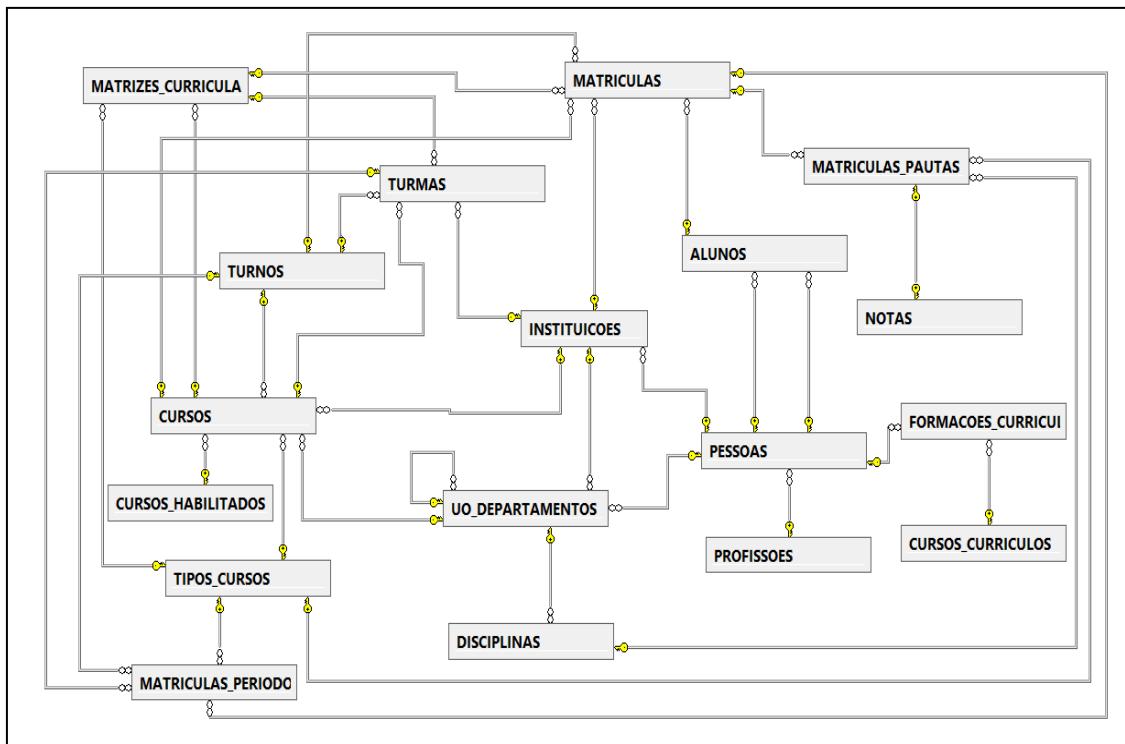


Figura 11. Diagrama Banco de Dados do sistema acadêmico.

Fonte: Autor

3.4. Modelo de Dados Dimensional (MDD)

Então, foi construído um modelo dimensional (na nossa arquitetura, figura 10, o **DDS**) para conter as informações dos fatos e dimensões. O DDS será composto de diversos data marts. Cada data mart irá tratar de um fato em específico e, isso será determinado pela granularidade de cada assunto. Por exemplo, quando estiver analisando dados sobre boletim escolar, temos um fato, pois a granularidade aqui, é baseada nos lançamentos dos dados no boletim do aluno. Neste caso, será analisado o índice da evasão e da repetência em função dos dados contidos no boletim escolar dos alunos. Por outro lado, quando estivermos analisando os dados sociais dos alunos, teremos um outro fato, com outra granularidade, pois temos outras informações diferentes das informações do boletim do aluno. Portanto, para cada tipo de informações lançada no sistema teremos um fato (Boletim, histórico, dados sociais, forma de ingresso nos cursos e assim por diante).

Um data mart é composto de tabela de fato e dimensões. Os fatos contêm as medidas, ou seja, os totais, e as dimensões discriminam essas medidas [Kimball, 2002].

O primeiro fato a ser trabalhado é para analisar a situação dos alunos em relação a evasão, o cancelamento de matrículas e o total de alunos que concluíram seus respectivos cursos, totalizados por campus, curso, ano letivo e período letivo. Para isto, foram criadas as dimensões e tabela de fatos, mostrada na Tabela 3.

Tabela 3. Dimensões e tabela de fatos.

Dimensão	Comentários
dimCursos	Dados referentes aos cursos ministrados no IFRN.
dimInstituicao	Cadastro de todos os campi.
dimTempo	Dados referentes ao ano ou semestre letivo.
dimEvaJubPorCampusCurso	Dados referente aos alunos que evadiram ou foram jubilados dos cursos.
dimCanceladoPorCampusCurso	Dados referentes aos alunos que cancelaram suas matrículas nos cursos.
dimConcluidosPorCampusCurso	Dados referentes aos alunos que concluíram seus cursos.
FatoEvaCanCon	Tabela fatos com os totais de alunos evadidos, que cancelaram matrícula e total de concluintes.

A Figura 12 mostra o Data Mart gerado para este fato. Devo lembrar que as tabelas dimEvaJubPorCampusCurso, dimCanceladoPorCampusCurso e dimConcluidosPorCampusCurso têm a mesma granularidade que é “Campus, curso, ano letivo e período letivo”.

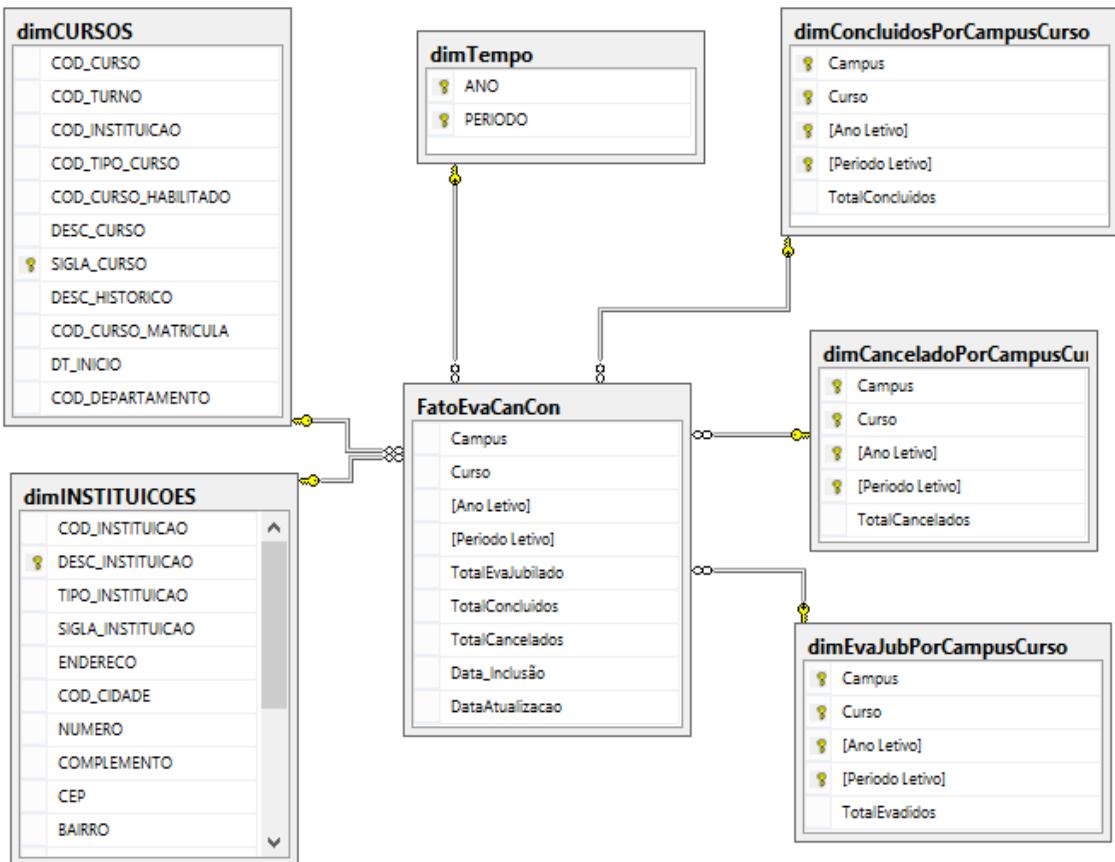


Figura 12. Data mart para as medidas de Evasão, cancelado e concluídos.

Fonte: Autor

Criado o modelo dimensional, o próximo passo, de acordo com a arquitetura da Figura 10, é a extração, a transformação e a carga dos dados da fonte de dados para o modelo dimensional. Para esta tarefa, será utilizada consultas SQL, para extrair os dados da base de dados acadêmica para as dimensões e em seguida, será feita a carga na tabela de fatos, também usando uma consulta SQL. A Listagem 3.1 a seguir, exemplifica a extração e carga do sistema acadêmico para a tabela dimEvaJubPorCampusCurso do modelo dimensional, usando uma consulta SQL. As demais dimensões usam código SQL semelhante.

Listagem 3.1: SQL para popular a dimensão dimEvaJubPorCampusCurso

```

INSERT INTO dimEvaJubPorCampusCurso
(Campus,
  Curso,
  [Ano Letivo],
  [Periodo Letivo],
  TotalEvididos)
SELECT desc_instituicao AS [Campus],

```

```

sigla_curso AS [Curso],
ano_let AS [Ano Letivo],
periodo_Let AS [Periodo Letivo],
sum(QTD_ALUNO) AS TotalEvadidos
FROM vJuntandoEvaDeslCon
WHERE Situacao_Da_Matricula = 'Evasão' OR
      Situacao_Da_Matricula ='Jubilado'
GROUP BY DESC_INSTITUICAO, Sigla_Curso,ANO_LET, PERIODO_LET

```

A Listagem 3.2 a seguir, exemplifica a população da tabela de fato, usando uma consulta SQL.

Listagem 3.2: SQL para popular a tabela de fatos FatoEvaCanCon

```

INSERT INTO FatoEvaCanCon (Campus,Curso, [Ano Letivo], [Periodo Letivo],
                           TotalEvaJubilado,
                           TotalConcluidos,
                           TotalCancelados,
                           Data_Inclusão,
                           DataAtualizacao)
SELECT coalesce(EJ.Campus,null), coalesce(EJ.Curso,null),
       coalesce(EJ.[Ano Letivo],null), coalesce(EJ.[Periodo Letivo],null),
       EJ.TotalEvadidos, CO.TotalConcluidos,
       CA.TotalCancelados, getdate(), getdate()
FROM dimEvaJubPorCampusCurso EJ inner Join dimConcluidosPorCampusCurso
CO
ON EJ.campus = CO.Campus and EJ.Curso = CO.Curso AND
   EJ.[Ano Letivo] = CO.[Ano Letivo] AND EJ.[Periodo Letivo] =
   CO.[Periodo letivo] inner Join dimCanceladoPorCampusCurso CA
ON CO.campus = CO.Campus and EJ.Curso = CA.Curso AND
   CO.[Ano Letivo] = CA.[Ano Letivo] AND
   CO.[Periodo Letivo] = CA.[Periodo letivo] inner Join dimTempo T
ON CA.[Ano Letivo] = T.[ANO] AND
   CA.[Periodo Letivo] = T.[PERÍODO]

```

A carga dos dados para as demais tabelas e fatos, seguem os exemplos das listagens 3.1 e 3.2 acima.

O segundo, data mart a ser modelado é para analisar repetência e a evasão escolar, em função dos dados sociais dos alunos. Com esse data mart, consigo saber o total de alunos reprovados por curso, renda familiar, escola de origem, sexo, etnia, disciplina e assim por diante e, ainda agrupados por anos e semestre letivo. A Tabela 4, lista as dimensão e fatos que foram definidos.

Tabela 4. Lista de dimensões e tabelas de fatos.

Dimensão ou tabela de fato	Comentários
dimData	Identifica do tempo onde ocorreu o fato.
dimCursos	Identifica o curso onde ocorreu o fato.
dimDisciplinas	Identifica a disciplina na qual ocorreu o fato.
dimDadosSociaisAlunos	Esta tabela contém o cadastro do aluno e seus dados sociais (Renda familiar, escola de origem, etnia, sexo, e assim por diante).
FatoAprRepCursoRendaDS	Total de Aprovados e reprovados, agrupados por curso e renda familiar.
FatoAprRepCursoEtniaDS	Total de Aprovados e reprovados, agrupados por curso e etnia.
FatoAprRepCursoSexoDS	Total de Aprovados e reprovados, agrupados por curso e sexo.
FatoAprRepCursoEscolaOrigemDS	Total de Aprovados e reprovados, agrupados por curso e escola de origem.
FatoAprRepCursoDisiplinaDS	Total de Aprovados e reprovados, agrupados por curso e disciplina.
FatoAprRepCursoAnoDS	Total de Aprovados e reprovados, agrupados por curso e ano letivo.

A Figura 13 mostra o diagrama do modelo dimensional ou *data mart* para para analisar repetência e a evasão escolar, em função dos dados sociais dos alunos.

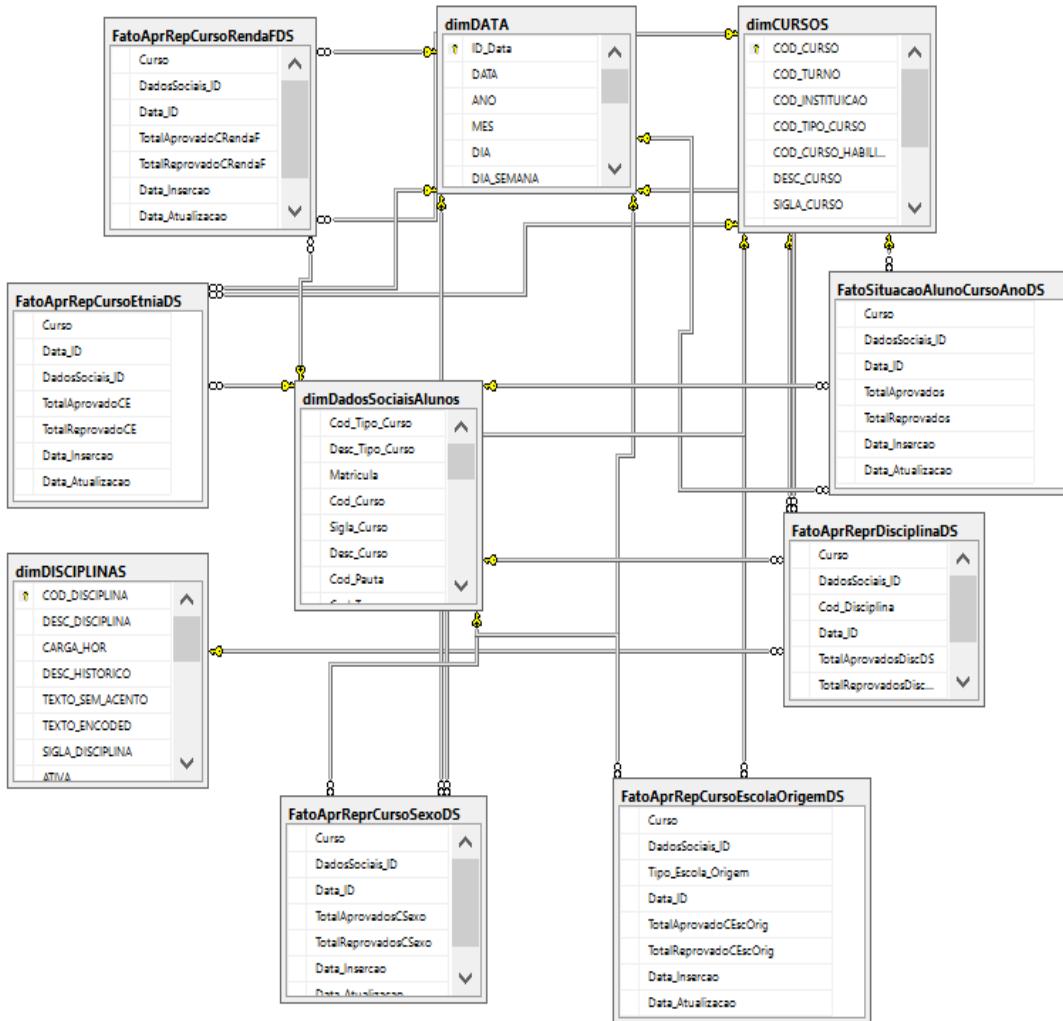


Figura 13. Modelo dimensional em função dos dados sociais dos alunos.

Fonte: Autor

O terceiro cenário, está relacionado, aos dados de entrada e saída dos alunos nos cursos. Pretende-se com esse fato, fazer uma projeção da evasão escolar, de cada curso em cada campus, ao longo dos anos. Tenho também, através desse fato, a possibilidade de calcular quantos alunos terminam os seus cursos no prazo determinado. Isso me dá a possibilidade de calcular o gasto extra, que foi feito com esses alunos repetentes. A Tabela 5, lista as dimensões e tabela de fato.

Tabela 5. Dimensões e fato para o “FatoEvaEntradaSaida”

Dimensão ou fato	Descrição
dimCursos	Cadastro de curso

dimData	Cadastro de datas
dimInstituicoes	Cadastrado de campus
dimAlunosSituacaoDadoIngresso	Esta dimensão contém os dados referentes a sua situação no curso.
FatoEvaEntradaSaida	Esta tabela de fato contém as medidas total de evadidos e concluídos.

A Figura 14 mostra o diagrama do modelo dimensional ou *data mart* para este cenário.

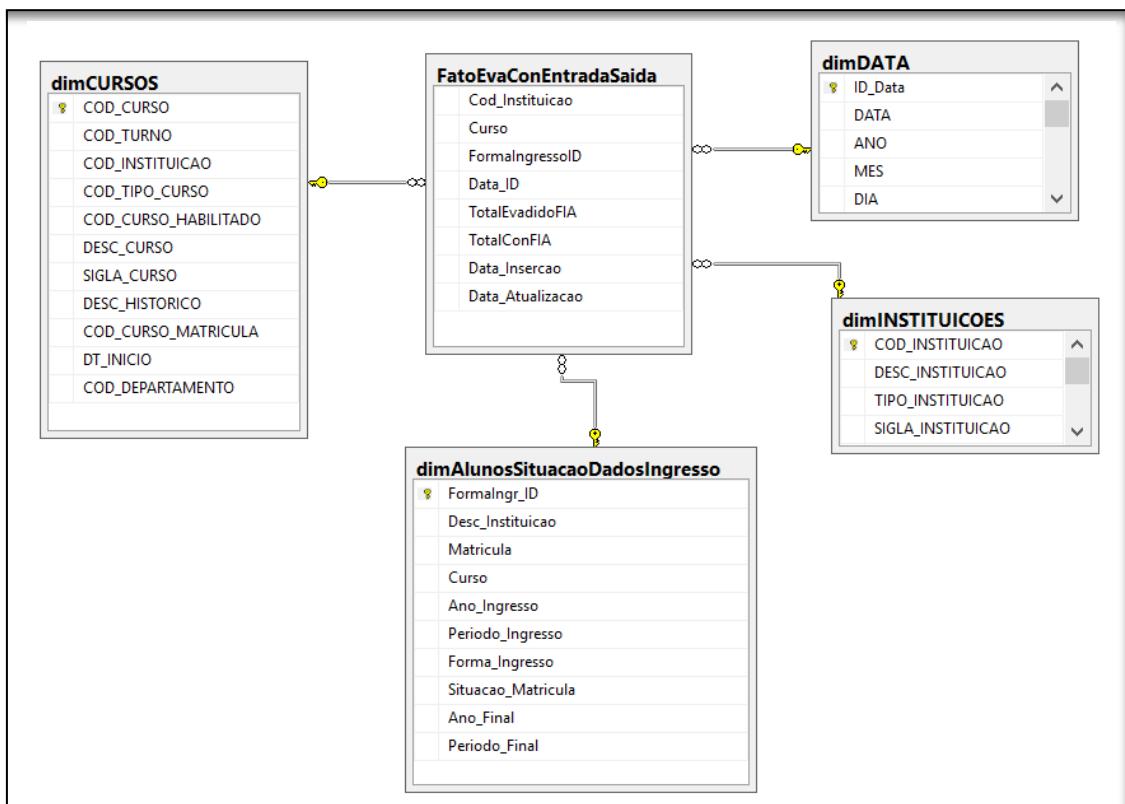


Figura 14. Data mart para o fato “FatoEvaEntradaSaida”.

Fonte: Autor

Um quarto cenário é sobre os dados dos boletins escolares. Com esses dados, será possível fazer um acompanhamento, bimestre a bimestre, da taxa de aprovação e reprovação dos alunos e também analisar a situação da turma.

Os dados do boletim escolar são extraídos do sistema acadêmico em formato de planilhas eletrônicas. A Figura 15 mostra o diagrama do pacote **ETL** para extrair os dados das planilhas Excel para o banco de dados dimensional.

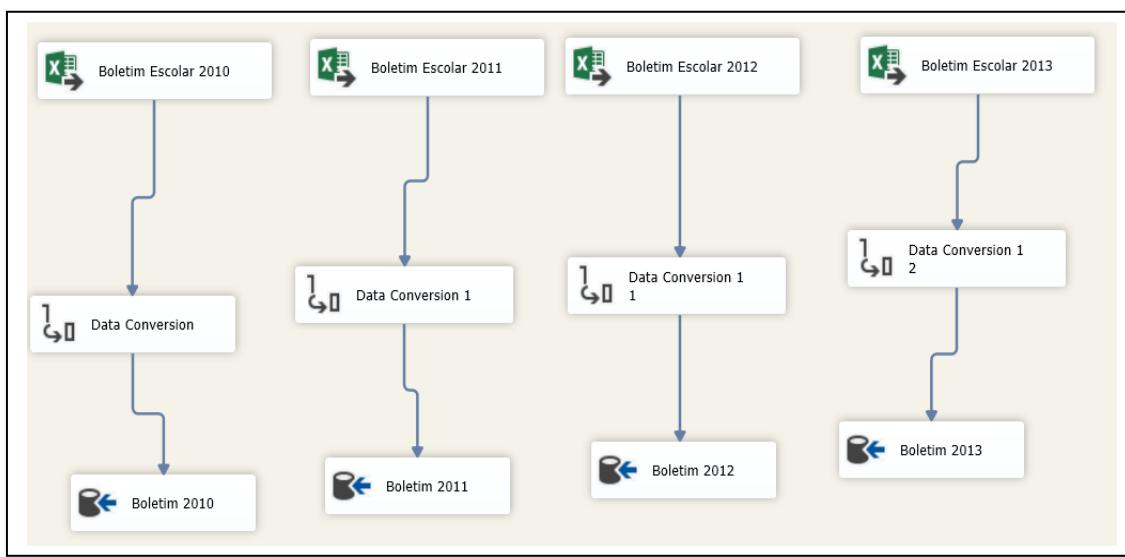


Figura 15. Pacote ETL para extrair os dados das planilhas Excel para o banco de dados dimensional.

Fonte: Autor

A Figura 16 mostra o modelo dimensional em função dos dados do boletim escolar dos alunos.

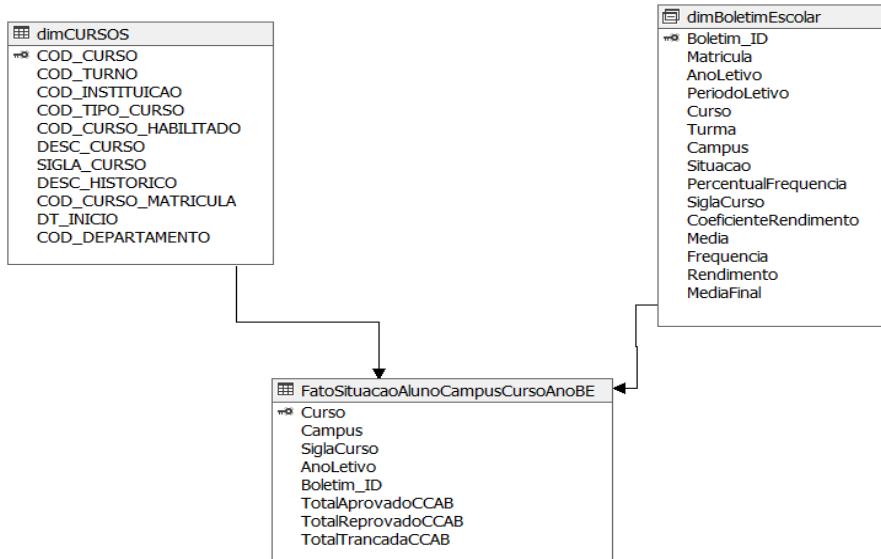


Figura 16. Data Mart para os dados relacionados ao boletim escolar.

Fonte: Autor

No próximo cenário, será analisado a taxa de repetência escolar por curso, disciplina e ano. A Figura 17 mostra o modelo dimensional para esse cenário.

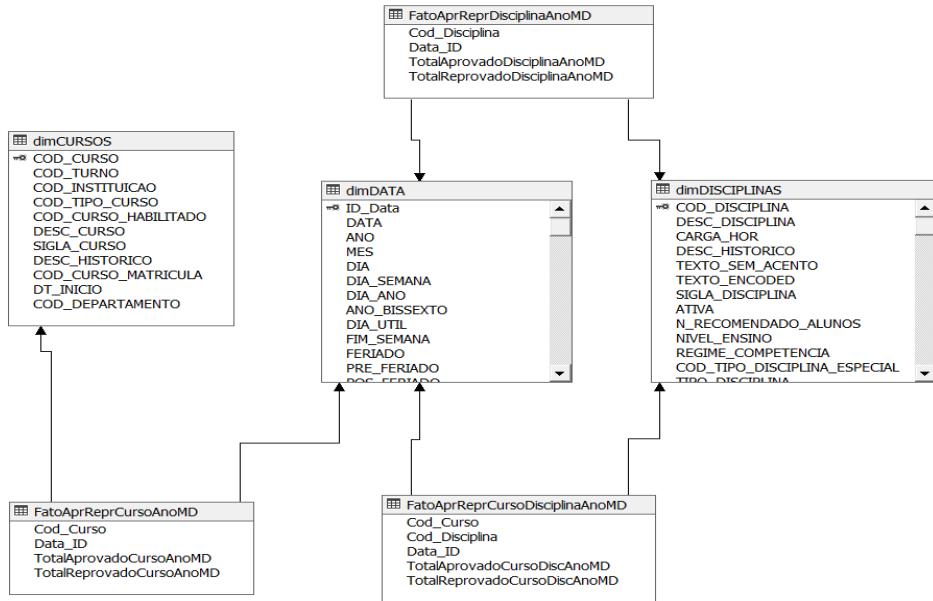


Figura 17. Data mart para análise de repetência escolar agrupado por curso, disciplina e ano letivo.

Fonte: Autor

No próximo cenário, será analisado a taxa de repetência escolar por professor. A Figura 18 mostra o modelo dimensional para esse cenário.

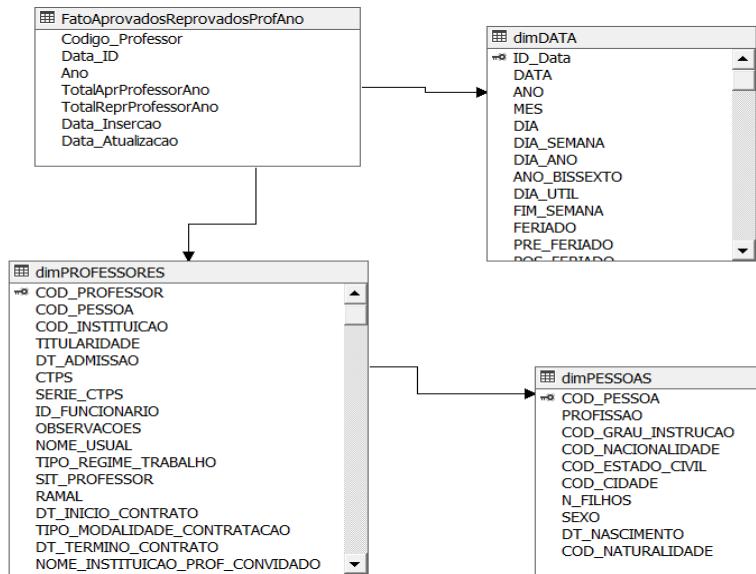


Figura 18. Data Mart para dados relacionados a repetência escolar agrupados por professor. **Fonte:** Autor

3.5. Modelo Multidimensional

Um banco de dados multidimensional é uma forma de banco de dados, onde os dados são armazenados em células e a localização de cada célula é

definida por hierarquias chamadas de dimensões. Cada célula representa um evento de negócio, e os valores das dimensões mostram quando e onde este evento ocorreu [Rainardi, 2008].

A estrutura multidimensional, armazena as agregações geradas, bem como os dados da base de dados, em formato de matriz, em vez de tabelas relacionais. Os valores das agregações são pré-calculadas, summarizadas em função dos dados da base de dados [Rainardi, 2008].

Banco de dados multidimensionais são tipicamente usados para **BI**, especialmente para **OLAP** e mineração de dados. A Figura 19 mostra a estrutura de cubo para o fato FatoEvaCanCon.

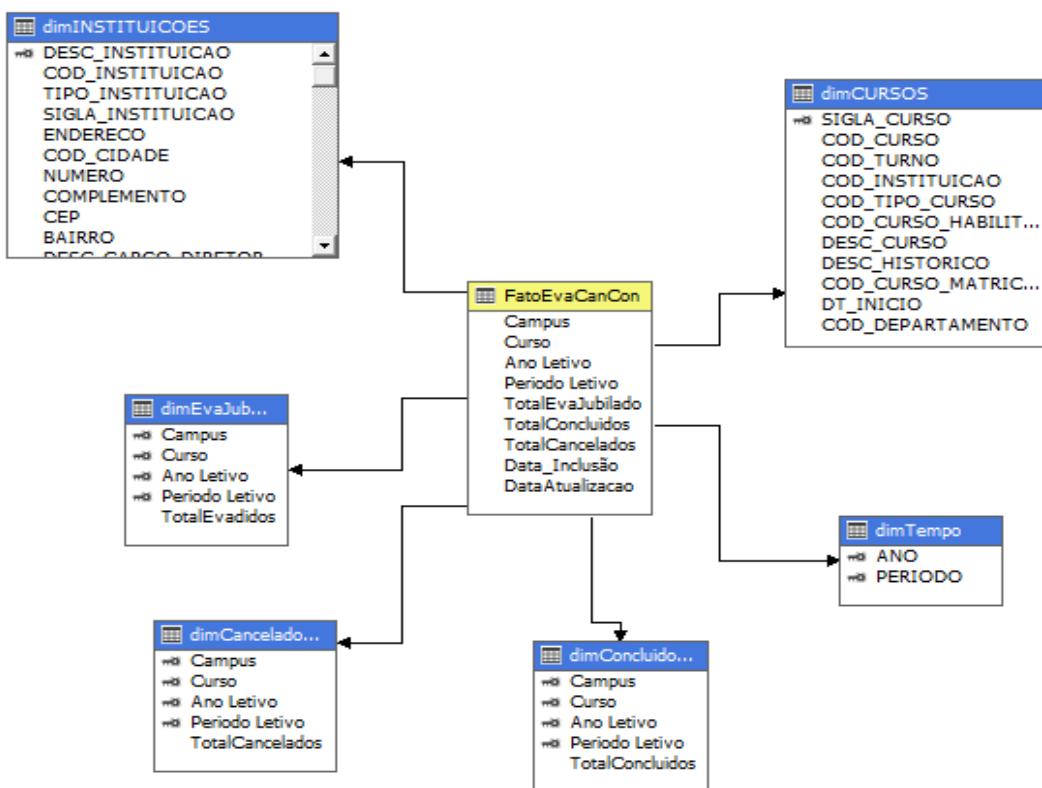


Figura 19. Estrutura do Cubo para o fatoEvaCanCon. **Fonte:** Autor

Já a Figura 20 mostra o modelo multidimensional para os fatos relacionados aos dados sociais dos alunos, tais como renda familiar, tipo de escola de origem, sexo e etnia. Este cubo servirá para a análise da repetência escolar. Como se pode observar, este modelo é uma constelação de fatos [Kimball, 2002].

A principal dimensão, nesse cubo é a dimensão dimDadosSociaisAlunos, pois a mesma contém os lançamentos de onde serão calculados os totais de alunos reprovados e ou aprovados, agrupados ora por etnia, ora por renda familiar e assim por diante. Esta tabela, contém aproximadamente 1 milhão de registros, se for considerado apenas os anos de 2010 até 2014.

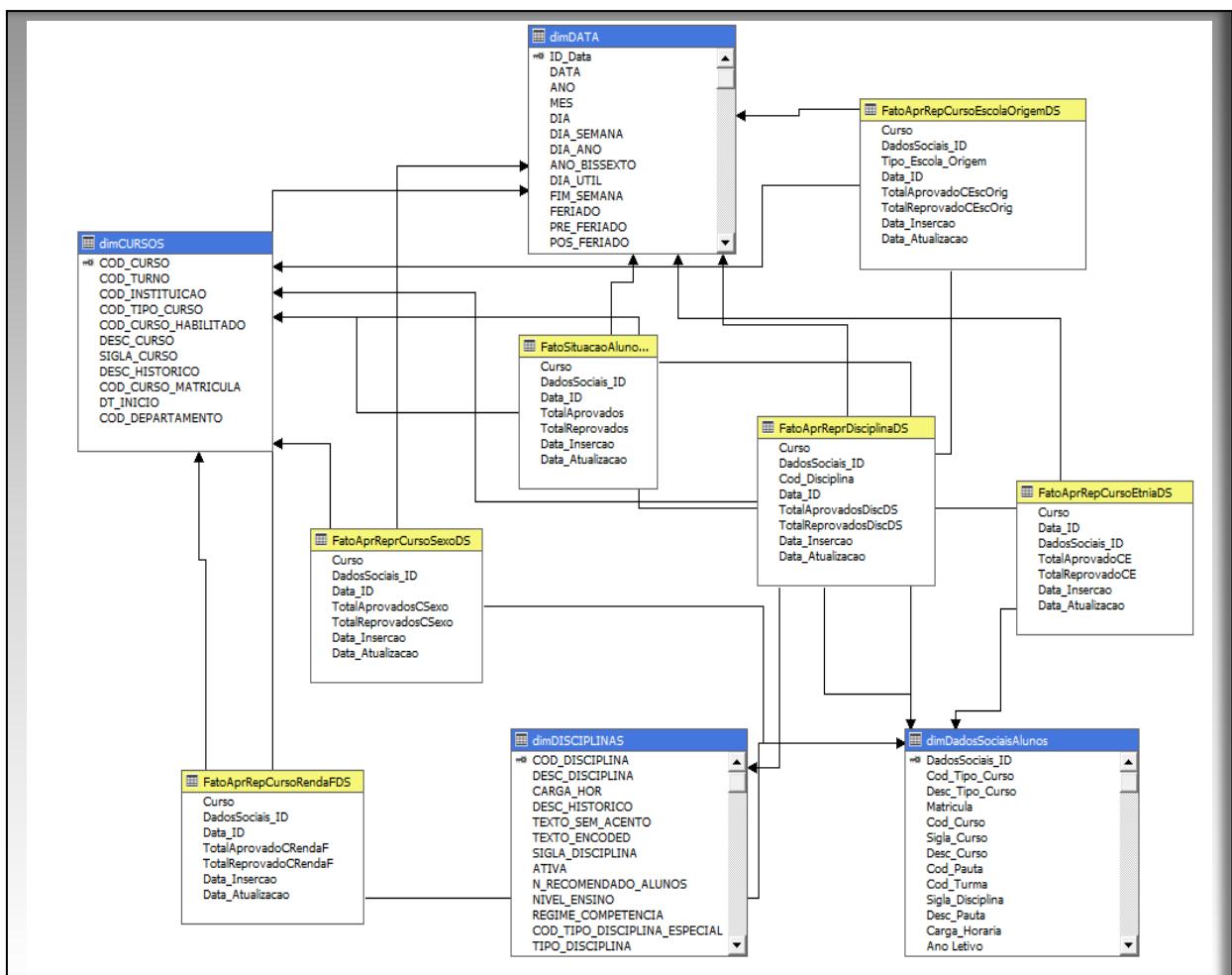


Figura 20. Estrutura de cubo criada para os dados sociais dos alunos.

Fonte: Autor

A Figura 21 mostra o modelo multidimensional para o fato “FatoEvaEntradaSaida”, onde encontramos as informações sobre a evasão escolar.

Já a Figura 22 mostra o modelo multidimensional para o fato “FatoAprovadosReprovadosProfAno”, com dados relacionados a repetência escolar por professor, podendo ser representado por semestre e ano.

A Figura 23 mostra o modelo multidimensional para os fatos relacionados a repetência escolar em função das médias anuais das disciplinas nos respectivos cursos.

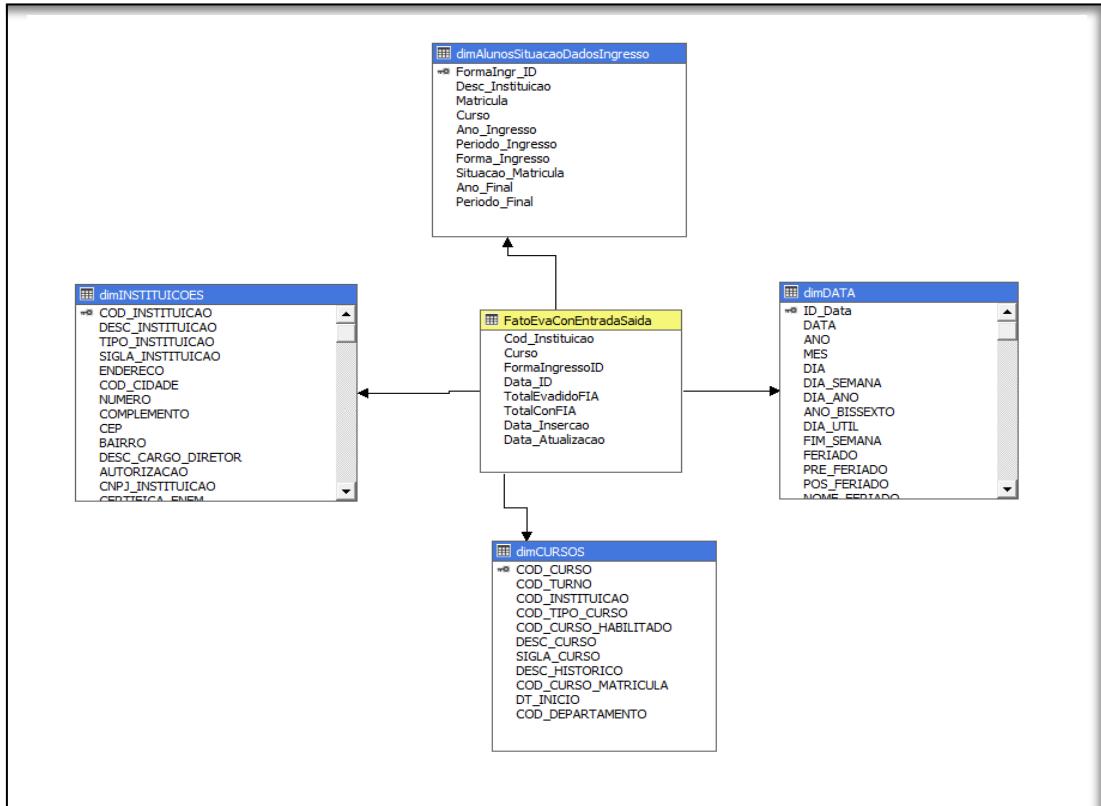


Figura 21. Estrutura de cubo criada para o fato “FatoEvaEntradaSaida”.

Fonte: Autor.

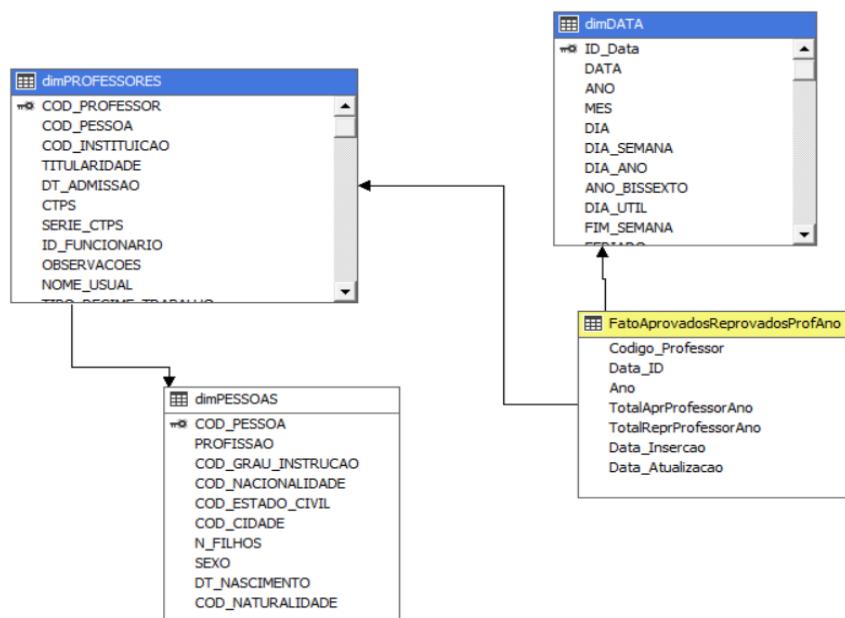


Figura 22. Estrutura de cubo para o fato “FatoAprovadosReprovadoProfAno”.

Fonte: Autor

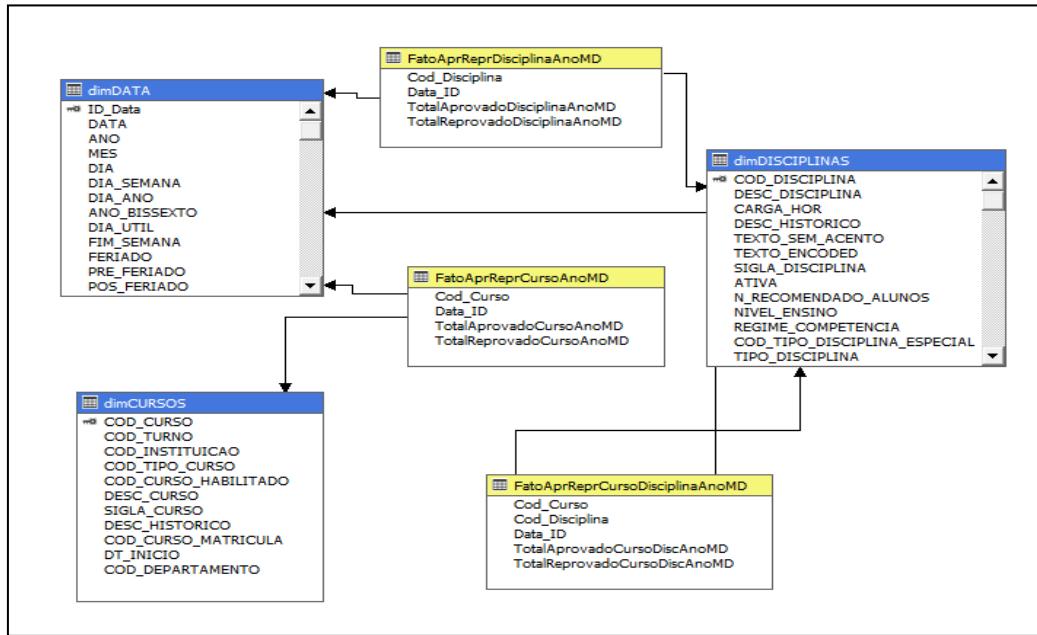


Figura 23. Estrutura de cubo para as medidas aprovados e reprovados por disciplina e curso.

Fonte: Autor.

Com todos os cubos criados, temos o nosso sistema multidimensional pronto para uso. Então, a partir desse ponto já se pode realizar a análise de dados, bem como gerar os relatórios Ad Hoc de acordo com as demandas solicitadas pelos usuários finais da aplicação.

Nas seções seguintes, serão implementados alguns recursos no intuito de melhorar a performance das consultas e também facilitar as montagens das consultas aos usuários finais.

3.5.1. Projetando Agregações e Hierarquias

Vamos, a partir desse ponto, considerar a possibilidade de otimização desses cubos, criando agregações e hierarquias de usuário.

As agregações melhoraram o desempenho das consultas sem aumentar excessivamente o espaço de disco necessário para armazenar dados no cubo. Por outro lado, as hierarquias permitem que os usuários naveguem por um cubo com mais eficiência, e também contribui para aperfeiçoar o desempenho da consulta [Harinath, 2012].

Um cubo **OLAP** parece conter cada valor totalizado possível por cada atributo de cada dimensão. Por exemplo, se um cubo contém as medidas “TotalEvadido” e “TotalConFIA” e, as dimensões dimCursos, dimAlunosSituacaoDadosIngresso, dimInstituicoes e dimTempo. Se você fizer uma consulta no cubo em um nível de detalhe mais baixo, por exemplo, qual é o total de evadidos do curso 01025 do campus CNAT em 2010, o cubo retornará um valor, como se estivesse armazenado diretamente nele. Ou seja, parece que o cubo contém todos os valores totalizados possíveis em todos os níveis de detalhes para cada dimensão. Supondo que a dimensão dimAlunosSituacaoDadosIngresso tenha 104060 registros e, que a dimensão dimCursos tem 696 registros e, que a dimensão dimInstituicoes tem 21 registros e que dimensão dimData tem 10959 registros. Isto nos dá um total de combinações da ordem de $10406 \times 696 \times 21 \times 10959 = 16.667.991.980.640$ combinações possíveis e, isso é denominado de explosão de dados. A explosão de dados é um dos principais problemas com os cubos **OLAP** e todos os produtos **OLAP** precisão lidar com isso de alguma maneira [Jacobson, 2005][Harinath, 2012].

De acordo com Jacobson (2005), a maneira mais fácil de evitar a explosão de dados é evitar armazenar as agregações juntas, calculando-as sob demanda. No entanto, quando o **DW** é realmente grande, essa opção afeta rapidamente o desempenho, porque solicitar um único valor totalizado de alto nível do cubo requer recuperar e somar centenas ou milhares de valores da fonte de dados. E como já foi falado, o desafio do **OLAP** é tornar as consultas mais rápidas possível evitando a explosão de dados.

Então o recurso usado neste trabalho para melhorar o desempenho das consultas feitas aos cubos **OLAP**, foi a criação de agregações.

As agregações são totalizações pré-calculadas de dados detalhados que habilitam o servidor de análise a responder consultas rapidamente. Isto acontece, pelo fato de que, quando for solicitado um valor de um cubo, o servidor de análise usa qualquer agregação que esteja disponível para recuperar o valor o mais rápido possível. Apesar de se poder criar cubos sem agregações – as agregações podem fazer uma imensa diferença no tempo de consulta de um cubo grande [Jacobson, 2005].

Um fato importante a ser observado aqui, é que, para todo cubo OLAP criado, é criada para o mesmo uma partição padrão, definido de acordo com sua granularidade.

Uma partição é um local físico de dados armazenados do cubo. E isto é importante, porque o servidor de análise executa mais rapidamente as consultas em um cubo particionado, pois o mesmo precisa apenas ler os dados das partições que contêm os dados que ele precisa para montar as consultas. Além dos mais, as consultas podem ser executadas ainda mais rápido quando a partição também armazenar agregações [Jacobson, 2005].

3.5.1.1. Projetando Agregações para o DW

O cubo da Figura 3.17, tem apenas a partição padrão, como mostra a Figura 24.

Portanto, deseja-se otimizar a performance dessa partição, para que a mesma tenha um ganho de performance de 30% em detrimento do espaço em disco a ser ocupado. Para isto, será utilizado o assistente de agregação, que após sua execução, exibe o resultado da Figura 25, onde pode-se ver que a partição será otimizada em 50% e o espaço em disco a ser ocupado é em torno de 18kbytes.

No gráfico da Figura 25, pode-se ver a relação entre o aumento de desempenho (no eixo y) e a quantidade de espaço em disco consumido (no eixo x) como resultado do projeto de agregação atual. Deve ser lembrado que, o objetivo da otimização da agregação é, obter o melhor ganho de desempenho, sem consumir espaço em disco desnecessários.

O cubo continua com apenas a partição padrão, no entanto, ele tem agora, uma agregação, como mostra a Figura 26.

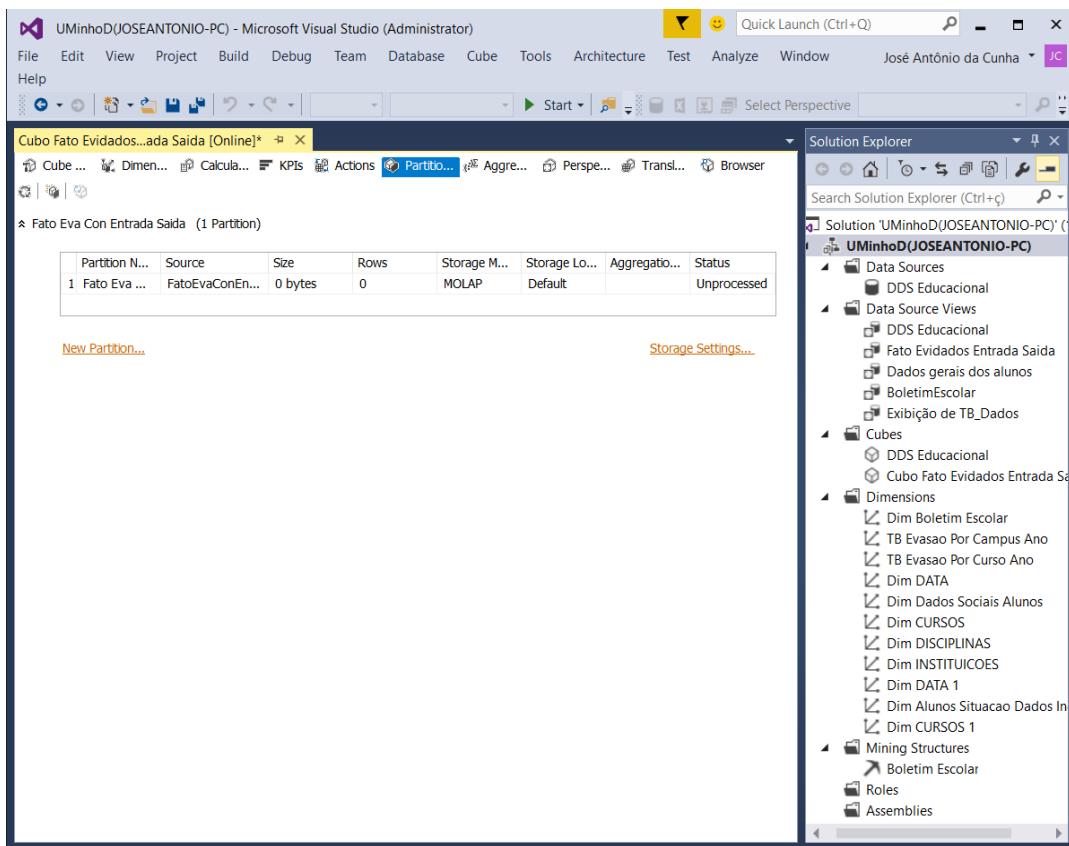


Figura 24. Partição padrão do cubo Fato Evadidos Entrada Saída.

Fonte: Autor

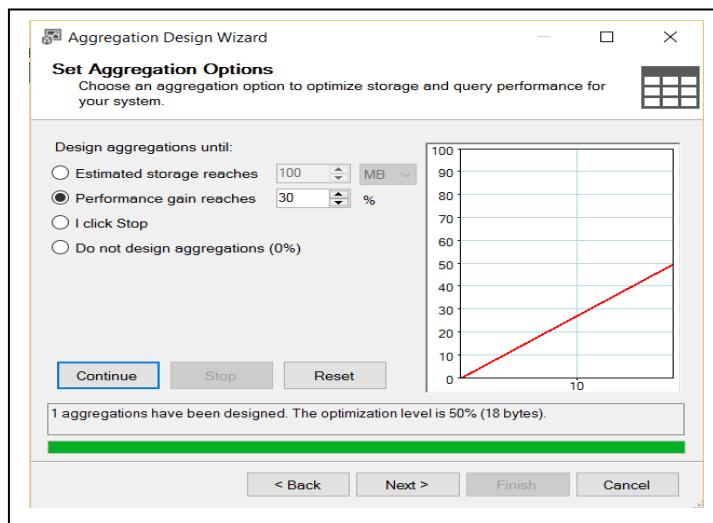


Figura 25. Calculando agregações para o cubo Fata Evadidos Entrada Saída.

Fonte: Autor.

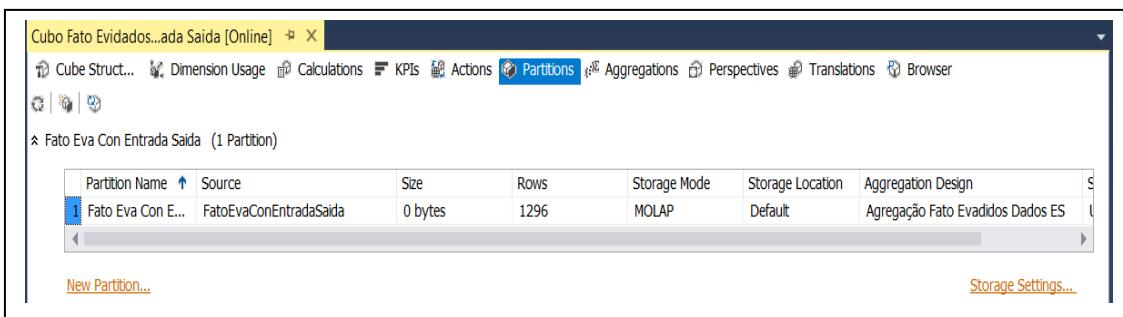


Figura 26. Partição padrão com a agregação calculada pelo assistente.

Fonte: Autor.

Observe na Figura 25 que a contagem de linhas (coluna Rows), calculada pelo assistente de agregações, foi estimado em 1296 linhas. Este valor reflete a contagem de linhas da tabela de fatos e, esse valor é um valor estimado pelo assistente e, possivelmente difere da contagem de linhas real. Entretanto, muitas vezes, para se obter um projeto de agregação mais preciso, o administrador do sistema deve fazer os ajustes para se obter um valor próximo do valor real. Deve ser observado que, as contagens são usadas apenas para o projeto de agregações, mas pode resultar em um projeto de agregação que esteja abaixo do ideal, se você subestimar a contagem de linhas.

Quando é feito ajustes na contagem de linhas pelo administrador do sistema, o assistente pode gerar outras agregações, para otimizar a performance e ocupação de espaço em disco. Então, a regra de negócio é quem vai determinar qual das configurações é a melhor para o sistema. Devo lembrar que, esses ajustes vão ocorrer durante todo o ciclo de vida do sistema.

Quando se cria uma agregação, por padrão, apenas os atributos de mais alto nível de hierarquia e de granularidade, estão disponíveis para consideração de agregação pelo assistente. No entanto, quando se tem certos atributos que são usados com frequência e, possui um número de membros relativamente pequena, é possível pensar em adicioná-lo ao pool de possíveis candidatos para a agregação. No nosso caso, tem um atributo da dimensão “dimAlunosSituacaodadosIngresso” denominado Situacao_Matricula que tem poucos membros e é bastante consultado. Portanto, esse atributo será configurado, neste cubo, para ser um atributo de agregação. Basta alterar a propriedade AggregationUsage para Full e depois recalcular as agregações.

Observe na Figura 27, que foi atribuído o valor 4 para a contagem de linha para o atributo “Situacao_Matricula”. Esse valor é devido o atributo só possui quatro membros, a saber Matriculado, Cancelado, Concluído e Evadido.

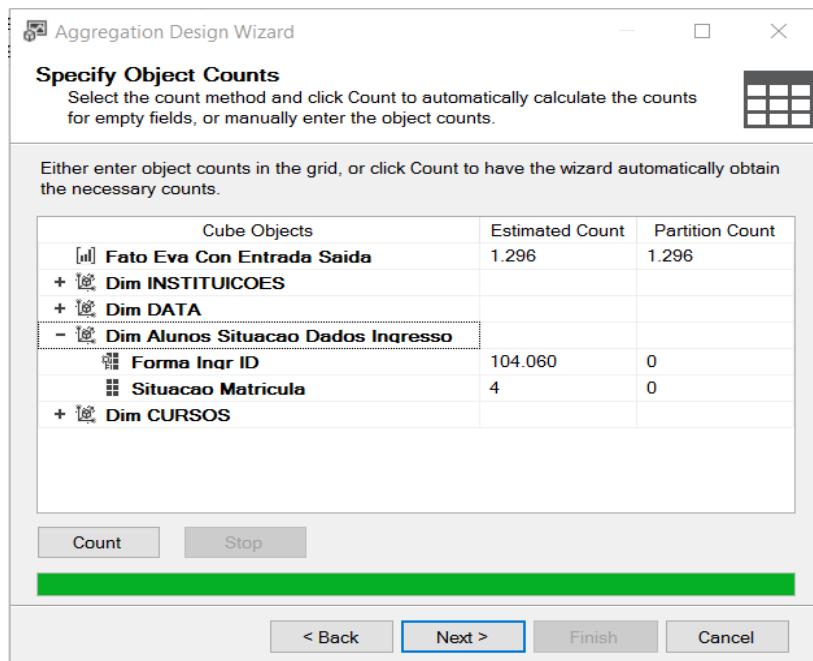


Figura 27. O atributo “Situacao Matricula” com o valor 4 estimado para a contagem de linhas.

Fonte: Autor.

Seguiu-se a mesma dinâmica para os demais cubos do sistema DW. Portanto, para o cubo relacionados aos dados sociais dos alunos, onde o mesmo é formado de vários fatos (um total de seis fatos), é criado por padrão, uma partição para cada um dos fatos desse cubo. Esse cubo, em particular, nesse projeto, tem um fator que pode ser preocupante, que é a explosão de dados, uma vez que, só a dimensão “dimDadosSoicia”, tem mais de milhão (precisamente 1074226) de registros.

Portanto, para este cubo, para se evitar a explosão de dados, foram projetar agregações, levando-se em conta o ajuste de contagem de linhas das dimensões e fatos, para um fator de otimização das consultas ajustado para 30% de performance para pouco espaço em disco.

A Figura 28 mostra a quantidade de agregações geradas para cada tabela de fato.

	Aggregations	Estimated Parti...	Partitions
[a] Fato Apr Rep Curso Escola Origem DS (1 Aggregat...	AggregationDesign	5	4013
[a] Fato Apr Rep Curso Etnia DS (1 Aggregat...	AggregationDesign	5	4514
[a] Fato Apr Rep Curso Renda FDS (1 Aggregat...	AggregationDesign	4	7364
[a] Fato Apr Repr Curso Sexo DS (1 Aggregat...	AggregationDesign	7	2664
[a] Fato Apr Repr Disciplina DS (1 Aggregatio...	AggregationDesign	13	24410
[a] Fato Situacao Aluno Curso Ano DS (1 Agg...	AggregationDesign	4	1400

Figura 28. Agregações criadas para cada tabela de fato do cubo dados sociais.

Fonte: Autor.

Deve-se ser observado que, o processo de otimização do desempenho das consultas, não é um processo estande, muito pelo contrário, é um processo que deve ser realizado em toda vida do sistema em produção. A todo instante você pode está adicionando novas medidas e criando novas hierarquias e, portanto, reprocessando o cálculo de agregações. Entretanto, existe o recurso de se fazer os ajustes das agregações, com base no log de consultas. Ou seja, o administrador do sistema, ativa a gravação em log, de todas as consultas realizadas pelos usuários e, então com base nas informações armazenadas no log, fazer a otimização das consultas.

3.5.1.2. Projetando Hierarquias

A finalidade das hierarquias é facilitar a navegação dos usuários finais pelo cubo, pois o usuário ao invés de arrastar cada atributo do cubo, para montar o relatório ou consulta, o mesmo pode apenas usar uma hierarquia, e isso, facilita em muito seu trabalho. Uma hierarquia é na verdade um empacotamento de atributos. Portanto, serão criadas algumas hierarquias de atributos para os cubos criados para o sistema DW.

Para exemplificar o uso das hierarquias, tome como exemplo a seguinte situação. Se o usuário desejar consultar o curso por código, sigla e descrição do curso. Então, pode-se empacotar esses atributos em uma hierarquia, como mostra a Tabela 6, a hierarquia da tabela Curso “CÓDIGO – SIGLA - CURSO”.

Tabela 6. Hierarquias criadas para o cubo dados sociais.

HIERARQUIA	ESTRUTURA DE ATRIBUTOS
TABELA DADOS SOCIAIS	
CURSO – SILGA	Desc_Curso → Sigla_Curso
RENDAS – ESCOLA – ETNIA	Renda_Familiar → Tipo_Escola_Origem → Etnia
CURSO – FORMA INGRESSO	Desc_Curso → Forma_De_Ingresso
TABELA DATA	
ANO – SEMESTRE	Ano → Semestre
TABELA CURSOS	
CURSO – SIGLA	Desc_Curso → Silga_Curso
CÓDIGO – SIGLA- CURSO	Cod_Curso → Sigla_Curso → Desc_Curso
CAMPUS – CURSO – DATA	Codigo_Instituicao → Desc_Curso → Data_Inicio
TABELA DISCIPLINAS	
DISCIPLINA – SIGLA	Desc_Disiplina → Sigla
TABELA DADOS INGRESSO	
CAMPUS – CURSO - ANO	Desc_Instituicao → Desc_Curso → Ano_Final
CURSO – ANO	Desc_Curso → Ano_Final
TABELA PROFESSORES	
PROFESSOR – REGIME TRABALHO	Nome_Usual → Tipo_Regime_Trabalho

As hierarquias da Tabela 3.8 são alguns exemplos, as demais hierarquias que provavelmente existirão, serão criadas da mesma forma e finalidade, que é a de facilitar a navegação pelo cubo pelos usuários finais.

3.5.2. Adicionando Medidas Calculadas ao Cubo

Se necessário, e quase sempre é, pode-se criar medidas calculadas em um cubo para estendem a capacidade analítica do mesmo. Isto significa que além das medidas definidas no cubo, você terá medidas extras que auxiliarão na

análise dos dados. Por exemplo, para o nosso caso, onde se estiver trabalhando com total de evadidos e conclusão, pode-se criar duas medidas calculadas para representar o percentual de evadidos e de concluídos.

O modelo multidimensional usa a linguagem (**MDX**) (um pseudo-acrônimo para Expressões Multidimensionais) como linguagem de consulta, para recuperar relatórios de um cubo e, como uma linguagem de expressões, usada para calcular valores. Portanto, será utilizada a **MDX** para criar as medidas calculadas, do percentual de evadidos e do percentual de concluídos. A Listagem 3.3 mostra essas medidas.

Listagem 3.3 Exemplo de medidas calculadas usando a linguagem **MDX**.

FatoEvaConEntradaSaida:

PercentualConclusão: Cálcula o percentual de alunos concluídos

MDX: $([\text{Measures}].[\text{Total Concluido}] * 100) / ([\text{Measures}].[\text{Total Concluido}] + [\text{Measures}].[\text{Total Evadido}])$

PercentualEvadidos: Cálcula o percentual de alunos evadidos

MDX: $(([\text{Measures}].[\text{Total Evadido}] * 100) / ([\text{Measures}].[\text{Total Concluido}] + [\text{Measures}].[\text{Total Evadido}]))$

Para o Sistema **DW** deste projeto, foram criadas várias medidas calculadas (todas usando a linguagem **MDX**) para estender a capacidade analítica dos cubos **OLAP**. A Listagem 3.1 é apenas um exemplo de como essas medidas são criadas.

3.5.3. Indicador de Desempenho (KPI)

Um **KPI** (Em inglês *Key performance indicator*) mede o progresso que uma empresa está tendo rumo ao cumprimento de suas metas. Nesse sentido, neste trabalho serão criados alguns KPI's, para acompanhar o progresso de alguns fatores ou índices. Um **KPI** mostra de forma visual (normalmente é usado um sinal luminoso de tráfego ou gráficos) a evolução ou o progresso daquele indicador de desempenho. A Tabela 3.9, lista alguns dos KPi's que serão implementado no sistema DW.

Para criara um **KPI** precisa-se desenvolver expressões que calculam o valor, meta, status atual e tendência do **KPI**.

- **Meta:** a expressão meta indica o objetivo a ser atingido.

- **Status:** o valor da expressão status normalmente é avaliado a partir de uma expressão **MDX**. Uma expressão de status válida precisa retornar um valor entre -1 e 1. O valor -1 atribuirá um status de sinal luminoso de trânsito vermelho ao **KPI**. Um valor 0 resultaria em um sinal luminoso de trânsito amarelo e um valor 1 resultaria em um sinal luminoso de trânsito verde.
- **Tendência:** também normalmente é definida como uma expressão **MDX** e deve retornar um valor entre -1 e 1. A finalidade da expressão de tendência é comparar um valor atual, definido pela expressão **Value** ao mesmo valor em um ponto anterior no tempo.

Tabela 3.9 KPI's para o sistema DW do IFRN

A Listagem 3.4 **MDX** mostra, como exemplo, as definições do **KPI** “Coeficiente de Rendimento Escolar”.

Listagem 3.4 Código MDX para o KPI Coeficiente de Rendimento Escolar

VALOR:

$(\text{SUM}(\text{Measures}).[\text{Media}]) * 100 / (\text{SUM}(\text{Measures}).[\text{Total Disciplinas}])$

META:

$(\text{SUM}(\text{Measures}).[\text{Media}]) * 100 / (\text{SUM}(\text{Measures}).[\text{Total Disciplinas}]) \geq 60$

STATUS:

Case

When $(\text{SUM}(\text{Measures}).[\text{Media}]) * 100 / (\text{SUM}(\text{Measures}).[\text{Total Disciplinas}]) < 30.0$ Then -1

When $((\text{SUM}(\text{Measures}).[\text{Media}]) * 100 / (\text{SUM}(\text{Measures}).[\text{Total Disciplinas}])) \geq 30.0$ AND

$(\text{SUM}(\text{Measures}).[\text{Media}]) * 100 / (\text{SUM}(\text{Measures}).[\text{Total Disciplinas}]) \leq 50.0$) Then 0

When $((\text{SUM}(\text{Measures}).[\text{Media}]) * 100 / (\text{SUM}(\text{Measures}).[\text{Total Disciplinas}])) \geq 60.0$ Then 1

End

TENDÊNCIA:

Case

When

```
((SUM(Measures].[Media])*100/([Measures].[Total  
Disciplinas]) >=  
((SUM(Measures].[Media])*100/([Measures].[Total  
Disciplinas]),[Data].[Calendar].PrevMember) ) then 1
```

When

```
((SUM(Measures].[Media])*100/([Measures].[Total  
Disciplinas]) <  
((SUM(Measures].[Media])*100/([Measures].[Total  
Disciplinas]),[Data].[Calendar].PrevMember) ) then -1
```

End

A Listagem 3.5 **MDX** mostra, como exemplo, as definições do **KPI** “Média” em função de suas notas obtidas em cada bimestre. Este KPI foi projetado em função dos dados do boletim escolar.

Listagem 3.5 Código MDX para o **KPI** Média dos alunos.

VALOR:

Media = (Nota1*1 + Nota2*2 + Nota3*3)/(5])

META:

Media = 60

STATUS:

Se media >= 60 então aprovado

Se não Se media 30 <= media < 60 então recuperação

Se não se media >= 30 e m < 60 então alerta (chama aluno para conversa)

Se não se media < 30 então alerta geral (chama os pais para conversa)

A declaração de tendência, compara a porção do período de tempo atual com o período de tempo anterior na hierarquia calendário (Em inglês *calendar*). Por exemplo, o período de tempo anterior a janeiro de 2010 é dezembro de 2009. Como também, se o período de tempo for o primeiro semestre de 2010, o período anterior será o segundo semestre de 2009.

3.5.4. Gerenciando Partições em Cubos OLAP

Quando é criado um cubo **OLAP**, o sistema criar para cada tabela de medidas, uma partição como padrão. No entanto, para banco de dados muito grande, em muitas situações, é desejado o particionamento do banco de dados.

Um dos benefícios de criar várias partições é que, pode-se projetar armazenagens diferentes para partes diferentes do cubo. Por exemplo, você pode projetar uma partição com informações mais recentes. Se essas informações são acessadas com frequência, então, pode ser especificado o armazenamento **OLAP** multidimensional (**MOLAP**) com agregações para proporcionar um aumento de desempenho de 50%. Uma outra partição poderia ser criada para armazenar as informações de dois, três ou vários anos anteriores, por exemplo. As informações desses anos, provavelmente, são acessadas a nível de totalizações e ocasionalmente, assim pode-se especificar a armazenagem (**HOLAP**), com agregações para um nível de desempenho de 30%.

Um outro benefício muito importante das partições, é você poder processar uma partição independentemente do restante do cubo. Imagine você, um cubo muito grande, que tenha que ser processado a cada 10 minutos. Em raríssima exceção, teríamos tempo de processar um grande cubo a cada dez minutos. Então, uma saída seria partitionar o cubo, de forma que as informações que precisam ser processadas a cada dez minutos, ficassem em uma partição independente, e então, aí sim, essa partição poderia ser processada isoladamente, a cada dez minutos, sem comprometer o desempenho do sistema.

Para este projeto, o cubo dados sociais da figura 3.13, tem vários fatos, formando uma constelação de fatos, sendo o maior cubo desse projeto. No entanto, cada fato não ultrapassa um total de 3000 linhas de dados. Dessa forma, não justifica ser fazer um particionamento desse cubo.

3.6. Mineração de dados

Devido à complexidade do processo de descoberta de conhecimento em base de dados (**KDD**), será adotada uma extensão ou adaptação da metodologia **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*), pois esse modelo foi um dos primeiros e bem aceitos para esse processo [Goldschmit, 2015].

A metodologia **CRISP** recomenda que a execução do processo ocorra de forma iterativa e interativa. Nessa execução, dependendo dos resultados alcançados, os especialistas em **KDD**, podem retornar a qualquer etapa realizada anteriormente para fazer refinamentos em busca de melhores resultados. Dessa forma, a metodologia requer uma documentação detalhada das ações realizadas e dos resultados produzidos.

O modelo **CRISP-DM**, utiliza dois conceitos de controles em sua aplicação: **sessão de KDD** e **ciclo de KDD**.

Cada sessão de **KDD** representa uma linha de raciocínio e de condução do processo. Toda sessão de **KDD** possui um, e somente um, objetivo a ser alcançado e, pode ser realizada em um ou mais ciclos de **KDD**. Cada ciclo envolve uma ou mais etapas de execução de um plano de ação e pertence a exatamente a uma sessão de **KDD**. Cada ciclo de **KDD** corresponde a uma iteração do processo de **KDD**.

A Tabela 7 mostra o formulário para o nosso primeiro estudo de caso que, a análise da repetência escolar.

Tabela 7. Formulário que documenta a execução do processo de KDD, no modelo CRISP.

Aplicação: Repetência Escolar	
Sessão: 1	
Objetivo: Previsão de comportamento de valores discretos e contínuos	Resumo: Esta sessão tem como objetivo extrair modelos baseados em dados do sistema acadêmico do IFRN , que consiga, através do relacionamento de seus atributos, traçar um perfil da repetência escolar
Tarefas de KDD: Árvore de Decisão	
Expectativa quanto ao modelo de conhecimento: <ul style="list-style-type: none">• Representação do modelo de forma transparente e fácil de interpretar• Representação em forma de árvore• Mapear os atributos que influenciam na repetência escolar	
Plano de ação: <ul style="list-style-type: none">• Utilizar o algoritmo Árvore de decisão• Aplicar os vários métodos de pontuação e comparar os resultados• Selecionar o melhor método de análise para ser usado pelo algoritmo árvore de decisão• Analisar o resultado obtido pelo método escolhido	

Ciclo nº 1			
Métodos			
Partição do BD em treino e teste	30% treino 70% teste		
Algoritmo: Microsoft Árvore de decisão			
Codificação continuo – categórica			
<u>Discretização:</u> quando a entrada é continua, elas são discretizadas automaticamente.			
Requisitos:			
<ul style="list-style-type: none"> Uma única coluna chave (key): Cada modelo deve conter uma coluna de texto ou numérica que identifique exclusivamente cada registro. Não são permitidas chaves compostas. Uma coluna previsível: Requer, pelo menos, uma coluna previsível. Você pode incluir vários atributos previsíveis em um modelo, e o atributo previsível pode ser de diferentes tipos, tanto continuo como discreto. Colunas de entrada: Requer colunas de entrada que podem ser discretas ou contínuas. 			
Métodos de análise:			
<ul style="list-style-type: none"> 1 Entropia de Shannon 3 Bayesian com K2 a priori 4 Bayesian Dirichlet com uniforme a priori (padrão) 			

O algoritmo Árvores de Decisão da Microsoft é um algoritmo de classificação e regressão fornecido pelo Microsoft SQL Server Analysis Services para uso em modelagens de previsão de atributos discretos e contínuos, da seguinte forma: [Bassan, 2014].

- **Atributo Discreto:** Para atributos discretos, o algoritmo faz previsões fundadas nas relações entre colunas de entrada em um conjunto de dados. Ele usa os valores, conhecidos como estados dessas colunas, para prever os estados de uma coluna que você define como previsível. Especificamente, o algoritmo identifica as colunas de entrada que são correlacionadas com a coluna previsível.
- **Atributo Contínuo:** No caso de atributos contínuos, o algoritmo usa a regressão linear para determinar onde uma árvore de decisão se divide.

Quando mais de uma coluna é definida como prevísil, ou se tiver uma tabela aninhada configurada como previsível, o algoritmo criará uma árvore de decisão separada para cada coluna previsível.

O modelo de mineração de dados gerado pelo algoritmo Árvores de Decisão da Microsoft é uma série de divisões/nós na árvore. Quando um atributo de entrada é considerado significativamente correlacionado a uma coluna prevísil, então é adicionado um nó ao modelo. A forma que o algoritmo determina uma divisão depende do fato de ele estar prevendo uma coluna contínua ou discreta.

O algoritmo Árvores de Decisão da Microsoft usa a *seleção de recurso* para guiar a seleção dos atributos mais úteis. A seleção de recurso é usada por todos os algoritmos de mineração de dados do Analysis Services para melhorar a análise e reduzir a carga de processamento. A seleção de recurso é importante para impedir que atributos sem-importância usem tempo do processador.

A seleção de recursos ajuda a resolver o problema, de ter dados demais de pouco valor, ou de ter pouquíssimos dados de alto valor.

A seleção de recurso, calcula uma pontuação para cada atributo e, em seguida, seleciona apenas os atributos com a pontuação mais alta. E isto sempre é feita antes do treinamento do modelo, para que, os atributos de um conjunto de dados com maior probabilidade de uso no modelo, sejam escolhidos de forma automática.

Segundo (Bassan, 2014), o algoritmo Árvore de Decisão da Microsoft, cria o conjunto de valores de entrada possíveis, ele executa feature selection para identificar os atributos e os valores que fornecem a maioria das informações e remove os valores considerados muito raros e, para otimizar o desempenho, ele agrupa esses valores em compartimentos, que podem ser processados como uma unidade.

Para criar uma árvores com as correlações das entradas e o resultado pretendido, o algoritmo, depois de ter correlacionado todos os atributos, identifica o único atributo que separa mais claramente os resultados. O ponto da melhor separação é medido com o uso de uma equação que calcula o ganho de informações. Portanto, o atributo escolhido é usado para dividir os casos em subconjuntos, que são analisados recursivamente pelo mesmo processo, até que não seja mais possível dividir a árvore.

A determinação da equação exata para avaliar o ganho de informações depende dos parâmetros definidos na criação do algoritmo, do tipo de coluna previsível e do tipo de dados de entrada.

O algoritmo Árvores de Decisão da Microsoft, oferece três fórmulas para pontuar o ganho de informação: **entropia de Shannon**, **rede bayesiana com antecedente K2** e **rede bayesiana** com uma **distribuição Dirichlet uniforme de antecedentes**. Neste trabalho, serão realizadas experiências com parâmetros e métodos de pontuação diferentes para determinar o método que fornece o melhor resultado.

3.6.1. Pontuação de Interesse

Por padrão, a pontuação de interesse é usada sempre que, a coluna contiver dados numéricos contínuos não binários [Bassan, 2014].

A medida de interesse usada no SQL Server Analysis Services baseia-se em entropia, ou seja, os atributos com distribuições aleatórias têm maior entropia e menor ganho de informações; sendo assim, esses atributos são menos interessantes. A entropia de qualquer atributo em particular é comparada à entropia de todos os outros atributos, como segue:

$$\text{Interestingness(Attribute)} = - (m - \text{Entropy(Attribute)}) * (m - \text{entropy(Attribute)}) \quad (\text{Equação 3.1})$$

Entropia central ou **m**, significa a entropia de todo o conjunto de recursos. Ao subtrair a entropia do atributo de destino da entropia central, é possível avaliar a quantidade de informações fornecida pelo atributo.

3.6.2. Entropia de Shannon

A entropia de Shannon mede a incerteza de uma variável aleatória para um resultado em particular. Dessa forma, a entropia pode ser representada como uma função da probabilidade de um evento ocorrer [Bassan, 2014].

O Analysis Services usa a seguinte fórmula para calcular a entropia de Shannon:

$$H(X) = -\sum P(x_i) \log(P(x_i)) \quad (\text{Equação 3.2})$$

3.6.3. Bayesiano com K2 a priori

Segundo Bassan (2014), o Analysis Services fornece duas pontuações para seleção de recursos que se baseiam em redes Bayesianas. Uma rede Bayesiana é um gráfico direcionado ou acíclico de estados e transições entre estados, ou seja, alguns estados vêm sempre antes do estado atual, alguns ocorrem depois, e o gráfico não se repete nem gera um loop. Por definição, as redes Bayesianas permitem o uso do conhecimento prévio.

O algoritmo **K2** usado com a rede Bayesiana foi desenvolvido por Cooper e Herskovits. É escalável e analisa diversas variáveis, mas requer ordenação das variáveis usadas como entrada [Bassan, 2014]. Esse método de pontuação está disponível para atributos discretos.

3.6.4. Bayesiano Dirichlet Equivalente com Uniforme a priori

O método de pontuação do **BDE** (Bayesiano Dirichlet Equivalente), foi desenvolvido por Heckerman e se baseia na métrica de BD desenvolvida por Cooper e Herskovits. A distribuição Dirichlet é uma distribuição multinomial que descreve a probabilidade condicional de cada variável da rede [Bassan, 2014].

O método **BDEU** (*Bayesiano Dirichlet Equivalente com Uniforme a priori*) assume um caso especial da distribuição *Dirichlet*, na qual uma constante matemática é usada para criar uma distribuição fixa ou uniforme de estados anteriores. A pontuação do **BDE** também assume equivalência de probabilidade, o que significa que não se pode esperar que os dados separem estruturas equivalentes. Em outras palavras, se a pontuação de **If A Then B** for igual à pontuação de **If B Then A**, não será possível distinguir as estruturas com base nos dados nem deduzir a causa [Bassan, 2014].

3.6.5. Personalizando o Algoritmo Microsoft Árvore de Decisão

O algoritmo Árvores de Decisão da Microsoft tem parâmetros que afetam o desempenho e a precisão do modelo de mineração resultante. A Tabela 8

descreve os parâmetros que você pode usar com o algoritmo Árvores de Decisão da Microsoft.

Tabela 8 Parâmetros do algoritmo Microsoft Árvore de Decisão

Parâmetro	Descrição
COMPLEXITY_PENALTY	Controla o crescimento da árvore de decisão. Um valor baixo aumenta o número de divisões e um valor alto diminui o número de divisões. <ul style="list-style-type: none"> • Para os atributos 1 a 9, o padrão é 0,5. • Para 10 a 99 atributos, o padrão é 0,9. • Para 100 ou mais atributos, o padrão é 0,99.
FORCE_REGRESSOR	Força o algoritmo a usar as colunas especificadas como regressores, independentemente da sua importância quando calculadas pelo algoritmo. Esse parâmetro é usado apenas para árvores de decisão que preveem um atributo contínuo.
MAXIMUM_INPUT_ATTRIBUTES	Define o número de atributos de entrada que o algoritmo pode manipular antes de invocar a seleção de recurso.
MAXIMUM_OUTPUT_ATTRIBUTES	Define o número de atributos de saída que o algoritmo pode manipular antes de invocar a seleção de recurso.
MINIMUM_SUPPORT	Determina o número mínimo de casos folha necessário para gerar uma divisão na árvore de decisão.

SCORE_METHOD

Determina o método o usado para calcular a pontuação da divisão. As seguintes opções estão disponíveis:

ID	Nome
1	Entropia
3	Bayesian com K2 a priori
4	Bayesiano Dirichlet Equivalente (BDE) com uniforme a priori

O padrão é 4 ou BDE.

SPLIT_METHOD

Determina o método usado para dividir o nó. As seguintes opções estão disponíveis:

ID	Nome
1	Binary: Indica que, independentemente do número real de valores do atributo, a árvore deverá ser dividida em duas ramificações.
2	Complete: Indica que a árvore pode criar tantas divisões quanto há valores de atributo.
3	Both: Especifica que o Analysis Services pode determinar se uma divisão binária ou completa deve ser usada para produzir os melhores resultados.

O padrão é 3.

3.6.6. Aplicar o Algoritmo Árvore de Decisão

Seguindo o plano de ação da Tabela 3.10, vamos executar o primeiro ciclo, para prever a situação da repetência escolar. Para esse primeiro caso, vamos utilizar os dados de uma tabela, cujos atributos estão descrevidos na Tabela 9.

Tabela 9 Relação de atributos da tabela onde será aplicado o algoritmo Árvore de Decisão para análise.

ATRIBUTO	DESCRIÇÃO	TIPO
AREA PROCEDENCIA	Urbana, Rural	Discreta
ESCOLA ORIGEM		
AULAS DADAS	Quantidade de aulas ministradas	Continuo
COEFICIENTE DE RENDIMENTO	Desempenho do aluno no curso	Continuo
ETNIA	Branca, Preta, Parda e amarela	Discreto
FALTAS	Total de faltas no semestre	Continuo
FORMA_INGRESSO	Exame seleção, ENEM, SISU a assim por dante.	Discreto
IDADE	Idade do aluno	Continuo
MEDIA FINAL	Média final do aluno	Continuo
RENDAA	Valor em real	Continuo
RENDAA_FAMILIAR	Valor textual: até 1 salário, de 1 até 2 salários e assim por diante.	Discreto
RESIDE	Com quem o aluno reside: com os pais ou não.	Discreto
SEXO	Masculino ou feminino	Discreto
SITUACAO	Aprovado, reprovado, jubilado, cancelado, evasão, etc.	Discreto
TIPO ESCOLA	Pública estadual, federal, municipal	discreto
ORIGEM	ou privada.	

Para este estudo de caso, será escolhido o atributo “Situacao” para ser o atributo previsível e os demais serão informados como atributos de entrada. O método usado para calcular a pontuação de divisão, foi escolhido a Entropia e o método especificado para dividir os nós foi o binário. A Figura 3.22 mostra a

árvore gerada para o atributo previsível “Situacao=Reprovado” e a Listagem 3.6 mostra o código **DMX** (Data Mining Extensions) usado para criar a estrutura de mineração de dados e define o modelo a ser usado pela estrutura.

Listagem 3.6 Código **DMX** para criar a estrutura de mineração de dados e os modelos (algoritmo) usados pela estrutura.

--Criando uma estrutura de mineração de dados

CREATE MINING STRUCTURE [Mine TB Juntando Dados Alunos]

(

[Codigo Curso] **LONG KEY**,
[Area Procedencia Escola Origem] **TEXT DISCRETE**,
[Aulas Dadas] **LONG CONTINUOUS**,
[Coeficiente Rendimento] **DOUBLE DISCRETE**,
[Disciplina] **TEXT DISCRETE**,
[Etnia] **TEXT DISCRETE**,
[Faltas] **LONG CONTINUOUS**,
[Forma Ingresso] **TEXT DISCRETE**,
[Idade] **LONG CONTINUOUS**,
[Media Final] **DOUBLE CONTINUOUS**,
[Renda Familiar] **TEXT DISCRETE**,
[Reside] **TEXT DISCRETE**,
[Sexo] **TEXT DISCRETE**,
[Situacao] **TEXT DISCRETE**,
[Situacao Ing] **TEXT DISCRETE**

)

WITH HOLDOUT (30 PERCENT or 20000 **CASES**)

--Após criar a estrutura de mineração de dados, adiciona-se o algoritmo (modelo) desejado.

--Adicionando o modelo de árvore de decisão

ALTER MINING STRUCTURE [Mine TB Juntando Dados Alunos]

ADD MINING MODEL [Decision Tree]

(

[Codigo Curso],
[Area Procedencia Escola Origem],
[Aulas Dadas],
[Coeficiente Rendimento],
[Disciplina],
[Etnia],
[Faltas],
[Forma Ingresso],
[Idade],
[Media Final],
[Renda Familiar],
[Reside],
[Sexo],
[Situacao] **PREDICT**,
[Situacao Ing] **PREDICT**

) **USING Microsoft_Decision_Trees** (algoritmo parâmetros)

WITH DRILLTHROUGH

--Adicionando um modelo cluster

ALTER MINING STRUCTURE [Mine TB Juntando Dados Alunos]

ADD MINING MODEL [Clustering]

USING Microsoft_Clustering

WITH DRILLTHROUGH

Para treinar o modelo da estrutura foi usado o código **DMX** da listagem 3.7.

Listagem 3.6 Treinando o modelo da estrutura criada na listagem 3.6.

--Treinando o modelo

INSERT INTO STRUCTURE [Mine TB Juntando Dados Alunos]

(

[Codigo Curso],
[Area Procedencia Escola Origem],
[Aulas Dadas],
[Coeficiente Rendimento],
[Disciplina],
[Etnia],
[Faltas],
[Forma Ingresso],
[Idade],
[Media Final],
[Renda Familiar],
[Reside],
[Sexo],
[Situacao],
[Situacao Ing]

)

OPENQUERY (DDSEducacional),

'SELECT [Codigo Curso], [Area Procedencia Escola Origem],
[Aulas Dadas], [Coeficiente Rendimento],
[Disciplina], [Etnia], [Forma Ingresso], [Idade], [Media Final], [Renda
Familiar], [Reside], [Sexo], [Situacao], [Situacao Ing]
FROM TB_JuntandoDadosAlunos')

A Figura 29 mostra a árvore de decisão criada para esse modelo. Como se pode ver ela tem duas ramificações. Clicando no nó “Renda familiar = Até 1 salário”, temos, como mostra a tabela “Mining Legend” que a taxa de Repetência é de 39,47%.

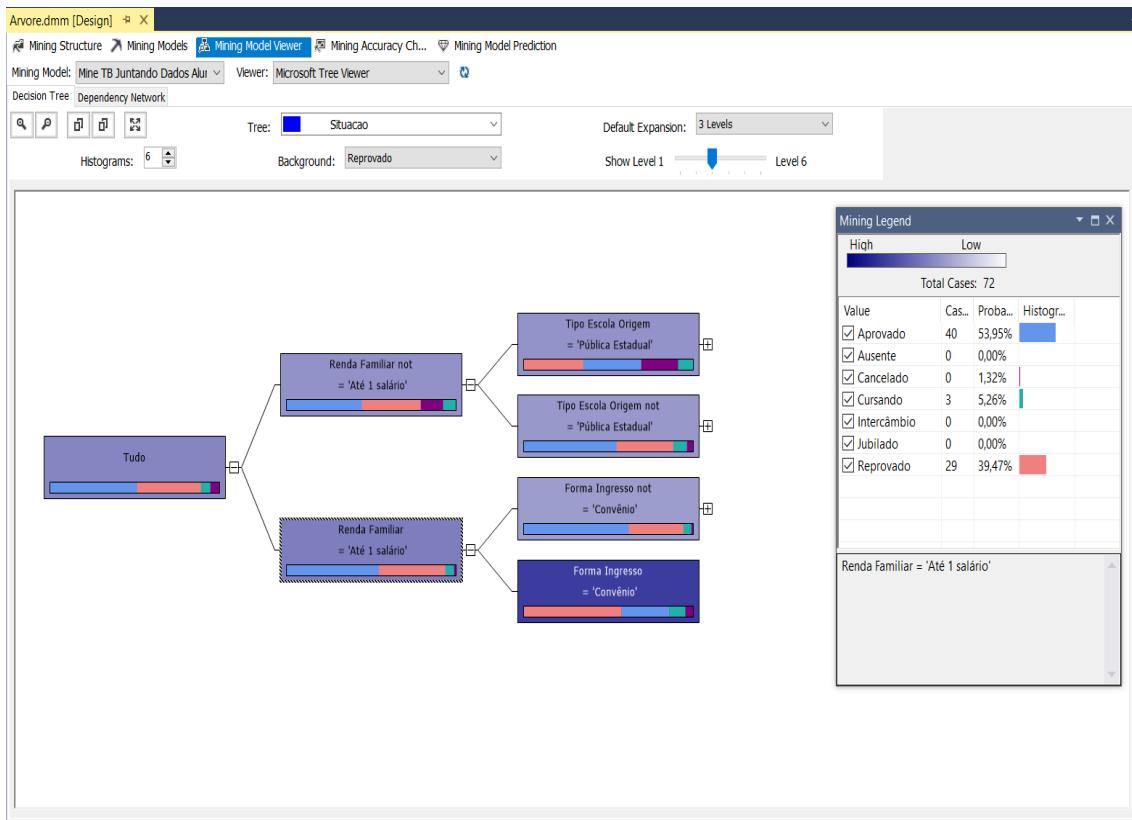


Figura 29. Árvore de Decisão para o atributo previsível “Situacao=Reprovado”.

Fonte: Autor.

O modelo de conhecimento, que se pode extraír dessa árvore, é mostrado no código **SQL** da Listagem 3.8:

Listagem 3.8 Código do modelo extraído da árvore de decisão da Figura 27

```

IF ([Renda Familiar] = 'Até 1 salário') Then
    Probabilidade de reprovação = 39,47%
    IF ([Forma Ingresso]='Convênio') Then
        Probabilidade de reprovação = 57,14%
    ELSE ([Forma Ingresso]!='Convênio') Then
        Probabilidade de reprovação = 25,53%
    IF ([Forma Ingresso] ='Seleção Geral Curso FIC') then
        Probabilidade de reprovação = 50%
    ELSE ([Forma Ingresso] !='Seleção Geral Curso FIC') Then
        Probabilidade de reprovação = 32%
        IF (Sexo = 'M') Then
            Probabilidade de reprovação = 23,8%
        ELSE (Sexo = 'F') Then
            Probabilidade de reprovação = 28%
        IF (Etnia = 'Branca') Tjen
            Probabilidade de reprovação = 33,33%
        EISE
            Probabilidade de reprovação = 21,43%
    ELSE
        Probabilidade de reprovação = 34,62%

```

```

IF ([Tipo Escola Origem] = 'Pública Estadual') Then
    Probabilidade de reprovação = 34,38%
IF ([Renda Familiar] = 'De 1 até 2 salários') Then
    Probabilidade de reprovação = 29,41%
ELSE
    Probabilidade de reprovação = 36,84%
ELSE ([Tipo Escola Origem] != 'Pública Estadual') Then
    Probabilidade de reprovação = 33%
IF (Reside = 'Com os pais') Then
    Probabilidade de reprovação = 28,57%
ELSE
    Probabilidade de reprovação = 35,71%

```

Observe que a árvore da Figura 29 tem duas ramificações, que estão representadas no modelo pelo **IF...ELSE**.

Como foi explicado anteriormente, a árvore gera para cada atributo de entrada, que tenha correlação significativa com o atributo previsível um nó, como mostrado na Figura 29. Portanto, resta agora verificar quais atributos apresentam correlação significativa com o atributo previsível, ou seja, quais atributos mais influenciam a formação de nós na árvore.

A Figura 30 mostra, através de um gráfico de correlação, os atributos que mais influenciaram na situação reprovado. Observe que, do lado esquerdo, tem uma barra de rolagem e que a mesma está posicionada no topo. Isto significa que o gráfico está exibindo todos os atributos de entrada que têm correlação significativa com o atributo previsível e, portanto, foram usados para gerar nós na árvore. Se deslocarmos a barra para baixo, ela vai desconectando os atributos que influenciam menos e, deixando aqueles que têm mais influência na formação da árvore. Então, posicinando a barra de rolagem além do centro, temos o resultado mostrado na Figura 31.

Na Figura 30, pode-se observar que, os atributos de entrada que mais influenciaram na situação (atributo previsível), foram os atributos forma de ingresso, faltas e media final.

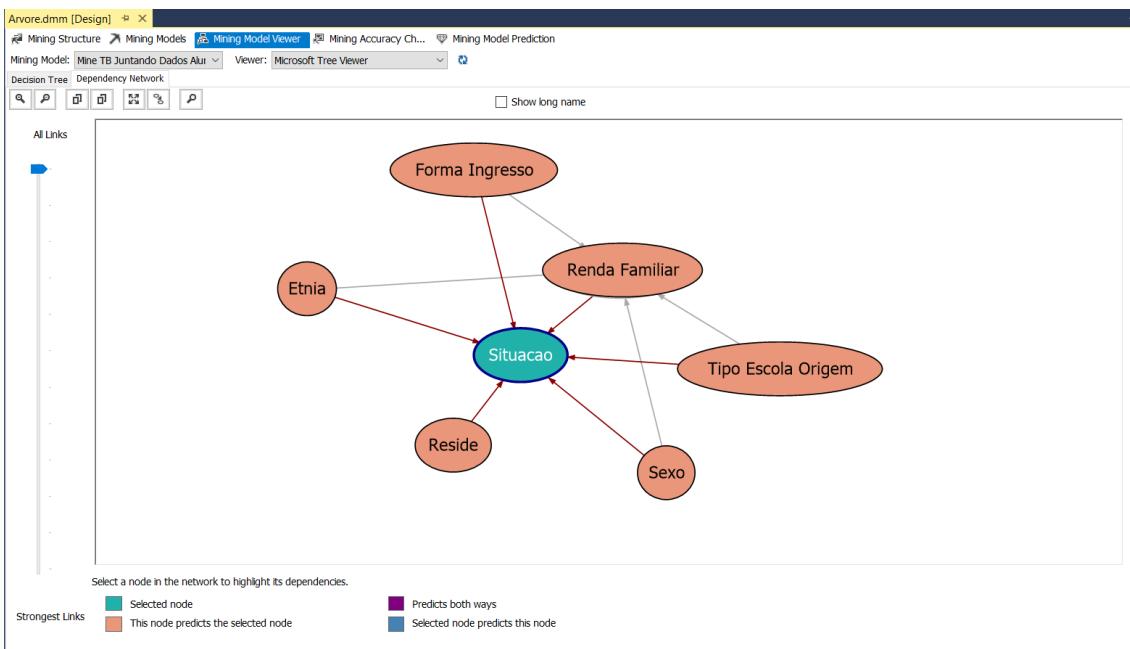


Figura 30. Rede de dependência dos atributos em relação ao atributo situação (Reprovado).

Fonte: Autor.

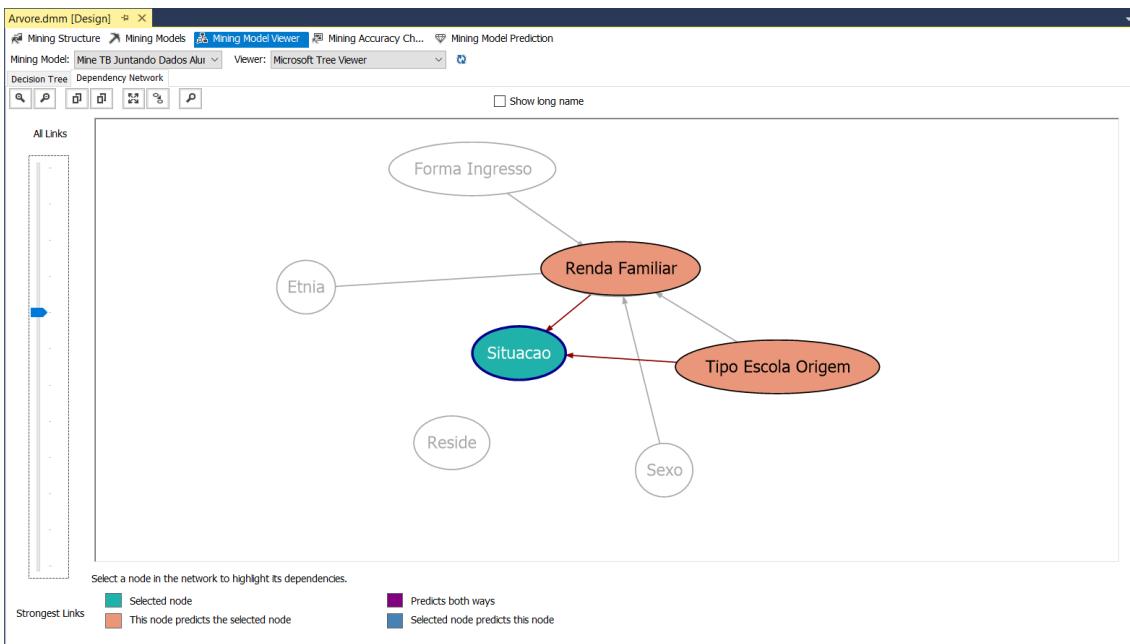


Figura 31. Rede de dependência dos atributos em relação ao atributo situação (Reprovado).

Fonte: Autor.

Usando o mesmo estudo de caso, foram criadas outras árvores de decisão, usando os mesmos valores da Tabela 3.10, apenas modificando o parâmetro “SCORE_METHOD”, que determina como os nós da árvore são

divididos e, observou-se que, para esse mesmo conjunto de dados, os resultados foram muito parecidos. Então, para que pudesse ser comprovada esta semelhança entre os resultados obtidos, foi verificada a acurácia entre os métodos selecionados. A Figura 32 mostra um gráfico, onde pode-se ver que, os métodos obteram resultados muito próximos um do outro.

Veja na Figura 32 que as linhas laranja, verde e roxa estão muito próximas uma da outra. Se você observar a tabela gerada pela ferramenta, você vai ver que a probabilidade é de 95,46% para o método padrão, 95,37% para o método de entropia e de 95,04% para o método K2 a priori. Portanto, não faz diferença em usar qualquer um dos métodos de pontuação, para este estudo de caso, é claro.

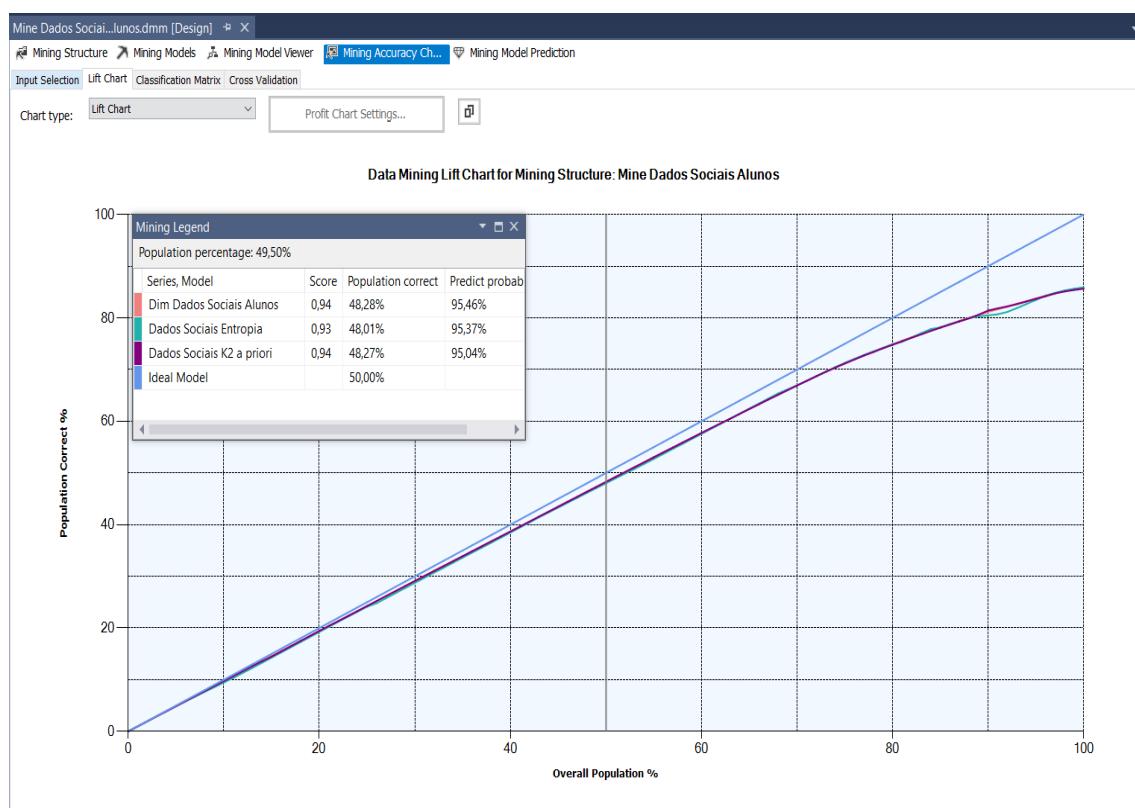


Figura 32. Gráfico de acurácia entre os métodos de divisão dos nós da árvore de decisão.

Fonte: Autor.

Na configuração do algoritmo Árvore de Decisão, foram definidos dois atributos como previsíveis, o atributo “Situacao” e o atributo “SituacaoIng”. Isto faz com que o algoritmo gere duas árvores de decisão, uma para cada atributo previsível. No caso, anterior, foi analisada uma das árvores criada, a que estava

relacionada a repetência escolar. A segunda árvore, que será analisada a partir desse momento, os atributos de entrada estam correlacionados a evasão escolar. A Figura 33 mostra a árvore de decisão criada para a situação evasão escolar.

O atributo previsível “Situacao Ing” possui vários estados, dentre eles pode-se citar Cancelado, Jubilado, Concluído, evasão, intercâmbio e ausente. No entanto, será analisada a situação evasão escolar. O método usado para calcular a pontuação de divisão, foi escolhido a Entropia e o método especificado para dividir os nós foi o binário.

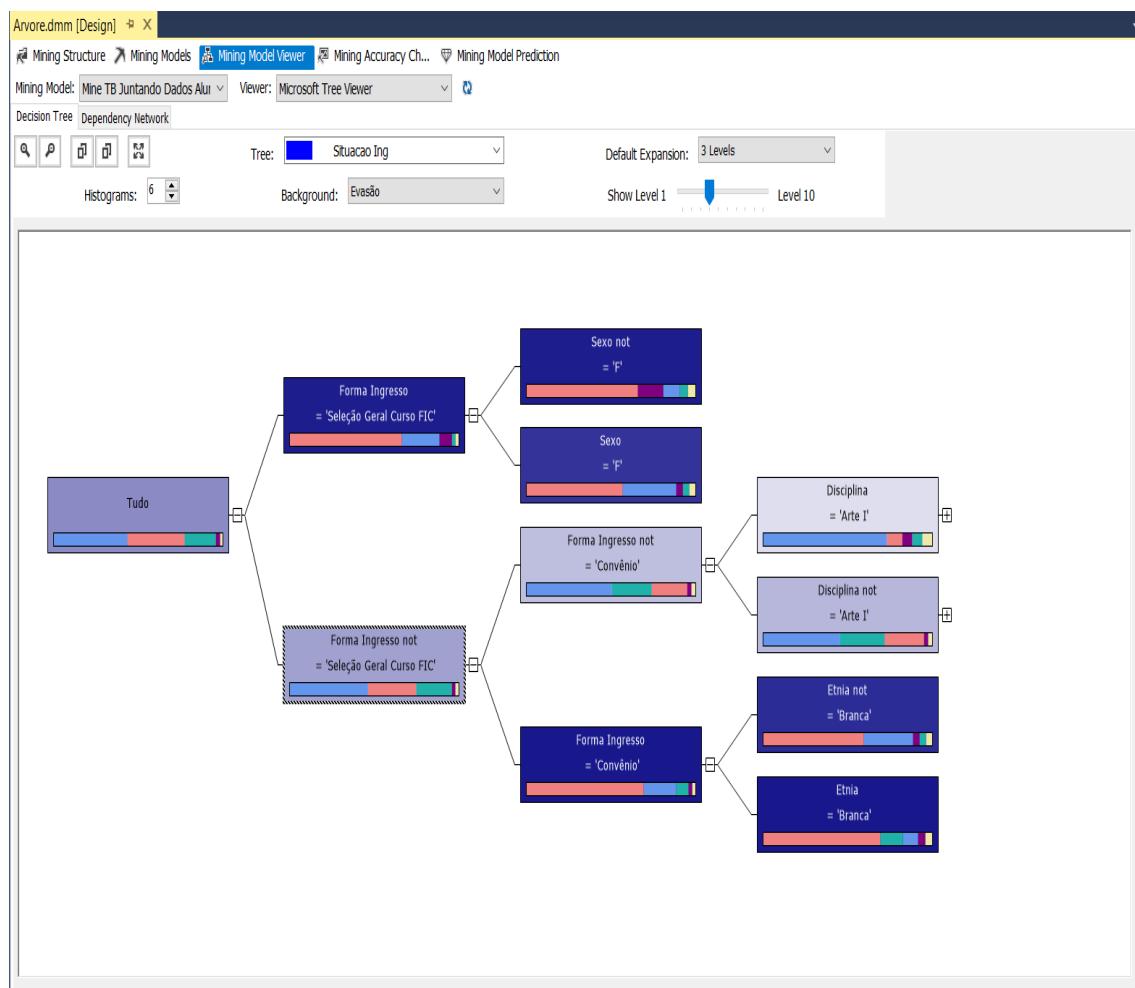


Figura 33. Árvore de Decisão usando o método de pontuação a Entropia.

Fonte: Autor.

O modelo de conhecimento, que se pode extraír dessa árvore é mostrado no código SQL da Listagem 3.9:

Listagem 3.9 Código do modelo extraído da árvore de decisão da Figura 31

```
IF ([Forma Ingresso] = 'Seleção Geral Curso FIC') Then
  IF (Sexo = 'F') Then
```

```

        evasão = 13%
    ELSE
        evasão = 12%
ELSE
    IF ([Forma Ingresso] = 'Convênio') Then
        IF (Etnia = 'Branca') Then
            evasão = 14%
        ELSE
            evasão = 13%
    ELSE
        IF (Disciplina = 'Arte 1') Then
            evasão = 2%
        Else
            evasão = 42%
        IF ([Forma Ingresso] = 'Seleção Dif. Téc. Subsequente') Then
            evasão = 8%
            IF (Sexo = 'F') Then
                evasão = 1%
            ELSE
                evasão = 7%
        ELSE
            evasão = 34%
            IF (Reside = 'Conjuge') Then
                evasão = 8%
                IF (Etnia = 'Branca') Then
                    evasão = 1%
                ELSE
                    evasão = 7%
            ELSE
                evasão = 26%
                IF ([Forma Ingresso] = 'Seleção Geral Téc. Sub.') Then
                    evasão = 2%
                ELSE
                    evasão = 24%
                    IF ([Forma Ingresso] = 'Seleção Dif. Grad. Vest.') Then
                        evasão = 4%
                    ELSE
                        evasão = 20%
                        IF (Etnia = 'Parda') Then
                            evasão = 9%
                            IF ([Tipo escola origem] = 'Pública Estadual') Then
                                evasão = 5%
                            ELSE
                                evasão = 4%
                        ELSE
                            evasão = 11%
                        IF ([Forma Ingresso] = 'Seleção Dif. Téc. Integ./PROITEC') Then
                            evasão = 1%
                        ELSE
                            evasão = 10%

```

O modelo da listagem 3.9 representa os dois ramos da árvore de decisão da Figura 33. Como se pode ver, a árvore gera para cada atributo de entrada, que tenha correlação significativa com o atributo previsível um nó.

A Figura 34 mostra, através de um gráfico de correlação, os atributos que mais influenciaram na situação evasão.

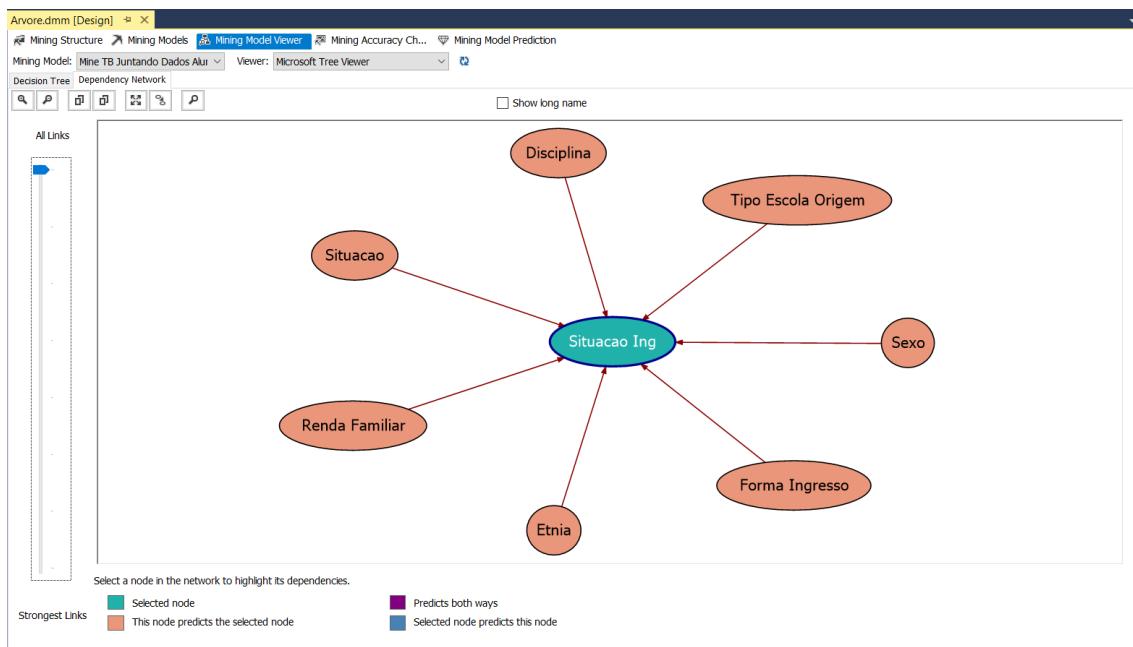


Figura 34. Gráfico de correlação dos atributos de entrada com o atributo previsível “Situacao Ing” destacado na cor verde.

Fonte: Autor.

Na Figura 34 os atributos que estão destacados na cor laranja são justamente os atributos de entrada que tem correlação com o atributo previsível.

O próximo passo é, utilizar o algoritmo de clustering para verificar o percentual de influencia de cada atributo de entrada, na formação do atributo previsível “Situacao Ing”.

3.6.7. Aplicando o Algoritmo Cluster

Vamos aplicar um algoritmo de clusterização sobre o mesmo conjunto de dados usado anteriormente, para tentar mapear o quanto cada atributo da base de dados selecionada, influencia na situação “Reprovado”.

O algoritmo Microsoft Clustering é um algoritmo de segmentação. Ele usa técnicas iterativas para agrupar casos em um conjunto de dados em clusters que contenham características semelhantes. Esses agrupamentos podem ser

usados para explorar dados, identificar anomalias nos dados e criar previsões [Bassan, 2014].

Diferentemente do algoritmo Árvore de Decisão, não é necessário designar um atributo privisível para poder criar o modelo de clustering. O mesmo treina o modelo a partir das relações existentes nos dados e a partir dos clusters que o algoritmo identifica.

O algoritmo do Microsoft Clustering, fornece dois métodos para criar clusters e atribuir pontos de dados aos clusters. O primeiro, o algoritmo **K-means**, um método de clustering rígido. Isso significa que um ponto de dados pode pertencer somente a um cluster e que, uma única probabilidade é calculada para a associação de cada ponto de dados nesse cluster. O segundo método, de *Maximização de Expectativa (EM)*, é um método de *clustering flexível*. Isso significa que, um ponto de dados sempre pertence a vários clusters e que uma probabilidade é calculada para cada combinação de ponto de dados e cluster [Bassan, 2014].

Pode-se escolher o algoritmo que será usado definindo o parâmetro **CLUSTERING_METHOD**. O método padrão de cluster é o **EM** evolutivo.

3.6.7.1. Cluster EM

No cluster **EM**, o algoritmo refina de modo iterativo, um modelo de clustering inicial para ajustar os dados e determina a probabilidade de um ponto de dados existir no cluster. O algoritmo termina o processo quando, o modelo de probabilidade ajusta os dados. A função usada para determinar o ajuste é a probabilidade de log dos dados de acordo com o modelo [Bassan, 2014].

Se clusters vazios forem gerados durante o processo ou se a associação de um ou mais clusters estiver abaixo de um determinado limite, os clusters com baixas populações serão propagados novamente em novos pontos e o algoritmo **EM** será executado mais uma vez [Bassan, 2015].

Os resultados do método de cluster **EM** são probabilidades. Isso significa que, cada ponto de dados pertence a todos os clusters, mas cada atribuição de um ponto de dados a um cluster, tem uma probabilidade diferente. Como o método permite a sobreposição dos clusters, a soma dos itens de todos os clusters pode ultrapassar o total de itens do conjunto de treinamento [Bassan, 2014].

3.6.7.2. Cluster K-means

O algoritmo **K-means**, atribui cada ponto de dados exatamente a um cluster e não permite incertezas na associação. A associação em um cluster é expressa como uma distância do centroide [Bassan, 2014].

De acordo com Bassan (2014), o algoritmo **K-means** é usado para criar clusters de atributos contínuos, onde calcular a distância até o centro é simples. No entanto, a implementação do Microsoft, adapta o método **K-means**, para atributos de distinção de cluster usando probabilidades. Para os atributos de distinção, a distância de um ponto de dados a partir de um cluster específico é calculada da seguinte maneira **P(data point, cluster)**.

3.6.7.3. Personalizando o Algoritmo Cluster

O algoritmo Microsoft Clustering dá suporte a vários parâmetros que afetam o comportamento, o desempenho e a exatidão do modelo de mineração resultante.

A Tabela 10 descreve os parâmetros que podem ser usados com o algoritmo do Microsoft Clustering. Esses parâmetros afetam o desempenho e a exatidão do modelo de mineração resultante.

Tabela 10 Especifica o método de cluster para o algoritmo a ser usado.

Parâmetro	Descrição	Padrão
CLUSTERING_METHOD	Especifica o método de cluster para o algoritmo a ser usado.	O padrão é 1.
	1 - EM evolutivo	
	2 - EM não evolutivo	
	3 - K-Means evolutivo	
	4 - K-means não evolutivo	
CLUSTER_COUNT	Especifica o número aproximado de clusters a serem criados pelo algoritmo.	O padrão é 10.

CLUSTER_SEED	Especifica o número de O padrão é propagação, usado apenas para 0. gerar clusters aleatoriamente para o estágio inicial de criação de modelo.
MINIMUM_SUPPORT	Especifica o número mínimo de O padrão é casos necessários para criar um 1 cluster. Se o número de casos do cluster for menor que esse número, o cluster será tratado como vazio e descartado.
MODELLING_CARDINALITY	Especifica o número de modelos de O padrão é exemplo construídos durante o 10. processo de cluster.
STOPPING_TOLERANCE	Especifica o valor usado para O padrão é determinar quando a convergência 10. é alcançada e, o algoritmo terminou de criar o modelo. A convergência é alcançada quando, a alteração geral nas probabilidades do cluster é menor do que a taxa do parâmetro STOPPING_TOLERANCE dividida pelo tamanho do modelo.
SAMPLE_SIZE	Especifica o número de casos, que O padrão é o algoritmo usará em cada 50.000. passagem, se o parâmetro CLUSTERING_METHOD for definido como um dos métodos de cluster evolutivo.
MAXIMUM_INPUT_ATTRIBUTES	Especifica o número máximo de O padrão é atributos de entrada, que o 255.

	algoritmo pode manipular, antes de invocar a seleção de recurso.
MAXIMUM_STATES	Especifica o número máximo de estados de atributo aos quais, o algoritmo dá suporte. Se um atributo tiver mais estados que o valor máximo, o algoritmo usará os estados mais populares e ignorará os demais estados.

A Tabela 11 mostra o formulário para o nosso primeiro estudo de caso que, é sobre a repetência escolar, para o segundo ciclo no modelo CRISP.

Tabela 11 Formulário que documenta a execução do processo de KDD.

Aplicação: Repetência Escolar	
Sessão: 1	
Objetivo: identificam as relações em um conjunto de dados	Resumo: Identificar as relações entre os dados do conjunto de dados e, gera uma série de clusters com base nelas. Fazer a representação dos clusters através de gráfico. Verificar observando os clusters, quais atributos mais influenciam a repetência escolar.
Tarefas de KDD: Cluster	Expectativa quanto ao modelo de conhecimento: <ul style="list-style-type: none"> • Representação do modelo de forma transparente e fácil de interpretar • Representação em forma de gráfico de dispersão • Representar os atributos que mais influenciam na repetência escolar
Plano de ação:	<ul style="list-style-type: none"> • Utilizar o algoritmo cluster • Aplicar os algoritmos EM e K-means e analisar os resultados obtidos • Selecionar o melhor resultado ou melhor algoritmo
Ciclo nº 2	
Métodos	
Partição do BD em treino e teste	30% treino 70% teste

Algoritmo: Microsoft Cluster

Codificação numérica	-			
categórica (Discretização)				

Parâmetros:

- CLUSTERING_METHOD = 1, 2, 3, 4
- CLUSTER_COUNT = 10
- MAXIMUM_STATES = 10

Nota: Os demais parâmetros foram definidos com o valor padrão.

Métodos de clusterização:

- 1 EM evolutivo (padrão)
- 2 EM não evolutivo
- 3 K-Means evolutivo
- 4 K-Means não evolutivo

A Figura 35 mostra o gráfico de dispersão, gerado pelo algoritmo cluster **EM Evolutivo**, para o conjunto de dados especificado, com a configuração dos parâmetros definidos na Tabela 3.14.

Observe na Figura 35, que os clusters com maior número de casos de reprovados são o cluster 3 e 6. A cor azul escura, indica exatamente o cluster onde ocorreram o maior número de casos para situação selecionada na opção “state”.

A Figura 36 mostra o gráfico gerado usando o método de clusterização o EM não Evolutivo. Para este modelo, os clusters com maior concentração de casos de repetência escolar, foram os clusters 6 e 9. O cluster 9 com 67% de reprovados e o cluster 3 com 57% de reprovados e o cluster 7 0% de reprovados.

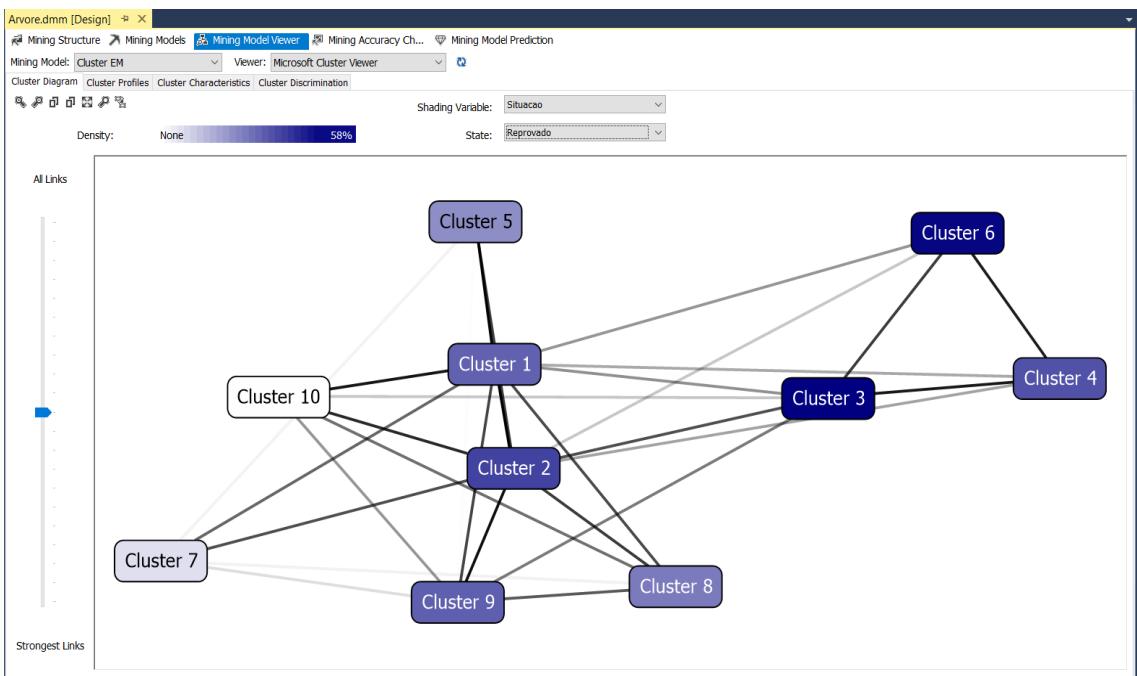


Figura 35. Gráfico dos cluster gerados para a situação “Reprovado” usando o método de clusterização o EM Evolutivo.

Fonte: Autor.

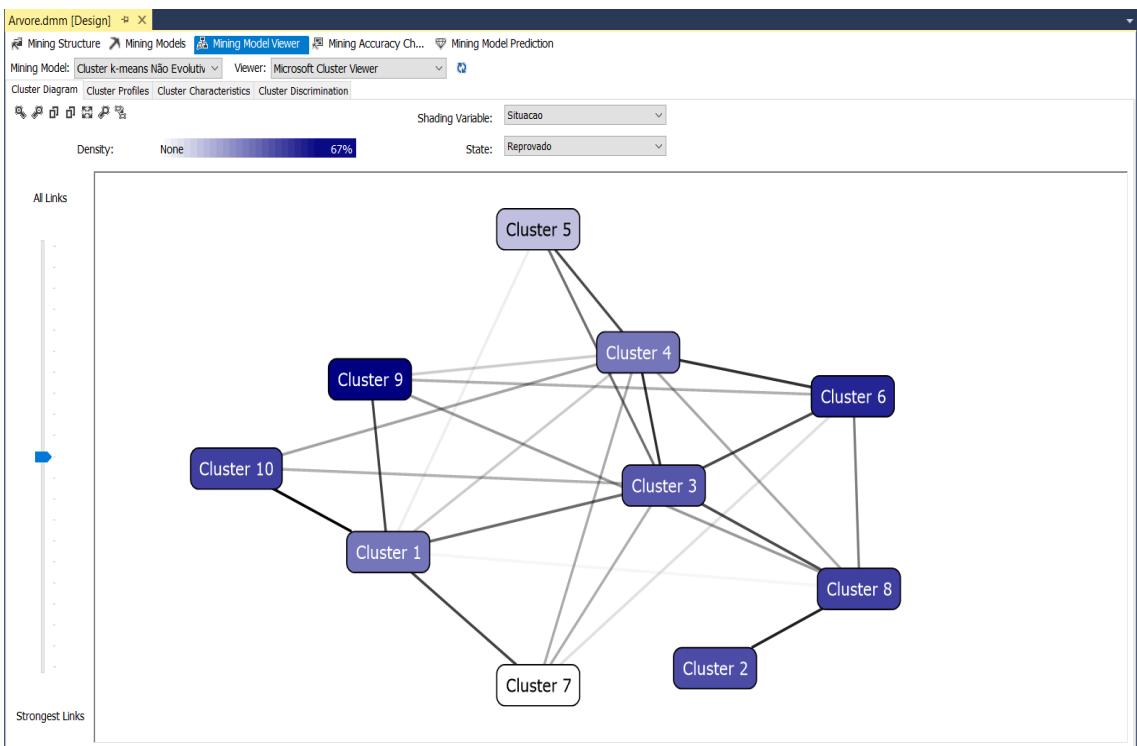


Figura 36. Cluster gerado com o método EM não Evolutivo.

Fonte: Autor.

Comparando com a situação anterior, onde o cluster 3 tem 58% de reprovados e o cluster 6 tem 57% de reprovados, pode-se perceber que os resultados obtidos pelos métodos **EM Evolutivo** e o **EM Não Evolutivo** são bastante próximos.

A Figura 37 mostra o gráfico gerado pelo método **K-means** Evolutivo para o mesmo DataSet e as mesmas configurações do métodos anteriores.

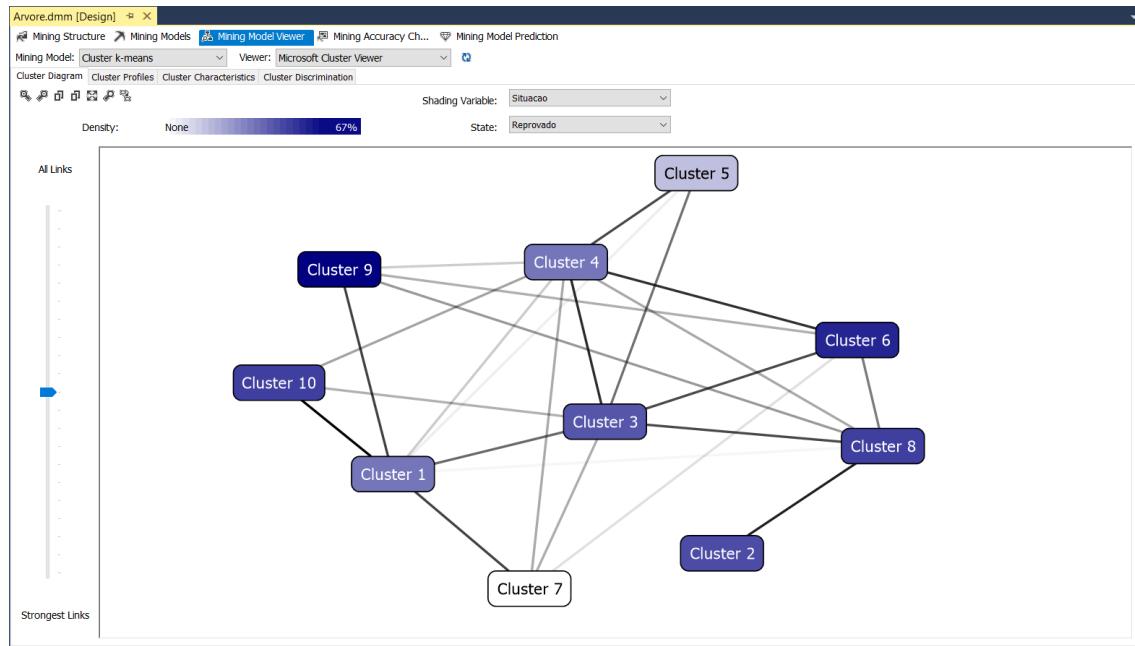


Figura 37. Gráfico de clusters gerado pelo método K-means Evolutivo.

Fonte: Autor.

Mais uma vez os clusters 6 e 9 aparecem com a maior concentração de casos de repetência escolar. Ness caso o cluster 9 tem 67% de reprovados e o cluster 6 57% de reprovados, enquanto o cluster 7 tem 0% de reprovados.

A Figura 38 mostra o gráfico gerado pelo método **K-means** Não Evolutivo para o mesmo DataSet e as mesmas configurações do métodos anteriores.

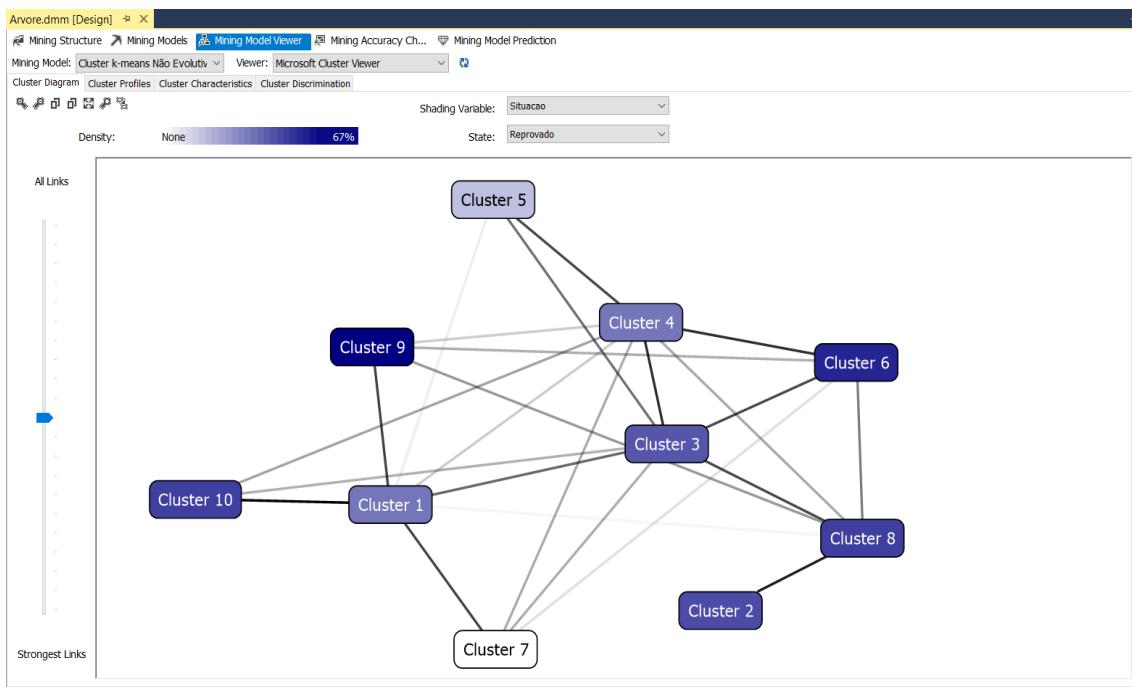


Figura 38. Gráfico de clusters gerado pelo método K-means Não Evolutivo.

Fonte: Autor.

Como se pode perceber, o resultado é praticamente o mesmo do gerado pelo método K-means Evolutivo. O cluster 9 tem 67% e o cluster 6 tem 57% dos casos de repetência escolar. Conclui-se que, os algoritmos K-means evolutivo e não evolutivo, produzem resultados semelhantes para esse DataSet avaliado.

A Figura 39, mostra o gráfico de acurácia entre estes algoritmos de clusterização. No gráfico da Figura 39 a linha azul representa a situação ideal. Então, os métodos geraram resultado longe do ideal. Porém, os quatro métodos geraram resultado muito próximo um do outro, como pode-se ver no gráfico de acurácia da Figura 39. A seguir será feita uma análise dos clusters 6 e 9, para se ter uma ideia o quanto cada atributo de entrada influencia na formação desses clusters.

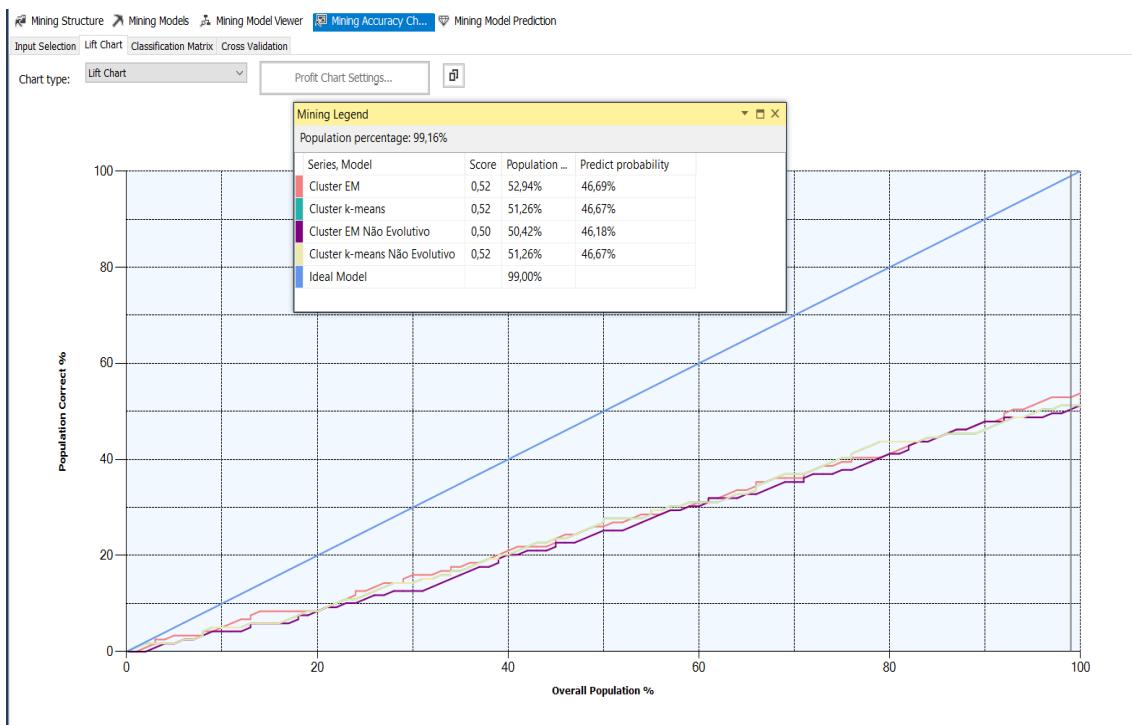


Figura 39. Gráfico comparativo entre os métodos de cluster.

Fonte: Autor.

A Figura 40 mostra o perfil de cada cluster, o seja, mostra exatamente o quanto cada atributo está influenciando a formação do cluster. A Tabela 12 detalha o quanto cada atributo participa percentualmente na composição dos cluster 6 e 9, justamente esses são os clusters estarem a maior concentração de reprovados.

Tabela 12 Tabela comparativa entre os clusters 6 e 9 da Figura 38.

CLUSTER	ATRIBUTO	VALOR
6	Escola Origem Pública Estadual	57%
6	Situação Reprovado	57%
6	Sexo “F”	57%
6	Reside com os pais	58%
6	Renda familiar até 1 mínimo	64,3%
9	Escola Origem Pública Municipal	66,7%
9	Situação Reprovado	67%
9	Sexo “F”	66%
9	Reside com os pais	60%
9	Renda familiar até 1 mínimo	68%

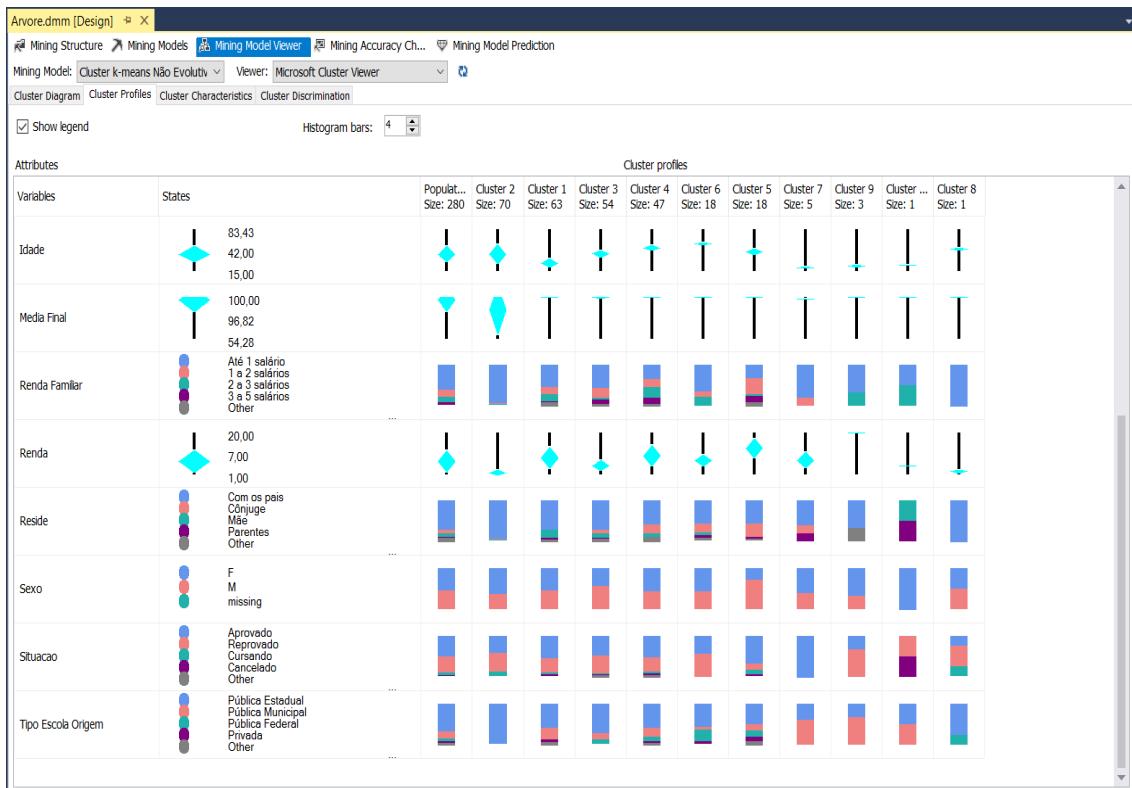


Figura 40. Perfis de Cluster algoritmo EM evolutivo.

Fonte: Autor.

Portanto, o perfil dos alunos com maior probabilidade de repetência escolar são aqueles oriundos das escolas públicas estaduais ou municipais, cuja família tem renda familiar de até 1 salário mínimo e residem com os pais.

A Figura 41 mostra o gráfico de dispersão, gerado pelo algoritmo cluster **EM Evolutivo**, em função do atributo previsível “Situacao Ing = evasão”.

O cluster 2 na cor azul mais escura, contém a maior concentração de casos de evasão escolar, com exatamente 79% dos casos. O cluster 5 com 59% e o cluster 7 com 49% dos casos de evasão escolar. O cluster 9, como se pode ver com 0% dos casos de evasão escolar.

Analizando o cluster 2, se pode comprovar que, os atributos que mais influenciam na formação do cluster 2 são: Renda Familiar de até 1 salário com 89%, Reside com os pais com percentual de 91%, situacao Reprovado com percentual de 62% e Tipo de escola de origem igual a Pública estadual com percentual de 85% dos casos de evasão escolar. A Figura 42 mostra de forma gráfica esses dados.

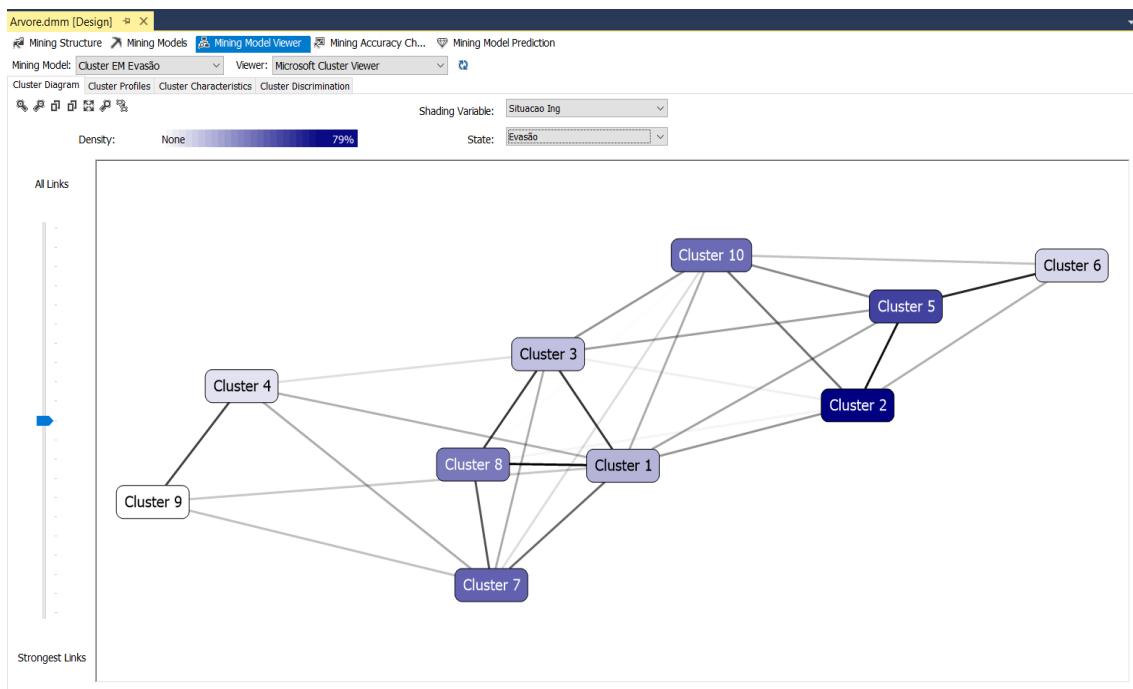


Figura 41. Cluster EM Evolutivo para o atributo previsível “Situacao Ing=Evasão”.

Fonte: Autor.

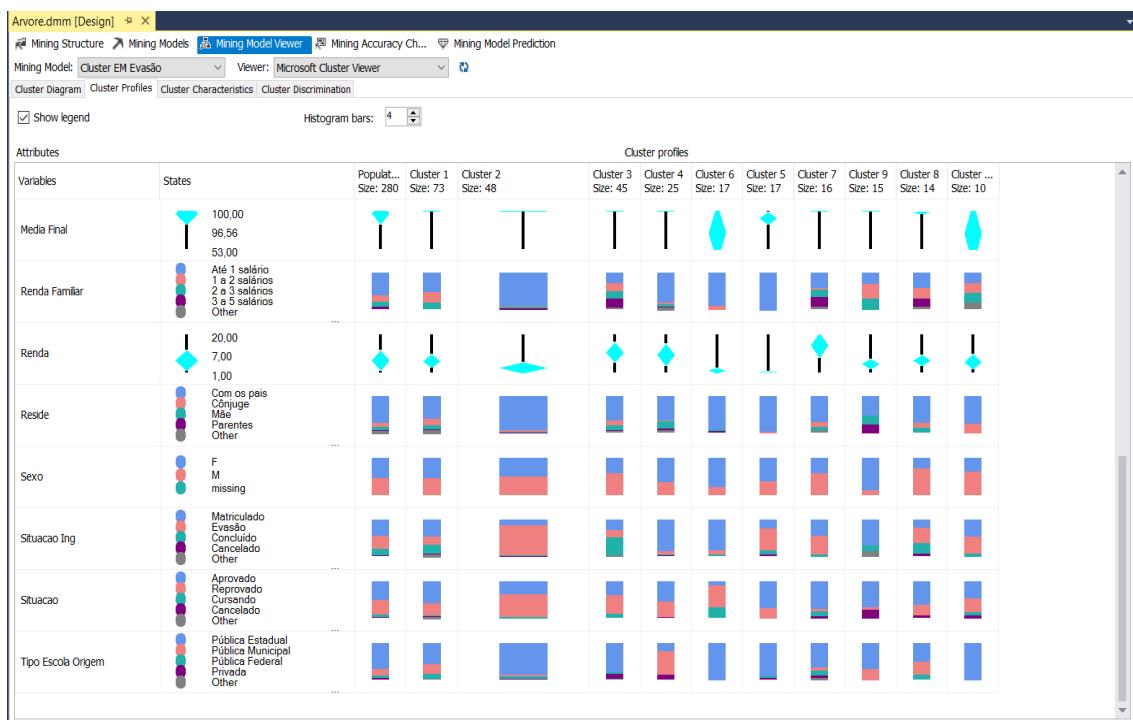


Figura 42. Gráfico tipo histograma mostrando o quanto cada atributo influencia percentualmente na formação de cada cluster.

Fonte: Autor.

3.6.8. Criando Consultas de Previsão

Na seção anterior, foram criados alguns modelos de mineração de dados. Então, irei usar esses modelos para fazer algumas previsões, como por exemplo, fazer provisões para identificar os possíveis perfis de alunos com maior probabilidade de repetência e evasão escolar.

3.6.8.1. Prevendo a Evasão e a Repetência Escolar

Nesta consulta pretende-se criar uma consulta para prever qual é o perfil dos alunos com a maior probabilidade de evasão escolar, tomando como base os dados as informações do modelo de mineração de Árvore de Decisões, explorado nesse trabalho e os dados da tabela base TB_JuntandoDadosAlunos.

A Listagem 3.10 mostra a consulta **DMX** criada para a previsão de evasão escolar em função dos atributos ‘Tipo Escola Origem = Pública Estadual’, ‘Renda Familiar = Até 1 salário’ e ‘Etnia = Branca’.

Listagem 3.10 Código de previsão escrito na linguagem **DMX**

```
SELECT  
    (PredictProbability([Mine TB Juntando Dados Alunos].[Situacao Ing])) as  
    [Probabilidade%],  
    [Mine TB Juntando Dados Alunos].[Situacao Ing],  
    t.[Tipo_Escola_Origem],  
    t.[Renda_Familiar],  
    t.[Etnia]  
From  
    [Mine TB Juntando Dados Alunos]  
PREDICTION JOIN  
OPENQUERY([DDS Educacional],  
'SELECT  
    [Tipo_Escola_Origem],  
    [Renda_Familiar],  
    [Etnia],  
    [Data_de_Nascimento],  
    [Idade],  
    [Forma_Ingresso],  
    [SituacaoIng],  
    [Coeficiente_Rendimento],  
    [Disciplina],  
    [Area_Procedencia_Escola_Origem],  
    [Media_Final],  
    [Aulas_Dadas],  
    [Faltas],  
    [Reside],
```

```

[Sexo],
[Situacao],
[Renda]
FROM
[dbo].[TB_JuntandoDadosAlunos]
') AS t
ON
[Mine TB Juntando Dados Alunos].[Data De Nascimento] =
t.[Data_de_Nascimento] AND
[Mine TB Juntando Dados Alunos].[Idade] = t.[Idade] AND
[Mine TB Juntando Dados Alunos].[Forma_Ingresso] = t.[Forma_Ingresso] AND
[Mine TB Juntando Dados Alunos].[Situacao_Ing] = t.[SituacaoIng] AND
[Mine TB Juntando Dados Alunos].[Coeficiente Rendimento] =
t.[Coeficiente_Rendimento] AND
[Mine TB Juntando Dados Alunos].[Disciplina] = t.[Disciplina] AND
[Mine TB Juntando Dados Alunos].[Area_Procedencia_Escola_Origem] =
t.[Area_Procedencia_Escola_Origem] AND
[Mine TB Juntando Dados Alunos].[Etnia] = t.[Etnia] AND
[Mine TB Juntando Dados Alunos].[Media_Final] = t.[Media_Final] AND
[Mine TB Juntando Dados Alunos].[Aulas_Dadas] = t.[Aulas_Dadas] AND
[Mine TB Juntando Dados Alunos].[Faltas] = t.[Faltas] AND
[Mine TB Juntando Dados Alunos].[Renda_Familiar] = t.[Renda_Familiar] AND
[Mine TB Juntando Dados Alunos].[Reside] = t.[Reside] AND
[Mine TB Juntando Dados Alunos].[Sexo] = t.[Sexo] AND
[Mine TB Juntando Dados Alunos].[Tipo_Escola_Origem] =
t.[Tipo_Escola_Origem] AND
[Mine TB Juntando Dados Alunos].[Situacao] = t.[Situacao] AND
[Mine TB Juntando Dados Alunos].[Renda] = t.[Renda]
WHERE
[Mine TB Juntando Dados Alunos].[Situacao_Ing] ='Evasão' AND
t.[Tipo_Escola_Origem] ='Pública Estadual' AND
t.[Renda_Familiar] ='Até 1 salário' AND
t.[Etnia] ='Branca'

```

Para obter todas as probabilidades, basta modificar os parâmetros dos filtros para os atributos Tipo_Escola_Origem, Renda_Familiar e Etnia. A Tabela 13 mostra os possíveis valores que esses atributos podem assumir.

Tabela 13 Valores dos atributos.

ATRIBUTO	VALORES
TIPO_ESCOLA_ORIGEM	Privada Pública Municipal Pública Estadual Pública Federal
RENDAS_FAMILIAR	Até 1 salário

	1 a 2 salários
	2 a 3 salários
	3 a 5 salários
	5 a 10 salários
	10 a 20 salários
	Mais de 20 salários
ETNIA	Branca
	Parda
	Preta
	amarela

Ao ser executada a consulta da Listagem 3.10, será obtido o valor de 57% de probabilidade de evasão escolar para o perfil especificado. E se mudarmos o perfil para ‘Tipo Escolar Origem = Privada’, será obtido o valor de 51% de probabilidade de evasão escolar. A Tabela 14 mostra os resultado da consulta com vários perfis.

Tabela 14 Previsão de evasão de acordo com o perfil especificado na tabela.

**PERFIL
(EVASÃO ESCOLAR)**

TIPO ESCOLA ORIGEM	RENDA FAMILIAR	ETNIA	Probabilidade
PÚBLICA ESTADUAL	Até 1 salário	Branca	57%
PRIVADA	Até 1 salário	Branca	51%

Se mudar o valor da situação de “Evasão” para “Reprovado” no código **DMX** da Listagem 3.8. Então, pode-se fazer previsão em relação a repetência escolar. A Tabela 15 mostra um resumo das previsões calculadas para reprovação.

Tabela 15 Prevendo repetência de acordo com o perfil especificado na tabela.

PERFIL
(REPETÊNCIA ESCOLAR)

TIPO ESCOLA	RENDAS	ETNIA	Probabilidade
PÚBLICA ESTADUAL	Até 1 salário	Branca	90%
PRIVADA	Até 1 salário	Branca	75%
PÚBLICA FEDERAL	Até 1 salário	Branca	88%
PÚBLICA MUNICIPAL	Até 1 salário	Branca	86%

Capítulo 4 – Resultados Obtidos

4.1 Portal de Acesso as Informações Administrativas

Como proposto nos objetivos apresentados neste trabalho, uma das metas, era a implementação de um protótipo de um sistema, que permitisse aos gestores da educação do Instituto Federal, ter acesso às informações geradas pelos Modelo Multidimensional (**OLAP**) e pelo Sistema Dimensional, desenvolvidos durante a pesquisa (o processo **BI**).

Este portal tem como objetivo externar as informações aos utilizadores finais, informações sobre repetência escolar em relação aos cursos, campus, disciplinas, e assim por diante. Mostra também a evasão escolar por diversos ângulos, como por exemplo: curso e campus. Além do mais, os utilizadores poderão através do portal, ter acesso a diversos tipos de relatórios ou consultas e KPIs disponíveis no portal.

A importância do portal, é justificada simplesmente pelo fato de que, nos últimos anos, se ter observado um aumento no índice de repetência e evasão escolar nos institutos, no entanto, não sabe com exatidão, de quanto são estes índices. O portal dá a medida exata destes índices, de forma clara e objetiva. Isto possibilitará que os gestores educacionais, equipe pedagógica e professores, tomem preventivas em prol da diminuição da repetência escolar e, consequentemente, a diminuição do índice da evasão escolar.

O processo de descoberta de conhecimento na base de dados, foi muito importante, pois foi através dela, que se pode traçar os perfis da repetência e da evasão escolar, onde ficou evidente quais os principais atributos que influenciam nestes fatores.

4.2 Visão Geral do Portal

Em entrevista realizada com ao Diretores de Departamento da Diretoria de Informática, Diretoria de Rede de Computadores, Diretoria de Serviços, com Equipe Técnica Pedagógica – **ETEP** e com os professores dessas diretorias, ficou evidente a necessidade de uma ferramenta, onde os diretores, os pedagogos e os professores, pudessem visualizar de forma fácil e rápida, informações

sobre o desempenho escolar dos alunos, bem como informações sobre os índice da repetência e da evasão escolar, por campus, curso. Então, foram mapeadas as demandas de relatórios/consultas que relatar estas informações. Devo salientar que, não existe nenhuma ferramenta no IFRN, que atenda este propósito. Portanto, ficou evidente a necessidade do desenvolvimento de uma ferramenta de apoio, neste contexto, acadêmico do **IFRN**.

4.3. Metodologia

Os resultados apresentados no portal, são consequência de consultas a base de dados dimensional e a base de dados multidimensional (modelo **OLAP**). As consultas feitas ao modelo dimensional, são realizadas através da tecnologia de acesso a dados conhecida como **ADO.NET** da plataforma .NET e, as consultas realizadas a base multidimensional, são feitas via a tecnologia **ADOMD.NET**.

- O **ADO.NET** é um amplo conjunto de classes do .NET que viabiliza a recuperação de dados, bem como a atualização de origens de dados, de várias maneiras diferentes oferece diversas classes que, permitem praticamente todas as tarefas relacionadas com o acesso e manipulação de dados. Além do mais, o **ADO.NET** permite a comunicação com qualquer fonte de dados, desde os já conhecidos gerenciadores de bases de dados relacionais (SGBD) como: **SQL Server**, **MySQL**, **FireBird**, **Oracle**, **SyBase**, **Access**, **XML**, arquivos de Textos, e assim por diante. A Figura 43, mostra a arquitetura **ADO.NET**.
- O **ADOMD.NET** é um provedor de dados do Microsoft .NET Framework criado para se comunicar com o Microsoft SQL Server Analysis Services. Os comandos podem ser enviados em **MDX** (Multidimensional Expressions), **DMX** (Data Mining Extensions), **ASSL** (Analysis Services Scripting Language) ou até em uma sintaxe limitada de SQL, e pode não retornar um resultado.

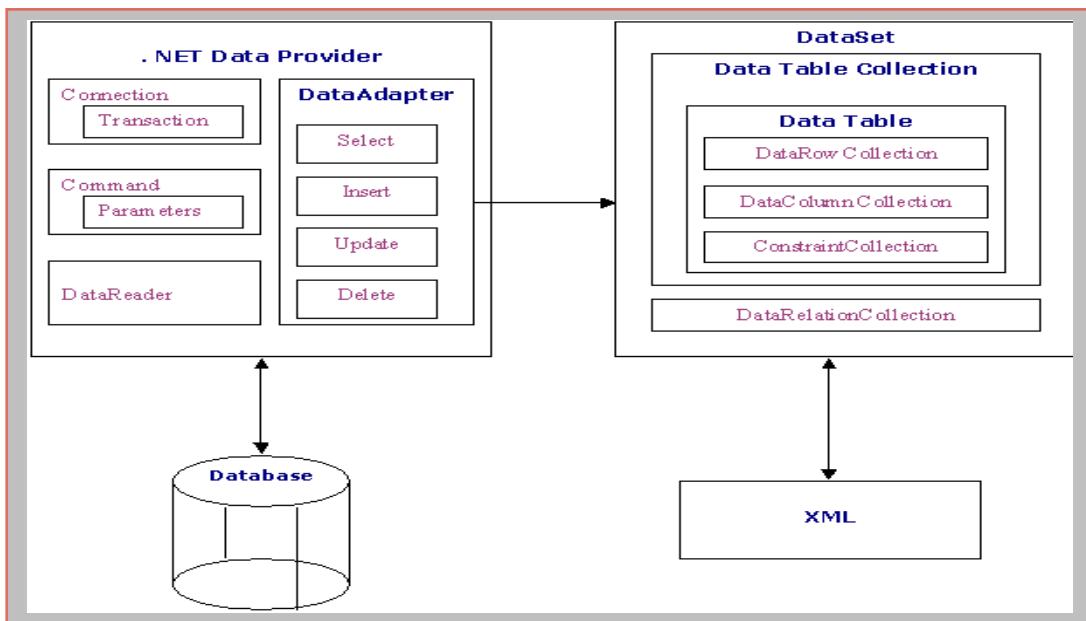


Figura 43. Arquitetura ADO.NET.

Fonte: Adaptado de [Dickinson et all, 2002]

No entanto, a tecnologia **ADO.NET** não permite manipular dados do modelo multidimensional. Então, para poder acessar os dados da base de dados OLAP, foi preciso utilizar outra tecnologia, conhecida como **ADOMD.NET**.

A programação do lado cliente para o modelo multidimensional é feita através ADOMD.NET cliente. Os componentes do ADOMD.NET cliente, oferecem a aplicativo cliente e da camada intermediária, a funcionalidade de consultar com facilidade dados e metadados de um repositório de dados analíticos, como o Microsoft SQL Server Analysis Services.

Uma instância do Analysis Services, é executada como serviço e a comunicação autônomos com o serviço, acontece por meio do XML for Analysis (XMLA), usando HTTP ou TCP.

A Figura 44, ilustra a arquitetura de componentes do Analysis Services, inclusive, todos os elementos principais executados dentro da instância do Analysis Services e todos os componentes de usuário que interagem com a instância. A figura também mostra que, o único modo de acessar a instância é usando o ouvinte do XML for Analysis (XMLA) ou usando HTTP ou TCP.

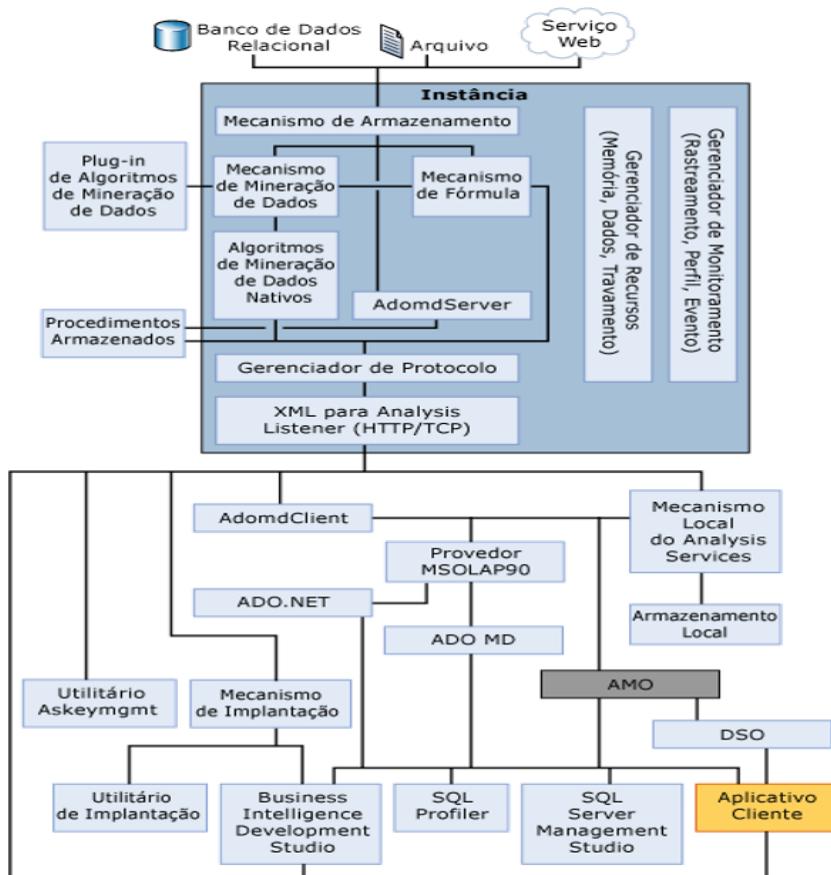


Figura 44. Arquitetura de Componentes do Analysis Services.

Fonte: Microsoft SQL Server 2005: Guia do Programador

Para processar e ou consultar cubos locais e os modelos de mineração locais, um aplicativo cliente pode chamar o OLE DB for OLAP 9.0 Provider (MSOLAP.3), conforme mostra a seguinte a Figura 45:

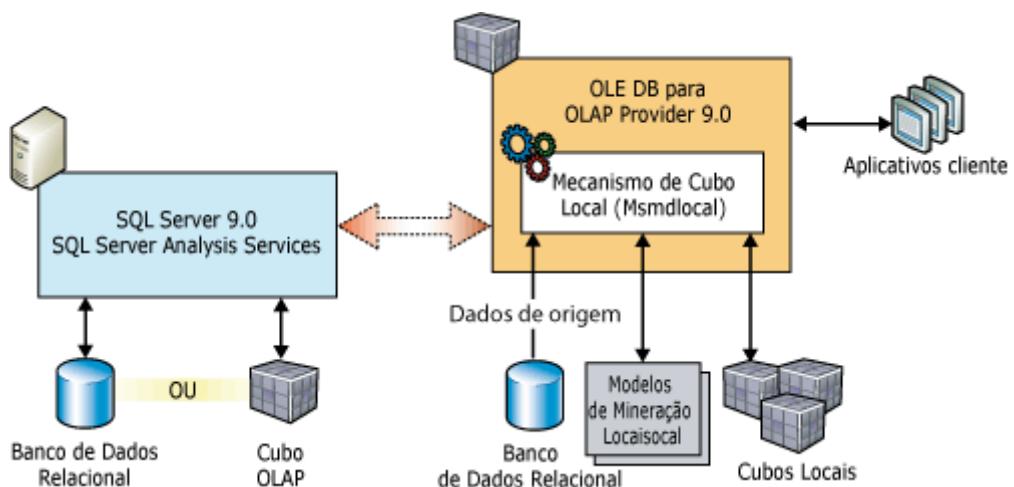


Figura 45. OLE DB for OLAP 9.0 Provider (MSOLAP.3).

Fonte: Microsoft SQL Server 2005: Guia do Programador

A Figura 46, mostra um exemplo de como se cria uma conexão no servidor local e, executa um comando nessa conexão, usando a tecnologia **ADOMD.NET**. A Figura 47, mostra o método para processar um cubo **OLAP**.

```
public List<CuboEvaConES> ADOMDGetCubeEvaConEntradaSaida(string query)
{
    List<CuboEvaConES> lista = null;

    using (AdomdConnection con = new AdomdConnection("Data Source=JoseAntonio-
PC;Integrated Security=SSPI;Initial Catalog=UMinhoD"))
    {
        try
        {
            con.Open();

            using (AdomdCommand cmd = new AdomdCommand(query, con))
            {
                using (AdomdDataReader dr = cmd.ExecuteReader())
                {
                    lista = new List<CuboEvaConES>();

                    while (dr.Read())
                    {
                        CuboEvaConES EvaConES = new CuboEvaConES();
                        EvaConES.Curso = dr[0].ToString();
                        EvaConES.Campus = dr[1].ToString();
                        EvaConES.Concluidos = Convert.ToInt16(dr[2]);
                        EvaConES.Evadidos = Convert.ToInt16(dr[3]);
                        EvaConES.TotalEvaCon = EvaConES.Concluidos +
EvaConES.Evadidos;
                        EvaConES.PerConclusao = Convert.ToDecimal(dr[4]);
                        EvaConES.PerEvadidos = Convert.ToDecimal(dr[5]);

                        lista.Add(EvaConES);
                    }
                }
            }
            return lista;
        }
        catch (AdomdException e)
        {
            throw new Exception("AdomdError: " + e.Message);
        }
        finally
        {
            con.Close();
        }
    }
}
```

Figura 46. Exemplo **ADOMD.NET** para se conectar ao Servidor do Analysis Services e executar uma consulta.

Fonte: Autor

```

public void ProcessarCubo(string banco)
{
    String ConnStr;
    ConnStr = @"Provider=SQLNCLI11.1;Data Source=JoseAntonio-
PC;Integrated Security=SSPI;Initial Catalog=UMinhoD";

    using (Server server = new Microsoft.AnalysisServices.Server())
    {
        server.Connect(ConnStr);

        Database database = server.Databases["UMinhoD"];

        foreach (Cube cube in database.Cubes)
        {
            cube.Process(ProcessType.ProcessFull);
        }
    }
}

```

Figura 47. Método **ADOMD.NET** para processar um cubo no Servidor do Analysis Services.

Fonte: Autor

Resumindo, as tecnologias **ADO.NET**, **ADOMD.NET** são tecnologias de manipulação de dados e objetos (metadados) no Servidor Analysis Services.

4.4. Arquitetura do Portal

Para a implementação do portal, foi adotada a arquitetura WCF RIA Service, ilustrada na Figura 48.

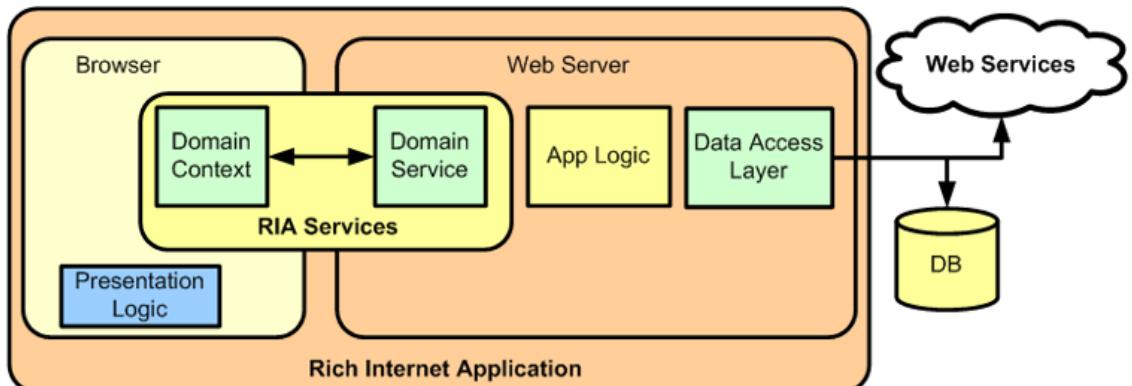


Figura 48. Arquitetura usada para implementação do portal.

Fonte: Adaptada de Building Business Applications with Microsoft Silverlight

A classe **Domain Context** é responsável por fornecer acesso aos dados, via uma série de métodos públicos, enquanto que, as regras de negócio são

implementadas no Web Services. No entanto, observe que, a tecnologia **WCF RIA Services** (Windows Communication Foundation), não especifica quais tecnologias o usuário deve usar. Ou seja, ela deixa que o implementador, use a tecnologia de manipulação de dados que bem desejar. Dessa forma, o utilizador pode usar **ADO.NET**, **ADOMD.NET**, **NHibernate** e assim por diante.

4.5. O Portal

Para exemplificar, a partir deste ponto serão apresentadas algumas telas do portal e, serão feitos comentários explicativos, sobre cada uma delas. Já que, este capítulo do trabalho, está reservado a apresentação dos resultados obtidos. No entanto, lembro que, as consultas realizadas sobre essas bases de dados, são muitas, quase inesgotáveis, elas surgem sobre demanda, então, serão apresentados alguns exemplos. Esses exemplos surgiram sobre demanda, ou seja, foram realizadas consultas na comunidade acadêmica, mais precisamente, junto os pedagogos, aos professores e alguns diretórios acadêmicos sobre suas demandas e, essas demandas foram implementadas sobre consultas no portal, para atender estes usuários.

A Figura 49, ilustra é a tela principal do sistema, onde temos duas opções principais “**Gestão**” e “**Análise**”. Na opção “Gestão”, o usuário terá acesso a todas as consultas e KPIs administrativos do portal e na opção “Análise”, o usuário terá acesso a todos os indicadores de desempenho, relacionados a repetência e a evasão escolar. As consultas são implementadas em função dos conhecimentos encontrados na mineração de dados nas bases de dados do IFRN. Ou seja, as consultas estam relacionadas aos índices de repetência ou de evasão escolar, de acordo com o perfil que foi encontrado pela mineração de dados.



Figura 49. Tela principal do Portal de Análise de dados.

Fonte: Autor

Como se pode observar, logo na tela principal do sistema está sendo apresentado um dashboard com os índices da evasão escolar nos cursos e ano, ou seja, o utilizador escolhe um ano e, o sistema mostra o total de alunos evadidos, em cada curso, naquele ano.

O usuário ao selecionar a opção do menu “**Gestão**”, o mesmo será direcionado para a tela da Figura 50. Nesta tela, temos cinco opções: situação (Concluídos, evasão, aprovados, etc.) por campus, repetência por curso, repetência por disciplina, evasão por campus e por curso.

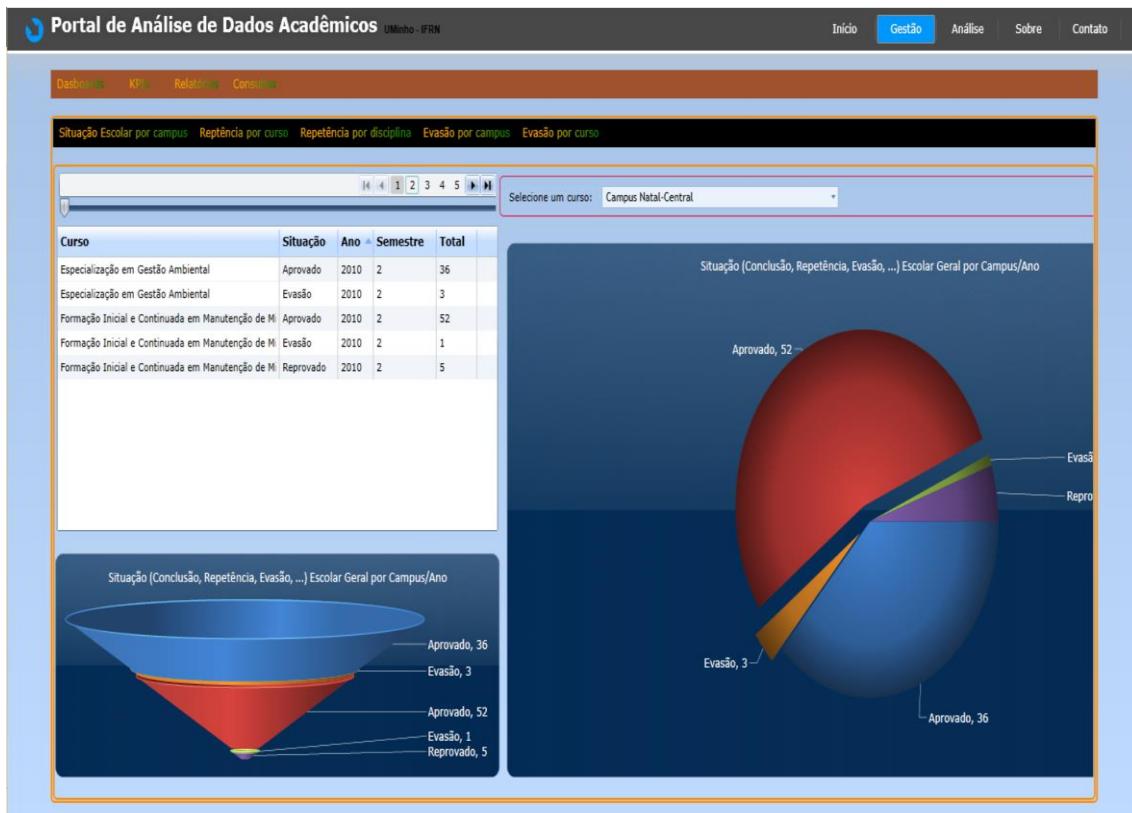


Figura 50. Dashboard Situação Escolar por Campus.

Fonte: Autor

A primeira opção do menu Dashboard é “**Situação Escolar por campus**”. Neste Dasborad, o utilizador escolhe o curso que deseja analisar e é exibido o total de aprovados, reprovados e evadidos em cada curso do campus. Para facilitar a leitura das informações, os dados são sempre mostrados de três formas diferentes, uma tabela e dois gráficos. Nesta tela, foram exibidos os dados do campus Natal central para o ano de 2010.

A Figura 51, mostra o dashboard da opção do menu “**Repetência por disciplina**”. O utilizador seleciona um curso e o anos, que deseja examinar e, os dados sobre a repetência nas disciplinas são exibidos. Os dados estão ordenados de forma decrescente pelo total de repetência escolar, para exibir as disciplinas que mais reprovação.

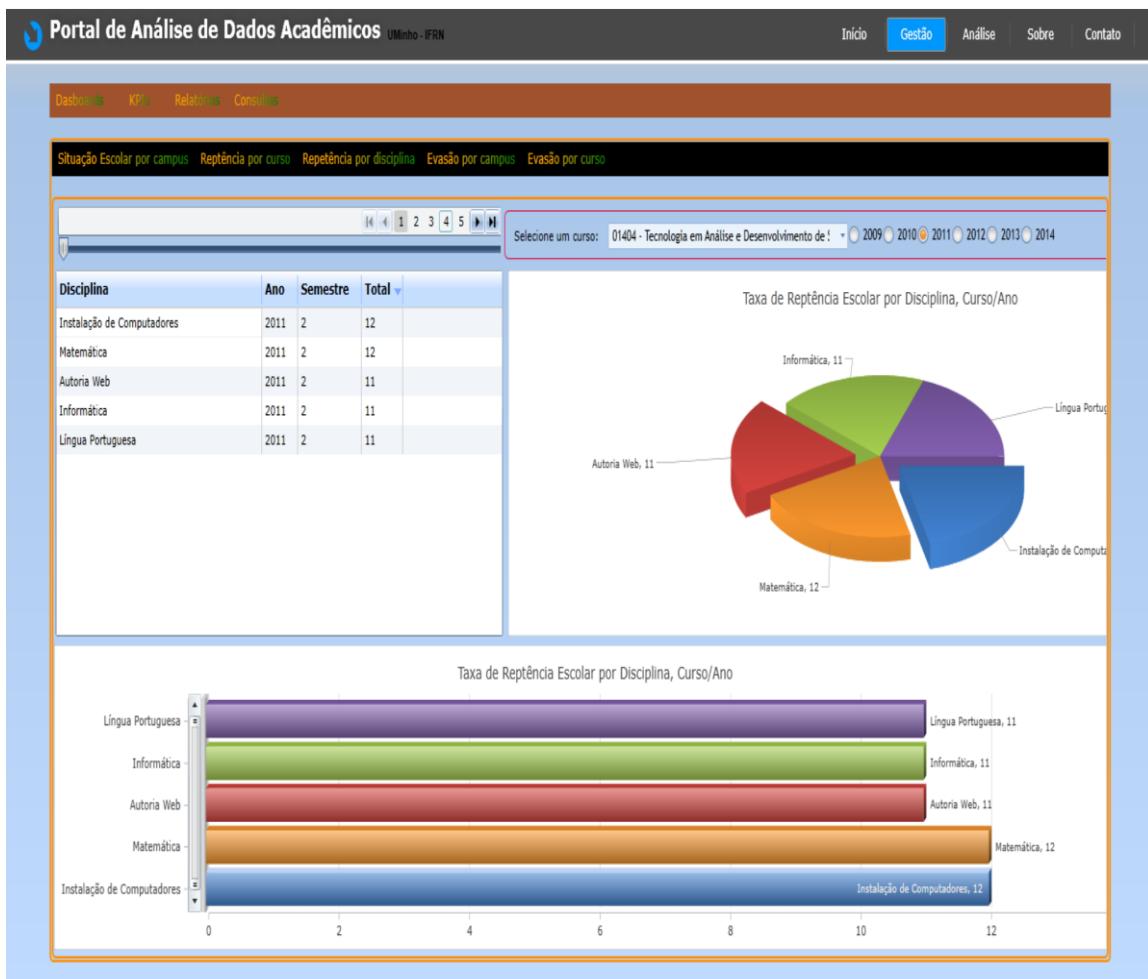


Figura 51. Dashboard Repetência por disciplina.

Fonte: Autor

Na tabela temos as 5 disciplinas com maior número de reprovações no curso selecionado. Veja que, para o curso selecionado “Tecnologia em Análise de Desenvolvimento de Sistemas” e ano 2011, as cinco disciplinas onde ocorreram o maior número de reprovações são: instalação de computadores, matemática, autoria web, informática e Língua Portuguesa.

A Figura 52, mostra o dashboard da opção do menu “**Evasão por campus**”. O utilizador escolhe o campus e o ano que deseja analisar e são apresentadas, informações tais como: total de concluídos, matriculados, evadidos e também o total de alunos que cancelaram os seus respectivos cursos. Então, neste dashboard temos as informações para o campus Natal Central e ano de 2010.

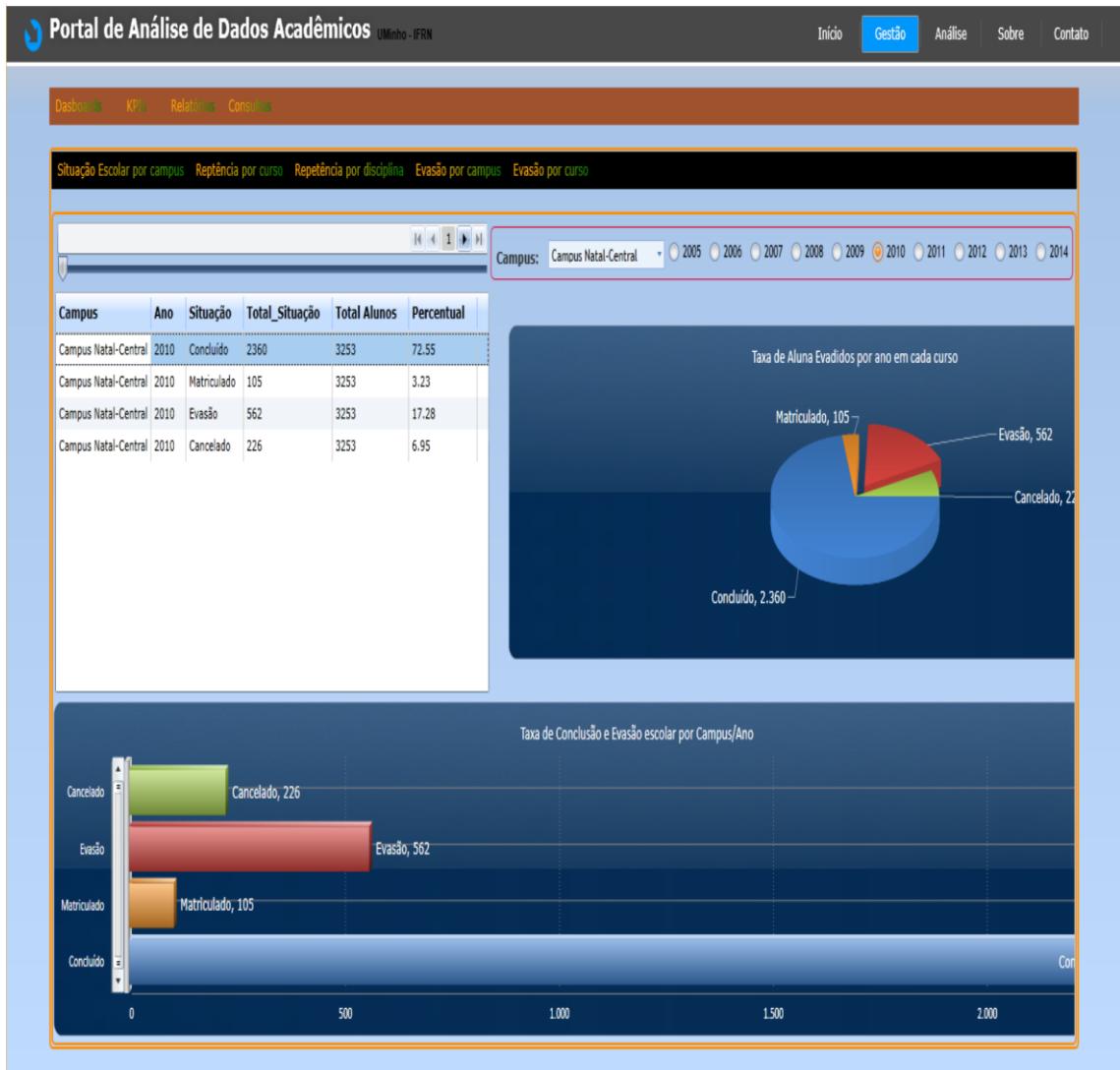


Figura 52. Dashboard Evasão por campus e ano.

Fonte: Autor

Na Figura 52 pode-se ver que, para o campus Natal central no ano de 2010, 72,55% dos alunos concluíram seus cursos com sucesso, 3,23% de alunos matriculados, ou seja, alunos que já atingiram o tempo de conclusão de seus cursos, no entanto, ainda não concluíram, nem cancelaram matrículas e não evadiram e, 17,28% é o total de alunos evadidos e 6,95% de alunos que cancelaram suas matrículas. Veja que se juntarmos os totais de evadidos com os de cancelados, temos $17,28\% + 6,95\% = 24,23\%$ de alunos que deixam os cursos sem concluir. Se considerarmos um contingente de 10.000 alunos, representaria um total de 2.423 alunos que, abandonaram seus respectivos cursos em 2010 no campus Natal central.

A Figura 53, mostra o dashboard da opção do menu “Evasão por curso”. O utilizador escolhe o ano que deseja analisar e, será exibido o total de evadidos por curso no ano selecionado.

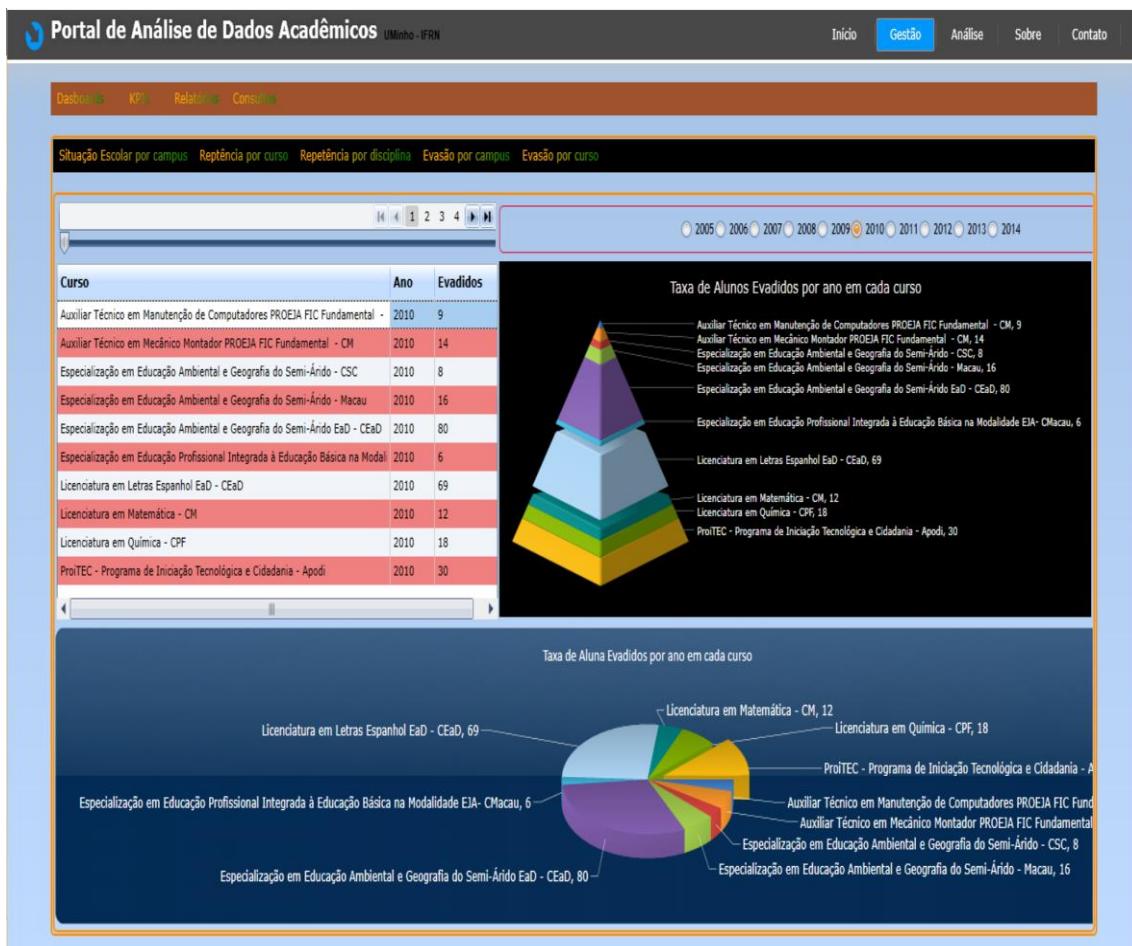


Figura 53. Dashboard Evasão por curso e ano.

Fonte: Autor

Na Figura 53, observa-se, por exemplo, que alguns cursos da Educação a Distância – **EAD**, apresentam alto índice de evasão escolar. O Curso “Especialização em Educação Ambiental e Geografia do Semi-Árido” no ano de 2010, 80 (oitenta) aluno evadiram do curso (no ano de 2010, esse curso tinha 2 turmas de 100 alunos cada uma delas). Isto representa 40% de alunos evadidos.

A segunda opção do menu “Gestão” é **KPI**, onde temos dois **KPI** a saber. O primeiro é o **KPI Desempenho do aluno**. Este **KPI** é baseado no coeficiente de desempenho do aluno, definido na seção **KIP**, no processo **BI**. O segundo **KPI**

está relacionado a média do aluno. Através desse **KPI**, os professores poderão acompanhar o desempenho dos alunos em cada disciplina. Isto é importante, pois permitirá que os professores tomem medidas preventivas, para tentar evitar a repetência dos alunos nas suas disciplinas.

A Figura 54, mostra o **KPI** “Desempenho do aluno”. Como já falado, esse **KPI** tem como base o coeficiente de desempenho do aluno. Este coeficiente é utilizado em diversas situação na vida escolar do aluno. Esse KPI é a soma de todas as médias obtidas pelo aluno, dividida pelo número de disciplinas cursadas. Por exemplo, se o aluno tiver cursado 8 disciplinas, o cálculo do KPI é a soma das oitos médias dividido por oito.



Figura 54. KPI Desempenho do Aluno.

Fonte: Autor

Nessa tela o utilizador seleciona a turma desejada e o coeficiente de rendimento do aluno é mostrado na tabela e também nos gráficos. A cor amarela significa alerta, coeficiente de rendimento entre 30 e 59 pontos, a cor vermelha significa coeficiente muito baixo, menor do que 30 pontos e a cor verde significa coeficiente acima de 60 pontos.

A Figura 55, mostra o KPI “Desempenho do aluno por disciplina”. Este KPI permite que os professores acompanhem, o desempenho dos alunos em suas disciplinas. O utilizador seleciona a turma e é exibida a situação do aluno, em função de sua média, em cada disciplina que o mesmo está cursando. A cor amarela significa que o aluno tem média entre 30 e 59 naquela disciplina. A cor vermelha, significa que o aluno está em situação de risco, média inferior a 30 e, a cor verde significa que o aluno está em boa situação na disciplina, média igual ou superior a 60. A importância deste KPI, é que o professor ou pedagogo, por exemplo, pode acompanhar o desempenho do em todas as disciplinas, que estam sendo cursadas por cada aluno da turma e, não somente na disciplina que o professor está ministrando. Pois, até o momento, o professor só podia acompanhar a situação dos alunos, somente nas disciplinas que ministras.

Isto é importante, pois, uma vez que, o professor toma conhecimento da situação dos alunos que, apresentam mau desempenho em várias disciplinas, ele pode imediatamente tomar medidas preventivas e, isto pode ser um fator que venha impactar na diminuição da repetência e da evasão escolar.



Figura 55. KPI Desempenho do aluno em cada disciplina.

Fonte: Autor

A última opção do menu “Gestão” é a janela de consultas, como mostra a Figura 56. Na janela de consultas, a primeira opção do menu é “Evasidos”. Nesta janela, o utilizador pode selecionar um ou vários campi que deseja analisar e, as informações sobre a evasão de cada campus, é organizada de forma gráfica em quadradinhos na janela, de forma que, as informações de cada campus, ficam um mesmo local da janela.

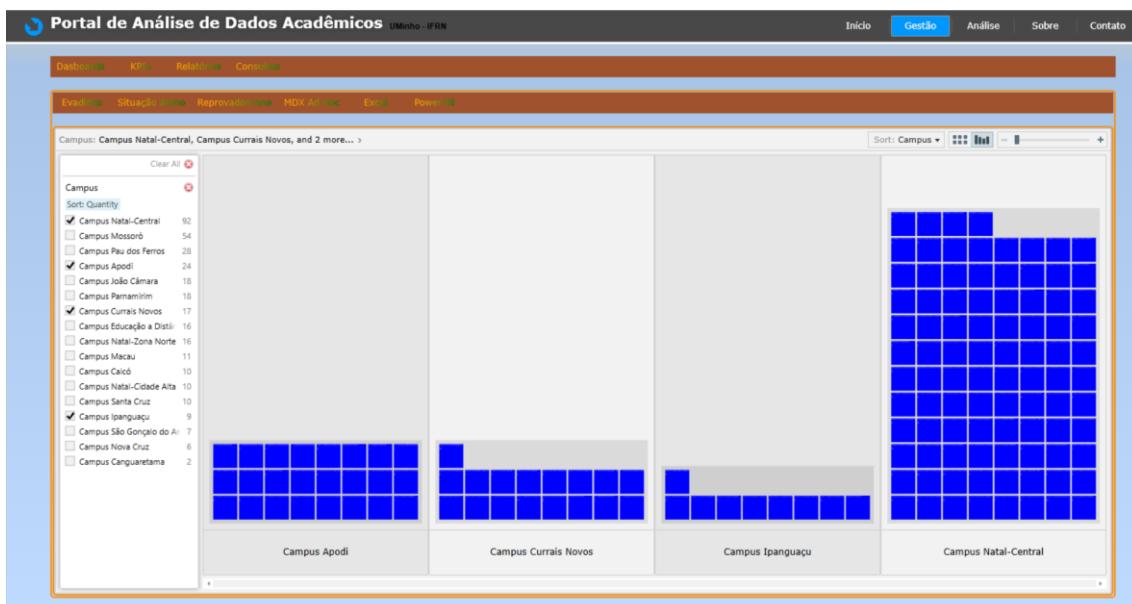


Figura 56. Janela Consultas sobre evasão escolar por campus. **Fonte:** Autor

Quando o utilizador clica em um quadradinho na cor azul, então as informações contidas no quadradinho são mostradas ao lado da janela. Veja a Figura 57.

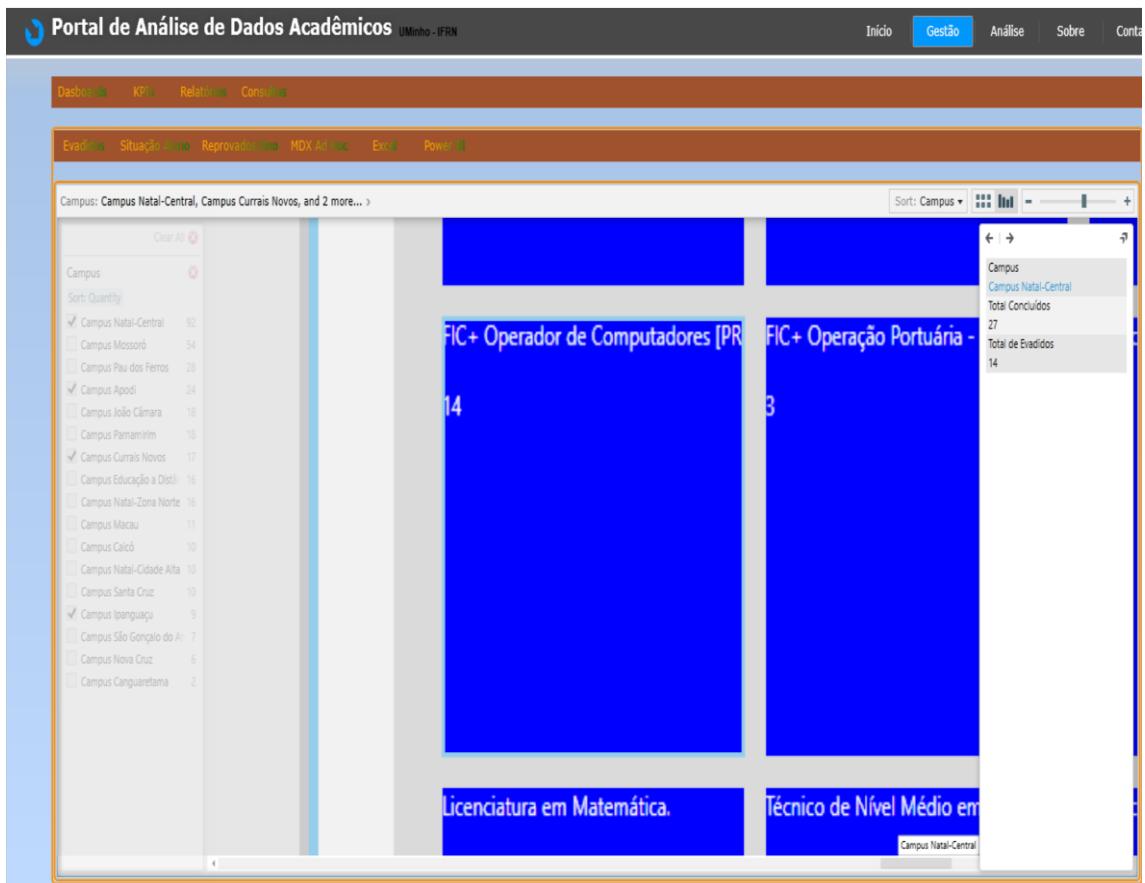


Figura 57. Resultado após o utilizador clicar no quadradinho azul. **Fonte:** Autor

Veja que do lado direito são exibidas as informações do Campus Natal Central, onde temos um total de 27 alunos concluídos e um total de 14 alunos evadidos, para o curso FIC Operador de Computadores. O utilizador também tem a opção de ver as informações, contidas nos quadradinhos, apenas arrastando a barra de slide, que fica no topo direito da janela.

Na janela de consultas temos a opção do menu “Reprovados/Ano”, que ao ser selecionada, mostra o total de alunos aprovados e reprovados, por curso e disciplina no ano. Veja a Figura 58.

The screenshot shows a web-based dashboard titled 'Portal de Análise de Dados Acadêmicos' from UMinho - IFRN. The top navigation bar includes links for Início, Gestão, Análise, Sobre, and Contato. Below the navigation is a menu bar with options like Dashboards, KPIs, Relatórios, Consultas, Evolução, Situação Ativa, Reprovados/Año, MDX Ad Hoc, Excel, and Power BI. A central table displays student grades across various courses and disciplines, with columns for Course, Discipline, Year, Approved Students, and Failed Students. A 'Excel' button is visible above the table, indicating it can be exported to a spreadsheet.

Curso	Disciplina	Ano Letivo	Aprovados	Reprovados
▲ Curso: Auxiliar Técnico em Operação de Computadores PROEJA FIC Fundamental (3 items)				
Auxiliar Técnico em Operação de Computadores PROEJA FIC Fundamental	Softwares Aplicativos II	2011	36	23
Auxiliar Técnico em Operação de Computadores PROEJA FIC Fundamental	Softwares Aplicativos III	2011	36	23
Auxiliar Técnico em Operação de Computadores PROEJA FIC Fundamental	Softwares Aplicativos III	2013	41	18
▲ Curso: Curso Certificador do ENEM (3 items)				
Curso Certificador do ENEM	Ciências da Natureza e suas Tecnologias	2010	3	1
Curso Certificador do ENEM	Linguagens, Códigos e suas Tecnologias	2010	3	1
Curso Certificador do ENEM	Redação	2010	3	1
▲ Curso: Ensino Médio - Campus Mossoró (9 items)				
Ensino Médio - Campus Mossoró	Biologia	2000	15	5
Ensino Médio - Campus Mossoró	Biologia	2001	126	11
Ensino Médio - Campus Mossoró	Biologia	2002	167	61
Ensino Médio - Campus Mossoró	Biologia	2006	104	1
Ensino Médio - Campus Mossoró	Desenho Básico	2000	12	5
Ensino Médio - Campus Mossoró	Desenho Básico	2002	98	30
Ensino Médio - Campus Mossoró	Educação Física	2001	122	17
Ensino Médio - Campus Mossoró	Educação Física	2002	227	3
Ensino Médio - Campus Mossoró	Educação Física	2006	100	1

Figura 58. Reprovados por Curso e Disciplina/Ano.

Fonte: Autor

Pode-se ver na Figura 58, o total de aprovados e reprovados em cada disciplina de determinado curso, por ano.

Observe que, no topo da consulta, tem um botão chamado “Excel”. Se o utilizador clicar neste botão, a consulta selecionada na tela, irá ser aberta imediatamente em uma planilha Excel, onde o utilizador poderá utilizar todos os recursos disponíveis na planilha, para manipular essas informações, como bem convier. Ele poderá gerar relatórios, gerar gráficos, fazer análise preditivas e assim por diante.

A opção seguinte do menu consulta é “MDX Ad Hoc”. Quando o utilizador selecionar esta opção, ele será direcionado para a janela mostrada na Figura 59, onde o utilizador terá a oportunidade de criar suas próprias consultas MDX. Esta opção, obviamente, só será utilizada por utilizadores avançados, que conheça a estrutura e instruções da linguagem MDX.

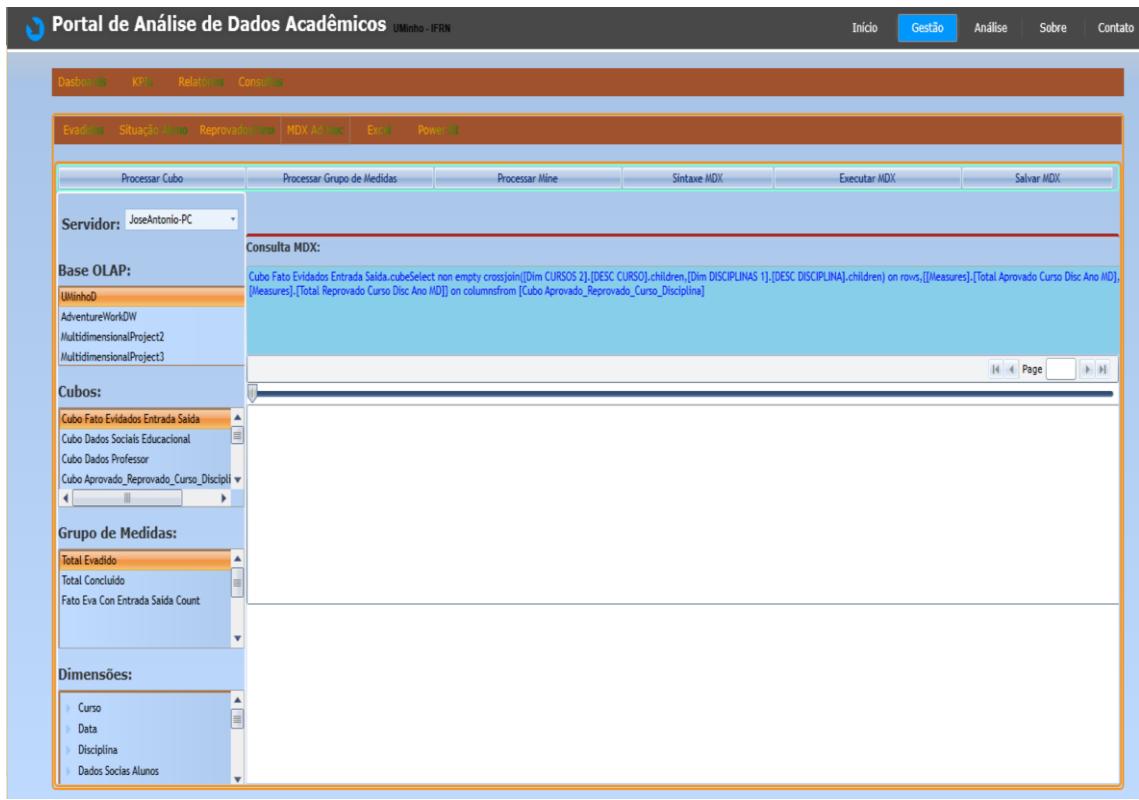


Figura 59. Constas MDX Ad Hoc.

Fonte: Autor

Do lado direito da janela, são exibidos os bancos de dados multidimensionais, os cubos, as medidas e as dimensões. Funciona da seguinte forma. O utilizador seleciona uma base de dados multidimensional, então, na aba cubos, são exibidos os cubos da base selecionada, na aba grupo de medidas, são exibidas as medidas do cubo que o utilizador selecionar e na aba dimensões, são exibidas as dimensões do cubo selecionado pelo utilizador. O utilizador então, escreve a consulta na caixa de texto “Consulta MDX” e clica no botão “Executar MDX” e os resultados são mostrados em uma tabela, na parte central da janela. Além disso, o utilizador tem a opção de antes de executar a consulta, checar se a sintaxe da consulta estar correta, clicando no botão “Sintaxe MDX”. Nessa mesma janela, temos ainda, os botões “Processar Cubo” e “Processar Medidas”, os quais ao serem clicados, processam os cubos e as medidas dos cubos, respectivamente, como seus próprios nomes sugerem.

Ainda na janela do menu consultas temos, foi disponibilizado para o utilizador as opções do mesmo poder abrir a planilha Excel ou o utilitário Power BI, apenas clicando em uma dessas opções.

Se o utilizador abrir o Excel, o mesmo poderá acessar um cubo do Analysis Services e manipular suas informações, ou seja, fazer as consultas Ad Hoc, sem precisar conhecer a linguagem **MDX**. Acessando um cubo **OLAP** via Excel, o utilizador cria suas consultas Ad Hoc, apenas arrastando os atributos para as áreas específicas da tabela dinâmica, pois é através das tabelas dinâmicas do Excel, que o utilizador pode acessar um cubo **OLAP**.

A Figura 60 mostra um exemplo de uma consulta Ad Hoc, feita no Excel, acessando um cubo **OLAP**.

	A	B	C	D	E	F	G
1	DESC INSTITUICAO	(Vários itens)					
2							
3							
4							
5							
6	Rótulos de Linha						
7	Ensino Médio - Campus Natal-Central						
8	Especialização em Educação Ambiental e Geografia do Semi-Árido (2010) - Campus Santa Cruz	21	21				
9	Especialização em Educação Profissional Integrada à Educação Básica na Modalidade Educação de Jovens (2010) - Campus Apodi	40	40				
10	Especialização em Gestão Ambiental [2011] - Campus Natal-Central			11	11		
11	FIC-Atualização em Educação Ambiental e Geografia do Semi-árido - Matutino (2005) - Câmpus Natal Central						
12	FIC-Desenhista Mecânico [PRONATEC 2013] - Câmpus Mossoró	90	90				
13	FIC-Operador de Computadores [PRONATEC 2012] - Câmpus Nova Cruz						
14	Licenciatura em Biologia (2009-2011) - Campus Macau						
15	Licenciatura em Química (2009-2011) - Campus Currais Novos						
16	ProTEC - Programa de Iniciação Tecnológica e Cidadania - Campus Caicó [2009]						
17	Técnico em Comércio Integrado EJA - Campus Zona Norte						
18	Técnico em Construção Civil Integrado [1995] - Câmpus Natal Central						
19	Técnico em Controle Ambiental Subsequente (2001-2011) - Campus Natal Central						
20	Técnico em Eletromecânica Integrado - Campus Mossoró						
21	Técnico em Eletromecânica Integrado - Câmpus Natal Central						
22	Técnico em Eletromecânica Subsequente EaD - Câmpus Natal Central						
23	Técnico em Eletrônica Integrado (2005-2011) - Campus Natal-Central						
24	Técnico em Geologia e Mineração Integrado [1995] - Câmpus Natal Central						
25	Técnico em Geologia e Mineração Integrado (1996-2003) - Câmpus Natal Central						
26	Técnico em Tecnologia Ambiental Integrado (1995) - Câmpus Natal Central						
27	Tecnologia em Comércio Exterior (2003) - Câmpus Natal Central						
28	Tecnologia em Desenvolvimento de Software (2002) - Câmpus Natal Central						
29	Tecnologia em Informática [1998-2008] - Câmpus Natal Central						
30	Total Geral	61	90	151	11	11	
31							

Figura 60. Consulta Ad Hoc com tabela dinâmica do Excel.

Fonte: Autor

Usando tabela dinâmica do Excel, o utilizador pode se conectar a um servidor Analysis Service, acessar um cubo e fazer as consultas que desejar. Veja na Figura 60, que os fatos, as medidas e as dimensões do cubo, ficam disponíveis para acesso, do lado direito da janela. O utilizador arrasta para a área de valores as medidas do cubo, que deseja ver e, arrasta para as áreas de colunas e linhas os detalhes das medidas. No exemplo, foram arrastados para a área de valores as medidas TotalConcluidos e TotalEvadidos, e para a área de linhas, foi arrastado a descrição do curso e para a área de colunas, foram

arrastados os atributos situação e ano. Essas consultas são dinâmicas, porque a qualquer momento, o utilizador pode retirar ou colocar novos atributos das áreas valores, colunas ou linhas. E são ad hoc, porque o utilizador pode fazer a consulta que imaginar, com os dados disponíveis. Com a consulta montada, o mesmo poderá então realizar previsões, ou construir gráficos para melhor visualização dos resultados.

Se o utilizador, selecionar a opção “Power BI”, o mesmo será direcionado para a ferramenta de BI do mesmo nome. Com esta ferramenta, o utilizador terá a opção de criar dashboards, KPIs, etc, acessando, tanto cubos **OLAP**, quanto tabelas do modelo dimensional. Uma observação, a ferramenta **Power BI** acessa inúmeras fontes de dados, não somente as citadas anteriormente. A Figura 61 mostra um exemplo do uso do Power BI, acessando um cubo **OLAP**.

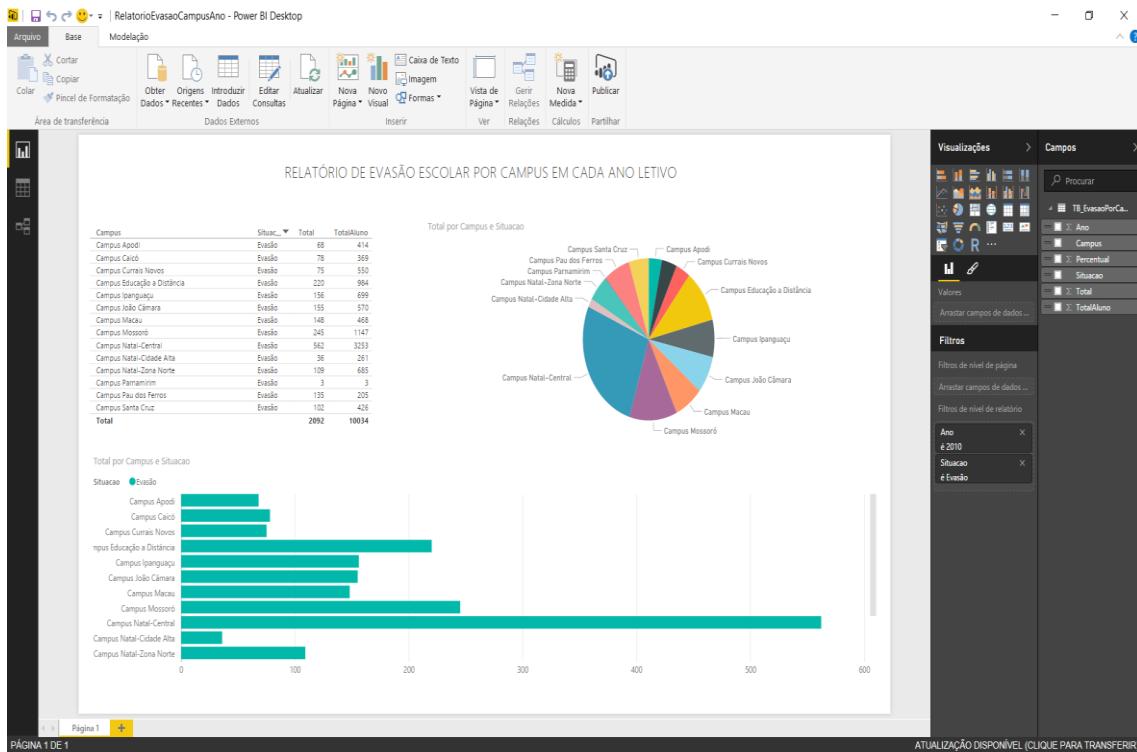


Figura 61. Um Dashboard exibindo a Evasão Escolar por campus do Power BI.

Fonte: Autor

Na janela do Power BI, do lado direito, tem-se os atributos da fonte de dados acessada, as visualizações disponíveis e a área de filtros. O utilizador seleciona a visualizada que deseja mostrar os dados, em seguida seleciona os atributos que deseja exibir na visualização e faz os filtros necessários, ou seja, monta o dashboard como desejar. O Power BI é uma ferramenta completa de BI. Além do mais, ele disponibiliza ao usuário, a opção, do mesmo publicar, os

resultados na web. Para isto, o utilizador basta clicar no botão “Publicar” na barra de ferramenta do Power BI, no topo da janela. Logicamente, que para poder publicar os resultados na web, o utilizador teria que ter uma conta no Power BI.

Para finalizar, observe que, na aba Visualizações temos, diversas formas de gráficos para visualização dos dados, além de tabelas e KPI.

Com isto, encera-se as opções do menu “Consulta”, vamos agora ver as opções disponíveis no menu “Análise” do menu principal.

A opção “Análise” do menu principal mostra uma série de índices relacionados, tanto a repetência, quanto a evasão escolar nos campi do instituto federal. A Figura 62 mostra as opções desse menu.



Figura 62. Dashboard de Índices Gerais de todos os campi do IFRN.

Fonte: Autor

Este dashboard mostra os índices de retenção, conclusão, evasão, reprovação (repetência) e de saída com êxito, dos dados levantados de todos os campi do IFRN, para os anos de 2005 a 2015. Veja que a taxa de evasão, para os dados levantados, é de 14,90%. A taxa de retenção representa os alunos que ficaram retidos no módulo. Por exemplo, se o aluno estar no 1º semestre e estar a cursar 8 disciplinas, e é reprovado em mais de 2 disciplinas, ele não avança

para o 2º semestre do curso, ele fica retido no 1º semestre. No entanto, se o aluno é reprovado em apenas 2 disciplinas, ele avança para o 2º semestre do curso e, cursa as disciplinas, que ele não obteve êxito, em turno inverso, ou seja, se seu curso é pela manhã, então, o ele irá cursar as duas disciplinas reprovadas, no período vespertino. Ou vice-versa, se o curso é vespertino, ele irá cursar as duas disciplinas reprovadas, no turno matutino.

A próxima opção do menu “Análise”, é indicadores de evasão escolar por curso no campus Natal-Central. A Figura 63 mostra o dashboard com os índices taxa de Retenção, taxa de Conclusão, Taxa de Evasão e taxa de reprovação.

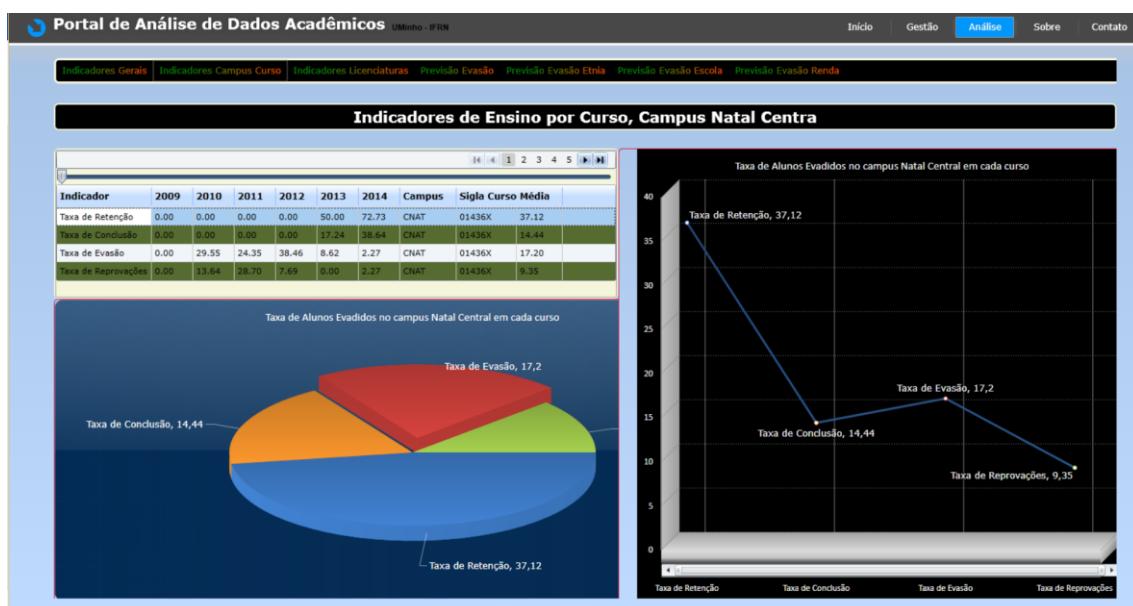


Figura 63. Dashboard Indicar de Ensino dos Curso no campus Natal-Central.

Fonte: Auto

Interpretando os dados apresentados na Figura 63, percebe-se que o curso 01436X, teve início em 2010 e que a retenção só começou a ocorrer em 2013, a taxa de evasão em 2010 já foi alta, em 2012 foi muito alta 38.46%, no entanto, diminui bastante em 2013, finalizando com uma taxa média de evasão de 17,2%.

Para analisar outros cursos, basta que o usuário movimente a barra de navegação acima da tabela ou o slide que fica abaixo da barra de navegação. A Figura 64 mostra outro curso, movimentando a barra de navegação para a página de número 16.



Figura 64. Indicador de Ensino do Curso 01434 do campus Natal Central.

Fonte: Autor

Previsão de evasão das licenciaturas é uma das opções do menu “Análise”. Este dashboard mostra a taxa de evasão dos cursos de licenciaturas ministrados pelo IFRN. Para este quadro foram somados todos os alunos de cada licenciatura de todos os anos da existência de cada curso. A Figura 65 mostra esses indicadores.



Figura 65. Indicadores de Evasão das Licenciaturas do IFRN.

Fonte: Autor

Neste quadro pode-se observar que a taxa de evasão nas licenciaturas, é bastante elevada, praticamente em todos eles. Este fato preocupante para o IFRN tentar resolver.

A opção seguinte do menu “Análise” está relacionada a previsão de evasão de cada aluno, com base no número de reprovações. A Figura 66, mostra este indicador. A probabilidade de sucesso ou insucesso foi calculada em função do número de reprovações que o aluno teve.

Para cálculo dessa probabilidade, foi utilizado a distribuição de Poisson para prever, o sucesso e insucesso dos alunos, com base no número de reprovações. A função de distribuição de Poisson expressa a probabilidade de uma série de eventos ocorrem num certo período de tempo se estes eventos ocorrem independentemente de quando ocorreu o último evento. A equação de Poisson utilizada foi a seguinte:

$$P(x) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (\text{equação 4.1})$$

Onde,

- λ um número real, igual ao número esperado de ocorrência num dado intervalo de tempos.
- k o número de ventos de sucesso.
- e base do logaritmo natural.
- $K!$ fatorial de k .



Figura 66. Probabilidade de Evasão em função do número de reprovações.

Fonte: Autor

As Figuras a seguir, mostram a probabilidade de evasão escolar com base no número de reprovações, classificados por etnia, tipo de escola de origem e renda familiar. Veja as Figuras 67, 68 e 69 respectivamente.



Figura 67. Probabilidade de Evasão em função do número de reprovações classificado por etnia.

Fonte: Autor

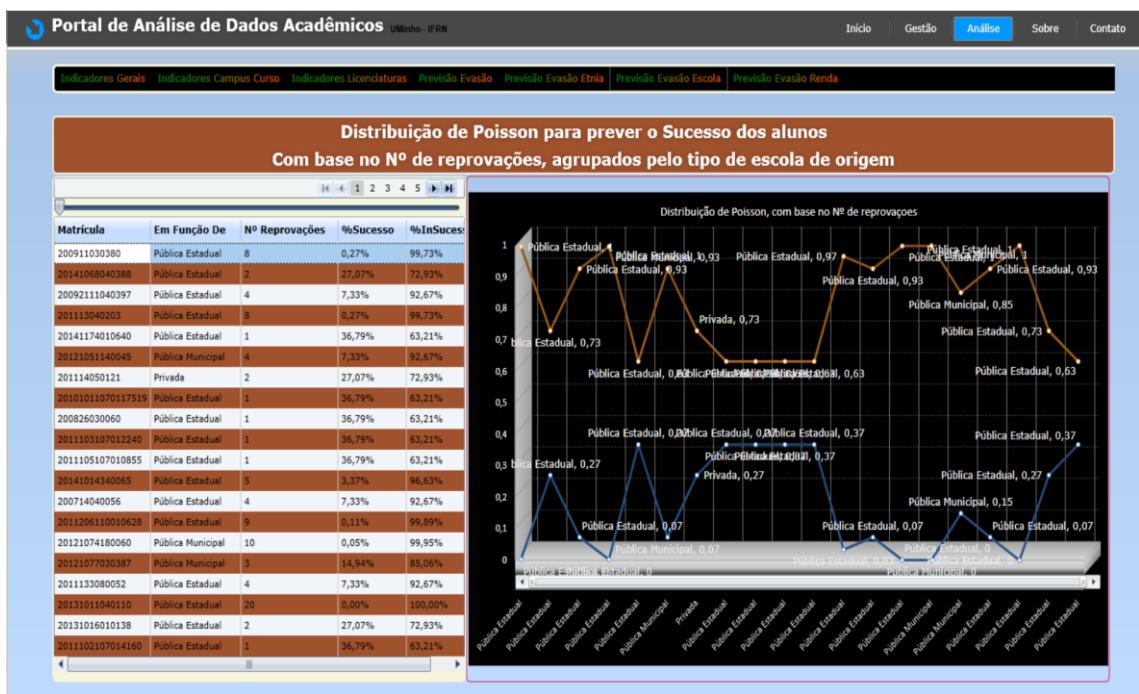


Figura 68. Probabilidade de Evasão em função do número de reprovações classificado por tipo de escola de origem.

Fonte: Autor



Figura 69. Probabilidade de Evasão em função do número de reprovações classificado por renda familiar.

Fonte: Autor

Na seção 4.1, foram apresentados os resultados da implementação do Portal de Análise, um dos objetivos específicos sugerido neste trabalho. Como se pode ver, o portal consta de diversos dashboards e KPIs, onde o usuário do sistema, tem acesso, os índices relacionados, tanto a repetência escolar, quanto a evasão escolar de várias maneiras e, também a evolução do aluno dentro de seu curso. A finalidade do portal é disponibilizar aos gestores da educação do **IFRN**, um mecanismo onde eles possam acessar as informações sobre repetência e evasão escolar em números reais, de forma visual fácil de entender, para que os mesmos possam tomar decisões de maneira mais rápida e precisa. Tudo isto, na tentativa de que, o **IFRN** possa diminuir os índices de repetência e evasão escolar.

4.2 Sistema de Raciocínio Baseado em Casos

Nesse módulo, serão utilizadas Raciocínio Baseado em Casos (**RBC**), para implementar um sistema especialista de aconselhamento pedagógico para o IFRN, aliando as técnicas do algoritmo de clustering, denominado de n-Vizinhos Mais Próximos (**kNN**) e à plataforma Web e suas tecnologias.

Portanto, O principal objetivo desse módulo é aplicar o Raciocínio Baseado em Casos (**RBC**), como técnica de Inteligência Artificial, no desenvolvimento de um sistema especialista que, auxilie uma equipe de acompanhamento pedagógico, na tarefa de aconselhamento de alunos, através da proposição de soluções aos problemas atuais, baseando-se no conhecimento adquirido com os casos anteriores que, por sua vez, foram inseridos na Base de Casos.

Então, o processo para desenvolver esse sistema, segue as seguintes tarefas:

- Desenvolver uma arquitetura ideal utilizando linguagem de programação e ambiente de desenvolvimento.
- Utilizar a plataforma web, a linguagem de programação Javascript e as notações JSON (Javascript Object Notation) e BSON (Binary Javascript Object Notation) no desenvolvimento do sistema proposto.
- Testar a aplicação com dados reais fornecidos pela equipe de Orientação Pedagógica do IFRN, campus Natal-Central, e analisar os resultados obtidos.

4.2.1. Visão do problema

Em entrevista realizada com a Equipe Técnica Pedagógica – **ETEP** da Diretoria Acadêmica de Gestão e Tecnologia da Informação do IFRN, onde foi exposto, como os procedimentos de atendimento pedagógico são realizados, e de que maneira são registrados esses casos de atendimentos, onde ficou evidente que, a dinâmica aplicada, não atingir a problemática central do atendimento que é a qualificação das informações e a transformação do conhecimento tácito de cada membro da equipe em conhecimento explícito de toda **ETEP**. Além dos mais, existe uma evolução do número de casos nos últimos

anos. Portanto, ficou evidente a necessidade do desenvolvimento de uma ferramenta de apoio, neste contexto, de aconselhamento pedagógico.

Em anos anteriores a 2012, o atendimento sequer contava com um documento padronizado para registro dos casos, sendo posteriormente criada uma ficha de atendimento, ou seja, todo registro dos casos de atendimento pedagógico são registrados, até o momento em fichas de papel. Mais recentemente, foi desenvolvido um formulário utilizando a plataforma Google Docs, que gera uma planilha compartilhada entre os membros da **ETEP**, com as principais informações cadastradas.

A equipe pedagógica relata na entrevista que:

"O acompanhamento dos alunos acontecia por verificação das notas nos boletins no final dos semestres, escrita de situações em livro ATA (registro das situações problemas) ou agenda da equipe (Até 2011). A partir de 2012, adotou-se um modelo de ficha individual que facilitava o acompanhamento da evolução dos encaminhamentos e das soluções e permitia identificar recorrências. Esse modelo foi importante, também, pela possibilidade que trouxe de partilhar informações entre as pedagogas da Diretoria sobre os atendimentos. [...] Verificou-se já um grande avanço com as fichas, mas começou-se a perceber que algumas situações de atendimento eram bem comuns (problemas que requeriam acompanhamento psicológico, por exemplo). E como verificar isso? Olhando de ficha em ficha a natureza do problema? Sem falar que requer um tempo para descrever a situação na ficha, descrever o problema, descrever o encaminhamento dado e outras informações importantes."

A equipe pedagógica, afirma ainda, a necessidade de uma ferramenta que, tornasse os dados mais "palpáveis" e mais "úteis", de forma que, fossem permitidos produzir relatórios e gráficos que exibissem as problemáticas mais comuns, a evolução dos casos e ainda a evolução dos encaminhamentos abordados e as melhores soluções ou ainda, as mais úteis. Dessa forma, foi apresentada a equipe pedagógica a proposta do **RBC**, onde pode-se, através dessa técnica, reutilizar soluções anteriores, em novas situações problema.

4.2.2. Metodologia

Como primeiro passo para a estruturação do sistema especialista, que utilize RBC, no auxílio de aconselhamento pedagógico aos alunos, foi realizar entrevista com a Equipe Técnico Pedagógica – **ETEP**, da diretoria Acadêmica de gestão e Informática, no sentido que fosse feito levantamento das principais situações que demandam atenção, bem como a classificação destas situações que se baseasse numa maior ou menor necessidade de atenção por parte dos

técnicos da **ETEP**. Como resultado deste levantamento, foram classificadas um total de 26 demandas.

Com as demandas identificadas, as mesmas foram distribuídas em cinco grupos, ou classes, de forma que possibilitou a determinação de um peso em ordem de relevância, de acordo com a avaliação dos especialistas da equipe pedagógica, onde o maior peso determina maior importância e necessidade de atenção.

A Tabela 16 mostra a classificação feita e seus respectivos pesos.

Tabela 16. Descriminação de classes e seus respectivos pesos, em ordem de relevância.

CLASSE	PESO
DIFÍCULDADE COGNITIVA/PSICOPEDAGÓGICA	5
CONFLITOS PSICOLÓGICOS	4
PROBLEMAS DE RELACIONAMENTO/COMPORTAMENTO NO ÂMBITO ESCOLAR	3
PROBLEMAS DISCIPLINARES	2
DIFÍCULDADE DE RELACIONAMENTO AFETIVO/FAMILIAR	1

Determinou-se os pesos das classes de 1 a 5, de forma que, as demandas que necessitam de atenção maior, por parte do profissional pedagogo esteja contida nas classes de peso maior, enquanto que demandas consideradas menos relevantes pertencem à classe que possuem peso menor.

Na Tabela 17, estão descritas todas as demandas que foram elencadas, no total de 26, que serão os atributos calculados na busca por similaridade entre os casos, relacionando estes atributos com os pesos que receberam de acordo com a análise pedagógica, pela classificação anteriormente determinada.

Tabela 17. Distribuição de demandas nas respectivas classes e pesos, conforme análise da especialista psicopedagógico por ordem alfabética.

DEMANDAS	CLASSE	PESO
CONFLITO FAMILIAR	E	1
PAIS EM SEPARAÇÃO	E	1
PROBLEMA DE RELACIONAMENTO COM A MÃE	E	1

PROBLEMA DE RELACIONAMENTO COM O PAI	E	1
PROBLEMA DE RELACIONAMENTO EM CASA	E	1
PROBLEMA DE ORDEM SOCIOECONÔMICA DO ALUNO OU DA FAMÍLIA	E	1
PROBLEMA DE ORDEM DISCIPLINAR GRAVE	D	2
PROBLEMA DE ORDEM DISCIPLINAR LEVE	D	2
PROBLEMA DE ORDEM DISCIPLINAR MÉDIO	D	2
PROBLEMA DE COMPORTAMENTO	C	3
PROBLEMA DE RELACIONAMENTO ALUNO X ALUNO	C	3
PROBLEMA DE RELACIONAMENTO PROFESSOR X ALUNO	C	3
BULLING	B	4
CONFLITO COM RELAÇÃO A OPÇÃO SEXUAL	B	4
CONFLITO DEVIDO A SITUAÇÃO RELACIONAL OU AFETIVA	B	4
DESMOTIVAÇÃO PELA OPÇÃO DE CURSO	B	4
DESMOTIVAÇÃO POR BAIXO RENDIMENTO	B	4
NECESSIDADE DE ORIENTAÇÃO SECULAR (PESSOAL/NÃO ACADÊMICA)	B	4
SITUAÇÃO DE ABUSO (MORAL, SEXUAL ETC.)	B	4
SITUAÇÃO DE EXCLUSÃO EM SALA	B	4
SITUAÇÃO RELACIONADA À TIMIDEZ	B	4
ATRASOS CONSTANTES	A	5
DESEQUILÍBrio OU PROBLEMA DE ORDEM PSICOLÓGICA	A	5
DIFICULDADE DE APRENDIZAGEM EM DISCIPLINA	A	5
MUITAS FALTAS	A	5
NECESSIDADE DE ORIENTAÇÃO PEDAGÓGICA	A	5

Seguindo-se no processo de desenvolvimento do sistema **RBC**, o passo seguinte é definir um método de cálculo de similaridade, ou dissimilaridade, entre os casos, de forma a tornar os processos de indexação e recuperação efetivamente funcionais. Foi decidido a utilização da Distância Euclidiana para calcular as distâncias entre os casos, onde é calculada a distância para cada

atributo que representa uma demanda entre os casos. A equação utilizada é a seguinte:

$$E(x, y) = \sqrt{\sum_{a=1}^n (x_a - y_a)^2}$$

(Equação 4.2)

A Equação 1 descreve o cálculo euclidiano para medidas de dissimilaridade, onde $E(x, y)$ é a distância e x e y são dois vetores com n atributos numéricos.

A Distância Euclidiana é uma medida de dissimilaridade, que representa um segmento de reta com a menor distância entre dois pontos, sendo escolhida para o desenvolvimento deste módulo por ser aplicável em espaços multidimensionais, com valores discretos ou contínuos. Os vetores a serem utilizados nos cálculos deste projeto, são o conjunto de demandas de cada caso registrado na base de casos, os atributos correspondem às demandas discriminadas no levantamento e classificação descrito anteriormente.

A Base de Casos de referência, foi populada com vários casos, todos bem definidos, completos e originados de experiências do mundo real. Para o levantamento dos mesmos, contamos com o apoio de profissionais especialistas em orientação pedagógica, para realizar o desenvolvimento da lista de demandas e principais soluções aplicadas aos casos, os quais tiveram conhecimento durante o primeiro período letivo do ano 2014, no Instituto Federal de Educação, Ciência e Tecnologia do RN – IFRN.

Na definição da Base de Casos, cada demanda está representada como um atributo do caso, sendo que, entre os casos elencados, alguns possuem mais de um atributo, porém, dados os limites da nossa fonte de informações, nem todos os atributos classificados estão contemplados na Base de Casos.

Estes atributos selecionados pelos especialistas, podem ser avaliados de forma condicional a estarem relacionados com o caso, e foram valorados numericamente com 0 (ausente) e 1 (presente). Ou seja, caso o atributo exista ele recebe o valor 1 (um), se não existir, o valor 0 (zero). Os atributos que não

forem respondidos ou que não são de ciência do orientador pedagógico recebem igualmente o valor 0.

Outros atributos também foram selecionados, de forma a dar uma maior especificidade aos casos e que servirão para outros propósitos, como a geração de relatórios e fichas de acompanhamento pedagógico e assim por diante. São atributos multivalorados que receberam código para seus valores, de acordo com a forma de classificação. Por exemplo, o atributo SEXO, recebe valor 0 para masculino e valor 1 para feminino. Dados com valores escalares absolutos, ou com valores textuais, como por exemplo, o percentual de rendimento acadêmico, ou a quantidade de salários da renda familiar, recebem seu valor nominal.

Essa distribuição de atributos, é importante, caso deseje-se utilizar, o processo de mapeamento, onde serão aplicadas funções para calcular a distância (dissimilaridade) entre os casos. Não deve se confundir esta função de mapeamento com o modelo de programação MapReduce, proposto por Dean e Ghemawat (2004), que está intrinsecamente ligado ao processamento e geração de dados massivos, sendo sua principal utilização em grandes massas de dados e processamento em clusters de computadores.

4.2.3. Implementação do Sistema RBC

De acordo com a definição de Raciocínio Baseado em Casos (**RBC**), é uma técnica de Inteligência Artificial, como um conjunto de atividades que auxilia na resolução de problemas, propondo soluções incontestavelmente utilizadas e documentadas, ao recuperar e adaptar experiências passadas – chamadas casos – armazenadas em uma Base de Casos. Um novo caso é resolvido com base na adaptação de solução de casos similares já conhecidos (Wangenheim *et al* 2013).

Portanto, é necessário que, exista uma base de casos e um mecanismo de inferência que, através de similaridade apresente soluções para o caso apresentado fazendo uso, quando necessário, de adaptação dos casos existentes, ao novo que foi apresentado. A Figura 4.19 do referencial teórico, descreve o processo cíclico que compõe as fases do RBC.

4.2.3.1. Casos

Como primeiro passo na construção da base de casos, está a aquisição e representação dos casos.

Depois de coletados os casos, deve-se definir a forma como serão representados. Não existe um consenso a respeito de que tipo de dados ou quais informações devem estar contidas em um caso, mas devemos ter em mente dois aspectos significativos destas informações: a facilidade de aquisição da informação e a funcionalidade desta.

Um caso ideal pode ser estruturado de forma a conter os seguintes dados, de forma estruturada: uma descrição dos aspectos relevantes do caso – para este projeto, as demandas que são objeto de atenção por parte da ETEP; o contexto no qual o caso está inserido – pode ser, um texto descritivo ou palavras-chave que, associam o caso com um contexto específico; a descrição da solução associada ao caso – neste projeto, foi descrito uma solução, como o conjunto de encaminhamentos dados a cada atendimento pedagógico, inserido na base de casos; e a avaliação da solução empregada. Na Figura 70, descreve-se a composição de um caso ideal com demandas e encaminhamentos.

```
{  
    nome: {type: String, required: true},  
    sexo: Number,  
    telefone: {type: String},  
    email: {type: String},  
    matricula: {type: String, required: true},  
    curso: {type: String, required: true},  
    periodo: {type: Number, required: true},  
    data: {type: String, required: true},  
    observacoes: {type: String},  
    demandas: {  
        desequilibrioPsicologico: Number,  
        orientacaoSecular: Number,  
        orientacaoPedagogica: Number,  
        muitasFaltas: Number,  
        problemaSocioEconomico: Number,  
        atrasosConstantes: Number,  
        bulling: Number,  
    }  
}
```

```

        desmotivacaoRendimento: Number,
        desmotivacaoCurso: Number,
        dificuldadeAprendizagem: Number,
        problemaDisciplinarLeve: Number,
        problemaDisciplinarMedio: Number,
        problemaDisciplinarGrave: Number,
        problemaComportamento: Number,
        conflitoFamiliar: Number,
        problemaRelacionamentoPai: Number,
        problemaRelacionamentoCasa: Number,
        problemaRelacionamentoAluno: Number,
        problemaRelacionamentoProfessor: Number,
        situacaoAbuso: Number,
        situacaoExclusao: Number,
        situacaoTimidez: Number,
        separacaoPais: Number
    },
    encaminhamentos: {
        psicologiaEscolar: Number,
        servicoSocial: Number,
        setorMedico: Number,
        chamarPais: Number,
        chamarProfessores: Number,
        acionarCoordenacao: Number,
        advertencia: Number,
        repreensao: Number,
        suspesaoEscolar: Number,
        trancarPeriodoCompulsorio: Number,
        trancarPeriodoVoluntario: Number,
        visitaDomiciliar: Number,
        atendimentoDomiciliar: Number,
        rotinaEstudo: Number,
        centroAprendizagem: Number,
        mudarTurno: Number,
        conversar: Number
    },
    contexto: {type: String, required: true},
    atendidoPor: {type: String, required: true},
    classe: {type: String, required: true}
}

```

Figura 70. Descrição de um caso ideal utilizando a notação **JSON**.

O que foi feito, foi o seguinte, pegarmos o domínio do sistema proposto, que é o aconselhamento pedagógico de estudantes, um caso ideal poderia ser descrito como, um documento onde existem registros com informações sobre cada estudante: nome, matrícula, turma, período e assim por diante. Estes seriam dados de contextualização.

Tem-se ainda, os aspectos relevantes do caso, geralmente nomeados de feição ou atributos, seriam as demandas de atenção, ou problemas que este aluno possa estar passando ou gerando para outrem. Ou seja, se o aluno está tendo problemas de relacionamento com outros alunos e/ou professores, se existe algum conflito familiar ou afetivo, se algum problema de ordem psicológica, econômica ou social está interferindo na aprendizagem e no convívio do aluno e assim por diante.

A descrição das soluções empregadas, incluiriam todas as ações tomadas pela orientação pedagógica, incluindo encaminhamentos e acompanhamento, por exemplo. Onde, Encaminhamentos podem ser: encaminhar efetivamente o aluno ao atendimento médico ou psicológico da escola, fazer visita à residência do aluno, elaborar rotinas de estudo, chamar os pais ou responsáveis para conversar e assim por diante.

A avaliação da solução empregada, poderia ser a descrição, ou consideração por parte da orientação pedagógica, sobre os resultados obtidos com a solução descrita e como estas contribuíram para o caso anterior, de forma a embasar a decisão de retomar esta como solução para o novo caso-problema.

4.2.3.2. Base de Casos

A estrutura de dados composta de casos, é chamada Base de Casos. E pode ser uma base de dados em forma de documentos, pares de chave-valor, objetos, predicados, redes semânticas, e assim por diante. Para o desenvolvimento da Base de Casos, nesse projeto foi utilizada a notação JSON, e sistema de banco de dados orientado a documentos para representação de casos. Retomando o exemplo do item anterior, poderíamos escrever o

documento que representa um exemplo de caso ideal, utilizando a notação JSON, da maneira descrita na Figura 70.

4.2.3.3. Indexação

Existem vários tipos de indexação. Indexação por medidas ou especificações, baseadas em diferenças, por similaridade e métodos indutivos de aprendizado são alguns exemplos que podemos utilizar na indexação. Para este projeto será utilizado, o mesmo algoritmo do k-Vizinho mais Próximos (neste trabalho abreviado para kNN, do inglês *k-nearest neighbor*) na indexação e na recuperação de casos.

4.2.3.4. Recuperação

A recuperação deve utilizar um algoritmo que verifique os clusters que tem maior similaridade com o caso-problema apresentado e, permita otimizar a reutilização e revisão do caso.

Entre os algoritmos mais relevantes e divulgados na recuperação de casos estão: Recuperação de Padrões, Algoritmo de Indução e Algoritmo de Vizinhança ou k-Vizinhos mais Próximos. Este último será o utilizado nesse projeto.

4.2.3.5. Reutilização

Raciocínio Baseado em Casos aplica-se na solução de problemas pela possibilidade de que, caso haja conhecimento do domínio aplicado ou das constantes existentes neste, pode-se aplicar uma técnica interpretativa, através da qual, podemos reutilizar uma solução conhecida de um caso semelhante, ou adaptar uma ou mais soluções conhecidas, para propor a solução para o caso apresentado.

Retomando-se o exemplo do domínio educacional, e supondo-se que uma das soluções propostas seja o encaminhamento para orientação psicológica, com o psicólogo da escola. Agora imagina-se que a escola em questão não possua um profissional disponível, deverá então, ser feita adaptação na solução para que, o aluno, seja encaminhado a um psicólogo que fará o acompanhamento externo ao ambiente escolar.

Qualquer que seja, este trabalho conta com a interação do usuário especialista do domínio que, irá determinar e irá realizar os ajustes, críticas ou adaptações nas situações que julgar necessárias.

4.2.4. Desenvolvimento do Sistema RBC

O sistema RBC proposto, faz uso do algoritmo intitulado Algoritmo **KNN**. O fluxograma simplificado do algoritmo é descrito na Figura 71.

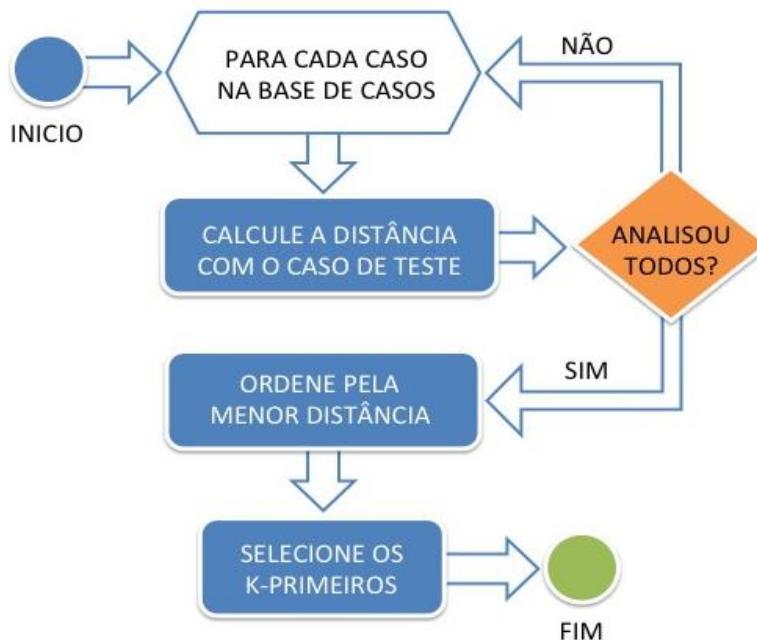


Figura 71. Fluxograma simplificado do algoritmo kNN.

Em princípio, o usuário fornece um conjunto de descrições do problema atual. Então, aplica-se uma função de cálculo para a medida de similaridade escolhida, à todas as instâncias de conhecimento contidas na base de casos. Em seguida, ordena-se os casos por ordem de similaridade, retornando um ou mais casos, até o limite estipulado por k , entre os mais próximos (Wangenheim *et al* 2013).

Uma desvantagem clara dessa abordagem, é o alto valor computacional do algoritmo, pois, é necessário comparar os dados de testes de cada novo caso apresentado, com todos os outros presentes na base de casos, recalculando as distâncias cada vez que é utilizado.

As principais características que, levam a utilizar-se este algoritmo, se encontra na sua facilidade de implementação e, na possibilidade de adaptação para diferentes entradas com múltiplas dimensões, inclusive.

Uma etapa natural deste algoritmo, é a classificação da entidade testada, inferindo a que grupo o conjunto de dados de teste pertence, com base na classe dominante, à qual pertencem os vizinhos mais próximos. Neste trabalho, esta etapa do algoritmo é descartada, em vista de que, optou-se por agrupar os casos, de acordo com a classe à qual pertence a demanda de maior peso presente no caso.

A Figura 72 exibe uma implementação de parte deste algoritmo, com ênfase no cálculo euclidiano de distâncias.

```
// Realiza o cálculo euclidiano para distâncias
var calcularDistancias = function (caso, vizinho) {
    var pesos = {
        atrasosConstantes: 5,
        desequilibrioPsicologico: 5,
        dificuldadeAprendizagem: 5,
        muitasFaltas: 5,
        orientacaoPedagogica: 5,
        bulling: 4,
        conflitoOpcionalSexual: 4,
        conflitoRelacionalAfetivo: 4,
        desmotivacaoCurso: 4,
        desmotivacaoRendimento: 4,
        orientacaoSecular: 4,
        situacaoAbuso: 4,
        situacaoExclusao: 4,
        situacaoTimidez: 4,
        problemaComportamento: 3,
        problemaRelacionamentoAluno: 3,
        problemaRelacionamentoProfessor: 3,
        problemaDisciplinarGrave: 2,
        problemaDisciplinarLeve: 2,
        problemaDisciplinarMedio: 2,
        conflitoFamiliar: 1,
        separacaoPais: 1,
        problemaRelacionamentoMae: 1,
        problemaRelacionamentoPai: 1,
        problemaRelacionamentoCasa: 1,
        problemaSocioEconomico: 1
    };
    // insere o peso nas demandas
```

```

var c = [];
var v = [];
for (demanda in caso) {
    c[demanda] = caso[demanda] * pesos[demanda];
    v[demanda] = (vizinho[demanda] != undefined) ?
(vizinho[demanda] * pesos[demanda]) : 0;
}
// calcula a soma dos quadrados das diferenças
var soma = 0;
for (demanda in c) {
    soma += Math.pow(c[demanda] - v[demanda], 2);
}
return Math.sqrt(soma / 26); // retorna a distância normalizada
}

```

Figura 72. Implementação do Cálculo de distância dos Casos na linguagem Javascript.

O primeiro passo, é definir o peso de cada entrada, conforme classificação da Tabela 4.1. Em seguida, percorre-se os vetores de entrada, ou seja, o caso-problema (caso) e um caso presente na Base de Casos (vizinho), multiplicando cada atributo pelo peso correspondente. Segue-se então a soma dos quadrados das diferenças de todos os atributos e por fim divide-se pelo intervalo de demandas e faz-se o cálculo da raiz quadrada, retornando seu valor que, conforme o cálculo euclidiano normalizado, é a distância entre os casos. Repete-se o cálculo para cada entrada registrada na Base de Casos.

Uma outra particularidade deste trabalho, é o fato de que, todos os casos deverão ser armazenados, já que deverão ser registradas, todas as atividades de atendimento pedagógico atendidas pela ETEP. Para tanto, será utilizada uma segunda coleção de registro de casos, esta última, apenas para fins estatísticos e administrativos, sendo que, estes aspectos da aplicação, não serão exploradas no presente trabalho.

4.2.4.1. Arquitetura do sistema RBC

A arquitetura de um sistema pode ser definida como a descrição da organização dos diversos componentes do sistema e como estes, e o sistema, interagem entre si e com o ambiente no qual estão inseridos. A arquitetura

selecionada para o desenvolvimento deste trabalho é o de cliente-servidor, comum em ambiente web. A Figura 73 demonstra essa arquitetura.

Na arquitetura cliente-servidor, conforme descrito na Figura 73, o usuário através da aplicação cliente realiza uma requisição e envia os dados ao servidor, que processa e responde ao cliente com as informações (Mendes 2002). Os dados são persistidos na base de casos e podem ser processados pelo servidor.

Esse modelo arquitetural é especialmente entendido como um modelo em camadas, onde permite-se compreender o papel de cada camada do sistema de forma isolada.



Figura 73. Modelo arquitetural de aplicação cliente-servidor.

O sistema proposto é compreendido em três camadas: 1) interface com usuários – é a camada que os usuários utilizam para interagir com o sistema; 2) camada de negócio – presente no lado servidor, corresponde a lógica de domínio da aplicação; 3) camada de persistência – nesta camada os dados da aplicação são guardados em um repositório ou banco de dados.

4.2.4.2. Interface com o Usuário

A camada mais superficial da aplicação é a interface com o usuário. A interface do sistema RBC foi desenvolvida usando um conjunto de documentos e formulários escritos utilizando HTML5 e algumas bibliotecas de script Javascript que podem ser interpretadas pelo navegador de internet.

Entre as bibliotecas utilizadas, destaca-se **AngularJS**, por ser uma biblioteca escrita em Javascript, que tem seu foco na manipulação de interface,

através de uma tecnologia nomeada *two-way data binding*. Esta tecnologia permite que, a interface seja alterada conforme a lógica do sistema mude, sem a necessidade de recarregar a aplicação para visualizar as novas informações, assim como, também reflete as ações do usuário na lógica do sistema em tempo de execução (Google Inc. 2010). Ou seja, se relacionarmos os campos de um formulário a um objeto que pode ser manipulado com AngularJS, a cada interação que o usuário tenha no preenchimento deste, o objeto é modificado simultaneamente, permitindo, por exemplo, que os dados preenchidos possam ser verificados mesmo antes do usuário decidir encerrar a interação.

Com o uso desta tecnologia, o modelo passa a ser o recurso único da verdade (**SSOT** – *Single Source of Truth*, em inglês). **SSOT** refere-se a prática de estruturar a informação, de maneira que exista apenas uma fonte de informação daquele dado, quaisquer outras formas de acesso aos dados, se dá por referência apenas (Google Inc. 2010). A Figura 74 ilustra o funcionamento do *two-way data binding*.

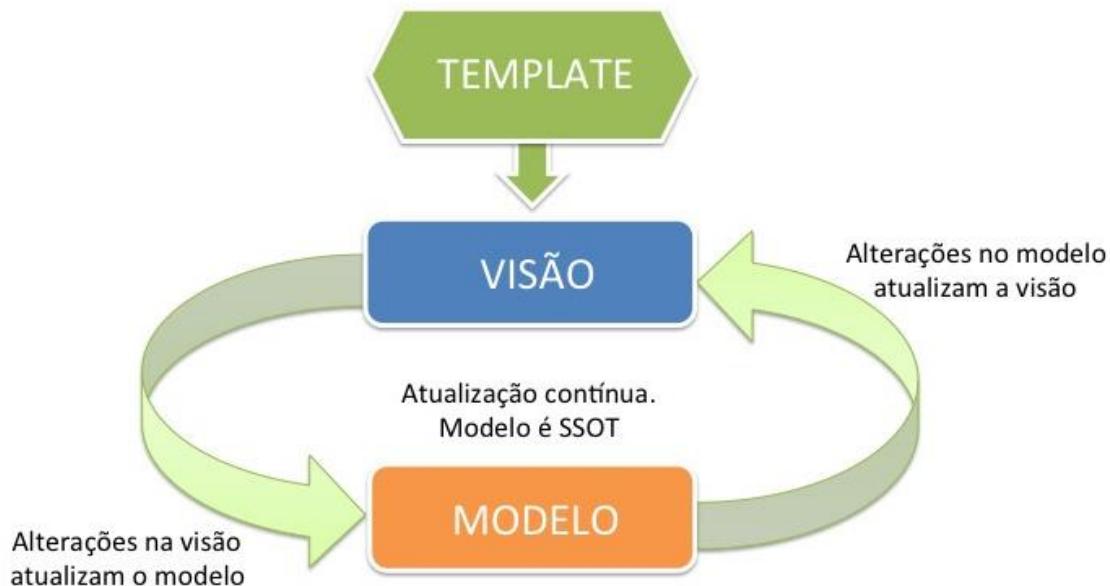


Figura 74. Esquema de *Two-way data binding* na biblioteca AngularJS.

Fonte: Adaptado de Google Inc, 2010.

4.2.4.3. Servidor de Aplicações

A segunda camada da aplicação fica alocada no servidor, responsável por receber as solicitações da interface com o usuário e, responder de acordo com o resultado do processamento da requisição. Esta camada intermediária é executada, em um ambiente de servidor desenvolvido para este fim, através da tecnologia usada para processar **Javascript** no lado do servidor chamada **NodeJS**.

NodeJS é uma plataforma de processamento, que utiliza a tecnologia direcionada a eventos, com o modelo de **I/O** não-bloqueante, leve e eficiente. Geralmente utilizada em aplicações de tempo real, com intensa transmissão de dados e, que funcionam como sistema distribuído. **I/O** é a sigla, em inglês, para o processo de entrada e saída de informações no servidor (Joyent Inc 2014).

Dizer que, o modelo de entrada e saída é não-bloqueante, significa que o servidor, quando receber uma requisição, irá processá-la e em algum momento ele poderá responder aquela requisição, porém, para tanto, não deixará de receber e responder outras requisições que porventura cheguem até o mesmo.

Uma das características que permite o **NodeJS** ser não-bloqueante é que este é *single-thread*, ou seja, a aplicação terá somente uma instância de cada processo, sendo, no entanto, possível a criação de clusters, como no caso da biblioteca **Mongoose**, uma biblioteca escrita em Javascript para facilitar o trabalho do desenvolvedor que deseja conectar a aplicação com o **SGBD MongoDB** (LearnBoost 2010).

Outra particularidade, é a forma como se dá, a orientação a eventos do **NodeJS**. Enquanto que, nos navegadores de internet, os eventos aos quais o Javascript, estão relacionados serem, por exemplo, o clique do mouse, o pressionar de uma tecla etc., o servidor de aplicação **NodeJS** não responderá a estas ações. Os eventos a que estão relacionadas as atividades dele são, entre outros, *connect* (conexão com banco de dados), *open*, *read*, *close* (operações em arquivos em geral), *data* (como em transmissão de dados) e assim por diante.

A tecnologia de *callback* do Javascript tem um papel importante em todas as etapas de processamento do **NodeJS**, já que este é um item crucial para o gerenciador de eventos do servidor. É através dela que, podemos mesmo com a limitação de ser *single-thread*, realizar operações paralelas assíncronas de forma não-bloqueante (Joyent Inc 2014).

O mecanismo utilizado para o gerenciamento destes eventos é o *Event-Loop*. O *event-loop*, na verdade, é um ciclo infinito que percorre, a cada iteração, toda a sua lista de eventos e verifica se algum evento foi emitido. Quando identificado um evento, ele o executa e o encaminha para a fila de executados. O poder deste mecanismo de processo está na possibilidade de, enquanto um evento é processado, podermos definir funcionalidades que serão disparadas, e descrever a lógica destes. A Figura 75, exemplifica a execução do mecanismo de *Event-Loop* do NodeJS.

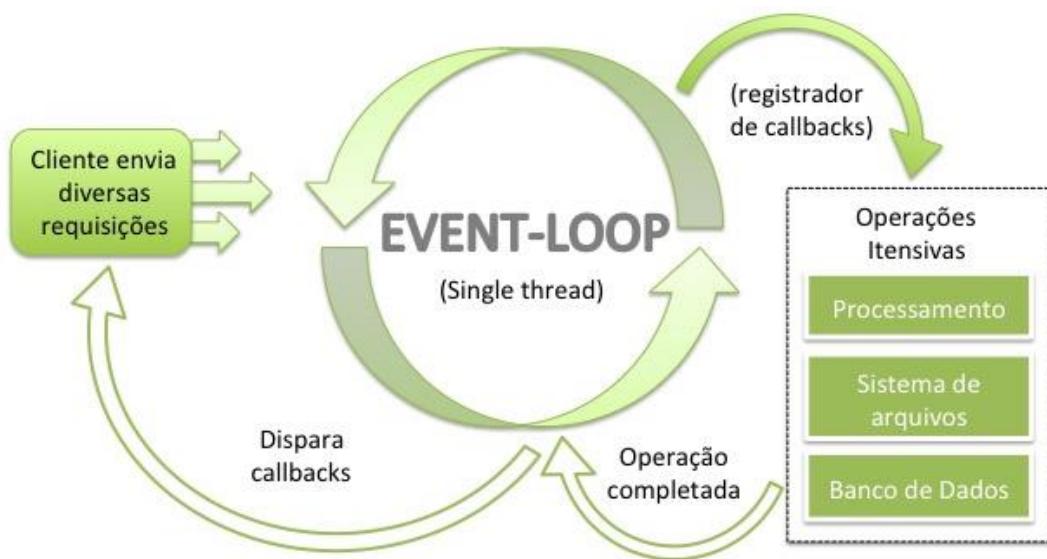


Figura 75. Mecanismo Event-Loop do NodeJS.

Fonte: Adaptado de Joyent Inc., 2014

4.4.4. Persistência de Dados

A terceira e última camada, é a camada de persistência dos dados. Foi utilizado um SGBD (Sistema Gerenciador de Bancos de Dados) que nos permitisse armazenar os casos de forma otimizada e mantivesse a mesma estrutura de documentos utilizada nas camadas superiores da aplicação, fazendo assim, com que os dados sofressem o mínimo de alterações possíveis e que somente as transformações necessárias fossem realizadas.

Como um fator importante ao **RBC**, a base de casos ser persistida em um banco de dados, que mantém sua estrutura na forma de coleção de documentos,

e ainda permite a utilização nativa de algoritmos de classificação, nos auxiliando a obter resultados de forma nativa, tende a ser altamente eficiente.

Foi utilizado, para desenvolvimento do projeto objeto deste trabalho, um **SGBD NoSQL**, orientado a documentos, de código aberto e uso livre, de nome **MongoDB** (MongoDB Inc. 2014).

O **MongoDB**, utiliza também Javascript como linguagem nativa para suas rotinas e aplicações, além de armazenar os dados, em um formato semelhante à notação de objetos Javascript (**JSON**), só que em binário ao invés de texto pleno. A Figura 76 ilustra a comparação entre um documento representado em texto pleno e sua representação em binário utilizando a notação Javascript.

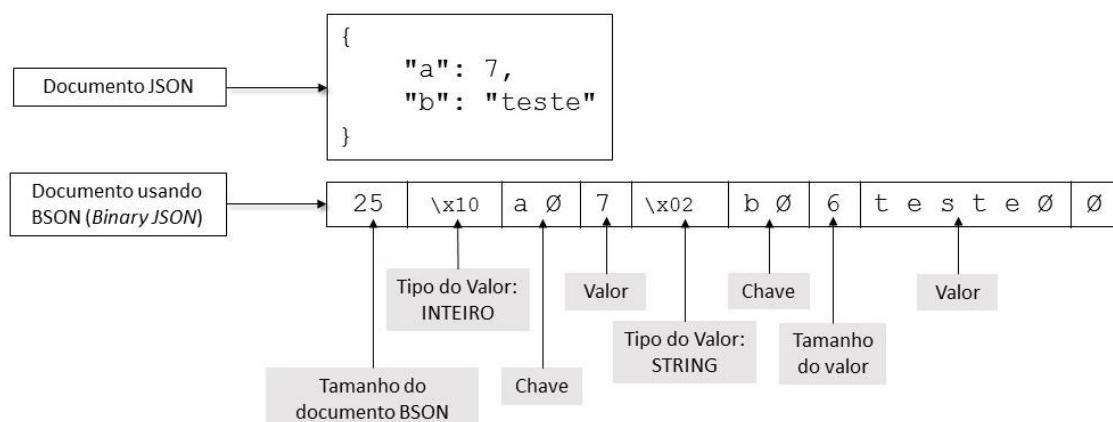


Figura 76. Comparação da representação de um mesmo documento usando **JSON** em texto pleno e na sua representação em binário, **BSON**.

Fonte: Adaptado de Especificação BSON

4.2.5. Resultado

Como meio de disponibilização, e visando a possibilidade de futuras contribuições da comunidade, escolhemos criar um repositório público de código usando a plataforma web, na comunidade de desenvolvedores GitHub.com. O endereço **URL** de acesso público é <https://github.com/embits/educase>. Esta plataforma, além de permitir a publicação, possui uma ferramenta de versionamento de código, que se mostrou bastante útil durante o período de desenvolvimento.

A Figura 77 representa uma das telas do sistema, mais especificamente, o cadastro de informações de contextualização e discriminação do caso, com informações sobre o aluno, curso, informações de contato.

The screenshot shows a web-based application titled 'Atendimento Pedagógico'. At the top, there's a navigation bar with links for 'Atendimento', 'Relatórios', 'Sobre o projeto', and 'Contato'. On the right side of the header, it says 'Olá, usuário@ifrn.edu.br' and has a 'sair' (Logout) link. The main content area is titled 'Informações pessoais do aluno'. It contains several input fields: 'DATA *' with value '00/00/0000', 'MATRÍCULA *' (empty), 'NOME *' with placeholder 'Nome Completo do Aluno', 'CURSO *' and 'PERÍODO *' both set to 'Selecione', 'SEXO *' set to 'Selecione', 'TELEFONE' with value '(84)0000-0000', and 'EMAIL' with value 'email@ifrn.edu.br'. Below these is a section titled 'DADOS SIGILOSOS' with a note about entering sensitive information related to the case. A large text area labeled 'CONTEXTO *' is present with a placeholder text about describing the context of the case. At the bottom right of the form area, there's a green button labeled 'DEMANDAS >'. To the right of the form, there's a decorative graphic of a brain made of circuit boards.

Figura 77. Interface com o usuário do sistema. Passo 1 – Informações pessoais do aluno.

Fonte: Autor

Através de um formulário interativo, é realizado o cadastro do atendimento pedagógico em 3 passos, sendo que estes passos incluem as etapas de recuperação, reutilização, revisão e retenção do **RBC**, a partir das informações fornecidas e da consulta na base de dados, conforme descrito a seguir.

No primeiro passo, são fornecidas as informações pessoais do aluno, como nome, matrícula, curso, sexo, data do atendimento, entre outras. Nesta tela também deve ser fornecido uma descrição do contexto no qual o caso está inserido, descrevendo locais, pessoas e situações, que poderão servir inclusive em um momento posterior, quando este novo caso já fizer parte da base de casos e vier a ser sugerido como uma das prováveis soluções.

No segundo passo, o usuário deverá elencar as demandas, dentre a lista de 26 demandas elencadas pela equipe da **ETEP** e descrita na Tabela 4.2 supracitada. Após seleciona as demandas que estão relacionadas a este novo caso, o usuário deverá clicar no botão “Encaminhamentos”, de cor verde, ilustrado na Figura 78, que realizará uma requisição assíncrona ao servidor informando as demandas que foram selecionadas. O servidor por sua vez

calculará, dentre todos os casos registrados na Base de Casos, os vizinhos mais próximos com base no cálculo euclidiano de distância e irá responder a solicitação sugerindo as soluções dos casos mais similares. Este é o primeiro passo do processo de Raciocínio Baseado em Casos, a recuperação.

The screenshot shows a web-based user interface for selecting demands for a new case. On the left, there is a vertical sidebar with a grey header and a white body containing a list of items. Each item consists of a small grey button on the left labeled 'NÃO' or 'SIM' and a text label on the right. The text labels describe various social and family issues. To the right of the sidebar is a large, stylized graphic of a human brain with a circuit board pattern, symbolizing knowledge or computation. At the bottom of the screen, there is a green footer bar with white text. The main content area has a light grey background.

Opção	Descrição
NÃO	SITUAÇÃO DE EXCLUSÃO EM SALA DE AULA
NÃO	SITUAÇÃO DE ABUSO (SEXUAL, MORAL ETC)
NÃO	NECESSIDADE DE ORIENTAÇÃO SECULAR (NÃO-RELACIONADO À VIDA ACADÉMICA)
NÃO	PROBLEMA DE COMPORTAMENTO
SIM	PROBLEMA DE RELACIONAMENTO ALUNO - PROFESSOR
NÃO	PROBLEMA DE RELACIONAMENTO ALUNO - ALUNO
NÃO	PROBLEMA DE ORDEM DISCIPLINAR LEVE
NÃO	PROBLEMA DE ORDEM DISCIPLINAR MÉDIO
NÃO	PROBLEMA DE ORDEM DISCIPLINAR GRAVE
SIM	PROBLEMA DE RELACIONAMENTO COM O PAI
NÃO	PROBLEMA DE RELACIONAMENTO COM A MÃE
SIM	PROBLEMA DE RELACIONAMENTO EM CASA
NÃO	PAIS EM SEPARAÇÃO
SIM	CONFLITO FAMILIAR
NÃO	PROBLEMAS DE ORDEM SÓCIO-ECONÔMICA DO ALUNO OU FAMÍLIA

Trabalho de Conclusão de Curso - 2014.1 - IFRN - Diretoria Educacional de Gestão e Informática © Elionai Moura

Figura 78. Passo 2 – Seleção de demandas do novo caso.

Fonte: Autor

Quando o servidor responde à solicitação, o usuário poderá optar por, aplicar a solução que melhor julgar conveniente à situação do novo caso, sendo esta a etapa de reutilização, ou ainda verificar uma ou mais soluções sugeridas e adaptar ao cenário do novo caso. A tela de sugestões de solução também permite que seja visualizado o contexto no qual aquelas sugestões estão relacionadas. Esta é a etapa de revisão do **RBC**.

Por último, é feito o cadastro do caso novo, sendo enviada, nova requisição ao servidor, de forma que este deverá calcular se, o conjunto formado das demandas e dos encaminhamentos (soluções) apresentados, forma uma nova unidade de conhecimento, a qual deverá ser inserida à base de casos, encerrando o ciclo **RBC** com a fase de retenção, ou aprendizado.

O processo de cadastro de um novo caso, no último passo, possui algumas particularidades que merecem uma maior atenção, inclusive, porque aqui são realizadas algumas etapas importantes que, irão influenciar não

somente nossa base de dados, mas também os registros gerais dos atendimentos pedagógicos e os problemas enfrentados pela **ETEP** do **IFRN**, conforme apontado pela pedagoga entrevistada.

Quando o servidor recebe a requisição, para persistir o registro de um caso, a primeira ação que este deve tomar é a de classificar o novo caso. Como exposto nessa seção, onde falamos sobre o algoritmo **kNN**, este é comumente utilizado para classificação, é através desse algoritmo que feita a proposição de classes com base na classe dominante entre os vizinhos mais próximo ao caso dado.

O **kNN** foi utilizado neste trabalho, tanto para fins de indexação, quanto para recuperação de casos na Base de Casos, através do uso do cálculo de distância e da seleção dos vizinhos mais próximos, porém na fase de classificação, utilizou-se o critério de classificar conforme a classe do atributo mais significativo presente no caso-problema, visto que, de acordo com a metodologia de enfrentamento de problemáticas que a equipe técnico-pedagógica utiliza, classificar os casos com base não em seus vizinhos mais próximos, mas sim com base na classe mais significativa entre as demandas que o compõe, facilitará a produção de dados estatísticos de forma a elencar mais evidentemente os problemas que necessitam de maior atenção.

Isto se dá pelo fato de que, muitas vezes, apesar de a classificação sugerida pelo **kNN** indicar determinado agrupamento, como a classificação foi inicialmente realizada em cima de cada uma das demandas especificamente, ocorre de serem sugeridas, às vezes, classificar o novo caso dentro de um grupo, ou classe, na qual este não possui nenhuma demanda relacionada.

Identificou-se que, este fenômeno ocorre devido a classe dominante estar presente em casos, que possui uma variedade de demandas, que influenciam aqueles casos que, de outra forma, estariam agrupados de maneira distinta à sugerida pelo algoritmo do vizinho mais próximo.

Outro ponto importante de abordar, na persistência da informação, tem a ver com a produção de conhecimento *versus* a necessidade de acumular informações dos atendimentos pedagógicos. A Base de Casos deve ser constituída de informações únicas. Cada caso ali presente, deve representar um conjunto isolado de demandas e soluções, de maneira que, ao ser solicitada a persistência de um caso que não produza um exemplar único e realmente novo

de conhecimento a ser agregado, este deverá ser descartado, de forma a manter a integridade e a performance da técnica de **RBC**. Por outro lado, a atividade da **ETEP** necessita que todos os atendimentos sejam registrados, de forma a criar um banco de informações que reflitam as situações do mundo real. A solução encontrada para que, ambas condições sejam satisfeitas, foi a utilização de duas coleções distintas no **SGBD**. Uma será nossa base de casos, onde as solicitações de recuperação, reutilização, revisão e retenção serão cumpridas e calculadas com base nos princípios da Inteligência Artificial, com unidades distintas de conhecimento, da forma descrita anteriormente.

A outra coleção, irá persistir todas as solicitações de retenção de casos, mesmos aqueles que não gerem uma nova unidade de conhecimento para o **RBC**, mas que, registra e representa o conjunto global de atendimentos realizados. Esta segunda coleção irá auxiliar, entre outras coisas, na produção de relatórios, como demonstrado na Figura 79, que ilustra o relatório de demandas mais atendidas. Este tipo de relatório será um ferramental importante para a ETEP realizar atividades de prevenção e cuidados básicos, tornando as informações colhidas durante os atendimentos “mais palpáveis e mais úteis”.



Figura 79. Relatório de Demandas mais atendidas.

Fonte: Autor

O Sistema Inteligente RBC desenvolvido, faz parte de um dos módulos proposto neste trabalho e, será acessado pela equipe pedagógica do IFRN para solucionar problemas relacionados aos alunos, em função da demanda elencada na Tabela 4.2.

Este sistema já foi testado e aprovado pela equipe pedagógica do IFRN. Apesar de ainda ser um protótipo, o mesmo está tendo grande aceitação pela comunidade IFRN, tanto é que o diretor geral, me procurou para sugerir acrescentar mais informações ao sistema, como por exemplo, dados do setor psicológico e, isto significa acrescentar mais variáveis as demandas e, possivelmente aos encaminhamentos. Estas mudanças poderão acarretar em ajustes na formula de indexação.

4.3 Gamificação como uma Ferramenta de Ensino

O principal foco desta seção, é olhar para elementos de jogos digitais que, possibilitem potencializar o processo de ensino e aprendizagem. A Gamificação, emprega a tentativa de melhorar o envolvimento dos usuários, na aplicação de elementos de jogos em atividades de não jogos. Ela pode ser um diferencial no processo de ensino e aprendizagem, pois proporciona condições de aprendizagem atuais, mais conectadas com o mundo real. Por meio dela, buscamos levar a inovação e o desenvolvimento na maneira de ensinar.

Constitui-se como problema da pesquisa, buscar alternativas para educar estudantes utilizando recursos tecnológicos, em tempos, onde estes recursos já estão tão presentes em nosso cotidiano, diferentemente do cenário onde alguns anos atrás, a maioria dos professores atuais foram formados. Buscamos utilizar a gamificação como uma ferramenta de ensino que, seja capaz de combater a falta de interesse, ou evasão do aluno na disciplina, de forma que, esse objeto de ensino, seja um auxílio para que os estudantes utilizem como um recurso a mais na sua aprendizagem.

Pretende-se envolver o aluno no ensino de uma forma iterativa, por meio de jogos digitais. Sabemos o poder e a influência que os jogos têm sobre nós, e se nesse trabalho, pretende-se usar essa atratividade, com intuito de engajar cada vez mais os alunos no desempenho de algumas disciplinas, que se apresentaram, no decorrer da pesquisa, como aquelas em que os alunos têm mais dificuldades de cursar.

As instituições de ensino por meio da educação, nos preparam para o futuro, e é um ambiente onde a inovação precisa estar presente, e que os profissionais que atuam na educação, precisam estarem preparados para aceitar e utilizar a inovação tecnológica. Atualmente, disponibilizamos de recursos tecnológicos aos quais, podemos fazer uso para inovar na maneira de ensinar, e oferecer ferramentas de auxílio a aprendizagem motivadoras e engajadoras. Podemos fazer hoje, o que a alguns anos atrás, seria pouco provável em termos de tecnologia, e possuímos ferramentas apropriadas para auxiliar as dificuldades dos estudantes nas instituições de ensino.

A ideia é aumentar a eficácia no processo de ensino e aprendizagem, oferecendo um canal onde o aluno possa buscar auxílio sem se sentir intimidado, e além disso, onde a forma de ensino seja atrativa e ele possa recorrer ao mesmo

assunto inúmeras vezes. O objetivo é usar a gamificação como ferramenta de ensino, para tentar estimular o estudante a buscar o conhecimento e também, usar a ferramenta de maneira produtiva na solução de problemas.

4.3.1. Motivação

Um percentual muito alto, dos alunos que ingressam no Instituto Federal, são oriundos de escolas pública estaduais e, como ficou evidente nos dashboards apresentados na seção anterior, essas escolas não preparam os seus alunos adequadamente, e como consequência, esses alunos ao ingressarem em nossos cursos, principalmente nos cursos da área tecnológica, sentem muita dificuldade ao buscar o conteúdo necessário para ajudá-los, e isso fica muito evidente nas disciplinas que envolvam matemática. Essa deficiência acaba gerando alto índice de reprovação ou evasão escolar.

O professor, é a pessoa que está posicionada na linha de frente dessa interação, é o principal responsável por fazer funcionar as estratégias pedagógicas de ensino e aprendizagem. Deste modo, nos questionamos sobre, como fazer para auxiliar os estudantes em disciplinas que eles tenham dificuldades, em um espaço que se organize de maneira diferente daquele onde a maioria foi formada. A ideia é tornar a aprendizagem de conteúdos desgastantes para os alunos, em uma forma divertida de aprendizagem.

Portanto, a gamificação tem essa característica de aumentar a motivação e o engajamento dos usuários, e dessa maneira, pode ser encarrada, como uma alternativa para despertar o interesse nos alunos a aprender de forma dinâmica e a tornar mais agradáveis disciplinas consideradas tediosas ou repetitivas. Portanto, é proposto neste trabalho, disponibilizar uma ferramenta interativa de auxílio ao aluno, que seja capaz de motivá-lo a aprender um conteúdo específico, sem expor a sua dificuldade aos demais, o que gera desconforto na maioria das vezes. Além disso, contribui para o desenvolvimento de novas tecnologias a serem utilizadas como recurso de aprendizagem na educação.

4.3.2. Justificativa

Pretende-se com o uso de jogos, trabalhar a dificuldade dos alunos de forma individual, mas que atinja um amplo grupo de alunos que tenham essa dificuldade em comum, e que usem a plataforma como um recurso. A incorporação de elementos de jogos, em cenários de sala de aula, é uma forma de proporcionar aos alunos, a oportunidade de tomar a iniciativa na busca pelo conhecimento. Os elementos de jogos, são uma linguagem comum a todos os alunos, ou seja, é um canal adicional através do qual os alunos vivenciam uma experiência real de ensino por meio de jogos.

Atualmente, nos encontramos em uma sociedade imersa na cultura digital, estamos vivenciando essa mudança a cada dia que se passa, e precisamos modernizar os meios de ensino de forma dinâmica para os alunos, proporcionando um espaço para que, possa ser adquirido conhecimento, e assim possamos aumentar a motivação e aprendizado. O ensino por meio de jogos, cria novos cenários para transmissão de conhecimento e, proporcionam mais absorção do conteúdo abordado de forma divertida, a liberdade para falhar e tentar novamente sem repercussões negativas e fornece ao aluno um conjunto gerenciável de tarefas, e motiva os estudantes para a aprendizagem.

Nossa intenção é apoiar a construção de práticas pedagógicas inovadoras e assim, contribuir para os processos de ensino e aprendizagem em sala de aula.

A gamificação traz inovações à Educação, principalmente no estímulo à aprendizagem experencial (baseada em experiência do mundo real) e à Educação a Distância. Devemos buscar metodologias estimulantes para o ensino, e nos adaptar com a incorporação de novas tecnologias que visam essa melhoria no processo de ensino aprendizagem.

4.3.3. Objetivos dessa Seção

Nesta seção, pretende-se aplicar técnicas de Gamificação, para desenvolver jogos educativos, com a finalidade de criar atividades engajadoras, na tentativa de incentivar os alunos a melhorarem os seus desempenhos escolares nas disciplinas e, dessa forma, tentar diminuir o alto índice de repetência e evasão escolar no Instituto Federal de Educação, Ciências e Tecnologia do Rio Grande do Norte – **IFRN**. Portanto, esta pesquisa tem também

como foco, o uso de técnicas de **Gamificação**, para o desenvolvimento de jogos educativos.

4.3.3.1. Objetivos específicos

- Pesquisar e selecionar quais são as disciplinas e assuntos específicos que os alunos encontram mais dificuldades;
- Aplicar as técnicas de Gamificação para modelar cenários para os assuntos das disciplinas selecionadas na pesquisa.
- Desenvolver os jogos educativos aplicando as técnicas da Gamificação (para ambientes web e móvel);
- Testar os jogos em situações reais de utilização.

4.3.4. Metodologia

Esta sessão tem como finalidade descrever o modo como será desenvolvida esta pesquisa, em relação aos objetivos definidos. Primeiramente foi feita uma pesquisa bibliográfica relacionada à Gamificação, ou seja, foi pesquisado o estado da arte da Gamificação, apresentado no referencial teórico desta pesquisa e, em seguida o desenvolvimento do projeto, que compreende a análise das disciplinas onde os alunos têm mais deficiência e, então serão criados os cenários dos assuntos a serem implementados nos jogos. Será feito também um estudo das tecnologias a serem utilizadas no desenvolvimento do jogo.

4.3.5. Desenvolvimento do Jogo

O primeiro passo no desenvolvimento do jogo é identificar quais disciplinas, dentre as ministradas no **IFRN**, são as que reprovam mais e, então selecionar uma delas para ser a contemplada, no desenvolvimento do primeiro protótipo. Um dos nossos dashboards do projeto de **BI** (Figura 51 Dashboard Repetência por Disciplina, página 162), lista as 5 disciplinas com maior número de reprovações (Instalação de Computadores, Matemática, Autoria Web, Informática e Língua Portuguesa).

Para este primeiro protótipo foi escolhida a disciplina de matemática. A justificativa para a escolha da matemática é que, a matemática é uma disciplina

chave para todos os cursos da área tecnológica, tais como os cursos de informática, mecânica, eletrotécnica, geologia, mineração, e assim por diante. Além do mais, foi feito um estudo, por parte de um professor de matemática do campus Natal central, onde o mesmo analisou, as provas do exame que dá acesso aos alunos aos nossos cursos e, constatou que, muitos dos alunos aprovados no exame, acertam duas questões de matemática, de um total de 15 questões. Isto é preocupante, pois a matemática é de fundamental importância, principalmente nos cursos da área tecnológica.

Portanto, o jogo a ser implementado neste primeiro protótipo, será um jogo baseado em alguns assuntos da matemática. Optei também em começar a trabalhar com assuntos mais básicos, separados por níveis, de forma que, à medida que o usuário do jogo for avançando de nível, a matemática vá também ficando mais avançada.

4.3.6. Projeto do Jogo

O protótipo desenvolvido está relacionado a aprendizagem baseada em desafios, a partir de temáticas levantadas pelo sistema de BI, também desenvolvido nesse projeto. Este protótipo foi desenvolvido com a aplicação dos conceitos de Gamificação, onde são estabelecidas algumas formas de recompensas pelas atividades desenvolvidas pelo aluno. Deve-se salientar que este protótipo utiliza as mecânicas de jogos mais comuns, tais como níveis, metas e pontos e quadro de resultados. Estrutura semelhante a definida por Nicholson (2012) como **BLAP** da Gamificação (Badges, Leaderboards, Achievements and Points) que correspondem a Emblemas, Quadro de Líderes, Conquistas e Pontos.

Entretanto, é importante ressaltar que os conceitos de gamificação implementados no protótipo proposto, vão além dos elementos **BLAP**. O próprio Nicholson (2012) introduz a definição de “Gamificação com significado” (do inglês, *meaningful gamification*). Segundo este conceito, em uma atividade não lúdica é importante criar um nível de envolvimento semelhante ao que se pode obter com jogos, gerando uma experiência com significado que não depende apenas de recompensas extrínsecas.

Com o propósito de promover a motivação do estudante e envolver-lo na dinâmica propostas pelas recompensas do jogo, foi adotada a visão de

Zichermann (2011) que propõem uma estrutura que lida com a maneira com que o jogador valoriza a recompensa, conhecida pelo acrônimo SASP: Status, Acesso, Poder e Coisas (do inglês, *Status, Access, Power, Stuff*). O Status permite que o jogador (estudante) veja o seu progresso através de níveis e emblemas e medalhas virtuais. O Acesso significa ter o acesso a determinadas funcionalidades do jogo. O Poder é a acapacidade de ter acesso a algumas funcionalidades do jogo, como por exemplo, poder trocar pontos por medalhas. Por último, as Coisas são os objetos reais obtidos pelo jogador e que tem valor, como por exemplo as medalhas adquiridas pelo jogador durante o jogo.

4.3.6.2. Descrição do Jogo

Na janela principal disponibiliza as disciplinas a serem trabalhadas nos jogos e algumas informações adicionais. Selecionada a disciplinas, o jogador será direcionado a uma tela de mapa do jogo, onde cada assunto da disciplina selecionada é um jogo. Os assuntos a serem trabalhados no nível 1 da disciplina de matemática, são os seguintes:

- Conjuntos numéricos;
- Operações matemáticas;
- Fatoração;
- Razão e proporção;
- Sequências e progressões;
- Princípios de contagem;
- Relação e função;
- Funções algébricas;
- Equações e inequações;
- Triângulos.
- Matemática financeira e estatística.

As regras básicas do jogo são as seguintes: Para cada jogo (assunto), o aluno terá que resolver 10 questões. As questões serão divididas por níveis de dificuldades fácil, médio e difícil. Cada questão a ser resolvida terá um tempo para solução. A Tabela 18, lista o tempo para cada nível.

Tabela 18 Tempo para a ser disponibilizado para a solução de cada questão.

NÍVEL	TEMPO
FÁCIL	60 segundos
MÉDIO	120 segundos
DIFÍCIL	180 segundos

Cada questão do jogo terá uma pontuação, obedecendo o seguinte critério: questão fácil vale 5 pontos cada, questão do nível médio, vale 10 pontos cada e questão difícil, vale 15 pontos cada. Portanto, como no nível 1 temos 10 jogos e cada jogo 10 questões, então, o jogador poderá acumular 500 pontos, ao concluir todo nível 1. A Tabela 19, faz um detalhamento da pontuação por nível.

Tabela 19 Pontuação de jogo por nível.

NÍVEL DO JOGO	Nº DE JOGOS	Nº QUESTÕES	VALOR DA QUESTÃO	TOTAL PONTOS
1	10	10	5	500
2	10	10	10	1000
3	10	10	15	1500

A pontuação obtida pelo jogador, poderá ser trocada por medalhas, obedecendo o seguinte critério: 100 pontos por ser trocado por uma medalha de bronze, 300 pontos por uma medalha de prata e 500 pontos, por uma medalha de ouro. Veja a Tabela 20, mostra o quadro de medalhas.

Tabela 20. Quadro de medalhas.

MEDALHA	PONTUAÇÃO
OURO	500
PRATA	300
BRONZE	100

4.3.6.3. Plataforma do Jogo

Este jogo foi desenvolvido para ser jogado nas plataformas a partir do navegador da internet e dispositivos móveis.

4.3.6.4. Regras do Jogo

1. Na tela principal do sistema, o jogador escolhe o jogo que deseja jogar.
2. O jogador é direcionado para a janela do jogo.
3. Cada jogo tem 10 questões para serem resolvidas e um tempo para solução. Se o tempo limite for ultrapassado, o jogador é direcionado para a próxima questão e a questão anterior fica como status de não resolvida. Observação, as 10 questões serão selecionadas aleatoriamente da base de questões.
4. A medida que o jogador for avançando, o mesmo vai acumulando pontos, e o mesmo poderá trocar por medalhas, obedecendo o seguinte critério, a cada 100 pontos uma medalha de bronze, a cada 300 pontos uma medalha de prata e, a cada 500 pontos uma medalha de ouro.
5. Se o jogador tiver 3 medalhas de bronze, ele pode trocar as medalhas por uma de prata. Se tiver uma medalha de prata e 2 de bronze, ele pode trocar tudo por uma de ouro. Se ele tiver 5 medalhas de bronze, ele pode trocar por uma medalha de ouro.
6. O jogador poderá usar as medalhas para acessar uma dica, obedecendo o seguinte critério. Uma medalha de bronze dá acesso a uma dica simples, uma medalha de prata uma dica melhor e uma medalha de ouro uma dica muito melhor (quase a solução da questão).
7. Se o jogador usar uma medalha para acessar uma dica, então será subtraída uma medalha de seu quadro do tipo usada.
8. Para que os dados do jogador fiquem acumulados, o mesmo tem que se locar no sistema.
9. Se o jogador concluir o jogo, o jogo receberá o status de finalizado com sucesso (marcado com cinco estrelas). Se não o mesmo ficar com o status em aberto (menos de cinco estrelas, significa jog em aberto).
10. Se o jogador conectado, sair do jogo sem ter finalizado (resolvidos as 10 questões), então, se ele retornar novamente ao jogo, ele será direcionado para a próxima questão não resolvida.
11. O jogador só poderá jogar novamente o mesmo jogo se ele tiver completado com sucesso. Por exemplo, se ele começar o jogo conjunto numéricos e não finalizar (resolvido como sucesso as dez questões propostas), então, ele não poderá iniciar esse jogo, só se concluir.
12. Ao terminar um jogo, o jogador terá a opção de ver um relatório completo de seu desempenho.

4.3.7. O Jogo

Apresento a seguir, como exemplo, algumas telas do jogo, apenas as principais. A Figura 80 mostra a tela principal do jogo.

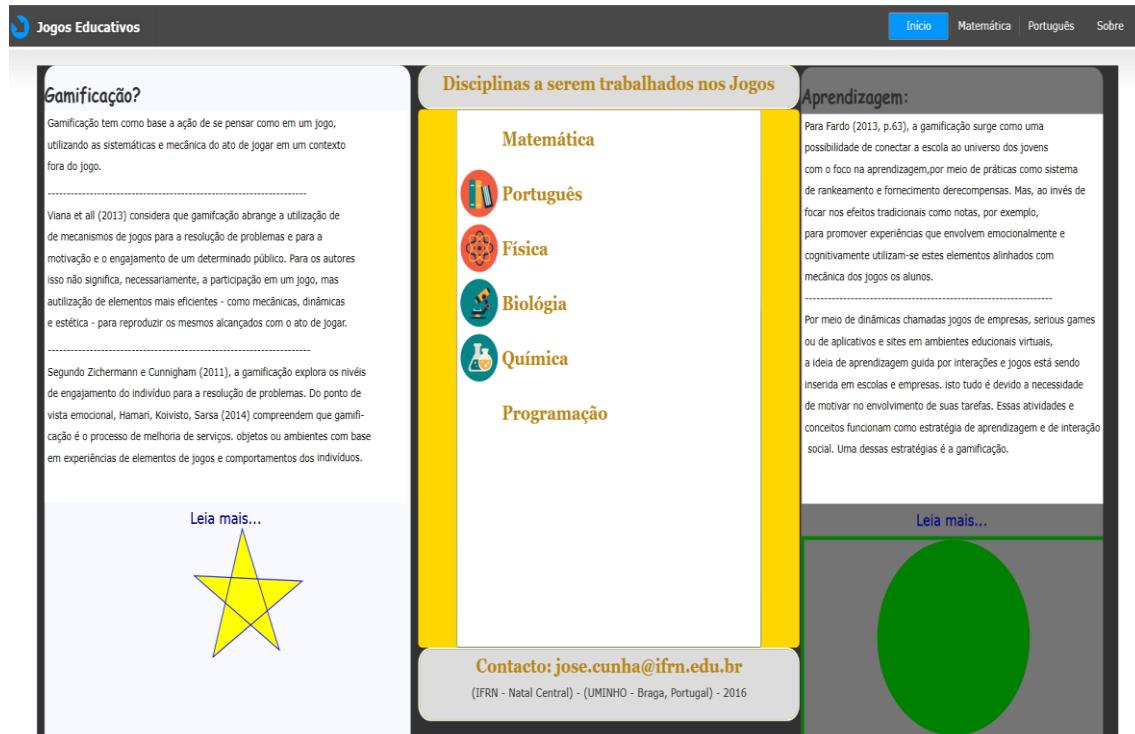


Figura 80. Tela Principal do Jogo. **Fonte:** Autor (Visual Studio 2015).

A Figura 81 mostra a tela do mapa do jogo, onde é apresentado o nível 1 para a disciplina de matemática. Cada assunto é um jogo. Os assuntos desse nível, já foram listados na seção “Descrição do Jogo”. Nesta janela, os links indicam cada jogo disponível para cada assunto e as estrelas indicam o status do jogo, se concluído ou não. Jogo concluído, o ícone será desabilitado.



Figura 81. Mapa de Jogos. **Fonte:** Autor.

A Figura 82, mostra a tela do Jogo propriamente dita. Neta tela, além do jogador poder jogar, ele pode ver a sua pontuação, as medalhas que o mesmo já obteve, pode ambém trocar os pontos obtidos por medalhas ou se o mesmo estiver com dúvidas para resolver a questão proposta, o mesmo poderá utilizar as medalhas para acessar uma dica. Além do mais, o jogador poderá recorrer a ajuda e revisar o assunto que está sendo trabalhado no jogo.

Do lado esquerdo da tela é apresentada o problema ou questão com as alternativas e, do lado direito as possíveis soluções. O jogador poderá digitar a solução diretamente na caixa de soluções ou apenas clicar nas opções de soluções apresentadas. Enquanto, o jogador está se decidindo sobre a solução do problema o relógio está marcando o tempo gasto. Se o tempo ultrapassar o limite estipulado pelo sistema, o jogador será automaticamente encaminhado ao problema seguinte e a questão anterior fica com o status de não resolvida. Se o jogador clicar no botão “Submeter” e a solução digitada for a solução correta, então, aparece o símbolo (check) informando que o mesmo acertou no resultado, caso contrário aparecerá o símbolo informando que o mesmo não acertou na solução do problema proposto.

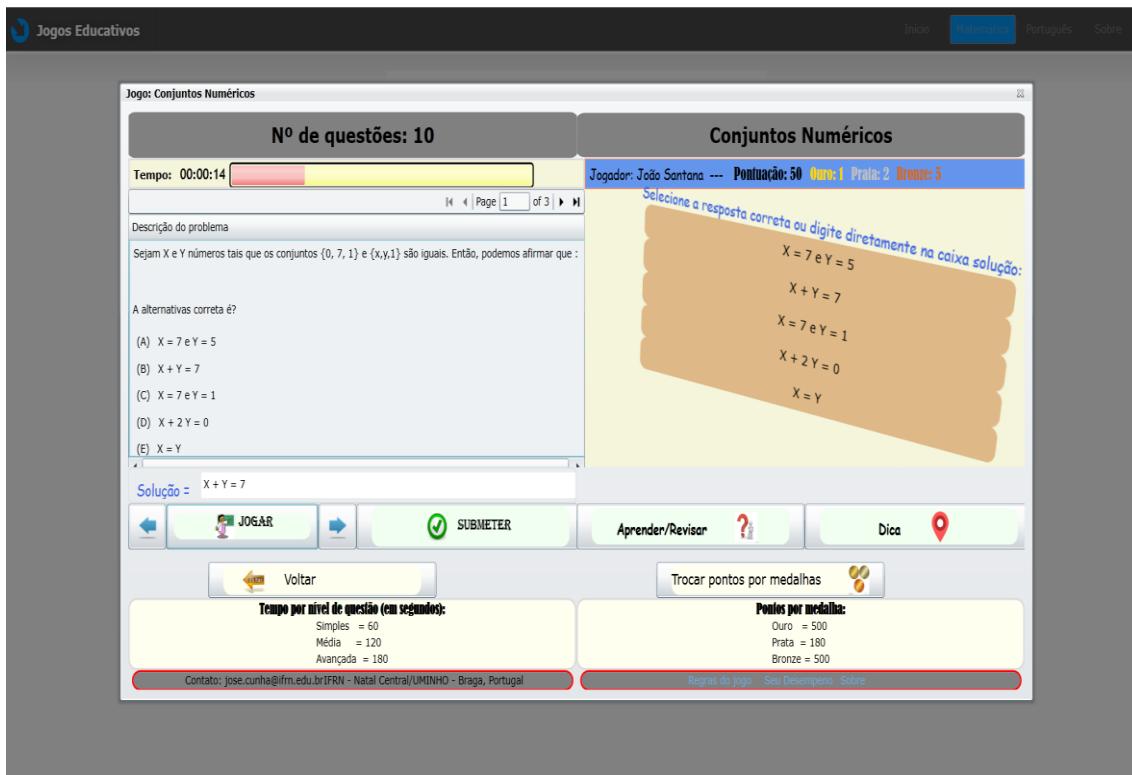


Figura 82. Tela do Jogo. **Fonte:** Autor.

O botão “Dica” leva o jogador para a tela de dicas. Nessa janela, o jogador vai poder usar suas medalhas para ter acesso a uma dica, dependendo das medalhas que o mesmo possuir. O botão “Aprender/Revisar” encaminha o jogador para a tela onde o mesmo poderá fazer uma revisão rápida, sobre o assunto do jogo. O botão “Trocár pontos por medalhas”, direciona o jogador para uma nova janela, onde o mesmo poderá fazer a troca dos pontos obtidos por medalhas, segundo as regras do jogo. O botão “Jogar” inicia o jogo e a partir desse momento, o relógio começa a contar e, a barra de tempo, mostra visualmente esse progresso. O botão “Submeter”, submete a resposta fornecida pelo jogador ao sistema, que verificar se a mesma está correta ou não, retornando um feedback ao jogador. E por fim, o botão “Voltar” que retorna para a tela de mapa de jogos.

4.3.8 Considerações finais sobre o Jogo

Busca-se através desse jogo a motivação ou engajamento do aluno, nas disciplinas identificadas como aquelas, onde os alunos apresentam maior grau de dificuldades em cursá-las. Com essa visão, os questionamentos foram

contextualizados na forma de desafios e conceitos da gamificação, como Status, Acesso, Poder e Coisas, incorporados a partir da definição de pontuações, ranking (estrelas que indicam se um jogo foi ou não concluído), recompensas (medalhas), níveis (um assunto poderá ter vários níveis a serem trabalhados) e debloqueios de funcionalidades (dicas).

Capítulo 5 - Conclusões e Trabalhos Futuros

5.1. Conclusões

O problema da evasão e da reprovação escolar, nos institutos federais, tem gerados alguns desafios a serem superados. O alto índice relacionados a estes fatores, tem sido vivenciado na experiência prática de todos os educadores, que fazem a educação nestas instituições. Sabe-se diante mão que, a evasão e a reprovação escolar estão, associados a fatores, tais como: áreas de conhecimento dos alunos, níveis de ensino e metodologias específicas de ensino aprendizagem. Portanto, para investigar, as possíveis causas relacionadas a esses problemas, aplicou-se técnicas de Mineração de Dados, na base de dado acadêmica, afim de mapear os fatores que estão associados a evasão e a reprovação escolar.

O problema da evasão escolar no Brasil, não é um problema recente, mais sim reincidente. É um dos fatores que preocupa os educadores e responsáveis pelas políticas públicas, em nosso país. De acordo com o Ministério da Educação e Cultura (**MEC**), a evasão escolar atinge 6,9% no Ensino Fundamental e 10% no Ensino Médio (3,2 milhões de crianças e jovens, segundo dados de 2015). São mais 2,9 milhões de alunos [9] que abandonam as aulas num ano e retornam no seguinte, engrossando outro índice preocupante, o da distorção de série_idade.

Com esta preocupação, o Ministério da Educação criou um grupo de trabalho para entender as causas e propor soluções para a evasão escolar em cursos técnicos. Segundo portaria publicada na edição de 25/11/2013 do "Diário Oficial da União", assinada pelo secretário de Educação Profissional e Tecnológica do **MEC**, Marco Antonio de Oliveira, o grupo terá 120 dias para concluir o trabalho [4].

Neste mesmo ano, ou seja, em abril de 2013, o Tribunal de Contas da União (**TCU**) realizou uma auditoria na rede federal de educação profissionalizante, científica e tecnológica, que apontou que os índices de evasão escolar, atingiram 24% do total de alunos matriculados nos cursos do Programa Nacional de Integração da Educação Profissional, com a Educação

Básica na Modalidade Educação de Jovens e Adultos (Proeja), além de 19% nos cursos médios subsequentes [4].

De acordo com [9], muitos dos problemas de evasão se deve à "expansão histórica", que está acontecendo na rede de educação profissionalizante. Durante um século tivemos 140 unidades, em pouco mais de 10 anos saltamos para 440 campi. Trata-se de uma expansão histórica, de larga escala e em alta velocidade, o que gera um descompasso, disse o secretário de educação na época.

Como foi frisado anteriormente, a evasão escolar no Brasil é reincidente, em 1995 foi fomentado pelo MEC [8] um amplo estudo sobre o desempenho das universidades públicas brasileiras em relação aos índices de diplomação, retenção e evasão dos estudantes de seus cursos de graduação.

Propõe-se, nesta investigação, aplicar técnicas de Mineração de Dados na base dado disponível, para detectar quais os atributos que mais influenciam a evasão escolar e, desta forma, traçar um perfil dos fatores que mais implicam na evasão escolar. Sabe-se que, alguns fatores influenciadores da evasão escolar, são externos ao ambiente escolar, tais como relacionamentos com os pais, famílias desajustadas, e assim por diante, mas o perfil traçado aqui poderá ser utilizado, juntamente com outros fatores para que se tenha uma análise mais precisa do problema em questão.

A justificativa para a necessidade de um processo, para se analisar os fatores que estão associados a evasão e a reprovação escolar, utilizando os recursos disponíveis na Mineração de Dados, está relacionado ao fato de que, o sistema acadêmico, ter uma base de dados grande, tornando difícil para o humano fazer análise sobre ela, sem a utilização de ferramentas adequadas. Além do fato de que, a Mineração de dados, utiliza algoritmos de Aprendizagem de Máquina, que favorecem a descoberta de conhecimento em grandes bases de dados. Justifica-se também a utilização das técnicas de Mineração de Dados, na identificação precoce e dinâmica, de informações precisas sobre a evasão e a reprovação escolar, que possam produzir resultados, que possam ser utilizados, para orientar ações pedagógicas eficientes, e que de alguma forma, esses resultados, posam fazer parte de um processo mais amplo.

No entanto, antes de tudo, foi feito um apanhado histórico sobre a evasão escolar no **IFRN**, de 2000 até 2013. A Figura 83 mostra o percentual de evadidos neste período.

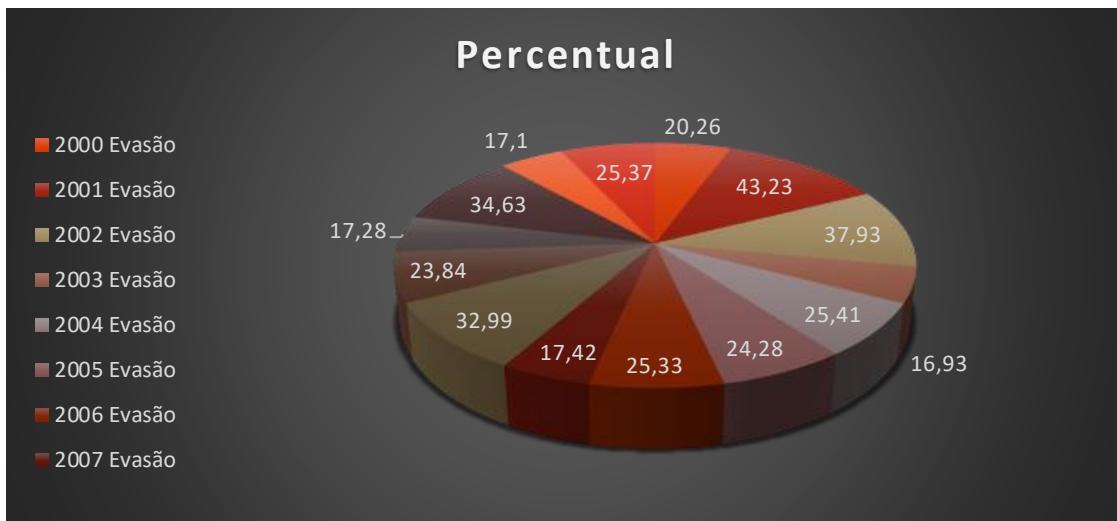


Figura 83. Gráfico mostrando a Evasão escolar no Campus Natal-central de 2000 a 2013.

Fonte: Autor.

O gráfico da Figura 83, mostra que, para o campus Natal-Central, no ano 2000 o índice de evasão escolar era de 20,26%, em 2001 era de 43,23%, em 2005 era de 34,63%. Observa-se que o percentual de evasão no campus Natal-Central está sempre acima de 17% e, em alguns anos mais elástico. Sem dúvida nenhuma, é um índice alto, e preocupante, e que merece um estudo detalhado sobre o mesmo. No entanto, tem outro dado que também merece ser observado. É o cancelamento de matrículas nos cursos da rede federal de ensino. A Figura 5.2 mostra o índice de cancelamento em matrículas no campus Natal-Central, entre 2000 e 2013.

A Figura 84 mostra o total de alunos evadidos e o total de alunos que cancelaram a matrícula para os anos entre 2000 e 2013. Analisando o ano 2007, temos o total de alunos que cancelaram a matrícula nos cursos, foi em torno de 240, e o total de alunos que evadiram dos cursos, foi em torno de 650 alunos. Se somarmos os dois fatores teremos em 900 alunos que desistiram de seus respectivos cursos. Portanto, o índice de cancelamento de matrícula também deve ser levado em conta, no processo de avaliação de o ensino aprendizagem.

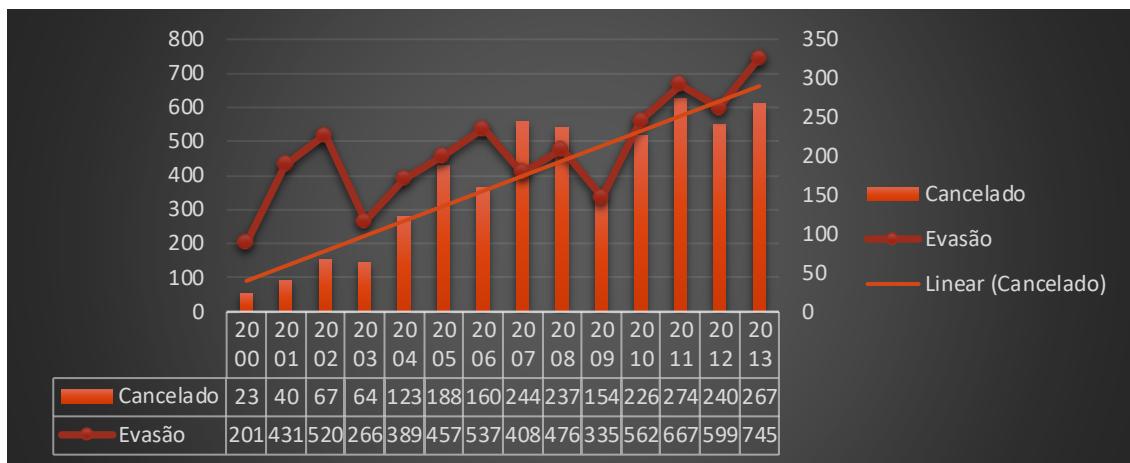


Figura 84. Cancelamento de matrículas no Campus Natal-Central entre 2000 e 2013. **Fonte:** Autor

A Tabela 21 mostra o percentual de reprovação por disciplina no ano de 2010. Foi feito um filtro nos dados para mostra apenas percentuais de reprovação entre 40% e 60%. Então, veja que, muitas disciplinas têm índice de reprovação muito alto. E como foi falado anteriormente, o índice de reprovação tem implicação na evasão escolar, pois muitas vezes, pode desestimular o aluno a continuar no curso.

Com base nos dados mostrados nos gráficos das Figura 83 e 84, temos a comprovação dos altos índices de reprovação e evasão escolar no IFRN, campus Natal-Central. Portanto, foi aplicado sobre essa base de dados, alguns algoritmos de Mineração de Dados, com o objetivo de encontrar alguma relação nos atributos da base de dados, que possa traçar um perfil das situações de reprovação e evasão de nossos alunos.

A Figura 85 mostra uma rede obtida da aplicação do algoritmo Árvore de Decisão, utilizando a ferramenta Analysis Services. Para formação dessa rede foi fornecido como atributo preditivo o atributo situação do aluno e, os demais atributos foram definidos como atributos de entrada, para o algoritmo.

Tabela 21 Índice de reprovação em Disciplinas no IFRN para o ano de 2010.

DISCIPLINA		ANO	REPROVADO	ALUNOS	%
PRÁTICA CURRICULAR	COMO	COMPONENTE	2010	17	34
LÍNGUA ESTRANGEIRA - INGLÊS			2010	22	44
PSICOLOGIA DO TRABALHO			2010	63	127
ALGORITMOS E PROGRAMAÇÃO ORIENTADA A OBJETOS			2010	58	118
EQUAÇÕES DIFERENCIAIS			2010	20	41
CONSERVAÇÃO DE ENERGIA			2010	17	35
AUTORIA WEB			2010	228	473
TÉCNICAS DE LABORATÓRIOS DE ALIMENTOS			2010	66	138
BIOLOGIA CELULAR			2010	18	38
INFORMÁTICA I			2010	61	129
CÁLCULO DIFERENCIAL E INTEGRAL II			2010	80	170
BIOLOGIA AMBIENTAL			2010	23	49
ARQUITETURA TCP/IP			2010	37	79
QUÍMICA GERAL E EXPERIMENTAL I			2010	79	170
SISTEMAS ELÉTRICOS			2010	26	56
MECÂNICA DOS SOLOS			2010	117	252
BIOLOGIA			2010	45	97
ESTRUTURA DE DADOS			2010	24	52
ADMINISTRAÇÃO DE SISTEMAS ABERTOS			2010	29	63
ELEMENTOS DE FÍSICA			2010	89	195
ÓPTICA			2010	26	57
ELETRICIDADE			2010	232	509
ELETRÔNICA DIGITAL			2010	81	180

A Figura 85, mostra todos os atributos que influenciam na evasão escolar. Portanto, podemos traçar um perfil para a evasão escolar, analisando cada um desses atributos. O atributo “tipo de escola de origem”, pode assumir os valores escola privada, pública e filantrópica. O atributo “Renda” representa a renda familiar do aluno, o atributo “Coeficiente de rendimento”, mede o desempenho do aluno no curso, e os atributos “Media e faltas” representam o desempenho escolar dos alunos. O atributo “Forma de ingresso” indica como o aluno entrou no curso (ENEM, exame de seleção, transferência, e assim por diante).

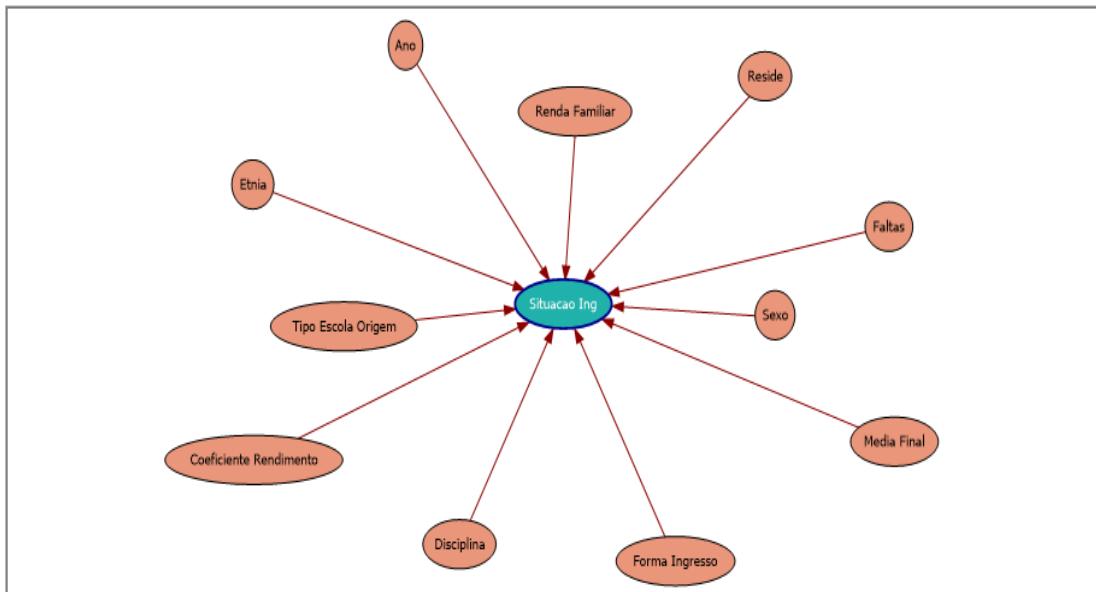


Figura 85. Árvore de Decisão mostrando a relação entre os atributos da base de dados Acadêmica.

Fonte: Obtida pela Mineração de Dados (Microsoft Analysis Services).

A partir da árvore da Figura 85, foi utilizar o algoritmo de cluster para agrupar alunos com características semelhantes em um mesmo grupo e, em seguida analisar cada cluster para identificar o grau de influência de cada atributo de entrada mostrado na Figura 85, em relação ao atributo preditivo “**Situacao=Evasão**”.

A Figura 86, mostra o gráfico gerado pelo algoritmo cluster. Na configuração do algoritmo cluster foi selecionada a situação “Evasão”, o cluster com maior quantidade de casos de evasão, aparece no gráfico com a cor azul intenso e, os clusters com menos casos de evasão com a cor mais clara.

Sendo o cluster 5 o de cor azul mais escura, significa que, esse cluster é detentor do maior número de evasão escolar e, o cluster 1 apresenta o menor número de casos de evasão escolar. O cluster 1 apresenta o maior número de aprovados.

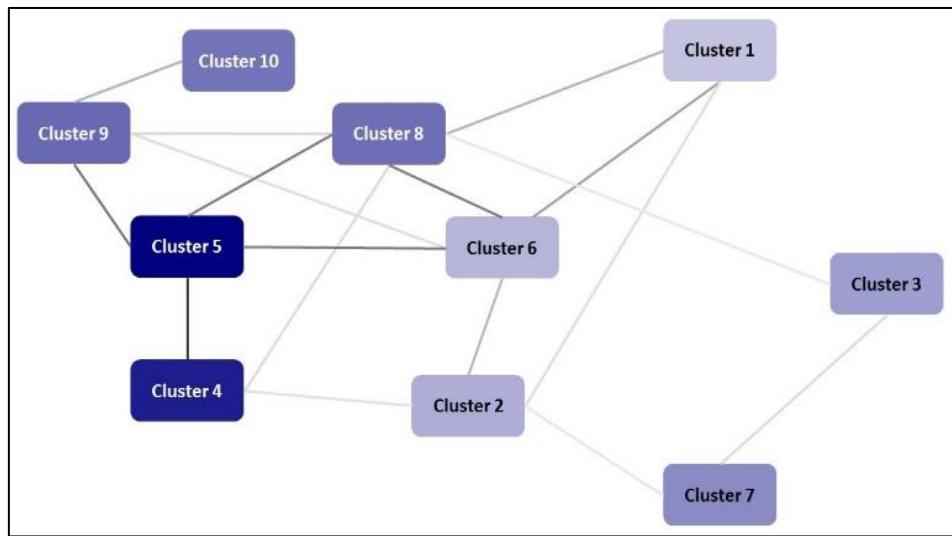


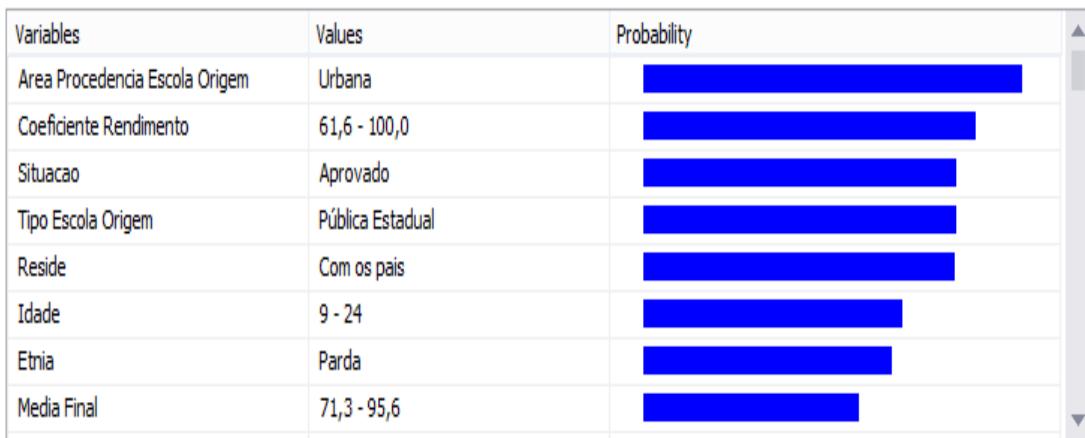
Figura 86. Gráfico de cluster gerado pelo Analysis Service.

Fonte: Autor

Vamos analisar os cluster 1 e 5 para ver quais foram os atributos de entrada que influenciaram na composição dos mesmos. A Figura 87 mostra as características do cluster 1, onde está concentrado o maior número de aprovados. A Figura 88 mostra as características do cluster 5, onde está concentrado o maior número de evadidos.

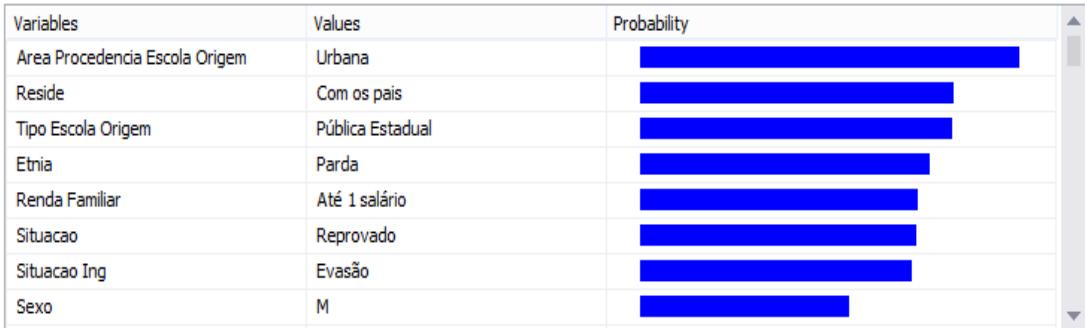
O cluster 1 mostra que, o perfil dos alunos aprovados são os com média final acima de 70, coeficiente de rendimento acima de 60, oriundos de escola pública estadual, residem com os pais e de etnia parda. Observando as características do cluster 5, percebe-se que, na formação do cluster 5, a renda familiar (de até 1 salário) e a situação (reprovado), aparecem como fatores que influenciam a evasão escolar. Justificando, dessa forma, a presença desses atributos, na relação de atributos que influenciam a evasão escolar, como mostram as Figura 87 e 88.

Characteristics for Cluster 1

**Figura 87.** Características do cluster 1. **Fonte:** Autor.

Ferramenta utilizada para gerar este gráfico foi o Microsoft Analysis Services

Characteristics for Cluster 5

**Figura 88.** Características do cluster 5. **Fonte:** Autor.

Ferramenta utilizada para gerar este gráfico foi o Microsoft Analysis Services

Os gráficos das Figuras 87 e 88, não mostram o quanto cada atributo influencia na formação de cluster, de forma percentual. Para isto, apresenta os dados da Figura 89, que mostra esta influencia de outra maneira, também gráfica, mas de uma forma mais fácil de interpretar. Para perceber o quanto cada atributo influencia na formação dos cluster, é só procurar o atributo na coluna status (cada status do atributo tem uma cor) e, observar na coluna do cluster a altura que essa cor representa na formação do histograma. Por exemplo, olhando o cluster 5 na Figura 89, se pode perceber que, para o atributo “Renda Familiar” a cor azul=“Até 1 salário”, preenche quase todo histograma. O atributo “Reside”, também cor azul=“Com os pais”, preenche quase todo histograma. O atributo “Tipo escola Origem”, cor também azul=“Publica

Estadual”, preenche quase todo histograma. Dessa forma, pode afirmar que os atributos “Renda Familiar”, “Tipo escola origem” e o atributo “Reside”, são os atributos com maior influencia na evasão escolar para o cluster 5.

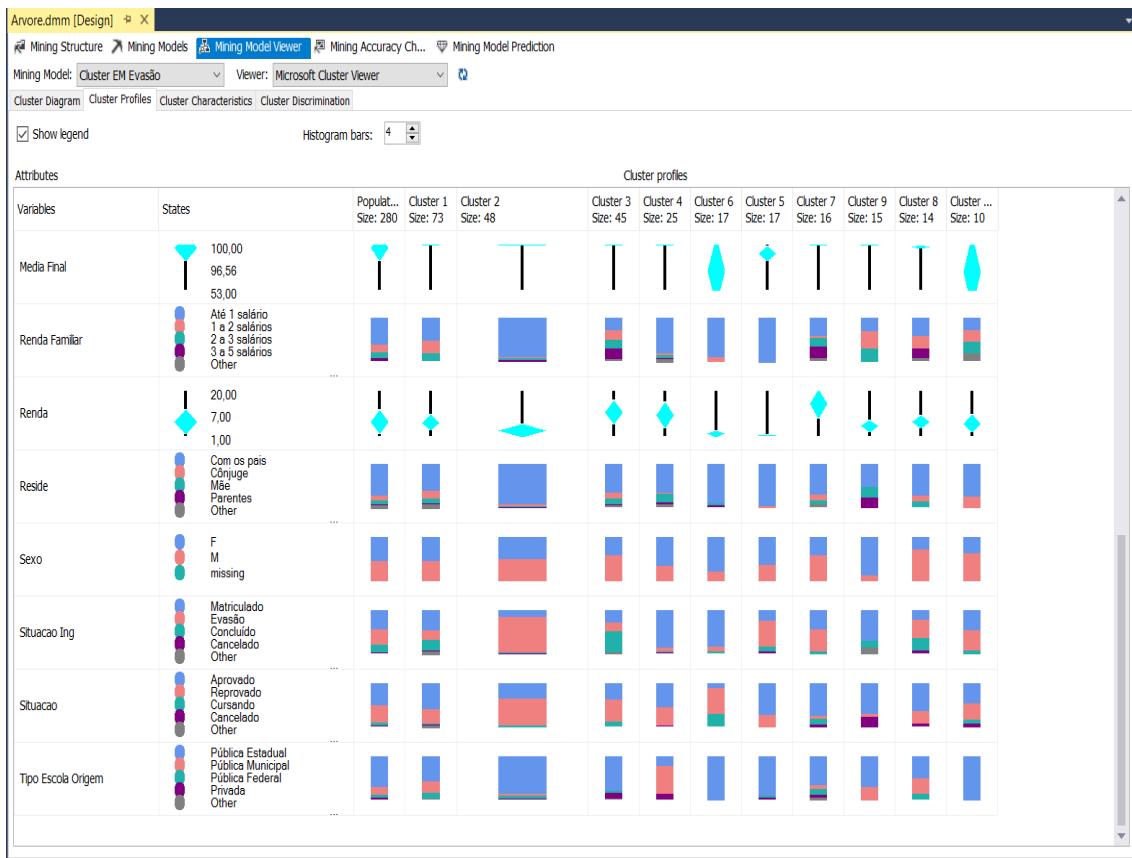


Figura 89. Representação dos clusters em forma de histograma.

Fonte: Auto (gráfico gerado pela ferramenta Microsoft Analysis Service).

Analizando os gráficos obtidos pelos algoritmos de Mineração de Dados, em um primeiro momento, pode-se traçar um perfil para a evasão escolar, como sendo de alunos oriundos de escolas públicas estaduais, com renda familiar de até um salário mínimo, que moram com os pais, consequentemente, são desempregados ou são menores de idade, de cor parda, com média final e coeficiente de rendimento muito baixo. Sabe-se na prática que os alunos que entram no IFRN vindos das escolas públicas, chegam com o conhecimento muito aquém do desejado nas disciplinas básicas, como matemática e português, que são fundamentais, para que o mesmo, tenha um bom desempenho em nosso curso. Esses alunos enfrentam muitas dificuldades para cursar disciplinas que

contenham lógica, abstração e ou matemática avançada, tais como as disciplinas técnicas da área tecnológica.

A Figura 90, mostra graficamente que, o perfil dos alunos com maior probabilidade de evasão escolar são, os alunos onde a renda familiar é muito baixa, oriundos de escola públicas estaduais, residem com os pais, tem muitas faltas, com média muito baixas e, por conseguinte, muitas repetências.

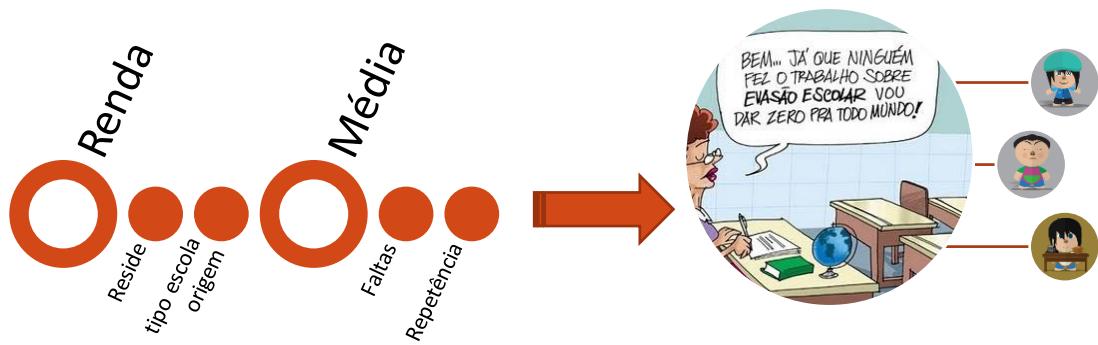


Figura 90. Perfil do aluno com maior probabilidade de evasão escolar.

Fonte: Autor

Esta análise preliminar, teve como base as informações retiradas das consultas geradas pelo processo **BI**, e pelos conhecimentos adquiridos no processo de Mineração de Dados. A partir desse momento, traçarei um quadro comparativo entre a evasão escolar no Brasil e entre outros países no mundo, para podermos ter uma real situação da evasão escolar nas instituições federais de ensino brasileira.

A evasão escolar no Brasil, segundo o **PNUD** (Programa das Nações Unidas para o Desenvolvimento), o Brasil, com uma taxa de 24,3%, tem a 3^a maior taxa de evasão escolar entre 100 países com maior **IDH** (Índice de Desenvolvimento Humano), só a Bosnia Herzegovina (26,8%) e as ilhas de São Cristovam e Névis, no Caribe (26,5%), têm taxas superiores.

Na América Latina, só Guatemala (35,2%) e Nicarágua (51,6%) tem taxa de evasão superiores.

Anida segndo o **PNUD**, algumas das principais causas da “evasão escolar” no Brasil são a pobreza, a dificuldade de acesso a escola, a

necessidade de trabalho e, principalmente, o desinteresse pelos estudos. Em nossa investigação, também ficou evidente que, a renda familiar e tipo de escola de origem, que também está ligada a baixa renda familiar, são fatores que influenciam muito na evasão escolar.

A Tabela 22, mostra dados estatísticos do relatório **PNUD**, com o ranking de um alguns países, em relação a taxa de evasão escolar e outros dados. Onde se pode observar que, o Brasil apresenta uma taxa de evasão escolar muito superior a outros países, inclusive, países da América do Sul e Latina.

Tabela 22. Dados relativos à Educação no Relatório do **PNUD**.

País	Ranking	IDH	População Alfabetizada	Ensino médio completo	Taxa de evasão
Noruega	1	0,955	1	0,952	0,5%
Austrália	2	0,938	1	0,922	-
Estados Unidos	3	0,937	1	0,945	6,9%
Holanda	4	0,921	1	0,889	-
Alemanha	5	0,92	1	0,965	4,4%
Chile	40	0,819	0,986	0,74	2,6%
Argentina	45	0,811	0,978	0,56	6,2%
Uruguai	51	0,792	0,981	0,498	4,8%
México	61	0,775	0,931	0,539	6,0%
Brasil	85	0,73	0,903	0,495	24,3%

Fonte: <http://educacao.uol.com.br/noticias/2013/03/14/brasil-tem-3-maior-taxa-de-evasao-escolar-entre-100-paises-diz-pnud.htm> ,acessado em 15/08/2016.

Como mostra os dados da Tabela 22, a taxa de evasão escolar no Brasil é alta, na ordem de 24,3%, segundo o relatório **PNUD**. Maior do que a taxa de evasão de alguns países da América Latina. Apesar de que, no **IFRN**, a taxa de evasão está por volta dos 19%, inferior a taxa nacional, mesmo assim, ainda é bastante elevada e merecedora de preocupação.

Na Europa, foi desenvolvido no período de dezembro de 2011 à novembro de 2013, um estudo sobre o abandono escolar precoce (Em inglês *Early School Leaving – ESL*). Esta pesquisa foi desenvolvida pelo “Grupo de Trabalho Temático” e, teve como resultado um relatório sobre "abandono escolar precoce". Este grupo de trabalho incluía peritos nomeados pelos 31 países europeus e organizações de partes interessadas e, foi assistido pelos consultores da Comissão, Anne-Marie Hall e Ms Ilona Murphy[http://ec.europa.eu/education/schooleducation/leaving_en.htm, 2013].

A Tabela 23, mostra um quadro do abandono escolar precoce, gerado pelo Grupo de Trabalho Temático, para países da União Europeia (**EU**).

Tabela 23 Taxa de Abandono Escolar Precoce na Europa.

	2009	2012			2020
	Total	Masculino	Feminino	Média	Meta
EU	14.3	14.5	11.0	12.7	< 10.0
Bélgica	11.1	14.4	9.5	12.0	9.5
Bulgária	14.7	12.1	13.0	12.5	11.0
República Checa	5.4	6.1	4.9	5.5	5.5
Dinamarca	11.3	10.8	7.4	9.1	<10.0
Alemanha	11.1	11.1p	9.8p	10.5p	<10.0
Estônia	13.9	14.0	7.1	10.5	9.5
Irlanda	11.6	11.2	8.2	9.7	8.0
Grécia	14.5	13.7	9.1	11.4	9.7
Espanha	31.2	28.8	20.8	24.9	15.0
França	12.2	13.4	9.8	11.6	9.5
Croácia	3.9	(4.6)	(3.6)	4.2	4.0
Itália	19.2	20.5	14.5	17.6	15.0
Chipre	11.7	16.5	7.0	11.4	10.0
Letônia	13.9	14.5	6.2	10.5	13.4
Lituânia	8.7	8.2	(4.6)	6.5	<9.0
Luxemburgo	7.7b	10.7p	5.5p	8.1p	<10.0

Hungria	11.2	12.2	10.7	11.5	10.0
Malta	27.1n	27.5	17.6	22.6	-
Países Baixos	10.9	10.2p	7.3p	8.8p	<8.0
Áustria	8.7	7.9	7.3	7.6	9.5
Polônia	5.3	7.8p	3.5p	5.7p	4.5
Portugal	31.2	27.1	14.3	20.8	10.0
Romênia	16.6	18.0	16.7	17.4	11.3
Eslovênia	5.3	5.4	(3.2)	4.4	5.0
Eslováquia	4.9	6.0	4.6	5.3	6.0
Finlândia	9.9	9.8	8.1	8.9	8.0
Suécia	7.0	8.5	6.3	7.5	<10.0
Reino Unido	15.7	14.6	12.4	13.5	-
Montenegro	:	:	:	:	-
Islândia	21.3	23.6	16.5	20.1	-
MK*	16.2	11.1	12.3	11.7	-
Sérvia	:	:	:	:	-
Turquia	44.3	36.1	43.0	39.6	-
Noruega	17.6	17.6	11.9	14.8	-
Suíça	9.1d	5.7	5.3	5.5	-

Source: Eurostat (LFS). Intermediate breaks in time series for NL (2010) and LV (2011). Notes: "b" = break in time series; "p" = provisional; "(" = Data lack reliability due to small sample size; ":" = data either not available or not reliable due to very small sample size; "n" = national data. *MK: The former Yugoslav Republic of Macedonia.

Os dados da Tabela 23, mostra que, a taxa de evasão escolar, também é preocupante em muitos países da Europa. No entanto, os países da União Europeia, estabeleceram uma meta para que, no ano de 2020, a taxa de evasão escolar fique abaixo de 10%. Pode-se destacar na Tabela 23, a taxa de evasão escolar, no ano de 2012, da Espanha, Malta e Portugal, que eram respectivamente de 24,9%, 22,6% e 20,8%, segundo a fonte pesquisada. São taxas de evasão escolar próxima à da nossa rede federal de ensino, segundo o relatório PNUD. A diferença é que, no Brasil, nenhuma medida ainda foi tomada nessa direção.

O que está acontecendo no momento, é que o Tribunal de Contas da União (**TCU**), identificou, através de suas análises que, o índice de evasão

escolar na rede federal estava em torno de 24% e, pediu ao Ministério de Educação e da Cultura do Brasil (**MEC**) uma explicação do fato e exigiu que o MEC tomasse medidas para solucionar o problema e, solicitou ao MEC, que o apresentasse um relatório com dados comprovatórios da evasão na rede federal brasileira.

A evasão escolar é um problema preocupante, por vários fatores: primeiro é um problema social, segundo é um problema econômico. É um problema social, porque são alunos que deixam de concluir seus estudos. É um problema econômico, porque gera despesas elevadas sem retorno. Basta, analisar o caso do campus Natal central (com um contingente de alunos, por volta de 8000), onde encontramos uma evasão na ordem de 19% (taxa anual de evasão escolar). Isto representa, mais ou menos 1520 alunos evadidos por ano. Se imaginarmos que, um aluno custo em média R\$ 1200,00 mês, para os cofres da união, então, se pode concluir que, a instituição paga um preço muito elevado pela evasão escolar.

Apesar dos programas destinados a assistência estudantil, na tentativa de manter os estudantes em atividades de ensino, pesquisa e extensão, além de buscar a garantia de um percurso exitoso de formação acadêmica e profissional, no sentido de garantir reais condições de permanência, mesmo assim, se constata a necessidade de atenção às taxas de evasão e de retenção nos cursos ofertados pela Rede Federal. A Tabela 24, mostra dados de relatório emitido pelo **TCU** (Tribunal de Contas da União) em 2012.

Tabela 24. Alunos evadidos, por tipos de cursos, de ciclos de matrícula iniciados a partir de 2004 e encerrados até dezembro de 2011.

Nível	Tipo de Curso	Taxa de Evasão	Taxa de Retenção	Taxa de Conclusão
Educação Básica	Técnico integrado para estudantes em idade própria	6,40%	44,42%	46,80%

	Técnico Integrado e concomitante na modalidade EJA*	24,00%	37,99%	37,50%
	Técnico Subsequente	18,90%	49,34%	31,40%
Educação Superior	Licenciatura	8,70%	64,53%	25,40%
	Bacharelado	4,00%	68,09%	27,50%
	Tecnólogo	5,80%	50,82%	42,70%

Fonte: TCU (2012).

O mapeamento encontrado do perfil do estudante realizado pela mineração de dados, na base de dados acadêmica do IFRN¹, sinaliza o atendimento de um percentual significativo de uma população socioeconomicamente vulnerável, constituída marcadamente por estudantes de baixa renda, que residem com os pais, que faltam muito e consequentemente, tem muitas repetências nas disciplinas.

Dessa forma, entender a evasão e a retenção escolar como fenômenos que envolvem fatores multidimensionais (culturais, sociais, institucionais e individuais), e relacionar esse entendimento à complexidade da Rede Federal no cumprimento da sua função social, implica em articular ações que deem conta do atendimento a um público diversificado que, em sua maioria, é socioeconomicamente vulnerável e egresso de sistemas públicos de ensino em regiões com baixo índice de desenvolvimento educacional.

No estudo realizado, ficou evidente, a necessidade premente de implementação de planos estratégicos de superação desses fenômenos de modo a possibilitar a realização de diagnósticos apurados em relação às causas da evasão e da retenção escolar, e a definição de políticas institucionais e a adoção de ações administrativas e pedagógicas que contribuam para o

¹ Esse mapeamento foi realizado por meio da Mineração de Dados, em um processo de Descoberta de Conhecimento em Bases de Dados (Em inglês Knowledge Discovery Databases – KDD).

enfrentamento da evasão e retenção escolar em todos os níveis e modalidades da oferta educacional.

Outro objetivo proposto nessa pesquisa, foi desenvolvido um sistema de aconselhamento pedagógico. Este sistema tem como finalidade, auxiliar a equipe pedagógica, no atendimento aos alunos com problemas de varias ordem: disciplinares, de relacionamento com os colegas, de relacionamento com os professores e assim por diante.

O sistema de aconselhamento é uma ferramenta, que possibilita a realização de diagnósticos apurados em relação às causas da evasão e repetência escolar, relacionados a problemas disciplinares, socioeconômicos e ou psicológicos apresentados pelos alunos.

É muito importante que a equipe pedagógica, tenha a sua disponibilidade, um instrumento que possa de maneira rápida e eficiente, diagnosticar esses problemas e apresentar as soluções para os mesmos. A principal função deste sistema, é fazer com que a equipe pedagógica, realize os atendimentos aos alunos mais rapidamente e eficientemente.

O sistema de aconselhamento irá servir de apoio ao ensino aprendizado, uma vez que, agilizando-se o atendimento ao aluno, se pode tomar medidas preventivas antecipadamente. E dessa forma, diminuir os índices de repetência e consequentemente, os índices da evasão escolar.

O sistema desenvolvido para o aconselhamento pedagógico, foi um sistema baseado em Raciocínio Base em Casos. Este sistema está documentado no capítulo de resultados, na seção 4.2.

O sistema de aconselhamento é baseado em casos anteriores. Estes casos são compostos de atendimentos, demandas e encaminhamentos, como mostra a Figura 91. Demadas são os problemas identificados em cada aluno. Os problemas foram mapeados pela equipe pedagógica, num total de 25 tipos ou variáveis. Um encaminhamento é uma solução para o problema identificado.

As demandas de cada caso, são problemas externos a base de dados acadêmica, ou seja, esses dados não estam na base de dados acadêmica. Na seção 4.2, Tabela 4.2, são apresentadas as demandas. Aqui elenco algumas delas que, podem influenciar no desempenho do aluno, em sua vida acadêmica, tais como: **desequilíbrio psicológico, muitas faltas, problemas sócio econômicos, atrasos constantes, desmotivação do curso, conflito familiar,**

separação dos pais, dentre outros. Portanto, quando um aluno se encontra dentro de uma situação que envolva estes fatores, seu desempenho fica prejudicado, e o quanto mais rápido for diagnóstificado este problema, mais rápido será o encaminhamento para uma solução.

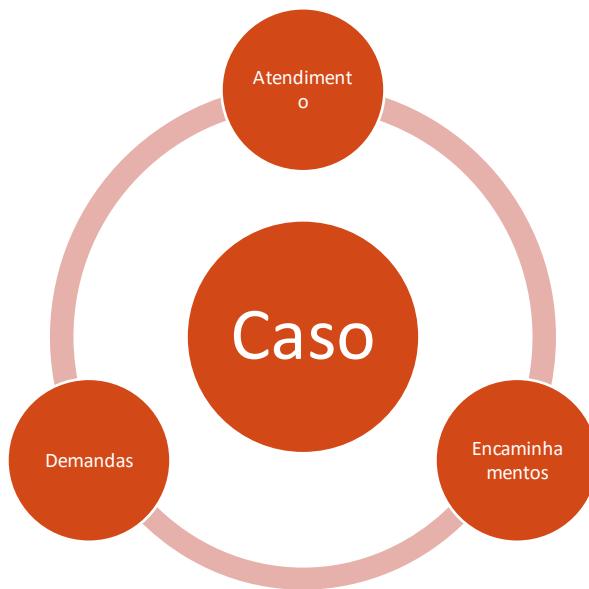


Figura 91. Composição de um Caso do Sistema de Aconselhamentos.

Fonte: Autor.

O sistema de aconselhamento não identifica o problema, no entanto, ao ser identificado o problema, o encaminhamento da solução pode ser muito rápido, usando esta ferramenta.

O processo de Mineração de Dados, traçou o perfil dos alunos com probabilidade de repetência e evasão escolar. O Sistema de Aconselhamento Pedagógico, através de seu desenvolvimento, identificou as variáveis que compõem as demandas dos problemas dos alunos. Percebe-se claramente, uma correspondência entre as informações do perfil determinado pela Mineração de Dados, e as demandas do Sistema de Aconselhamento. Veja que, a variável problemas sócios econômicos das demandas, está relacionada aos atributos renda familiar, tipo de escola de origem e, até mesmo atrasos constantes. Uma família com renda muito baixa, implica diretamente em fatores como conflito familiar, relacionamentos com os pais, e até mesmo desencadear desequilíbrio psicológico, dentre outros.

Portanto, fica evidente que os problemas elencados pelas demandas do Sistema de Aconselhamento Pedagógico, comprovam que o perfil descoberto pela Mineração de Dados, é correspondente aos fatos relacionados os índices de repetência e evasão escolar em nossa instituição. Olhando a Figura 92, percebe-se uma relação entre os atributos do perfil e as varáveis da demanda.

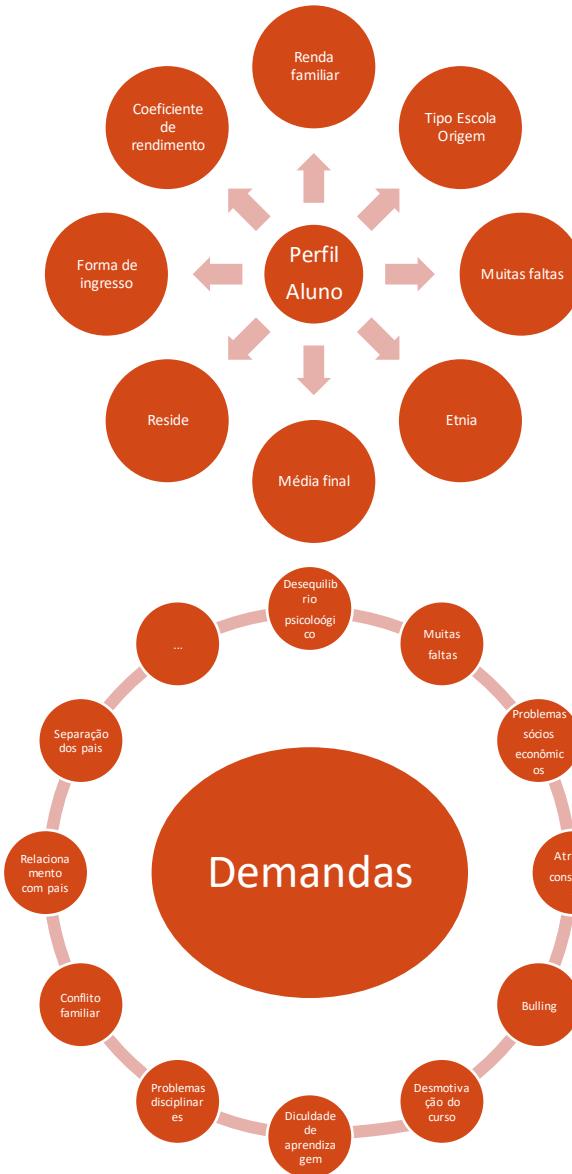


Figura 92. Relação dos atributos do perfil e variáveis da demanda.

Fonte: Autor

A repetência e a evasão escolar, estam relacionados a vários problemas, como os identificados pelo perfil do aluno e pelas demandas do sistema de aconselhamento. Um destes fatores identificados é a desmotivação do aluno pelo curso.

Por sua vez, esta pesquisa propõe como mais um dos objetivos, o desenvolvimento de um jogo na forma de gamificação, com o propósito de motivar os alunos a estudarem os assuntos relevantes para o seu desempenho, de forma divertida. Principalmente nas disciplinas a onde os alunos geralmente tem mais dificuldades de cursar.

Através do Portal desenvolvido na seção 4.1, pode-se identificar as disciplinas com maiores índices de reprovação e, dentre elas está a matemática. A matemática foi a disciplina selecionada para o desenvolvimento do jogo, pelo fato de ser uma disciplina de grande relevância, em muitas outras disciplinas da área tecnológica, que são a maioria dos cursos ministrados pelo IFRN. Portanto, foram selecionados alguns assuntos da matemática para implementação do jogo. Estes assuntos foram selecionados através de consultas feitas junto aos professores, onde os mesmos apontaram as maiores dificuldades dos alunos na disciplina alvo.

Os assuntos selecionados, foram classificados em níveis. No nível 1, ficaram os assuntos mais básicos como conjunto numéricos, operações aritimética, fatoração e assim por diante. No nível 2 os assuntos sobre potência e raízes, funções, equações, inequações, logaritmos e assim por diante. E no nível 3 os cálculos.

O jogo foi implementado seguindo as características voltada a gamificação, onde elementos da gamificação foram utilizados para tornar o jogo divertido e, com isso, fazer com que o aluno sinta vontade de adquirir novos conhecimento de forma fácil e divertida.

Resumindo, o sistema **RBC** mostrou que uma das demandas ou problemas encontrados nos alunos, é a desmotivação dos alunos pelos seus respectivos cursos. Portanto, espera-se que o jogo seja uma forma engajadora/estimuladora, que faça com que, aqueles alunos que se sintam desmotivados pelos cursos, voltem a se estimular. Se o jogo realmente conseguir seu propósito, os alunos demotivados, voltarem a se motivar pelo curso, isto deve implicar na diminuição do índice de reprovações e, será refletido também, no índice de evasão escolar.

5.2. Trabalhos Futuros

Além da pesquisa efetuada neste trabalho, foram propostos alguns produtos e, estes produtos devem servir ao propósito de vários utilizadores diferentes (gestores, professores, equipe pedagógica, alunos, etc.).

Os três projetos implementados, o Portal, o sistema de aconselhamento e o game, tem funcionalidade que são implementadas sob demanda. Por exemplo, os dashboard e KPIs não se esgotaram neste Portal, outros surgiram no futuro com certeza. O sistema de aconselhamento pedagógico, com certeza, sofrerá atuações futuras e no jogo, com certeza, serão trabalhados outros assuntos de outras disciplinas.

No entanto, observou-se no decorrer do trabalho que, apesar do IFRN ter um Data center, sobre seu domínio, não existe uma consolidação dos dados. Ou seja, existem diversas bases de dados armazenadas no sistema, porém, quando se deseja fazer uma pesquisa estatística sobre esses dados, é necessário muito esforço para localizar esses dados, pois as bases de dados, geralmente são muito normalizadas, dificultando a seleção dos dados pretendidos.

O IFRN tem atualmente 21 campi e, todos os dados produzidos sobre a vida acadêmica dos alunos e servidores, estão armazenadas nas bases de dados do sistema central (Data Center). E toda essa massa de dados são gerenciadas por gerenciadores de bancos de dados relacionais, no nosso caso o SQL Server da Microsoft. Os bancos de dados relacionais não foram projetados para serem executados em clusters. O Microsoft SQL Server, por exemplo, funciona baseado no conceito de um subsistema de disco compartilhado. Ele utiliza um sistema de arquivos que reconhece clusters e grava em um subsistema de disco com alta disponibilidade.

Apesar da massa de dados do sistema acadêmico, não ser muito grande, hoje (2016) na ordem de 60GB, e ser controlada por um sistema relacional, que não foi projetado para executar em clusters, é importante desde já, se pensar em um modelo de **persistência poliglota** – utilizar diferentes armazenamentos de dados em diferentes circunstâncias, dependendo da natureza dos dados. Isto resulta em uma mistura de tecnologias de armazenamento de dados para diferentes circunstâncias.

Então, mesmo que não exista a necessidade de escalar para além de uma máquina, é importante pensar em tecnologias que lide com acesso a dados cujo

tamanho e desempenho demandem cluster. Tudo isso, aliada a necessidade de se fazer análise de dados em grande massa de dados, aponta para a necessidade de se adotar as tecnologias de armazenamento de dados **NoSQL** e para as tecnologias de **Big Data**.

Os bancos de dados NoSQL, apresentam características comuns, tais como: não utilizam o modelo relacional, tem uma boa execução em clusters, seu código é aberto (open source), não tem esquema e o acesso aos dados é rápido.

Por outro lado, o Big Data utiliza soluções com algoritmos matemáticos, que capturam e cruzam dados de diferentes formatos.

Segundo Hurwitz et all (2016), o Big Data funciona baseado em três pilares:

- Volume de dados extremamente grande;
- Velocidade de dados extremamente alta;
- Variedade de dados extremamente ampla.

Aliada aos três pilares temos, integração da informação, a veracidade e a governança da informação e a segurança da informação.

Então para que o IFRN, possa ser capaz de realizar análise avançada de dados, proponho como trabalho futuro “**o desenvolvimento de um projeto de Big Data, utilizando a infraestrutura de Clusters do Data Center, consolidando os dados importantes para análise, em bases de dados NoSQL**”.

Atualmente, os aplicativos se comunicam diretamente de forma independente com o banco de dados relacional, para o qual foi projetado. A nossa proposta como trabalho futuro “**é de encapsular os bancos de dados em serviços (APIs Rest)**”, permitindo que as aplicações apenas se comuniquem com os serviços. Isto irá permitir que os bancos de dados dentro dos serviços se desenvolvam sem que alguém tenha que alterar os aplicativos dependentes.

Referencial Bibliográfico

ABEL, Mara; CASTILHO, José Mauro Volkmer (1996). **Um Estudo Sobre Raciocínio Baseado em Casos.** UFRS – Porto Alegre, RS.

Agrawal, R.; IMIELINSKI, T.; SWAMI, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. ACM SIGMOD Conference Management of Data.

Alves, Flora (2014). Gamification: como criar experiências de aprendizagem engajadoras: um guia completo do conceito à prática / Flora Alves. 1. Ed. – São Paulo: DVS Editora.

Crockford, D. (2006) **RFC 4627** – 2006. The Internet Society – Acessado em: 23/9/2014. Disponível em: <http://www.ietf.org/rfc/rfc4627.txt>.

Dean, Jeffrey; Ghemawat, Sanjay (2008). MapReduce: Simplified Data Processing On Large Cluster. CUMMUNICATIONS OF THE ACM.

Dong, G. & J. LI (1998). Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. Lecture Notes in Artificial Intelligence, pp. 72-86.

Dickinson, Paul; Ferracchiati, F. C.; Hoffman, Kevin; Joshi, Gipin; Mack, D. ny; McTainsh, John; Milner, Mathew; Narkiewicz, Jan; Seven, Doug (2002). Profissional ADO.NET Programando. Editora Alta Books. Rio de Janeiro-RJ.

Elman, Jeffrey L (1993). *Learning and Development in Neural Networks: The Importance of Starting Small.* Cognition, 48(1993), pp.71-99. Web: <http://crl.ucsd.edu/~elman/>
Ftp: <ftp://crl.ucsd.edu/pub/neuralnets/cognition.ps.Z>

Elmasri, Ramez; Navathe, Shamkant B. (2005). Sistemas de Banco de Dados. São Paulo: Addison Wesley.

Engels. R. (1996). Planning tasks for knowledge discovery in databases: Performing Task-Oriented User-Guidance. Proceeding of the International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press.

Engels, R.; LINDNER, G.; STUDER, R. (1997). A Guided Tour Through the Data Mining Jungle. Proceeding of the Third International Conference on Knowledge Discovery in Databases. Newport Beach.

Fayyad, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTH, Padhraic (Nov de 1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM, pp.27-34.

FLANAGAN, David. (2004) **Javascript: O guia definitivo** – Tradução de Edson Furmankiewicz. Editora Bookman – Porto Alegre, PR.

FERREIRA, Silvio (2013). **Guia Prático de HTML5**. Editora Universo dos Livros – São Paulo, SP.

Fiesler, E. (1994). *Neural Networks Formalization and Classification*. Computer Standard & Interfaces, Special Issue on Neural Networks Standards, John Fulcher (Ed.). V.16, N.3. Elsevier Sciences Publishers, Amsterdam. Web: <http://www.idiap.ch/idiap-networks.html>.

Freitas A. A. (1998). A multi-criteria approach for the evaluation of rule interestingness. Em Proceedings of the International Conference on Data Mining. Rio de Janeiro, RJ, pp. 7-20.

Freitas A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems* 12(5-6), 309-315.

Goldschmidt, R.; Passos, E.; Vellasco, M.; Pacheco, M. (2003). Task Definition Assistance in KDD Applications. CLEI'03 – XXIX Conferência Latino Americana de Informática. La Paz.

Han, Jiawei; Kamber, Micheline (2006). *Data Mining: Concepts and Techniques*. Second Edition. Elsevier. San Francisco, CA.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2001). *Data Mining: Concepts and Techniques*. Third Edition. Elsevier. San Francisco, CA.

Haykin, Simon (2001). Redes neurais: princípios e práticas/Simon Haykin; trad. Paulo Martins Engel. – 2.ed. – Porto Alegre: Bookman.

Hurwitz, Judith; Nugent, Alan; Halper, Fern; Kaufman, Marcia. (2016). Big Data para Leigos. Editora Alta Books. Rio de Janeiro.

Hussain F.; Liu H.; Suzuki E.; Lu H. (2000). EXCEPTION RULE MINING WITH RELATIVE INTERESTINGNESS MEASURE. PAKDD; pg 86-97.

Inmon, Bill & Chuck Kelly (1994). The Twelve Rules of Data Warehouse for a Client/Server World, Data Management Review.

JEFFREY, Dean; GUEMAWAT, Sanjay (2004). **MapReduce: Simplified Data Processing on Large Clusters** – Google Inc. Acessado em 14/9/2014. Disponível em:
<http://static.googleusercontent.com/media/research.google.com/pt-BR/archive/mapreduce-osdi04.pdf>.

KAPP, Karl (2012). The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education. Pfeiffer.

Kimball, Ralph (2002). Data Warehouse toolkit: o guia completo para modelagem multidimensional /Ralph Kimball, Margy Ross; tradução Ana Beatriz Tavares, Daniela Lacerda. Rio de Janeiro: Campus.

Kolb, Jason; KOLB, Jeremy (2013). The Big Data Revolution. The Tricks That Competitors Don't Want You To Know By Jason Kolb and Jeremy Kolb. AppliedData Labs. Plainfield, IL.

Kohonen, Teuvo (1987). *Self-Organization and Associative Memory*. Springer-Verlag Series in Information Science.

Kolodner, J. L. (1998). Proceedings of the DARPA Case-Based Reasoning Workshop. San Francisco: Morgan Kaufmann Publishers

Liu, B. & W. Hsu. Post-analysis of learned rules. AAAI 1, 828-834, 1996.

Mayer-Schönberger, Viktor; Cukier, Kenneth (2013). Big Data. A Revolution That Will Transform How We Live, Work and Think. First published in Greta Britain. John Murray (Publishers) an Hachette UK Company.

MENDES, Antônio (2002). **Arquitetura de Software: desenvolvimento orientado para arquitetura** – Editora Campus. Rio de Janeiro – RJ.

Morik, K. (2000). The Representation Race- Preprocessing for Handling Time Phenomena. Proceedings of the European Conference on Machine Learning 2000, Lecture Notes in Artificial Intelligence 1810. Berlin: Springer Verlag.

NAGY, Heba Mohammed; ALY, Walid Mohamed; HEGAZY, Osama Fathy (2013). **An Educational Data Mining System for Advising Higher Education Students**. International Journal of Computer, Information Science and

Engineering, Vol:7, N:10, 2013. World Academy of Science, Engineering and Technology.

Negnevitsky, Michael (2005). Artificial Intelligence: a guide to intelligent Systems/Michael Negvitsky. Pearson Education Limited. Edinburgh Gate.

Nicholson, S. (2012) A User-Centered Theoretical Framework for Meaningful Gamification. Paper Presented at Games+Learning+Society 8.0, Madison, WI.

Oliveira, C. (2007). EDACLUSTER: Um Algoritmo Evolucionário para Análise de agrupamentos Baseados em Densidade e Grade, Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal do Pará.

Osório, Fernando (2011). Redes Neurais – Aprendizado Artificial. Forum de I.A. “99 – pg.13”. Rosa, João Luís Garcia. Fundamentos da Inteligência artificial /João Luís Garcia Rosa. Rio de Janeiro: LTC.

Passos, Emanuel; GOLDSCHMIDT, Ronaldo (2005). Data Mining: Um guia prático. Editora Campos. Rio de Janeiro.

Pazzini, M. J. (2000). Knowledge discovery from data? IEEE Intelligent Systems.

Piatetsky-Shapiro, G & C. J. Matheus (1994). The Interestingness of deviations. Em Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 23-36.

Raj, Subu (2013). BIG DATA – AN INTRODUCTION. Kindle Ver 1.1.

Rezende, Solange Oliveira (2003). Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda. Barueri, SP.

Riesbeck, C. K., and Schank, R. (1996). Inside Case-Based Reasoning. Northvale, NJ: Lawrence Erlbaum Associates.

Rob, Peter; Coronel, Carlos (2009). Database Systems: Design, Implementation, and Management by Peter Rob and Carlos Coronel 8th Edition. Thomson Place, Boston, Massachusetts.

Robt, Peter (2011). Sistemas de Banco de Dados: Projeto, implantação e gerenciamento / Peter Rob, Carlos Coronel. São Paulo: Cengage Learning.

Sadalage, Pramod J; Fowler, Martin (2013). **NoSQL Essencial, Um guia conciso para o mundo emergente da persistência poliglota**. Novatec Editora – São Paulo, SP.

Silberschatz, A. & Tuzhilin (1995). On subjective measures of interestingness in knowledge discovery. Proceeding of the First International Conference on Knowledge Discovery and Data Mining 1, 275-281.

Stonebraker, Michael, Abadi; Daniel, DeWitt; David J.; Madden, Sam; Paulson, Erik; Pavlo, Andrew; Rasin, Alexander (2001). MapReduce complements DBMSs since databases are not designed for extracttransform-load tasks, a MapReduce specialty. COMMUNICATIONS OF THE ACM, pp 71.

Tiwary, Shashank (2011). **Professional NoSQL, a hands-on guide to leveraging NoSQL databases**. John Wiley & Sons.

Utgolff, P. (1996). Shift of Bias for Inductive Concept Learning. Machine Learning: an Artificial Intelligence Approach, v.3, São Francisco: Morgan Kaufmann.

Watson, Ian D. (1997). Applying case-based reasoning: techniques for enterprise systems. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Watson, Ian D. (2003). Applying Knowledge Management: Techniques for Building Corporate Memory. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Wangenheim, Christiane G. VON; Wangenheim, A. V. (2003). Raciocínio Baseado em Casos. São Paulo: Manole Ltda.

Witten, Ian H.; Frank, Eibe; Hall, Mark A. (2005). Data Mining. Practical Machine Learning Tools and Techniques. 2nd ed. Morgan Kaufmann Publishers is an imprint of Elsevier.

Witten, Ian H.; Frank, Eibe; Hall, Mark A. (2011). Data Mining. Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufmann Publishers is an imprint of Elsevier.

Zichermann, Gabe. (2011) Gamification by Design. ISBN 1449397670. 150 pages. O'Reilly.

_____. **AngularJS** (2010). Google Inc. – Acessado em: 14/8/2014. Disponível em: <http://docs.angularjs.org>

_____. **BSON Specs** (2014). Creative Commons (sem direitos autorais). – Acessado em: 6/8/2014. Disponível em: <http://bsonspec.org/spec.html>

_____. **MongoDB** (2014). MongoDB Inc. – Acessado em: 14/8/2014. Disponível em: <http://www.mongodb.com/what-is-mongodb>

_____. **Mongoose** (2010). LearnBoost. – Acessado em: 14/8/2014. Disponível em: <http://mongoosejs.com/docs/guide.html>

_____. **NodeJS** (2014). Joyent Inc. – Acessado em: 10/6/2014. Disponível em: <http://nodejs.org/documentation>

_____. **Standard ECMA-262** (2011). ECMA International – Acessado em: 23/9/2014. Disponível em: <http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-262.pdf>

Leituras Complementares

Boente, A.N.P. (2006). Descoberta de Conhecimento em Bases de Dados. Iowa, Tese de Doutorado – Departamento de Informática, AWU – American World University.

Horst, P. S. (1999). Avaliação do conhecimento adquirido por algoritmos de aprendizado de máquina utilizando exemplos. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação. São Paulo, SP – Brasil.

Larson, Brain; Davis, Mark; English, Dan; Purigton, Paul (2012). Visualizing Data with Microsoft Power View. McGraw Hill companies.

Loukides, Mike (2011). What is Data Science? The future belongs to the companies and people that turn data into products. O'Reilly.

Mayer-Schönberger, Victor; Curier Kenneth (2013). A Revolution That Will Transform How We Live, Work and Think. First published in Great Britain.

Milani, Cristian Simioni, Carvalho, Deborah Ribeiro (2013). PÓS-PROCESSAMENTO EM KDD. Revista de Engenharia e Tecnologia. PUCPR, ISSN 2176-7270, <http://www.revistaret.com.br/ojs-2.2.3/index.php/ret/article/viewFile/170/182>, pp. 153-155.

Noren, Allen (2011). Big Data Now. Current Perspectives from O'Reilly Radar. O'Reilly Strata Making Data Work.

Özu, M. Tamer (2001). Princípio de sistemas de banco de dados distribuídos/M. Tamer Özu, Patrick Valduriez; tradução [da 2. Ed. Americana] Vandenberg D. de Souza – Rio de Janeiro: Campus.

Payandeh, Fari (2013). BI vs. Big Data vs. Data Analytics By Example. <http://bigdatastudio.com/2013/08/24/bi-vs-big-data-vs-data-analytics-by-example/>. Acessado em 27 de setembro de 2013.

Patil, DJ (2011). The skills, Tools, and Perspectives Behind Great Data Science Groups. Building Data Science Teams. O'Reilly Strata Makin Data Work.

Rhoton, John; Haukioja, Risto (2011). Cloud Computing Architected. Solution Design HandBook. Recursive Press.

Schank, R. (1982). Dynamic Memory: A Theory of Learning in Computers and People. New York: Cambridge University Press.

Schroeder, Christine da Silva (2005). Critérios e indicadores de desempenho para sistemas de treinamento corporativo virtual: um modelo para medir resultados. Dissertação de mestrado – Universidade Federal do Rio Grande do Sul, Escola de Administração, Programa de Pós-Graduação em Administração.

Silberschatz, Abraham; Korth, Henry F. e Sudarshan, S. (1999). Sistema de Banco de Dados. São Paulo: Makron Books.

Sosinsky, Barrie (2011). Cloud Computing Bible. Wiley Publishing, Inc.

Velte, Anthony T.; Velte, Toby J.; Elsenpeter, Robert (2012). Cloud Computing: Computação em Nuvem: Uma Abordagem Prática. ALTA BOOKS. Rio de Janeiro.

Zikopoulos, Paul C.; Eaton, Chris; deRoos, Dirk; Deutsch, Thomas; Lapis, George (2012). Understanding Big Data. Anaylsis for Enterprise Class. Hadoop and Streaming Data. McGraw-Hill.