Projeto da Disciplina Ciência de Dados
Fase de Entendimento dos Dados

Relatório de Coleta, Descrição, Exploração e Avaliação da Qualidade dos Dados

Carlos Eduardo Nascimento Cajado

Análise e Predição de Alunos com Risco de Baixo Desempenho Acadêmico

Histórico de Revisões do Documento

Revisão	Descrição	Modificado por	Status	Data
1.0	Levantamento, coleta e exploração dos dados.	Carlos Eduardo Cajado	Em desenvolvimento	23/10/2024

1 Coleta dos Dados

A coleta de dados foi realizada a partir da extração de um banco de dados SQL Server de uma instituição acadêmica específica, referida como “Instituição A” localizada na cidade de são Luís do maranhão do nível básico ao médio, limitando-se exclusivamente às três séries do ensino médio. Escolhida pela relevância em relação ao número de alunos, e por representar a maior instituição cliente.

O foco da coleta foi a junção de várias tabelas com a tabela de notas, que contém, além do conjunto de notas de cada disciplina, informações sobre a situação acadêmica final do aluno (Tabela 1) ao longo dos anos letivos, referentes a cada disciplina cursada.

Os dados foram limitados ao período de sete anos, especificamente de 2018 a 2023. Na seção 1.1, são apresentadas as tabelas com suas respectivas descrições, juntamente com algumas problemáticas encontradas durante o processo de coleta.

1.1 Banco SQL-Server

A Tabela 1 apresenta a lista das tabelas do banco de dados utilizadas neste projeto. O objetivo é concatenar as colunas relevantes por meio de métodos e manipulações na linguagem SQL, visando a criação de um Dataset no formato CSV como saída.

Tabela 1: Tabelas extraídas da Base de Dados

Nome Tabela	Colunas Utilizadas	Descrição
Notas	• Matricula • idAluno • idGrade • idDisciplina • idSerie	Armazena as notas bimestrais e finais, bem como as informações sobre recuperações, além das chaves estrangeiras relacionadas a disciplinas, séries e classificações de ocorrência.
Classificação_Ocorrências	• idClassifica-cao_de_Ocorrencias • Descricao	Contém a descrição de cada classificação de ocorrência, como, por exemplo: 'REPROVADO', 'DESISTENTE' e 'BOM/APROVADO' (mais detalhes podem ser encontrados na Tabela 2, presente na Seção 2).
Disciplinas	• Descricao • IdDisciplina	Registra as informações relacionadas a disciplina como: Descrição, carga Horária, idDiario.
Alunos	• Sexo • dataNasc	Contém informações pessoais dos alunos, como data de nascimento e sexo.
Responsavel	• idEstadoCivil • EscolaridadeResponsavel	Inclui dados sobre o responsável pelo aluno, destacando informações relevantes, como escolaridade e estado civil.
Faltas_Alunos	• Aula1 – Aula8	Armazena as faltas dos alunos em cada disciplina.

Fonte: Proprio Autor

Além disso, é importante destacar a exclusão de dados espúrios e nulos, especialmente os presentes na tabela de notas, onde podem ser encontrados alguns registros com notas vazias.

Também é relevante mencionar a ausência de lançamentos de frequências para todos os registros das disciplinas.

2 Descrição dos Dados

Ao final do processo de coleta, visando a futura necessidade de converter as colunas de texto relacionadas a “Situacao_Disciplina” e “Sexo” para análises gráficas ou aplicação em um modelo preditivo, foram realizadas manipulações no ambiente *Jupyter Notebook*. Essas manipulações converteram os textos em valores inteiros, conforme detalhado na Tabela 2. O *dataset* final resultou em um arquivo CSV contendo 52.720 registros válidos de 1.118 alunos, abrangendo o período de 2018 a 2023, com as colunas e características descritas na Tabela 3.

2.1 Arquivo CSV

A conversão por mapeamento para a coluna Sexo é realizada da seguinte forma: o valor "F" é convertido para o inteiro 1, representando o gênero feminino, enquanto "M" é convertido para 2, representando o gênero masculino. Para a coluna Situação da Disciplina, a conversão foi feita conforme detalhado a seguir:

Tabela 2: Correspondência do mapeamento situação disciplina.

Situação Disciplina	Situação Disciplina Int
NÃO VAI CURSAR A DISCIPLINA	0
INTERCAMBISTA	
PONTUAÇÃO INSUFICIENTE	1
DESISTENTE	
REPROVADO	
TRANSFERIDO	2
APROVADO PELO CONSELHO	5
APROVADO	7
BOM/APROVADO	8
MUITO BOM/APROVADO	10
EXCELENTE/APROVADO	
BRILHANTE/APROVADO	

Fonte: Proprio Autor

Como resultado das adições tem-se as seguintes colunas e seus respectivos tipos, conforme apresentado na figura 1. A tabela 3 mostra uma breve amostra da disposição das colunas e seus valores.

Figura 1: Colunas do arquvivo.csv *dataset*.

Nome da coluna	Tipo de dados
Index	int64
situacaoDisciplina	string
anoNascimento	int64
Sexo	string
idSerie	int64
EscolaridadeResponsavel	flutuador64
idEstadoCivilResponsavel	float64
Total_Faltas	int64
Reprovacoes_%	flutuador64
AprovacoesComExcelencia_%	float64
Aprovacoes	float64
Diciplina	int64
sexo_int	int64
situacaoDisciplina_inteiro	int64

Fonte: Proprio Autor

As colunas "Index" (inteiro), que atua como contador para cada registro, além de "situacaoDisciplina" e "Sexo" (string). Também estão disponíveis dados numéricos, como "anoNascimento", extraído da data de nascimento, "idSerie" e "Total_Faltas" (inteiro). Os percentuais, como "Reprovacoes_%", "AprovacoesComExcelencia_%" e "aprovações" — que representam a porcentagem de disciplinas aprovadas ou reprovadas pelo aluno (float).

As colunas "EscolaridadeResponsavel" e "idEstadoCivilResponsavel (float), cujas correspondências estão detalhadas nas Tabelas 4 e 5. Além disso, as colunas "Diciplina", "sexo_int" e "situacaoDisciplina_inteiro" (inteiro).

Tabela 3: Disposição dos registos no dataset.

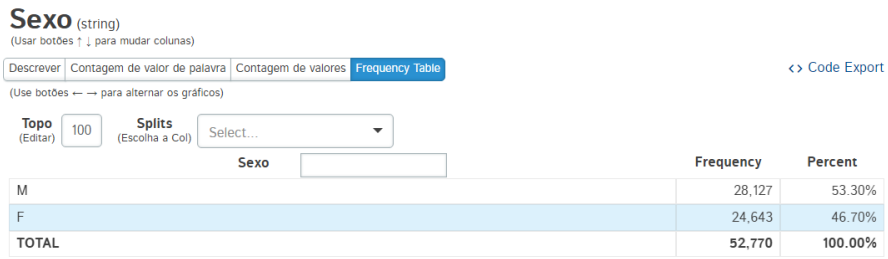
Index	situacaoDisciplina	anoNascimento	Sexo	idSerie	EscolaridadeResponsavel	idEstadoCivilResponsavel	Total_Faltas	Reprovacoes_%	AprovacoesComExcelencia_%	Aprovacoes	Diciplina	sexo_int
3	REPROVADO	2002	F	49	1.00	2.00	0	100.00	0.00	0.00	28	1
4	REPROVADO	2002	F	49	2.00	2.00	0	100.00	0.00	0.00	28	1
5	REPROVADO	2002	F	49	2.00	2.00	0	100.00	0.00	0.00	11	1
6	REPROVADO	2002	F	49	1.00	2.00	0	100.00	0.00	0.00	29	1
7	REPROVADO	2002	F	49	6.00	2.00	0	100.00	0.00	0.00	29	1
8	REPROVADO	2002	F	49	2.00	2.00	0	100.00	0.00	0.00	12	1
9	REPROVADO	2002	F	49	1.00	2.00	0	100.00	0.00	0.00	31	1
10	REPROVADO	2002	F	49	6.00	2.00	0	100.00	0.00	0.00	31	1

Fonte: Proprio Autor

3 Exploração dos Dados

Com o objetivo de realizar análises mais detalhadas, foram realizadas agregações simples e identificadas propriedades de subpopulações significativas, além de análises estatísticas básicas. Nas Figuras 2 e 3, foram utilizadas a biblioteca D-Tale do Python para analisar e visualizar a distribuição das instâncias utilizadas.

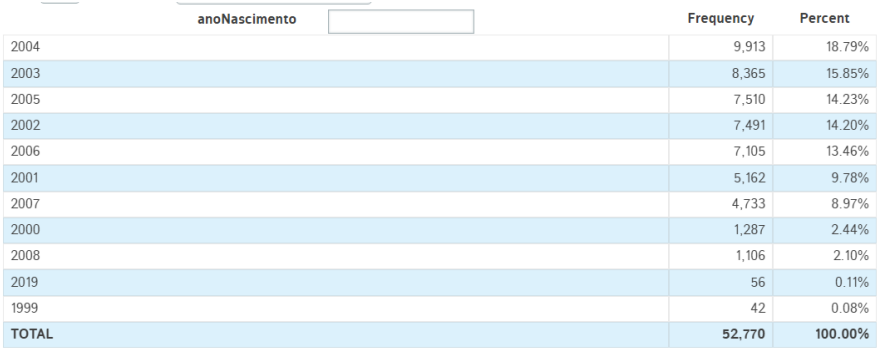
Figura 2: Disposição dos registos relacionado ao sexo.



Fonte: Proprio Autor

É notório que a distribuição dos sexos dos alunos é quase equivalente, com 53,30% do total composto por alunos do sexo masculino e 46,70% do sexo feminino. A Figura 3 apresenta a lista do ano de nascimento, incluindo a frequência e a porcentagem em relação ao total de registros.

Figura 3: Disposição dos registos relacionado a data de nascimento.

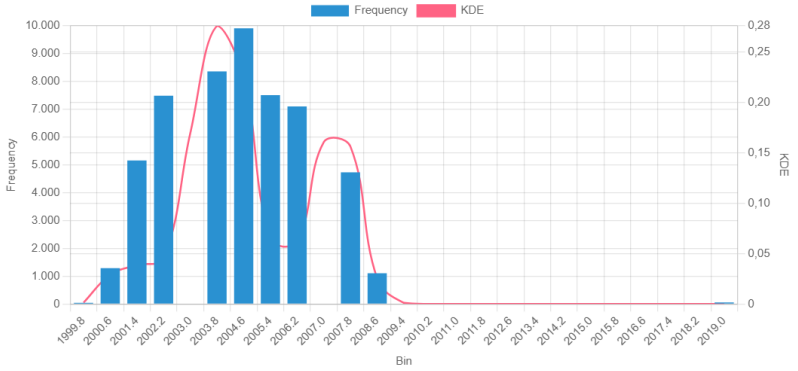


Fonte: Proprio Autor

É importante salientar que, além do intervalo de anos de 1999, que possui poucos registros (0,08%), até 2008, que apresenta 2,10%, a maior frequência foi observada em 2004, com 18,79%. Além disso, destaca-se a presença de um outlier no ano de 2019, já que é inviável cronologicamente que um grupo de alunos, precisamente 56, nascidos em 2019 esteja atualmente cursando uma das

três séries do ensino médio. Isso provavelmente se deve a registros incorretos por parte da instituição responsável. Para uma melhor visualização, a Figura 4 apresenta o histograma da coluna em questão.

Figura 4: Histograma ano de nascimento dos alunos.



Fonte: Proprio Autor

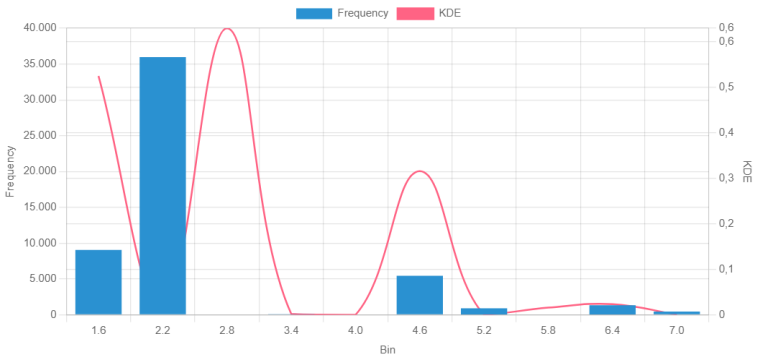
Na figura 5, é apresentado o histograma da escolaridade do responsável, este com a distribuição da escolaridade dos responsáveis, categorizada da seguinte forma, tabela 4:

Tabela 4: Descrição escolaridade responsável.

Id Escolaridade Responsável	Descrição Escolaridade
1	SUPERIOR COMPLETO
2	ENSINO MÉDIO COMPLETO
3	ENSINO MÉDIO INCOMPLETO
4	ENSINO FUNDAMENTAL COMPLETO
5	ENSINO FUNDAMENTAL INCOMPLETO
6	NÃO INFORMADA
7	SUPERIOR INCOMPLETO

Fonte: Proprio Autor

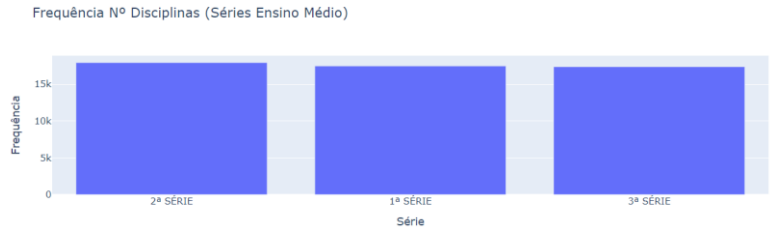
Figura 5: Histograma escolaridade dos responsáveis.



Fonte: Proprio Autor

Outra característica importante é a distribuição dos registros em relação à série do aluno, que abrange a 1ª, 2ª e 3ª séries, conforme mostrado na Figura 6. Além disso, a Figura 7 apresenta a situação das disciplinas dos alunos, conforme detalhado na tabela 2. Essas análises são essenciais para entender o desempenho acadêmico e identificar áreas que podem exigir atenção adicional.

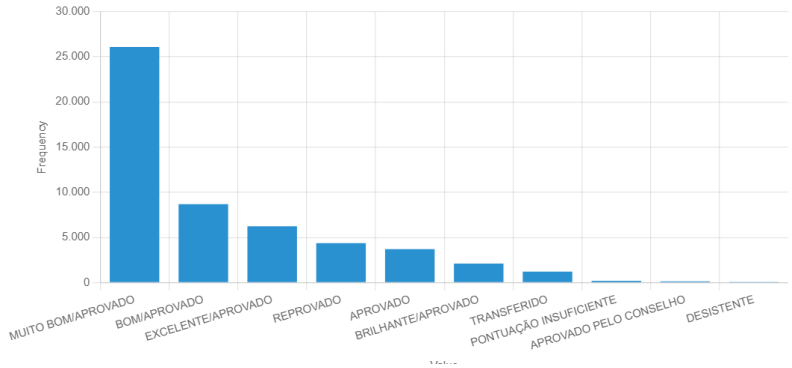
Figura 6: Disposição dos registros nas séries, grade ensino médio.



Fonte: Proprio Autor

Um fato relevante é a distribuição quase uniforme dos registros, mesmo ao considerar apenas o recorte da grade (ensino médio), sem qualquer seleção ou manipulação dos dados. Essa equidade na distribuição sugere uma amostra representativa e fortalece a validade das análises realizadas.

Figura 7: Disposição dos registos relacionado a situação final disciplina.



Fonte: Proprio Autor

Nesta figura, observa-se a frequência de cada situação final das disciplinas cursadas pelos alunos. Curiosamente, a grande maioria dos alunos é aprovada com excelência, recebendo classificações como “Muito Bom”, “Excelente” e “Brilhante”, totalizando cerca de 81,64%. Além disso, os alunos aprovados nas categorias “APROVADO” e “APROVADO PELO CONSELHO” somam mais 7,30%. A porcentagem de reprovados é de 8,28%.

Para uma análise mais aprofundada dos dados, apresentamos as Figuras 8 e 9. A Figura 8 mostra a relação entre o número de aprovações com excelência e a escolaridade do responsável, agrupando as informações por série.

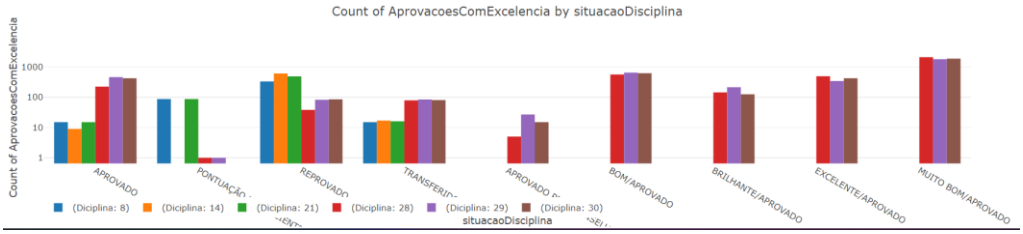
Figura 8: Sprovações com excelência X escolaridade do responsável.

Fonte: Proprio Autor

É indubitável que a maior parte dos registros de aprovações com excelência está associada ao grupo de responsáveis com escolaridade de "Ensino Superior Completo", conforme apresentado na Tabela 4. Em segundo lugar, destaca-se a categoria "Não Informado".

Por sua vez, a Figura 9 ilustra a proporção entre o número de aprovações com excelência e a situação das disciplinas, agrupadas pelas disciplinas.

Figura 9: Aprovações com excelência em relação à situação das disciplinas.

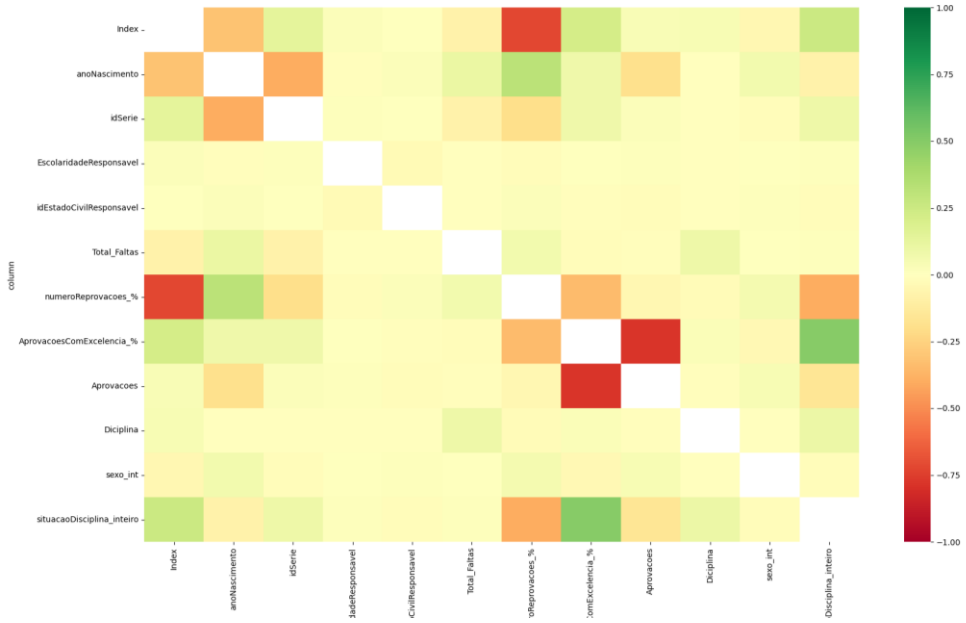


Fonte: Proprio Autor

Através da figura 9, são apresentadas informações sobre a frequência de aprovações com excelência em relação à situação das disciplinas, com alguns exemplos incluídos: 8 (PROJETOS PARA A VIDA), 14 (FILOSOFIA / SOCIOLOGIA), 21 (LÍNGUA PORTUGUESA), 28 (BIOLOGIA), 29 (FÍSICA) e 30 (QUÍMICA). Observa-se que as disciplinas com maior índice de reprovação e menos aprovações com excelência são PROJETOS PARA A VIDA, FILOSOFIA / SOCIOLOGIA e LÍNGUA PORTUGUESA. Por outro lado, as disciplinas com o maior número de aprovações com excelência são BIOLOGIA, FÍSICA e QUÍMICA.

Por fim, a base de dados utilizada apresenta as correlações mostradas na Figura 10. Observa-se correlações positivas fortes, como entre número de reprovações e o ano de nascimento, além de correlações negativas entre o percentual de aprovações com excelência e o percentual de aprovações.

Figura 10: Matrix de correlação.



Fonte: Proprio Autor

4 Qualidade dos Dados

Ao examinar os dados coletados, constatou-se que, embora a maioria dos registros esteja completa, alguns apresentam inconsistências e erros. A análise revelou a ausência de notas em determinadas disciplinas, o que pode comprometer a avaliação do desempenho acadêmico dos alunos. Para resolver a problemática, foram implementadas estratégias de limpeza de dados, incluindo a exclusão de registros incompletos.

Embora os dados incorretos sejam relativamente raros, foram identificados alguns casos, especialmente nas tabelas relacionadas às frequências e ao cadastro dos alunos, onde surgiram registros com valores nulos e erros, como datas de nascimento inconsistentes. Essas falhas podem prejudicar futuras aplicações em modelos preditivos, distorcendo as análises e impactando as decisões baseadas nas informações coletadas.