CENTRO UNIVERSITÁRIO DE JOÃO PESSOA – UNIPÊ
PRÓ-REITORIA ACADÊMICA - PROAC
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

**PEDRO HENRIQUE FONTES FEITOSA**

# UMA ANÁLISE DO DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR NO ENADE UTILIZANDO APRENDIZAGEM DE MÁQUINA

**JOÃO PESSOA – PB**

**2018**

**PEDRO HENRIQUE FONTES FEITOSA**

# UMA ANÁLISE DO DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR NO ENADE UTILIZANDO APRENDIZAGEM DE MÁQUINA

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação do Centro Universitário de João Pessoa - UNIPÊ, como pré-requisito para a obtenção do grau de Bacharel em Ciência da Computação, sob orientação do Prof. MS.c. Hugo Vieira Lucena de Souza.

**JOÃO PESSOA - PB**

**2018**

F311a Feitosa, Pedro Henrique Fontes.

Uma Análise no Desempenho de Instituições de Ensino Superior no ENADE Utilizando Aprendizagem de Máquina

Pedro Henrique Fontes Feitosa. - João Pessoa, 2018.

86f.

Orientador (a): Prof. MS.c. Hugo Vieira Lucena de Souza.

Monografia (Curso de Ciências da Computação)

**PEDRO HENRIQUE FONTES FEITOSA**

**UMA ANÁLISE DO DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR NO ENADE UTILIZANDO APRENDIZAGEM DE MÁQUINA**

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação do Centro Universitário de João Pessoa - UNIPÊ, como pré-requisito para a obtenção do grau de Bacharel em Ciência da Computação, apreciada pela Banca Examinadora composta pelos seguintes membros:

Aprovada em _____/_____/2018.

**BANCA EXAMINADORA**

_____

Prof. MS.c. Hugo Vieira Lucena de Souza (UNIPÊ)

_____

Prof. (título ex.: Dr./Ms./Esp.) Nome do professor Examinador (a) (UNIPÊ)

_____

Prof. (título ex.: Dr./Ms./Esp.) Nome do professor Examinador (a) (UNIPÊ)

Dedico a todos que desviaram minha atenção

quando estava no caminho errado.

## AGRADECIMENTOS

Quero agradecer à minha família por me dar tantas oportunidades, por me guiar e pelo esforço de estar presente, e por conta disso acabaram aprendendo junto a mim o que se estuda em Ciências da Computação.

Aos amigos pelas conversas e alegrias compartilhadas.

Aos professores e colegas de curso pela oportunidade de aprender juntos, dentro e fora da sala de aula.

A todos com que trabalhei, por me ajudarem a pôr em prática e validar a importância do que aprendi, e pela paciência e atenção dada neste processo.

Ao professor Hugo e à Coordenação por me ajudar nas dificuldades passadas no desenvolver deste trabalho.

## RESUMO

Ao final de cursos de graduação, os estudantes brasileiros de instituições públicas e privadas podem ser convidados a prestar o Exame Nacional de Desempenho de Estudantes (ENADE). Este exame avalia as competências e habilidades desenvolvidas ao longo das suas vidas acadêmicas, nas mais diversas áreas de conhecimento tais como as ciências humanas, ciências sociais, ciências exatas e da natureza, dentre outras. Anterior ao processo de aplicação da prova de desempenho, é realizada uma pesquisa socioeconômica com os participantes que tem por objetivo identificar os perfis dos egressos, coletando dados gerais e demográficos a respeito dos cursos e das instituições de ensino superior. Após a realização da prova de desempenho, o INEP tem o trabalho de calcular as médias dos cursos e, a partir destas, é atribuído o Conceito ENADE, uma nota que irá ponderar o nível de qualidade e excelência para cada curso. Entretanto, poucas instituições de ensino superior conseguem detectar quais são os fatores socioeconômicos do corpo discente, que causam um maior impacto na atribuição das notas, para que se obtenha um bom desempenho. A partir dos dados públicos disponibilizados pelo MEC é possível entender, estruturar e extrair as informações utilizando as técnicas provenientes da aprendizagem de máquina. Com uma base de dados adequada, torna-se praticável a criação de modelos preditivos que eventualmente poderão replicar as respostas para o problema citado. Portanto, este trabalho tem como propósito realizar um estudo de predição, com a base de microdados do ENADE 2015, para identificar quais aspectos sociais ou econômicos podem afetar o desempenho dos alunos em uma avaliação posterior utilizando as técnicas recomendadas pela aprendizagem de máquina. Com base nos resultados obtidos pelo estudo, instituições de ensino que procuram obter bons resultados podem ser melhor orientadas.



**Palavras-Chave:** Análise de desempenho escolar. Aprendizagem de Máquina. Mineração de dados. Mineração de Dados Educacionais.

## ABSTRACT

Before graduating college, Brazilian students from both public and private institutions may be invited to take the National Students Performance Exam (ENADE). This exam assesses the skills and abilities developed throughout their academic careers in the most diverse areas of knowledge such as human science, social science, exact sciences, science of nature, among others. Prior to its application, a socio-economic survey is carried out with the participants to identify the profiles of the graduates collecting general and demographic data about the courses and Higher Education Institutes (HEI). After performing the exam, INEP focuses on determining for each course its students average, and, from these, it is assigned the ENADE Concept, a grade which will measure the course's quality and excellence. However, few HEI are able to detect which are the socioeconomic factors which have a greater impact on the allocation of the student's grades, in order to obtain a good performance. From the public data provided by MEC it is possible to understand, structure and extract information using techniques derived from machine learning. Using a suitable database, it becomes feasible to create predictive models that may eventually replicate the answers to the problem mentioned. Therefore, the purpose of this paper is to perform a prediction study based on the microdata from 2015's ENADE to identify which social or economic aspects may affect students' performance in a later evaluation using techniques provided by machine learning. By the results obtained from this study, HEI whom seek help in obtaining good results in the ENADE may be better oriented.



**Keywords:** Analysis of school performance. Machine Learning. Data Mining. Mining of Educational Data.

**LISTA DE TABELAS**

**LISTA DE FIGURAS**

## LISTA DE GRÁFICOS

## LISTA DE ABREVIATURAS E SIGLAS

**AM** - Aprendizagem de Máquina

**CE** - Conceito ENADE

**CRISP-DM** - *Cross Industry Standard Process for Data Mining*

**CSV** - *Comma-Separated Values*

**ENADE** - Exame Nacional do Desempenho de Estudantes

**INEP** - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

**KDD** - *Knowledge Discovery in Databases*

**K-S** - Kolmogorov-Smirnov

**MDE** - Mineração de Dados Educacionais

**MEC** - Ministério da Educação

**NCc** - Nota dos Concluintes no ENADE do Curso de graduação c

**PI** - Procuradores Educacionais Institucionais

**ROC** - *Receiver Operating Characteristics*

**SQL** - *Structured Query Language*

**SUMÁRIO**

# 1 INTRODUÇÃO

Em um período que tem um intervalo de três anos, os alunos concluintes dos cursos de graduação de Instituições de Ensino Superior (IES) públicas e privadas são submetidos a uma prova que busca avaliar o desempenho em relação aos conteúdos programáticos, habilidades, e competências adquiridas em sua formação, o Exame Nacional do Desempenho de Estudantes (ENADE) (INEP, 2017).

Inicialmente, os estudantes passam por um questionário socioeconômico, que aborda as seguintes informações: o estado civil, a etnia, nacionalidade, informações sobre o grau de estudos dos pais, tipo de habitação e quantidade de pessoas nela, renda total familiar, relação de renda e incentivos por programas governamentais, situação de trabalho, auxílio de bolsas ou financiamento do curso, auxílio permanência, auxílio de bolsas acadêmicas, entre outras em que o aluno avalia critérios da estrutura dada pela IES e seus professores.

Já o exame é composto por questões de formação geral e de conteúdo específico ao seu curso, contendo questões objetivas e discursivas, além de outro questionário sobre a percepção do estudante na prova. O resultado individual é de grande importância para o estudante uma vez que fica em seu histórico e também, como um todo para a instituição de ensino. Unindo os resultados do exame é realizada uma série de avaliações, e a partir delas é obtida a Nota dos Concluintes no ENADE do Curso de graduação c (NCc). Com o NCc se dá o Conceito ENADE (CE) conforme descrito na Tabela 1.

Tabela 1 - Parâmetros de conversão do NCc em Conceito ENADE

| **Conceito Enade** (Faixa) | **NCc** (Valor Contínuo) |
|---|---|
| 1 | $0 \leq NCc < 0{,}945$ |
| 2 | $0{,}945 \leq NCc < 1{,}945$ |
| 3 | $1{,}945 \leq NCc < 2{,}945$ |
| 4 | $2{,}945 \leq NCc < 3{,}945$ |
| 5 | $3{,}945 \leq NCc \leq 5$ |

Fonte: Inep/Daes (2015)

Esta nota é um fator de preocupação para as instituições de ensino superior, pois o baixo rendimento pode levar a instituição a ser notificada pelo Ministério da Educação (MEC) uma vez que é considerado aceitável pelo ministério os conceitos maiores ou iguais a 3. Quando a nota é inferior a este valor a IES está propensa de ter a suspensão das atividades relacionadas para um determinado curso.

Após a divulgação do resultado final do exame a IES tem a dificuldade em identificar quais fatores sociais poderiam gerar melhorias no desempenho dos alunos. Este objetivo pode ser alcançado com estudo dos dados históricos que são divulgados no formato de microdados[1] pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

A área que tem se mostrado eficiente para atingir esse objetivo é conhecida como mineração de dados. Ela orienta a execução de um conjunto de etapas que possibilita a extração de conhecimento com o uso de algoritmos que utilizam mecânicas provenientes da Inteligência Artificial.

Uma destas mecânicas concentra-se na detecção de padrões de comportamento proporcionados por uma base de dados a partir do momento em que ocorre uma sequência de passos que envolvem a limpeza, a integração, a seleção e a transformação dos dados. Tais operações são possíveis de serem realizadas graças a uma área da IA conhecida por Aprendizagem de Máquina (AM). Portanto, o problema definido para esta pesquisa é o de identificar quais aspectos sociais e econômicos geram um efeito no desempenho dos alunos que realizam o ENADE com intuito de orientar IES em suas próximas avaliações inserindo as informações necessárias dos estudantes para predizer seus desempenhos.

## 1.1 RELEVÂNCIA DO ESTUDO

Há vários estudos, como os abordados na seção 3.5, que obtém resultados a partir de predições realizadas com o uso de bases de dados do ENADE de anos anteriores, dentre eles há pesquisas que buscam desde validar a aplicação de AM nos microdados, classificar para cada universidade as áreas do exame que seus estudantes passam por maior dificuldade, ou até demonstrar o impacto da pesquisa socioeconômica no desempenho estudantil.

Esse tipo de estudo se encaixa em uma área denominada Mineração de Dados Educacionais (MDE). A área abrange vários tipos de pesquisas que buscam analisar dados educacionais e existem diversos materiais que buscam orientar e regulamentar a sua aplicação em algumas de suas subáreas. Porém ela tem seu foco principal em um problema explicado por Costa et al. (2012) que surgiu do crescente uso de ambientes virtuais de aprendizagem, e que teve como consequência a geração de um grande número de dados a partir das mais diversas modalidades de interações com tais sistemas, exigindo cada vez mais a sua análise. Tal análise, feita com mineração de dados, pode extrair conhecimento a respeito do comportamento dos estudantes e sobre a forma como eles aprendem, servindo de apoio para a melhoria do desempenho dos estudantes.

1 Microdados: menor nível de desagregação de dados recolhidos por pesquisas, avaliações e exames realizados.

Porém na MDE não deixam de ser abordados problemas como o proposto por este trabalho. Observando os resultados de estudos relacionados, no qual serão abordados mais à frente na seção de trabalhos relacionados, torna-se possível direcionar as IES às características socioeconômicas dos estudantes que mais passam por dificuldades ao realizar o ENADE. Os dados fornecidos pelo INEP em 2015 contam com mais estudantes, mais cursos, e a continuidade das pesquisas podem ser validadas e atualizadas.

## 1.2 OBJETIVO GERAL

Realizar uma análise na base de dados do ENADE 2015, com o propósito de gerar um modelo preditivo que possibilite identificar quais fatores sociais e econômicos geram impacto no desempenho dos alunos para uma eventual avaliação em um futuro exame que outros estudantes poderão participar.

## 1.3 OBJETIVOS ESPECÍFICOS

Os objetivos específicos elencados para este trabalho são:

- Realizar o processo de limpeza, pré-processamento e tratamento da base de dados do ENADE de 2015.
- Gerar a modelagem do fluxo decisório, que definirá como será o processo de comparação e verificação dos dados utilizando a metodologia KDD.
- Construir um modelo preditivo que reúne as variáveis escolhidas que têm uma maior representatividade para verificar os efeitos no modelo de predição proposto.
- Realizar os testes de sensibilidade (Curva KS), testes de efetividade (Curva ROC), validação e treinamento do modelo gerado (K-folder) para verificar o grau de confiança e satisfação do modelo gerado.

## 1.4 INDICAÇÃO DA METODOLOGIA

Esta pesquisa é do tipo de levantamento de dados, onde se utiliza dados da população para que obtenha o conhecimento direto da realidade, utilizando-se da análise estatística (GERHARDT; SILVEIRA, 2009). Buscando conhecer as relações das características da população, para se utilizar da generalização constatada a partir da observação de exemplos concretos, será utilizado o método de abordagem indutivo, e para chegar nesta generalização, se utilizará o método de procedimento estatístico, analisando a probabilidade de uma população pertencer a um grupo, validando a probabilidade de acerto e a margem de erro (GIL, 2008). A técnica de documentação será direta extensiva e os dados serão coletados a partir de planilhas.

Será dividido em quatro etapas que buscarão obter resultados a partir de pesquisas:

A primeira etapa consiste em realizar as pesquisas bibliográficas e exploratórias, acerca dos conceitos de Aprendizagem de Máquina. Os estudos serão focados na pesquisa exploratória com objetivo de obter domínio sobre o que vai ser utilizado na mineração de dados.

Na segunda etapa, será aplicado o que foi pesquisado, obtendo os primeiros resultados a partir de pesquisas descritivas, que tem como objetivo o detalhamento das características de determinada população ou fenômeno ou estabelecimento de relações entre variáveis (GIL, 2008). Nela será feita a análise dos dados, que se dará através de apresentação de gráficos.

Na terceira etapa o foco voltará às pesquisas exploratórias, em que será feito um estudo mais detalhado sobre aprendizado de máquina, para tornar possível a exploração a fundo os dados resultantes da segunda etapa.

Na quarta e última etapa serão utilizadas as pesquisas do tipo quantitativa, que são indicadas para responder a questionamentos que passam por conhecer o grau e a abrangência de determinados traços em uma população (PEREIRA; ORTIGÃO, 2016). Construindo o modelo de AM, documentando os resultados e concluindo o trabalho.

## 1.5 ORGANIZAÇÃO DO TRABALHO

Após esse capítulo introdutório, o conteúdo deste trabalho organiza-se da seguinte forma:

- Capítulo 2 – EXAME NACIONAL DE AVALIAÇÃO DO DESEMPENHO EM INSTITUIÇÕES DO ENSINO SUPERIOR: apresentará o funcionamento e as regras aplicadas no exame, que serão importantes para a obtenção e análise de seus dados;
- Capítulo 3 – PRINCÍPIOS E TÉCNICAS DA APRENDIZAGEM DA MÁQUINA: introduzirá o conteúdo teórico acerca de aprendizagem de máquina, aprofundando a metodologia KDD e apresentando alguns algoritmos;
- Capítulo 4 – UMA ANÁLISE PREDITIVA DO DESEMPENHO NO ENADE: explicará o processo aplicado nos dados, os principais fatores que serão utilizados na análise e criação do modelo, realizando testes para verificar o desempenho do modelo gerado;
- Capítulo 5 – CONCLUSÃO: finaliza o trabalho apresentando os resultados obtidos, formas de utilizar o modelo para entender o desempenho dos alunos, discutindo os possíveis trabalhos futuros.

# 2 EXAME NACIONAL DE AVALIAÇÃO DE DESEMPENHO DE INSTITUIÇÕES DE ENSINO SUPERIOR

O ENADE, de acordo com o INEP (2015) é um dos pilares da avaliação do Sistema Nacional de Avaliação da Educação Superior (SINAES), este é composto pelos processos de Avaliação de Cursos de Graduação e de Avaliação Institucional. A avaliação ocorre anualmente dividindo os cursos das instituições em três grupos. No ano de 2015 foram aplicadas para as áreas de Ciências Sociais Aplicadas, Ciências Humanas e áreas afins e para os cursos que se encaixarem nos segmentos tecnológicos que envolvem Gestão e Negócios, Apoio Escolar, Hospitalidade e Lazer, Produção Cultural e Design.

Cada IES deve ser categorizada pelos seus coordenadores junto aos Procuradores Educacionais Institucionais (PI). Eles são os responsáveis por inscrever, acompanhar e principalmente orientar o corpo discente durante o processo, explicando suas regras e importância.

## 2.1 REGRAS DO ENADE

A partir dos resultados que os estudantes obtiveram, é feita uma série de cálculos para a obtenção do Conceito ENADE. Com a obtenção do desempenho médio dos concluintes nos exames de Formação Geral (FG) e no Componente Específico (CE), é calculada a média nacional da área de avaliação. Da média nacional é calculado seu desvio padrão, o afastamento padronizado e com eles, finalmente se calcula a NCc, abordada na Tabela 1.

O Conceito ENADE não é dado apenas pela avaliação da prova de desempenho. Em paralelo ocorrem avaliações da infraestrutura, do corpo docente, e também dos dados obtidos do Questionário do Estudante. As características principais das provas realizadas pelos alunos serão descritas adiante.

### 2.1.1 Questionário do Estudante

O primeiro passo para o estudante realizar a prova, é responder por meio do *site* do ENADE o Questionário do Estudante e caso não seja respondido o estudante corre risco de ficar irregular com o ENADE, ficando impossibilitado de obter o certificado de conclusão de curso. No questionário são abordadas características econômicas e sociais a respeito estudante, sua relação com o curso, o recebimento de auxílios governamentais, além de uma série de perguntas avaliando a estrutura dada pela IES.

### 2.1.2 Prova de Desempenho

Cerca de trinta dias após a abertura do questionário *online* os estudantes realizam a prova. Nela é avaliado o conhecimento geral do estudante contendo 2 questões discursivas e 8 questões objetivas, além do seu conhecimento específico à sua respectiva área, com um total de 29 questões, sendo 3 discursivas e 26 objetivas. Ao final da prova, o participante deve responder a um questionário de percepção do exame, que visa levantar sua opinião enquanto à qualidade e adequação deste. O resultado desta, junto ao Questionário Socioeconômico devem ser divulgados para que as instituições possam conferir o cálculo do conceito ENADE.

## 2.2 MICRODADOS DO ENADE

Seguindo a Lei de Acesso à Informação, o INEP divulga anualmente informações sobre as provas aplicadas com intuito de garantir a transparência do processo. Os microdados tratam de informações sobre os participantes, os cursos e as IES, tornando possível a análise dos dados que mais influenciam no desempenho dos alunos. São omitidos os dados pessoais que identificariam os estudantes, porém o fato não impossibilita a avaliação. Os arquivos são disponibilizados no formato **CSV**[2], junto ao questionário do estudante e ao dicionário de variáveis (Anexo A), que permite abstrair o que foi disponibilizado.

Na base de dados, que pode ser encontrada no site do INEP, para cada registro, são descritas informações de identificação de sua IES, algumas informações básicas do estudante, como idade e sexo, dados sobre término do ensino médio, início da graduação, turno que atende, indicadores de deficiências, o gabarito marcado pelo estudante, notas parciais por questões discursivas e objetivas, nota geral do estudante, tipo de presença na prova e no questionário, além das respostas dadas no Questionário do Estudante que se encontra no Anexo B.

No estado em que são disponibilizados, a extração do conhecimento acerca dessa base de dados se torna inviável. Este motivo é justificado pelo fato da necessidade da aplicação das técnicas abordadas no KDD discutidas na próxima seção.

2 CSV: do inglês, valores separados por vírgulas. Arquivo com um conjunto de dados.

# 3 PRINCÍPIOS E TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

O conceito de aprendizagem de máquina foi se estabelecendo aos poucos entre os anos oitenta e noventa do século passado, diante da necessidade de se analisar informações em larga escala devido ao crescimento e popularização da internet. Esta área evoluiu mediante o crescimento das tecnologias relacionadas à Inteligência Artificial, com foco de reconhecimento de padrões e aprendizagem computacional.

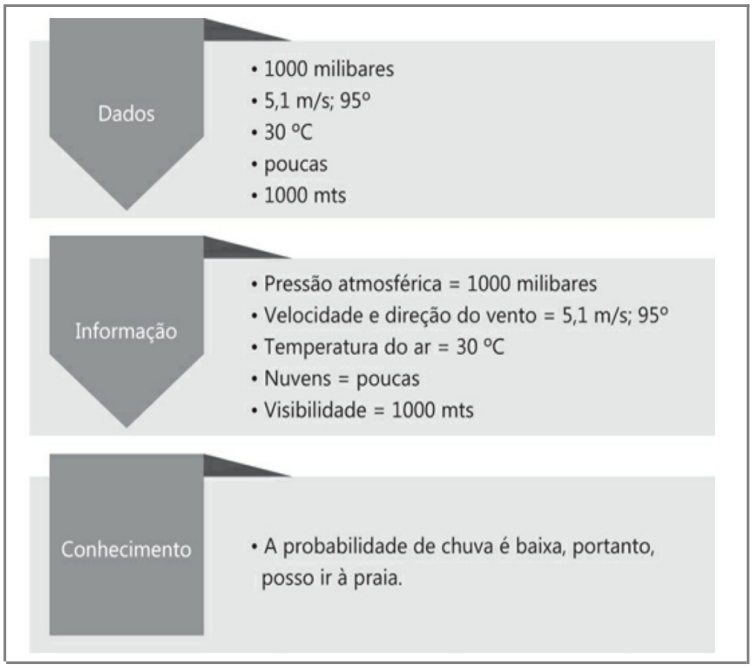
Ela pode ser aplicada nas mais diversas áreas da computação, como no desenvolvimento de robôs, na construção de agentes inteligentes, seja para comunicação, jogos, fornecimento de informações, entre outros; em sistemas de recomendações para redes sociais ou empresas de vendas. Também é utilizada na área de negócios para decisões, por exemplo, na liberação de crédito para clientes, ou ainda na análise de compra e venda de ações empresariais. O limite da AM se dá onde não tem dados, ou onde não há dados suficientes para construção de seus modelos.

A construção de modelos de aprendizagem de máquina, conforme explicam Castro e Ferrari (2016) se inicia com a obtenção de dados. Posteriormente, os dados em estudo passam por um processo de diagnóstico para que sejam identificados conforme indicam as nomenclaturas descritas abaixo:

- Não estruturados - quando não existe uma organização ou quando não se pode separar suas características, este caso ocorre caso o objeto de estudo sejam imagens, textos, áudio ou vídeo;
- Semiestruturados - sua sintaxe é entendível, mas alcançar a informação torna-se difícil para os algoritmos, como exemplo temos análise de códigos, ou de estruturas com *tags ou* linguagens de marcação;
- Estruturados - são dados que contém várias amostras, e para cada amostra, suas características, encaixando num modelo de banco de dados.

Os modelos de aprendizagem de máquina são altamente dependentes da organização dos dados em estudo, para que posteriormente sejam minerados. Esta mineração consiste em um conjunto de etapas, que serão abordadas ainda nesta seção. Quando se trabalha com dados, informação e conhecimento é importante saber suas diferenças. Unindo dados à sua descrição, denotam-se informações e a partir delas, junto à sua análise, se gera conhecimento, como é ilustrado na Figura 1.

Figura 1 - Regressão logística na classificação da espécie de flores




Fonte: Castro e Ferrari (2016, p. 12)

Para a construção desse tipo de modelo torna-se necessário saber qual tipo de aprendizagem de máquina será utilizado. Os tipos de aprendizagem de máquina devem ser decididos após analisar o modo em que os dados se apresentam, esses e alguns outros conceitos do tema serão apresentados a seguir.

## 3.1 CONCEITOS DE APRENDIZAGEM SUPERVISIONADA E NÃO SUPERVISIONADA

Para que se chegue em uma resposta, gerada pela obtenção de conhecimento a partir de dados, é necessário reconhecer um problema, e em qual categoria de aprendizagem ele deve se encaixar. A aprendizagem supervisionada é aquela em que se tem os rótulos que os dados devem receber e, com o modelo, será possível segmentar os dados em cada um deles. Uma abordagem citada pelo livro *Introduction to Machine Learning with Python* (MÜLLER; GUIDO, 2016), que trata da categorização de espécies de rosas. Com base nos dados acerca do tamanho de suas pétalas e sépalas. Desde a introdução do problema, são apresentadas as espécies que devem ser encontradas, configurando um problema de aprendizagem supervisionada, tipo de abordagem utilizada quando é conhecida a característica-alvo.

Quando se quer obter novas representações dos dados, ou seja, quando se busca identificar caraterísticas ou padrões desconhecidos, deve ser utilizado o aprendizado do tipo não supervisionado.

É importante saber que para agrupar dados, é possível utilizar os dois tipos de aprendizagem. Se os tipos ou rótulos em que os dados devem se encaixar são dados por intervenção humana, empregam-se algoritmos de categorização, pela aprendizagem supervisionada, e no caso contrário é utilizada a aprendizagem não supervisionada, utilizando algoritmos de *clustering*. Estes e outros tipos de algoritmos serão melhor abordados em seguida.

## 3.2 ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Obtendo uma situação para aplicação de AM, é preciso entender os conceitos de classificação, regressão e *clustering*. “Classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados” (AMO, 2004, p. 4). *Clustering* se diferencia de classificação por dois pontos principais, primeiramente por ser alcançado pelo aprendizado não supervisionado, ou seja, por não se ter conhecimento das classes ou categorias por que os dados serão divididos. Já a regressão é dada pela estimação de valores, buscando dar um valor numérico através de predição.

### 3.2.1 Algoritmo de Árvore de Decisão

A árvore de decisão é utilizada em tarefas de classificação ou regressão, e pode ser construída a partir de diversos algoritmos, sendo o mais atual o C4.5[3], e mostra os possíveis passos que a informação passa para chegar em cada resultado. A árvore torna fácil o entendimento e a visualização do modelo gerado. Nela é possível utilizar e predizer tanto dados categóricos como numéricos. É possível obter dois tipos de representações para a árvore de decisão, a representação em forma algorítmica, ou representação por árvore conforme é ilustrado na Tabela 2 e na Figura 2.

Tabela 2 - Regras condicionais obtidas a partir da árvore de decisão

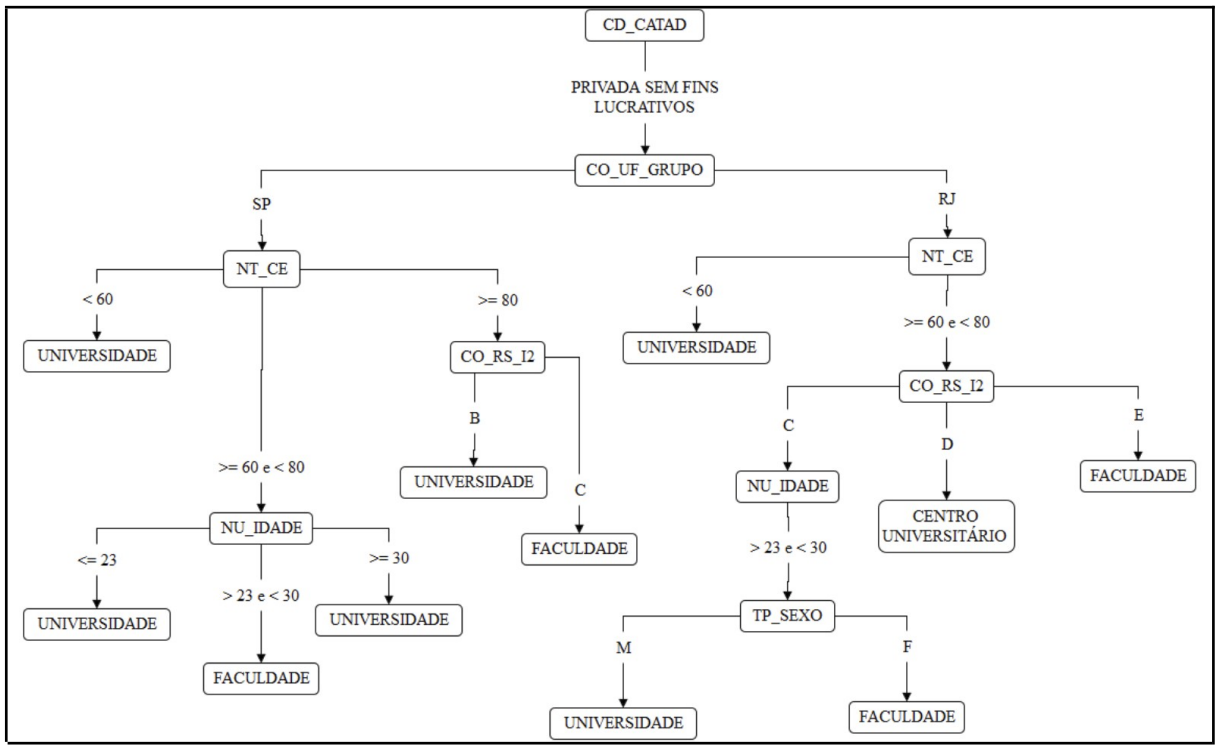
| |
|---|
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = D, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = E, **ENTÃO** NOTA >50 |

3 C4.5 - Algoritmo de classificação utilizado para gerar uma árvore de decisão

| |
|---|
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = B **E** co_rs_21=B **E** co_rs_38=B, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = B **E** co_rs_21=B **E** co_rs_38=C, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = B **E** co_rs_21=C **E** co_rs_38=D, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = C **E** co_rs_46=A, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = C **E** co_rs_46=B, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = C **E** co_rs_46=D **E** co_rs_s8=C, **ENTÃO** NOTA >50 |
| **SE** co_rs_s9=B **E** co_rs_s17=B **E** co_rs_s20 = C **E** co_rs_46=D **E** co_rs_s8=D, **ENTÃO** NOTA >50 |

Fonte: Nogueira e Tsunoda (2015)

Figura 2 - Árvore de decisão referente às instituições privadas sem fins lucrativos




Fonte: Cretton e Gomes (2016)

A Tabela 2 descreve o conjunto de regras algorítmicas de uma árvore de decisão, apresentando cada conjunto de entradas, e suas possíveis saídas. Já a Figura 2, mostra uma outra árvore em sua forma comum, sendo cada nó uma característica e os dados de entradas suas ligações, chegando a um nó final como saída ou resposta à um problema.

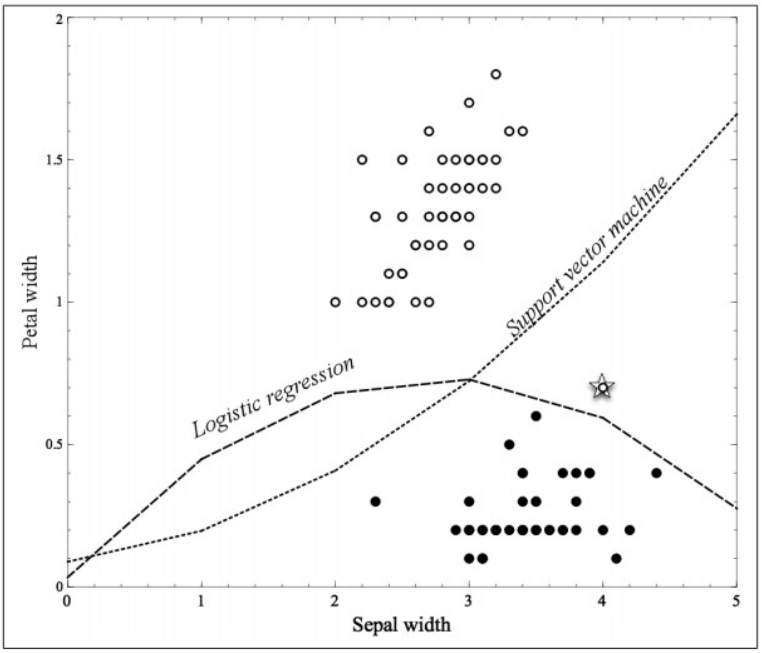
### 3.2.2 Algoritmo de Regressão Logística

Os algoritmos de Regressão Logística realizam a correlação de duas variáveis para a geração de um modelo de aprendizagem. O resultado final da regressão é uma equação denominada

equação de regressão, e a linha gerada pela equação é denominada linha de regressão. Gerando um gráfico com amostras junto à linha de regressão, será possível identificar que a linha toma a função de separar ou classificar os objetos.

O Gráfico 1 trata de um problema já abordado, que visa separar em categorias as rosas. Nela são exibidas o tamanho das pétalas e das sépalas, demonstrando a categorização e a facilidade de visualização dada pelo algoritmo.

Gráfico 1 - Regressão logística na classificação da espécie de flores




Fonte: Provost, Fawcett (2013, p.106)

A linha gerada por este fica no ponto médio da distância de cada variável, mostrando sua precisão na geração de um modelo preditivo para o problema em questão.

### 3.2.3 Algoritmo *Naive Bayes*

A classificação no *Naive Bayes* é dada pela probabilidade de um elemento pertencer a uma classe. O algoritmo utiliza do Teorema de *Bayes;* Ele recebe o nome de *naive* (ingênuo) pelo fato de considerar as características de forma independentes, de modo que a relação entre elas não gere impacto na predição. O teorema utiliza de uma hipótese *H*, para classificar um dado *x*, com intuito de calcular a probabilidade de ela acontecer: *P(H|x)* (CASTRO; FERRARI, 2016).

O modelo do algoritmo é representado por uma matriz composta pela probabilidade de cada característica representar sua classe. A partir de uma nova entrada no modelo, a matriz pode ser

utilizada para classificá-la a custo fixo, apenas multiplicando a probabilidade de suas características e chegando a um padrão de uma de suas classes.

## 3.3 APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS

Os algoritmos citados necessitam de entradas concisas, para que não ocorram erros na geração de aprendizagem a partir deles. Para chegar a tal ponto, os dados devem passar por um processo de entendimento, limpeza ou enriquecimento, em que se busca avaliar o problema e obter informação suficiente para gerar um modelo de aprendizagem.

A Mineração de Dados contém um conjunto de tarefas e técnicas que podem ser utilizadas para cada tipo de problema e também para cada tipo de base, cabe ao cientista de dados elencar quais encaixam na sua aplicação. Algumas das tarefas são essenciais, anteriores ao processo de mineração, na qual envolvem:

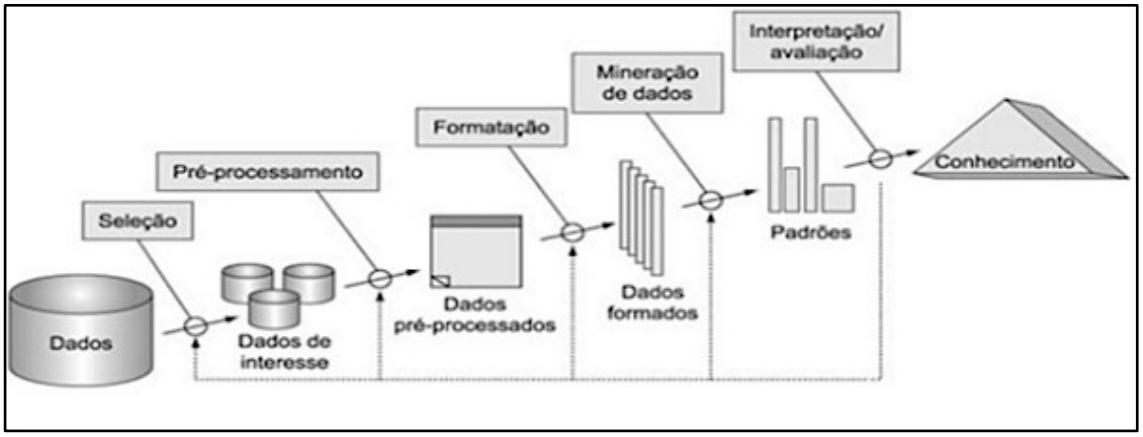
- Buscar os dados que serão utilizados;
- Entender como eles podem ajudar a resolver um problema;
- Inseri-los em uma ferramenta que facilite seu acesso e manipulação (como a partir de consultas SQL[4]);
- Reconhecer e filtrar os ruídos, ou seja, remover o que não será utilizado ou o que pode induzir o modelo a erros, seja por conta de valores nulos, inconsistentes ou que fogem da regra da aplicação. Os dados removidos nesta etapa são denominados *outliers*[5];
- Analisar a integração com outras bases ou até validar a possibilidade de extrair mais características que podem ser importantes para a formação do modelo.

Para a realização de tais tarefas pode ser empregada a metodologia KDD que tem como objetivo, como consta em seu nome: obter conhecimento a partir de bases dados. Amo (2004) apresenta o KDD como o conjunto das seguintes técnicas de mineração de dados, com base na figura 3:

4 *Structured Query Language*: Linguagem para bancos de dados

5 Dados que não se encaixam em uma amostra por inconsistências

Figura 3 - Fases do KDD



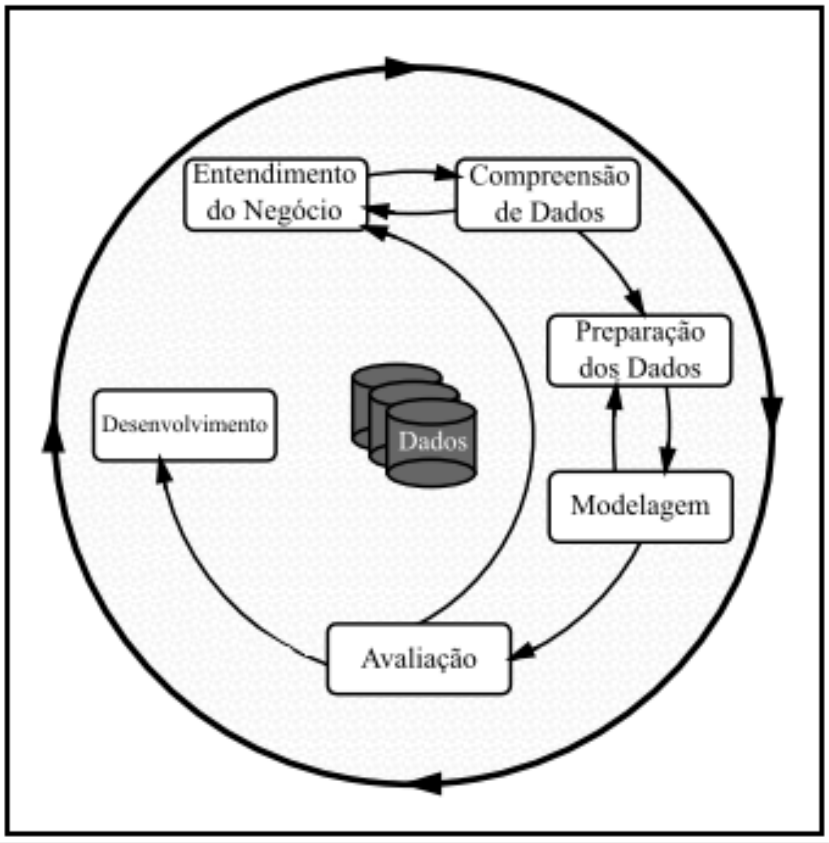

Fonte: Teofilo (2015)[6]

- Análise de Regras de Associação - um caso clássico dessa análise é a de compra de produtos, ao relacionar produtos distintos induzindo o cliente a adquirir ambos;
- Análise de Padrões Sequenciais - se analisa o tempo em que ações se repetem com mais frequência, examinando quando deve ocorrer uma próxima ação;
- Classificação e Predição - nas ocasiões onde os padrões são pré-determinados, emprega-se classificação, já quando se quer classificar quanto a um valor, predição;
- Análise de *Clusters* - esta análise ocorre quando se quer estudar as características que contém relações em dados não classificados, podendo assim formar grupos;
- Análise de *Outliers* - quando é desenvolvido um modelo, se dá atenção à um padrão, e o que ele pode gerar ou classificar. Esta etapa visa investigar as exceções, os piores casos ou as possíveis induções a erros.

O KDD é derivado de uma outra metodologia que visa regulamentar a aplicação de aprendizado de máquina, que tem sua importância e será útil para entender a regra de negócio e a aplicação prática, denominada *Cross Industry Standard Process for Data Mining* (CRISP-DM). A Figura 4 ilustra o funcionamento do *framework*.

6 Disponível em <https://danielteofilo.wordpress.com/2015/02/16/kdd-knowlegde-discovery-in-database/>. Acesso em: 16 out. 2017.

Figura 4 - Fases do CRISP-DM




Fonte: Adaptado de WIRTH e HIPP (2000)

As indicações dadas pelas metodologias citadas mostram a importância dos passos que antecedem a obtenção de conhecimento. Ainda na Figura 4 é notável que as etapas são passíveis de iteração, além de que não é possível prosseguir sem compreender os dados e entender o problema. A imagem trata também de uma etapa do trabalho, a avaliação de modelos, que deve se feita a partir de testes estatísticos que serão abordados em seguida

## 3.4 AVALIAÇÃO DE MODELOS PREDITIVOS

Durante o processo de construção de um modelo, a maneira mais usual para validação de sua eficácia inicia-se a partir da separação dos dados em dois conjuntos sendo um para treinamento, e outro para testes. Caso na validação fossem utilizados os mesmos dados utilizados no treinamento "não seria feita nenhuma avaliação de quão bem o modelo generalizaria casos que não foram vistos" (PROVOST; FAWCETT, 2013 p.113, tradução nossa).

Como os algoritmos necessitam de um bom número de dados para interpretação e avaliação de suas características, na maior parte dos casos, o conjunto de treinamento deve receber mais dados que o conjunto de testes. Para simular o uso do modelo, é utilizado o conjunto de treinamento, escondendo dele o valor real da variável-alvo, retornando à interpretação tomada (PROVOST; FAWCETT, 2013).

Unindo o resultado da predição de cada instância do conjunto, ao seu resultado real, é possível realizar uma validação que buscará dizer em quantos casos a hipótese gerada pelo modelo foi confirmada pelo seu real resultado, dando-se então o desempenho ou acurácia do modelo. A alta acurácia, em forma de porcentagem, serve de indicativo de um bom resultado do algoritmo de aprendizagem, existe então alguns testes estatísticos para confirmar tal resultados, como os seguintes.

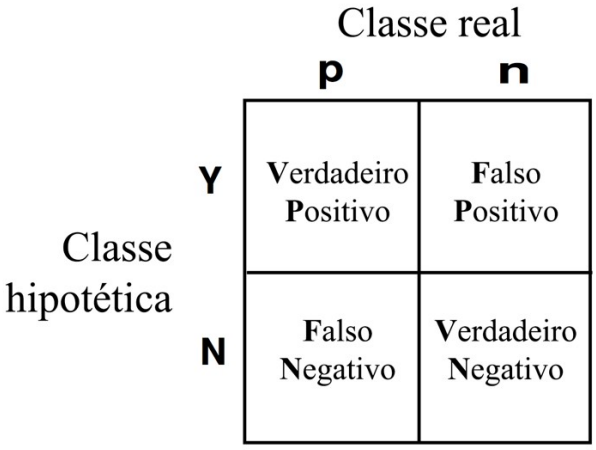
### 3.4.1 Matriz de confusão e Curva ROC

Abordando um problema de classificação binária,

> cada instancia ***I*** é mapeada para um elemento do conjunto {**p, n**} com rótulos para classes positivas e negativas[...]. Para diferenciar entre a classe verdadeira e a predita pelo modelo, utilizamos os rótulos {**Y, N**} para as classes preditas produzidas por um modelo. Dados um classificador e uma instância, existem quatro resultados possíveis. Se a instância é positiva e é classificada como positiva, ela é considerada como verdadeiro-positivo; se é classificado como negativo, é considerada como falso-negativo. Se a instância é negativa e classificada como negativa, ela é considerada verdadeiro-negativa; se é classificada como positiva, é julgada como falso-positiva. (FAWCETT, 2004, p. 2, tradução nossa).

Com as informações providas pelo autor, é possível contar a quantidade de instâncias em cada quadrante, gerando uma matriz de confusão, como observado na figura 5:

Figura 5 - Matriz de Confusão




Fonte: Adaptado de Fawcett (2004)

A partir dos dados da matriz de confusão, é possível calcular várias métricas, entre elas a precisão, acurácia, sensitividade, especificidade além das taxas de verdadeiro-positivos e falso-positivos. Ela recebe esse nome por demonstrar a quantidade de falso-positivos e falso-negativos,

somando-os é dada a taxa de confusão de um classificador. O cálculo de cada métrica é feito pelas fórmulas apresentadas na tabela 2:

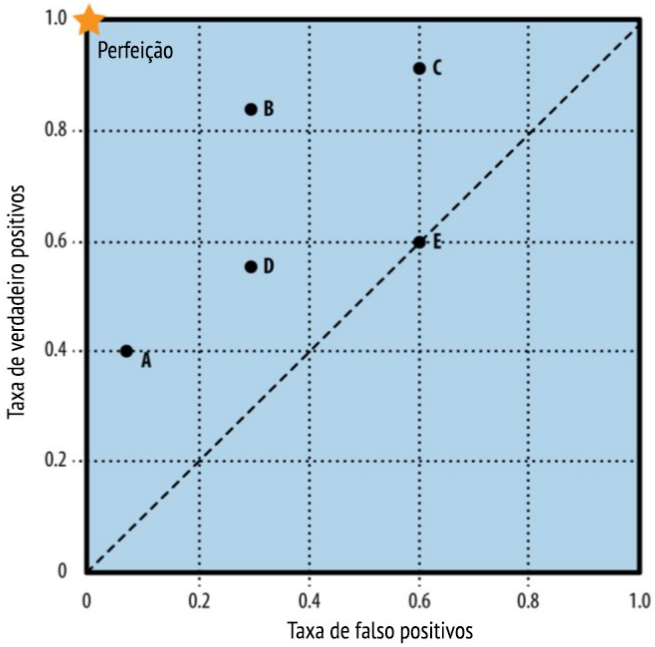
Tabela 3 - Métricas calculadas pelos dados da matriz de confusão

| | |
|---|---|
| $Precisão = \frac{VP}{VP+FP}$ | $Acurácia = \frac{VP+VN}{P+N}$ |
| $Sensibilidade = \frac{VP}{P}$ | $Especificidade = \frac{VN}{FP+VN}$ |

Fonte: próprio autor (2018)

A sensibilidade, como visto, é a proporção de valores realmente positivos que foram preditos como positivos, sendo a especificidade a proporção de negativos em relação aos erros encontrados. A partir dessas duas medidas é possível construir gráficos denominados *Receiver Operating Characteristics* (ROC). O espaço ROC combina a taxa de verdadeiro-positivos com a de falso-positivos, nele é possível observar a troca feita entre benefícios (verdadeiros-positivos) e custos (falso-positivos) (PROVOST; FAWCETT, 2013). Os autores explicam que também é possível utilizar os verdadeiro-negativos e falso-negativos, mas por convenção, utilizam os valores positivos para realizar a análise. O espaço ROC pode ser entendido ao analisar o Gráfico 2:

Gráfico 2 - Espaço ROC de 5 classificadores diferentes




Fonte: Adaptado de Provost e Fawcett (2013, p. 215)

Para interpretar o gráfico, necessita-se interpretar alguns de seus pontos. O ponto (0,0) mostra um classificador que não aponta nenhum caso positivo, e o ponto (1,1) sendo o oposto. O ponto (0,1) representa uma classificação perfeita, onde todos os casos são acertados. A linha

diagonal que une os pontos (0,0) a (1,1) denota os casos de um classificador que tenta adivinhar aleatoriamente suas classes (PROVOST; FAWCETT, 2013).

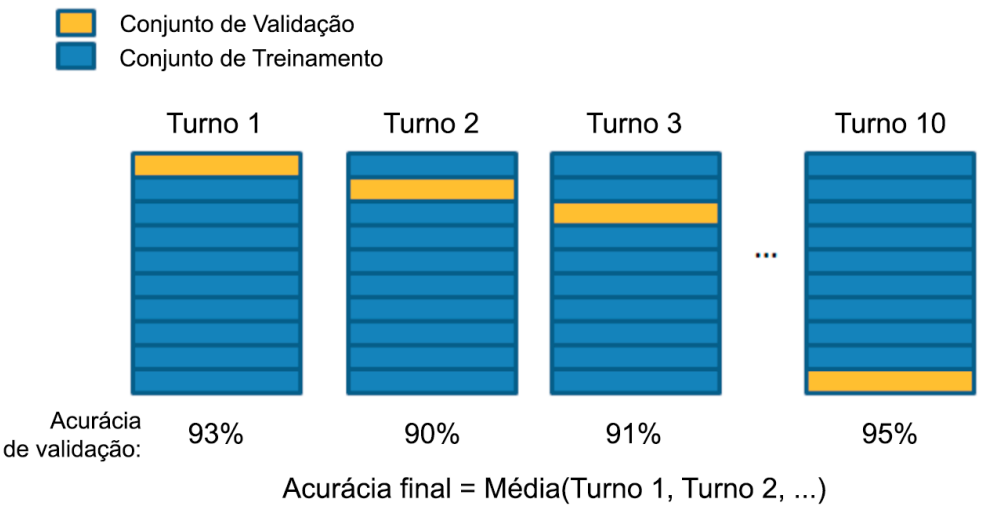
Com estes pontos, podemos interpretar que ao construir um gráfico ROC de um modelo, para que ele seja aceitável, deve-se iniciar partindo da posição (0,0), indo ao encontro do ponto (1,1), fazendo uma curva em direção ao ponto de classificação perfeita, e nunca estando abaixo da diagonal. Se nota que nessa abordagem não são tratados os casos com classe negativa, sendo mais um caso de avaliação em que é possível indicar o bom desempenho de um classificador.

### 3.4.2 Teste *K-Fold*

A abordagem anterior acerca da divisão dos dados em dois conjuntos para calcular a acurácia de um classificador, conhecida por *holdout*, apresenta problemas decorridos da omissão de alguns dos dados em seu treinamento, uma vez que “assumindo que a acurácia de um indutor aumenta à medida que mais instâncias são vistas, o método *holdout* é um estimador pessimista, pois apenas uma parte dos dados é dada ao indutor para treinamento” (KOHAVI, 1995).

Para contornar o problema, pode ser utilizada a validação-cruzada *K-Fold*, em que toda a base de dados é utilizada para treinamento, e também para testes, reduzindo qualquer possível viés gerado por apenas uma validação. Também conhecida por *random subsampling* (subamostragem aleatória) ou *rotation estimation* (estimação por rotação), nela “o método *holdout* é repetido ***k*** vezes, e a acurácia estimada é derivada a partir da média de cada turno. O desvio padrão pode ser estimado a partir do desvio padrão de cada repetição.” (KOHAVI, 1995, p. 2, tradução nossa). A Figura 6 apresenta a sequência de passos para elaborar o teste K-fold.

Figura 6 - Representação visual do K-Fold



Fonte: adaptado de Bronshtein (2017)[7]

Conforme exibido na Figura 6, o método consiste em dividir aleatoriamente os dados em ***k*** conjuntos em tamanhos aproximadamente iguais, e para cada iteração é feito o treinamento do classificador sem k, utilizando-o para realizar a validação, e por fim é calculada a média das acurácias de cada turno (KOHAVI, 1995). Para indicar que o teste obteve sucesso, deve-se avaliar se a variação das iterações foi baixa.

### 3.4.3 Teste de Kolmogorov-Smirnov

O teste de distribuição Kolmogorov-Smirnov (K-S) é "um teste não paramétrico e compara dois conjuntos de dados avaliando se são ou não significativamente diferentes" (ALBARDEIRO et al, 2014, p. 1). Neste se verifica "o grau de concordância entre distribuição de um conjunto de valores (escores observados) e alguma distribuição teórica, ou seja, verificar se os dados seguem a distribuição normal" (SCUDINO, 2008, p. 18).

Apesar da abordagem ser indicada para dados contínuos, Krzanowski e Hand (2009, apud ADEODATO; MELO, 2016, p. 2) afirmam que em "problemas de classificação binária, ela tem sido usada como medida de dissimilaridade para avaliar o poder discriminante do classificador medindo a distância que a sua pontuação produz entre as funções de distribuição acumulada (FDA) das duas classes de dados".

Antecedendo o desenvolvimento da pesquisa, foram buscados trabalhos que utilizam dados educacionais, como os do ENADE, exemplificando as várias possíveis abordagens, a importância deste tipo de estudo, e como se chegou a interpretações utilizando aprendizagem de máquina.

## 3.5 TRABALHOS RELACIONADOS

No artigo publicado por Gotti et al. (2012), foram aplicados algoritmos de Redes Neurais Artificiais (RNA) nos dados da prova do ENADE de 2009. O trabalho busca validar a aplicação de inteligência computacional a partir do desempenho dos estudantes, na avaliação das instituições. Os microdados do exame foram reduzidos em dez características para a criação da rede, a partir de dados de doze IES, com intuito de avaliar se o curso poderá obter o resultado esperado, cruzando informações históricas do exame.

Os autores conseguiram resultados positivos ao construir a rede neural, demonstrando as características que se destacaram na classificação da instituição, e concluíram que "os resultados apresentados com os dados escolhidos junto à analise com redes neurais treinadas com essas

7 Disponível em: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>. Acesso em: 05 abr. 2018

informações garantem uma rápida estimativa das notas para tomadas de decisões rápidas, por gerentes de uma instituição" (GOTTI et. al, 2012, p.1, tradução nossa).

Na pesquisa de Nogueira e Tsunoda (2015) foi analisada a relação da nota bruta com a pesquisa socioeconômica, com a mesma base de dados, porém do ano de 2012. Demonstram que apenas onze das cinquenta e quatro perguntas foram essenciais para verificar que um aluno pode ter um bom resultado nas provas. Para a criação do modelo preditivo, foram utilizados os algoritmos para obtenção de uma árvore de decisão.

Ao final da pesquisa, os autores sugerem "a expansão do período em que a análise foi realizada, considerando uma janela de tempo maior, com o objetivo de estudar como as variáveis se comportam no decorrer dos anos". Embora o estudo proposto seja importante, seria difícil realizar tal expansão apenas com os microdados do INEP, sem levar em consideração que podem haver alterações nos padrões de ensino seja na instituição de ensino médio ou superior.

O estudo de Cretton e Gomes (2016) utilizou-se dos dados do ENADE do ano de 2013 para gerar um modelo preditivo da nota geral do aluno. Este se diferenciou ao focar apenas nos cursos de medicina, mostrando as diversas aplicações possíveis nos dados fornecidos pelo INEP. Foi utilizado o algoritmo J48[8] para a geração de uma árvore de decisão.

Com o processo de KDD, chegaram a sete atributos para aplicar o algoritmo. O pré-processamento dos dados foi de importância extrema para o trabalho. Dentre as características selecionadas as que se destacaram foram a dificuldade que o aluno deu a prova, o estado onde estuda, o tipo de instituição, faixa etária e sexo, se distanciando do padrão que as pesquisas citadas anteriormente tendem a encontrar.

O estudo referido ressalta a importância da aplicação de aprendizagem de máquina, junto ao resultado da pesquisa, para as instituições de ensino, os estudantes participantes do exame, e também para os que buscam adentrar em uma IES. Apresentados a introdução ao trabalho, a revisão do conteúdo que será utilizado e a de pesquisas relacionadas, o trabalho segue ao detalhar as etapas para chegar aos objetivos propostos na seção seguinte.

8 Algoritmo que gera uma árvore de decisão a partir da análise dos dados e criação de regras

# 4 UMA ANÁLISE PREDITIVA DO DESEMPENHO NO ENADE

Revisitando o problema proposto, este trabalho deve analisar, a partir dos dados publicados pelo MEC no site do INEP acerca do ENADE, a possibilidade de criação de um modelo preditivo do desempenho dos alunos das IES com enfoque em seus dados socioeconômicos, com objetivo de orientar as instituições na preparação do exame. Para chegar ao objetivo final, foram utilizados um conjunto de ferramentas apresentadas a seguir.

## 4.1 FERRAMENTAS UTILIZADAS

As ferramentas para análise e visualização de dados, utilização de algoritmos de aprendizagem e realização de testes utilizadas durante o desenvolvimento da pesquisa dispuseram de características em comum, como o fato de serem amplamente empregadas, de se apresentarem como plataformas de código-aberto e pela fácil manipulação e entendimento do conteúdo desenvolvido.

A Tabela 4 mostra a popularidade e o crescimento, baseado nos questionários aplicados aos desenvolvedores que utilizam o site *Stack Overflow.* A *Stack Overflow Annual Developer Survey*[9] vem ocorrendo desde o ano de 2011, questionando o público sobre “tudo, desde suas tecnologias favoritas até pretensões de emprego” (*STACK OVERFLOW*, 2018, tradução nossa):

Tabela 4 - Ranking de popularidade das ferramentas a partir dos dados do *Stack Overflow*

| Ferramenta | Categoria | Posição (2016) | Posição (2017) | Posição (2018) |
|---|---|---|---|---|
| MySQL | Banco de dados utilizado | * | 1 | 1 |
| Python | Linguagem de programação mais desejada | 4 | 1 | 1 |

* Categoria inexistente no período.

Fonte: *Stack Overflow Surveys* (2016 – 2018).

### 4.1.1 MySQL e MySQL Workbench

A escolha de um banco de dados para tratar do problema proposto apresentou poucos requisitos essenciais, inicialmente demandado apenas a importação de dados a partir arquivos CSV. O *MySql* se mostrou um Sistema Gerenciador de Banco de Dados (SGBD) que o melhor atendeu, uma vez que além de facilitar a importação de dados a partir da ferramenta *MySQL Workbench*, ele é um SGBD relacional gratuito, de fácil configuração, *open source*, e com compatibilidade para diversos sistemas operacionais.

9 Disponível em: <https://insights.stackoverflow.com/survey/>. Acesso em: 19 abr. 2018.

O *MySQL Workbench* auxilia na realização de tarefas como a manutenção de vários arquivos ou *scripts* SQL, na visualização de tabelas, conexão e seleção de esquemas de banco dados, realização de backups, criação, exportação e importação de tabelas, etc.

### 4.1.2 Python e a Plataforma Anaconda

A linguagem *Python*, é uma linguagem de programação interpretada, de código aberto, que preza pela legibilidade do código escrito. Dentre suas características, a mais relevante à área de aprendizagem de máquina é a de que "*Python* é a linguagem de programação aberta de Ciência de Dados que cresce mais rapidamente, com um incrível ecossistema *open-source*" (CHAMBERS *et al*., 2016, tradução nossa). A linguagem "contém bibliotecas para importação de dados, visualização, estatística, processamento de linguagem natural, processamento de imagens e mais" (MULLER; GUIDO, 2016, tradução nossa).

A plataforma *Anaconda* mantém a instalação, manutenção e atualização de diversas ferramentas, desde uma distribuição personalizada da linguagem de programação, até diversas ferramentas integradas de desenvolvimento (*IDE*, do inglês). *Jupyter*, a IDE usada permite a criação de um editor visual de blocos de código, por padrão, ou de texto. Cada bloco permite a visualização de sua saída, seja ela em formato de texto, tabelas, imagens ou gráficos. Em um mesmo arquivo, os blocos se apresentam no mesmo contexto, preservando as bibliotecas importadas e o valor das variáveis utilizadas.

A biblioteca *Pandas*, em conjunto com a biblioteca para manutenção de números e listas, *NumPy* permite a visualização e importação dos dados. Assim como uma ferramenta SGBD, o *Pandas* permite a consulta, ordenação, agrupamento e integração das tabelas, ou como denominadas *dataframes*. Já o *NumPy* auxilia no tratamento de memória da estrutura dos dados, e realização cálculos estatísticos e matemáticos. A limitação deste conjunto é por armazenarem os dados em memória, e caso se recorra à um grande número de dados, pode se tornar necessário um tratamento prévio, ou a importação parcial deles. A Figura 7 exibe o funcionamento e a integração das bibliotecas, e da ferramenta de desenvolvimento.

Figura 7 - Exemplo de utilização do Pandas

```
#Carrega dados
dataframe = pipeline.carrega_csv()

#Imprime média da nota geral
print("Média  total= ", np.mean(dataframe['nt_ger']))

#Exibe as cinco primeiras linhas do dataframe
dataframe.head()
```

```
Média  total=  46.02439959652505
```

| | co_grupo | co_ies | co_categad | co_orgacad | co_munic_curso | co_uf_curso |
|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 10002 | 10028 | 5103403 | 51 |
| **1** | 1 | 1 | 10002 | 10028 | 5103403 | 51 |
| **2** | 1 | 1 | 10002 | 10028 | 5103403 | 51 |
| **3** | 1 | 1 | 10002 | 10028 | 5103403 | 51 |
| **4** | 1 | 1 | 10002 | 10028 | 5103403 | 51 |

Fonte: próprio autor (2018)

Outro conjunto de bibliotecas foi utilizado para visualização dos dados, o *Seaborn*, que é desenvolvido com base no *Matplotlib*. Ambos têm como objetivo a criação dos mais diversos tipos de gráficos, sendo o *Seaborn* um facilitador, contendo métodos explícitos para criação e exibição destes. Na Figura 8 é exemplificado o uso das ferramentas.

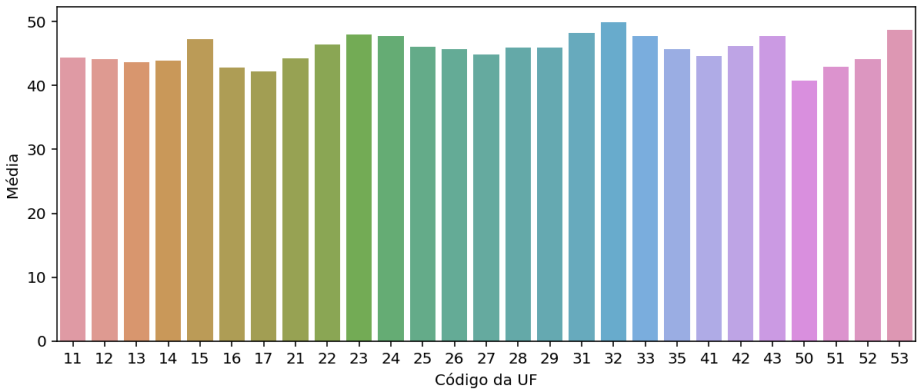
Figura 8 - Criando um gráfico de barra com o Seaborn e Matplotlib

```
#Lista médias por UF
medias_por_uf = dataframe.groupby(by='co_uf_curso').mean()['nt_ger'].sort_values()

#Altera tamanho da figura gerada. O Pyplot é parte do Matplotlib
fig = pyplot.figure()
fig.set_figwidth(10)

#Gera gráfico de barra e atribui legendas
grafico = seaborn.barplot(x=medias_por_uf.index, y=medias_por_uf.values)
grafico.set_xlabel("Código da UF")
grafico.set_ylabel("Média")
```

```
Text(0,0.5,'Média')
```




Fonte: próprio autor (2018)

Por fim, para a utilização de algoritmos de aprendizagem e criação de modelos, foi utilizada a biblioteca *Scikit-learn*, que disponibiliza vários algoritmos de aprendizagem, além de mecanismos para a realização de treinamento, testes e validação dos modelos, conforme exibido na Figura 9.

Figura 9 - Aplicação do Scikit-learn

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
# Atribui colunas para treinamento
X_columns = ['co_uf_curso', 'nu_idade', 'escol_paterna', 'escol_materna',
         'ren_total', 'uf_ens_medio', 'cat_esc_ens_medio', 'anos_egressao',
         'anos_em_curso', 'curso_bom_desempenho_participante_2012', 'turno']

# Separa característica-alvo(y) do restante dos dados(X)
X = dataframe[X_columns]
y = dataframe[pipeline.target_column]

# Cria classificador de árvore de decisão
classificador = DecisionTreeClassifier()

# Separa X e y em conjuntos para testes e para treinamento
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)

# Realiza treinamento dos dados
classificador.fit(X_train, y_train)

# Prediz as características do conjunto de testes
lista_predicoes = classificador.predict(X_test)

# Valida e imprime a acurácia do modelo ao comparar a lista de predicoes
# e o conjunto 'y_test'
print("Acurácia: ", accuracy_score(y_test, lista_predicoes)*100, "%")
```

```
Acurácia:  68.97623875630427 %
```

Fonte: próprio autor (2018)

A figura apresenta a criação de um modelo a partir do algoritmo Y, testando sua acurácia. A seguir será apresentado o desenvolvimento do trabalho e como foram utilizadas as ferramentas descritas.

## 4.1 ENTENDIMENTO DOS DADOS E LIMPEZA DA BASE

Como foi ressaltado anteriormente, o estudo está centrado nos dados socioeconômicos. A análise deve focar nas variáveis ou características que mais representam o problema. A metodologia KDD se dará ao compreender, avaliar o impacto no desempenho e por fim selecionar tais características para que se obtenha conhecimento a partir deles.

Os microdados são passíveis de visualização a partir de seu formato original, e podem ser entendidos com auxílio do dicionário de variáveis, que é disponibilizado em conjunto. Porém, com intuito de facilitar a visualização, manipulação e consulta dos dados, necessitou-se a adaptação e

importação do arquivo CSV para um SGBD. Para esta situação foi escolhido a utilização do *MySQL*, utilizado com auxílio da ferramenta *MySQL Workbench*[10].

A ferramenta dispõe de meios para a importação automática de arquivos como o cedido pelo MEC, porém foi notável a inconsistência da base de dados gerada a partir deles. Ao realizar a importação de forma automática, não só a quantidade de amostras se mostrava diferentes, como também ocorriam erros ou dados inconsistentes, em que não se demonstrava o mesmo no arquivo original. Além disso, a importação automática chegou a demandar muito desempenho e também de um longo tempo para finalização da tarefa.

Tal inconsistência foi gerada por uma particularidade da ferramenta de administração do SGBD selecionado. Ele não trata espaços vazios ou não preenchidos como nulos, então foi necessária a busca e substituição dos espaços vazios pela sequência de caracteres "**\N**". Como está se tratando de um arquivo CSV com um de grande tamanho, foi utilizada a ferramenta de linha de comando *SED*, acrônimo do inglês de editor de fluxo (*Stream Editor*), que tem como função particionar a busca e edição de texto pelos núcleos do processador. Uma de suas características é a utilização de linguagens regulares para busca de termos, que facilitou a resolução do problema.

Em seguida, para contornar o problema do tempo de importação, desta vez de forma manual, foi utilizada uma alternativa para importação dos dados por SQL, utilizando a instrução "*LOAD DATA INFILE*". A tentativa de importação automatizada não foi inutilizada, uma vez que a partir do arquivo fonte, a ferramenta facilitou também na criação da tabela. Neste passo foi reconhecido mais um erro durante a importação, que decorreu também das características do SGBD: os dados ainda não estavam na mesma quantidade do originário. Esse foi solucionado ao corrigir a sintaxe utilizada para o reconhecimento de quebras de linhas.

A partir deste ponto, os dados se encontram devidamente importados. O primeiro passo da análise dos dados, foi buscar entender o motivo de alguns valores que se encontrarem nulos. Observando o problema, notou-se que existiam colunas em que definiam a presença do aluno em cada passo importante do exame. Por não ser possível avaliar o desempenho do estudante que não participou da prova, os que se encontrarem nessa situação não devem ser selecionados nas etapas posteriores.

Iniciando a exploração dos dados, que será feita com ajuda do SQL, foi montada uma estrutura do que deverá ser descartado ou do que não se aplicará ao problema em estudo,

10 Ferramenta visual de arquitetura e desenvolvimento de bancos de dados

destacando-se então os *outliers*. Os *scripts* utilizados nesta seção estão documentados no Apêndice A.

Para que não sejam processados dados que ainda estão em análise, a exclusão e limpeza dos dados será consolidada apenas no final do processo de enriquecimento dos dados, ao criar a versão final da base a partir dos *scripts* citados, para que possa ser utilizada por outras ferramentas. Tendo as amostras que serão utilizadas e as que serão descartadas, a próxima etapa tende a explorar quais características daquelas se relacionam diretamente com o problema.

## 4.2 ENRIQUECIMENTO DA BASE DO ENADE

A base de dados no estado que se encontra antes do início deste segmento continua em seu estado original, após a importação. O passo seguinte deve retornar ao dicionário de variáveis, com intuito de não apenas entender os dados de uma forma geral, mas sim aplicado ao problema. São destacadas as características que se identificam com o problema, aqui denominadas de variáveis representativas, dando início a etapa de seleção do KDD.

Anterior à seleção, se deu uma parte importante que visou simplificar, melhorar, ou como utilizado anteriormente, enriquecer os dados. Os microdados disponibilizaram informações acerca do ano em que um aluno terminou o ensino médio e em que iniciou o ensino superior. A partir de tais informações, foi possível calcular o tempo de egressão, ou seja, o tempo em que o aluno levou para adentrar em uma IES. Foi calculado também o tempo relativo em curso de cada aluno, a partir do ano de ingresso ou matrícula, subtraindo-o do ano de realização da prova.

Enfim, se torna possível a pré-seleção dos dados. Esta etapa buscou selecionar previamente as características que podem ter maior representatividade direta no desempenho do estudante, então deve ser confirmada a sua representatividade mais adiante, ao aplicá-las em um algoritmo de aprendizagem. Alguma das características apresentadas nos microdados foram descartadas, seja por apresentar informações repetitivas, como cidade, estado e região; por não apresentar informações úteis para o tipo de análise proposta, como o gabarito assinalado pelo estudante; características que se mostraram desnecessárias como o tipo de presença e a situação do participante, uma vez que para realizar a prova e obter uma nota, o estudante deve se apresentar em todas as etapas e ter situação regular, dentre outras razões. Tais variáveis sem relação direta ao problema, ou que não caracterizam o objeto de análise, poderiam gerar inconsistências no modelo de aprendizagem, inclusive induzi-lo a erros.

A base de microdados original apresenta 146 colunas e, inicialmente 11 delas foram pré-selecionadas como variáveis representativas, que envolvem as seguintes características do

participante: o grau de escolaridade paterna, o grau de escolaridade materna, a renda total familiar, a situação de trabalho do estudante, o recebimento de algum auxílio de permanência durante a graduação, a categoria da escola em que cursou o ensino médio, a existência de familiares concluintes de ensino superior, a quantidade de horas de dedicação aos estudos, o tempo de egressão e o tempo em curso. Estas variáveis são apresentadas e detalhadas no Apêndice B. A pré-seleção foi realizada para que o problema seja analisado de forma descritiva, apresentando gráficos e estatísticas acerca do desempenho dos alunos, e ela foi feita baseada nas pesquisas relacionadas, assim como na de Souza et. al (2017).

Finalizando o enriquecimento dos dados, foi necessária a transformação dos dados que se encontravam no formato *CHAR* para o formato *INTEGER*, seguindo padrões adotados pelos algoritmos que serão utilizados posteriormente. Terminada esta fase, serão exploradas informações estatísticas acerca do estudo dos dados.

## 4.3 ANÁLISE DESCRITIVA DOS DADOS

A partir do que foi obtido nos processos anteriores, torna-se possível realizar uma análise breve dos dados, anterior a extração de conhecimento. Esta análise tem como objetivo aprofundar o entendimento no problema, reconhecendo a quantidade de dados e visualizando o resultado das melhorias proporcionadas pelo processo de seleção e transformação.

Os microdados do ENADE contam com 549.487 registros e com auxílio do SQL foi possível realizar a análise dos mesmos. Anterior à criação e seleção dos dados, foi detectado que 81,36% dos alunos estiveram presentes no exame, e que 54,1% estiverem presentes em todas etapas e não deixaram quesitos sem reposta.

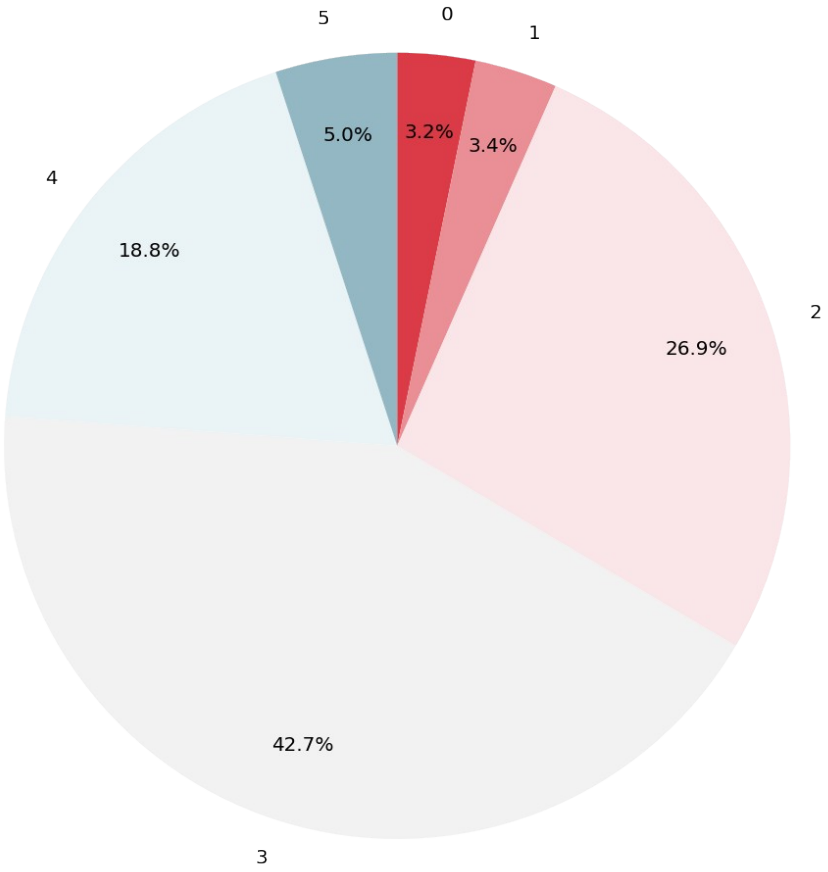
Foi observado que alguns dados acerca das provas que podem ser importantes para as próximas etapas por tornar possível a generalização e a obtenção de conhecimento do que é observado nos dados. Verificou-se que os cursos que tiveram mais alunos foram os de Administração (27,66% do total de alunos), Direito (22,73%) e Ciências Contábeis (11,92%). O estado que mais matriculou alunos foi São Paulo (26%), seguido do Paraná (11,8%) e Minas Gerais (9,22%).

A partir dos dados obtidos do site do INEP contendo o Conceito ENADE de cada instituição, foi possível observar que os alunos de Instituições Federais têm sua média de conceito (aproximadamente 3,75) superior à dos de Instituições Estaduais (2,98), ficando com a menor média as Instituições Privativas (2.76). As áreas que obtiveram melhores resultados, a partir do cálculo da

média foram as de Tecnologia em Design de Interiores, Tecnologia em Design de Moda e de Relações Internacionais, respectivamente.

Como o grupo de IES que este trabalho visa orientar é o que obteve o Conceito ENADE abaixo de três, foi notado que esse grupo é composto por 33,44% dos cursos. O Gráfico 3 agrupa a quantidade de cursos pelo Conceito ENADE. Os cursos que não obtiveram nota ou não foram avaliados são representados com nota 0.

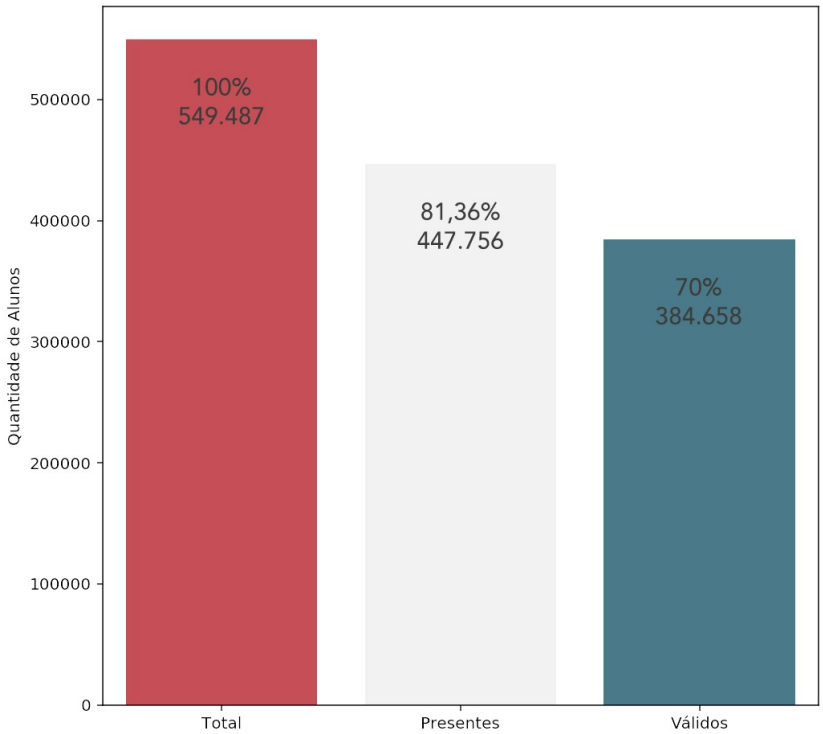
Gráfico 3 - Quantidade de Instituições Agrupadas por Conceito ENADE




Fonte: próprio autor (2018).

Com os dados das etapas anteriores foram considerados 71,6% dos alunos como aptos a serem utilizados nas próximas etapas, identificando os que fizeram ambas as provas e responderam ao mínimo todas as questões objetivas, apresentando ainda a possibilidade de não responder algumas das questões discursivas. Existiam ainda alguns dados que apresentaram valores nulos, diminuindo o número de dados válidos para 70%, como é exibido no Gráfico 4:

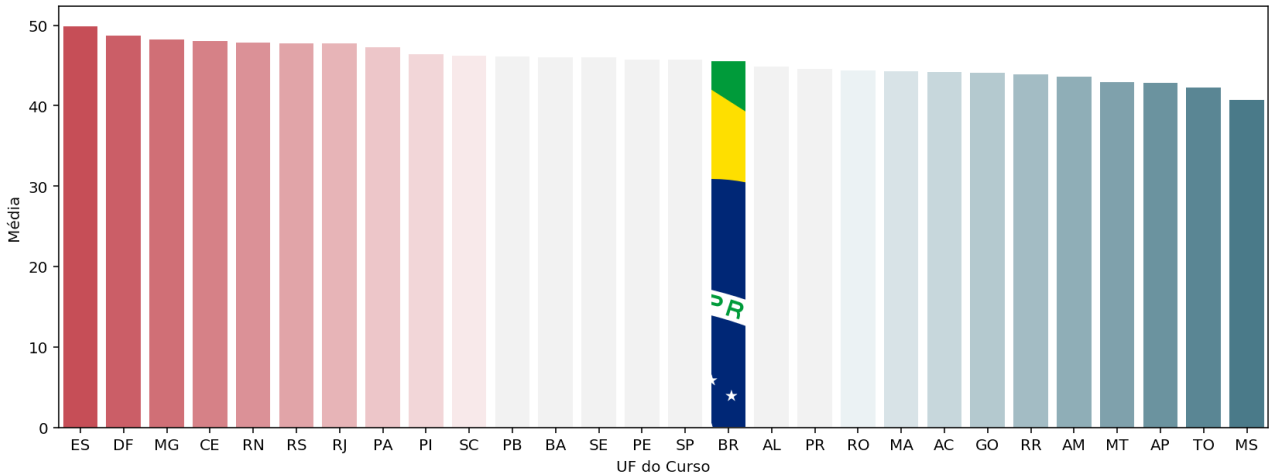
Gráfico 4 - Apresentação Gráfica da Quantidade de Dados




Fonte: próprio autor (2018)

No Gráfico 4 se observa no eixo vertical a quantidade de alunos, enquanto no horizontal, os grupos de observação. Os *outliers*, obtidos a partir da quantidade total de alunos subtraída da quantidade alunos que foram considerados válidos à pesquisa fizeram parte então de 30% da base em sua disposição original, e a partir deste ponto, serão estudadas as características que foram pré-selecionadas, analisando sua relação com o desempenho dos estudantes, utilizando apenas os dados considerados válidos. A avaliação da correlação das características com o desempenho, será iniciado pela média da nota geral dos alunos agrupados pelo estado, como é visto no Gráfico 5:

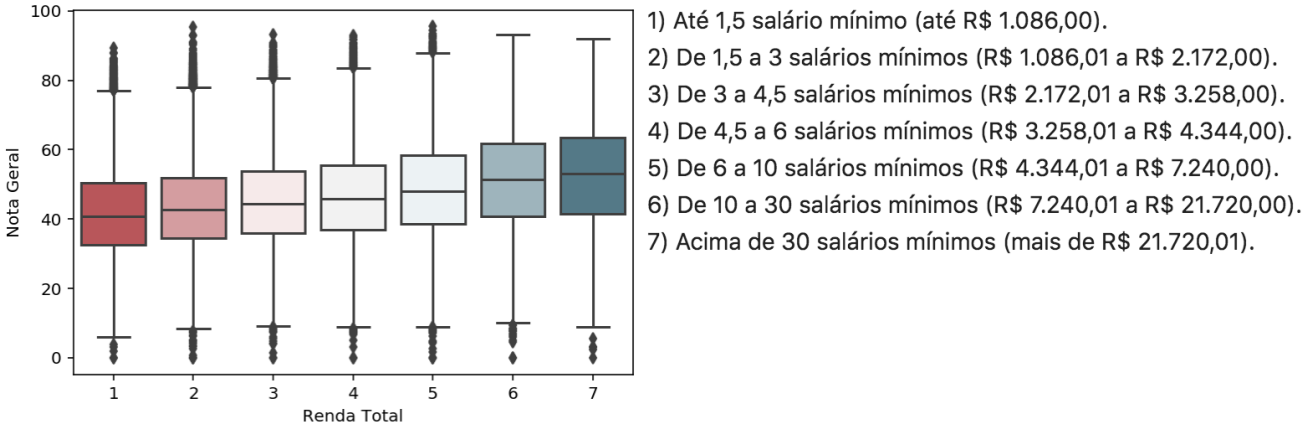
Gráfico 5 - Média da nota geral dos alunos por agrupadas por UF




Fonte: próprio autor (2018)

É notável a variação da média entre os estados, que pode servir de um bom indicativo na predição do desempenho dos estudantes, no gráfico foi destacada a média nacional. Outras medidas estatísticas foram úteis durante a análise, como os valores máximos e mínimos, mediana, moda, etc. No Gráfico 6 foi utilizada a distribuição das notas associadas à resposta dada no questionário sobre a renda total familiar.

Gráfico 6 - Distribuição das notas por renda total familiar

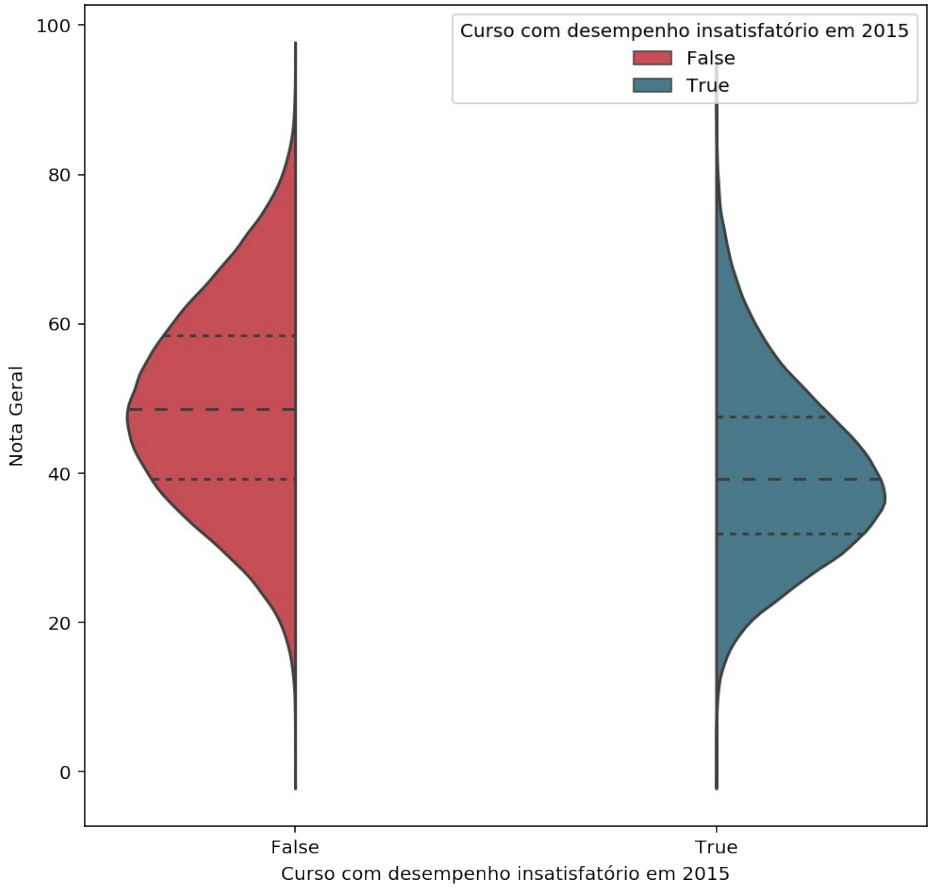



Fonte: próprio autor (2018)

O Gráfico 6 tornou possível visualizar a média, os valores-limite e os quartis da distribuição da nota geral dos alunos. O eixo vertical indica a nota geral dos alunos, enquanto que no horizontal, unido às informações do quadro à sua direita, exibe as alternativas da pergunta do questionário socioeconômico que trata sobre a renda total familiar do participante. A partir dele pode se interpretar que à medida que a renda sobe, são aumentadas as médias das notas, o que leva a concluir que o aluno com maior renda tem mais chances de obter um melhor resultado no exame.

Uma última interpretação buscada, é a relação do desempenho dos alunos com a nota dada a IES pelo MEC, o Conceito ENADE. A nota foi inserida como uma nova característica da base de dados a partir da junção dos dados por meio do identificador em comum tanto nos microdados, quanto nos dados que exibem o CE das IES: o código do curso. No Gráfico 7, foram divididos os alunos pelo Conceito ENADE, se a instituição teve a nota menor que a considerada satisfatória pelo INEP, o aluno será rotulado como *True*, e caso contrário, *False*.

Gráfico 7 - Agrupando notas dos participantes a partir da variável




Fonte: próprio autor (2018)

O eixo vertical representa a nota geral dos participantes, enquanto o horizontal exibe a categoria criada a partir da divisão pelo Conceito ENADE, que indica se o curso do aluno teve desempenho insatisfatório no ano de 2015. É visível a diferença do desempenho dos alunos em cada categoria, os que se classificaram como *True* se agrupam com notas inferiores aos que são rotulados como *False*. Nas próximas seções, serão exploradas as características analisadas nesta, a partir da criação de modelos, relatando como se chegou aos resultados apresentados.

## 4.4 MINERAÇÃO DOS DADOS PARA CONSTRUÇÃO DE MODELOS

Assim como foi apresentado na introdução à metodologia KDD, o processo de construção de um modelo de aprendizagem a partir de um algoritmo foi um processo iterativo e incremental, em que se buscou reduzir a dimensão dos dados e aumentar a eficiência dos modelos, ao passo que se adquiria mais experiência na utilização das ferramentas e mais conhecimento acerca dos dados.

O processo se iniciou ganhando familiaridade com os algoritmos apresentados na seção 3.2, ao trabalhar com os dados do problema e selecionar o que melhor condissesse com o problema.

Visto previamente que está se tratando de um problema de classificação, e que as características selecionadas como mais representativas eram em sua maioria categóricas, o algoritmo que teve maior afinidade foi o de árvore de decisão, por facilitar a visualização de suas regras ao criar a figura da árvore. As regras geradas pelo algoritmo devem servir de base para a orientação das IES que se proponham a analisar o perfil dos alunos, utilizando os resultados apresentados em pesquisas como esta.

As primeiras tentativas de criação de um modelo consistiram em categorizar a nota geral dos alunos pelo método *binning* ou encaixotamento, que tem como objetivo “distribuir os valores de um atributo em um conjunto de caixas (*bins*)” (CASTRO; FERRARI, 2016, p. 57), e foram utilizadas apenas as 11 variáveis pré-selecionadas. Iniciou-se então dividindo as notas em 5 categorias, divididas em espaços de 20, denotando um encaixotamento de mesma largura, em que “o intervalo de cada caixa tem o mesmo tamanho” (CASTRO; FERRARI, 2016, p. 57). Pelo resultado insatisfatório, com a acurácia se mostrando inferior a 50%, a tentativa de criação de um modelo de aprendizagem que apresentasse maior taxa de acerto foi repetida utilizando 4 e em seguida 3 categorias. Ainda com resultados ruins, outra abordagem de encaixotamento foi utilizada, dessa vez utilizando do encaixotamento de mesma frequência, em que “a quantidade de objetos em cada caixa é a mesma” (CASTRO; FERRARI, 2016, p. 57), buscando a normalização das categorias. Mais uma vez sem resultados aceitáveis, as tentativas seguintes tenderam a buscar novas interpretações, sejam elas acerca da variável-alvo ou das características que foram selecionadas anteriormente.

Examinando a abordagem utilizada na seção anterior de dividir os estudantes a partir do Conceito ENADE, a nova estratégia seguiu em tentar criar um classificador binário para predizer se o aluno tem o perfil socioeconômico de um participante que seu curso obteve o Conceito ENADE superior ou inferior ao limiar dado pelo INEP, ou seja, obtidas as notas do Conceito ENADE e inseridas nos microdados, foram separados os estudantes que tiveram nota maior ou igual a 3, dos que apresentaram resultado inferior. Como as notas são dadas a uma instituição ao todo, e não condiz com o desempenho individual de um estudante, foram omitidas as características que possam gerar viés a partir da identificação da instituição, uma vez que os dados relatam as características de estudantes.

A nova intepretação sobre a característica-alvo permitiu da criação de uma nova característica obtida a partir do cruzamento de dados do Conceito ENADE de anos anteriores. Como o exame só é realizado para os mesmos grupos de cursos a cada três anos, foram buscados os dados do Conceito ENADE do ano de 2012. O conjunto novo ao problema não apresentou a

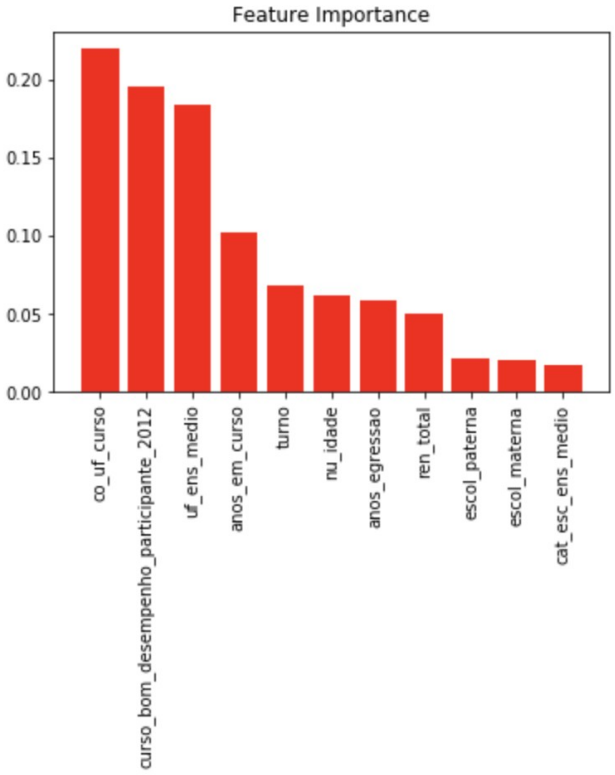
característica utilizada anteriormente para realizar a combinação dos dados, porém se tornou possível gerar um identificador único ao unir os códigos da área do curso, da IES e do município do curso. Foram verificadas que nem todas instituições existiam em ambas bases de dados, e para criar uma característica que houvesse impacto sem inconsistências, foi designada uma nova variável binária que identifica se o curso do participante participou da avaliação do ano de 2012 e teve um desempenho abaixo do esperado.

Utilizando os dados com a nova informação acerca da participação do curso na avaliação anterior, os primeiros modelos gerados obtiveram resultados superiores aos anteriores, e aos poucos foram sendo aperfeiçoados. Para se chegar em um resultado viável, a ferramenta permitiu realizar experimentos com as características selecionadas, visto que a ferramenta utilizada para criação de modelos permite a criação de classificadores de árvore de decisão, onde além dos dados que devem ser treinados, pode ser informada a quantidade limite de características que serão consideradas para que seja feita a melhor escolha nas diversas iterações que o algoritmo realiza. A ferramenta permitiu também a parametrização do grau da árvore, que indica a quantidade máxima de decisões que o algoritmo pode tomar. Como o grau escolhido impacta no resultado de um classificador, foram realizados vários testes, modificando o grau máximo para se chegar no melhor resultado possível.

Para o primeiro modelo, construído a partir do classificador de árvore de decisão da ferramenta *Scikit-learn*, que será denominado **Classificador A** e irá predizer se o perfil o aluno se encaixa em um que obteve o CE abaixo do esperado, nele foi atingida a acurácia de **75, 18%**, com grau máximo de 14. Devido ao tamanho da árvore, não foi possível gerar a figura em seu tamanho original, porém foi adaptada parte de árvore, que consta no **Apêndice C**. Apesar do impedimento, é possível visualizar a importância das características apresentadas nos nós iniciais da árvore.

No **Classificador A** foram utilizadas as seguintes características: UF do curso, idade, tempo de egressão, tempo em curso, turno, e as questões respondidas acerca da escolaridade paterna, escolaridade materna, renda total, categoria de escola que cursou no ensino médio, UF que cursou o ensino médio, juntos ao indicador de desempenho satisfatório do ano de 2012. O Gráfico 8 mostra a importância das variáveis para o classificador:

Gráfico 8 - Importância das características para o **Classificador A**

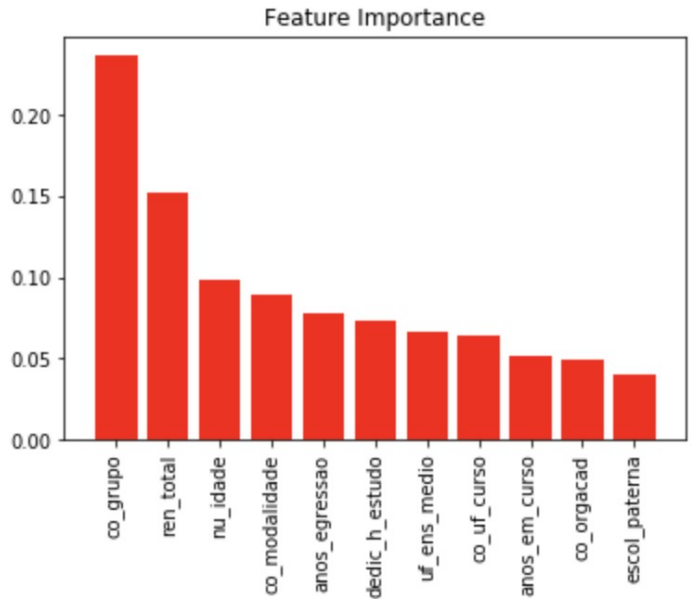



Fonte: próprio autor (2018)

Todavia, não foi descartada a proposição inicial de predizer o desempenho individual dos estudantes a partir da nota geral. Nesta abordagem não houve a limitação de omitir as características que identifiquem os cursos. Então, foi procurada uma estratégia para categorizar a nota dos participantes a partir de um limiar em que se encontrasse a melhor acurácia. Foram criadas colunas temporárias para análise da distribuição de cada uma delas, tendo os limites divisores entre 20 e 50. A partir das experimentações, o limiar com valor 37 se mostrou a melhor escolha, tendo dividido 27,22% das notas abaixo deste.

O **Classificador B,** desenvolvido a partir do mesmo algoritmo do anterior, obteve acurácia de 70,49%, com uma árvore de grau máximo 13, que pode ser observada parcialmente, no Apêndice D, pelos mesmos motivos já apresentados. Nele foram selecionadas 11 características: os códigos do grupo acadêmico, do tipo de organização acadêmica, de UF do curso, da modalidade do curso, e as informações de idade, escolaridade paterna, renda total, UF que cursou o ensino médio, quantidade de horas dedicadas ao estudo, tempo de egressão e tempo em curso. A seguir é exibida a ordem de importância das características selecionadas no Gráfico 9:

Gráfico 9 - Importância das características para a geração do **Classificador B**




Fonte: próprio autor (2018)

Cada abordagem teve suas especificidades quanto às características escolhidas como é possível observar. Apesar das diferentes limitações e abordagens em cada classificador, algumas das características se mostraram fundamentais em ambos casos, se mostrando como as características socioeconômicas que geram mais impacto no desempenho dos estudantes, como a UF do curso, o tempo de egressão, a idade, e informações acerca da escolaridade paterna e da renda total familiar. Na seção seguinte serão elaborados os testes propostos, para cada classificador criado.

## 4.5 VALIDANDO OS CLASSIFICADORES

Ambos os classificadores foram testados com os testes estatísticos apresentados. O primeiro passo se deu ao criar a matriz de confusão para cada modelo gerado, como é observado na Figura 10:

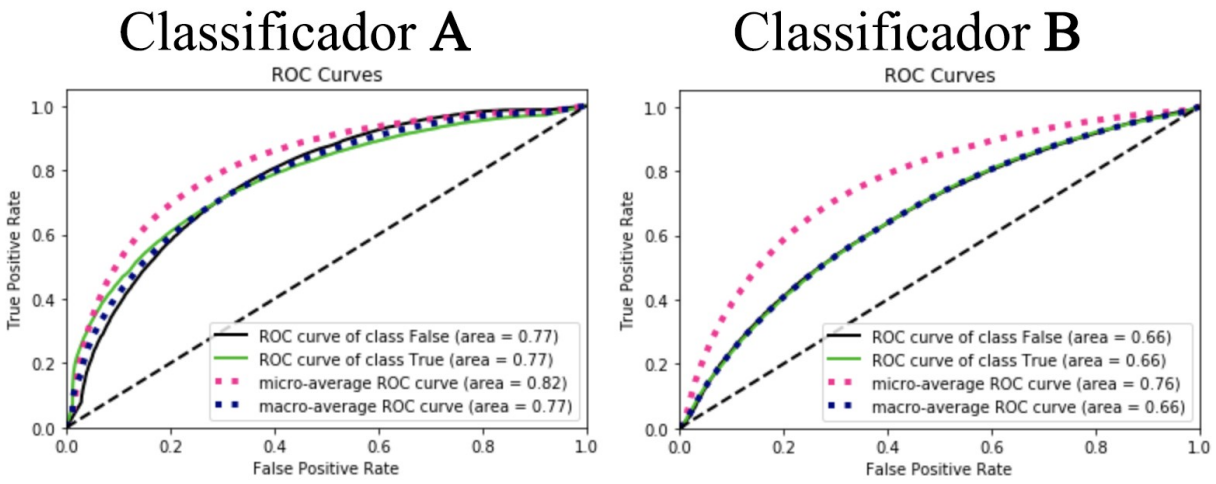
Figura 10 - Matrizes de Confusão dos classificadores A e B

| | Classificador **A** | | Classificador **B** | |
|---|---|---|---|---|
| | ***Positivo*** | ***Negativo*** | ***Positivo*** | ***Negativo*** |
| ***Verdadeiro*** | 17065 | 15069 | 23866 | 4823 |
| ***Falso*** | 57095 | 6936 | 62964 | 4512 |

Fonte: próprio autor (2018)

As matrizes mostram os erros e acertos dos modelos, é notável a melhor eficácia do **Classificador A**, uma vez que além de acertar mais casos, faz o mesmo quando se analisa apenas as classes verdadeiras, que indicam os estudantes que tiveram desempenho ruim. A partir das matrizes foram geradas as curvas ROC ilustradas no conjunto de Gráficos 10:

Gráfico 10 - Curva ROC dos classificadores A e B




Fonte: próprio autor (2018)

Ambas se situam acima da diagonal e fazem uma curva em direção ao ponto em que se acertariam todas as hipóteses, ou seja, o gráfico segue o que foi descrito anterior ao apresentar os gráficos ROC, mostrando um resultado esperado. O próximo teste realizado dirá sobre o quanto os modelos variam quando são tomadas diferentes amostras para realização de treinamento e testes, além de possibilitar utilizar toda a base para cada uma das duas tarefas, dividindo-as em 10 *folds*. Na Tabela 3 é exibido o resultado do K-Fold, detalhando a acurácia de cada *fold* criado:

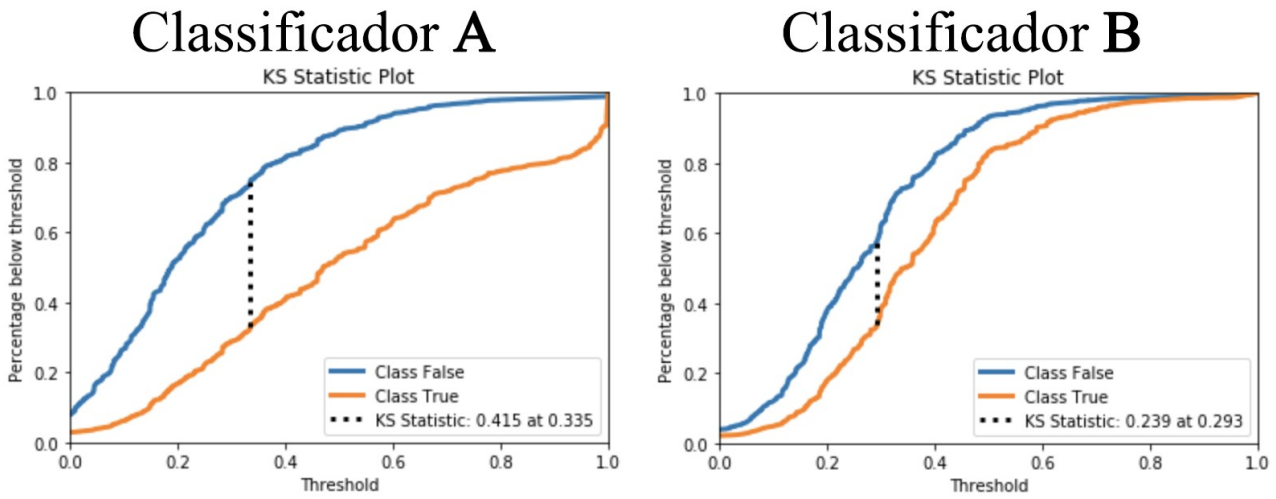
Tabela 5 - Testes KS dos classificadores A e B

| **Classificador A** | | **Classificador B** | |
|---|---|---|---|
| ***Fold*** | **Acurácia** | ***Fold*** | **Acurácia** |
| 0 | 63,03% | 0 | 58,63% |
| *1* | 64,78% | *1* | 57,46% |
| *2* | 63,09% | *2* | 57,90% |
| *3* | 68,22% | *3* | 62,49% |
| *4* | 61,35% | *4* | 59,23% |
| *5* | 66,77% | *5* | 61,37% |
| *6* | 60,06% | *6* | 55,30% |
| *7* | 65,01% | *7* | 60,35% |
| *8* | 62,01% | *8* | 59,42% |
| *9* | 62,70% | *9* | 67,46% |

Fonte: próprio autor (2018)

O **Classificador A** obteve uma acurácia média de 63,70%, enquanto o **B** obteve 59,96%. O último teste realizado, o KS, permitiu a geração do conjunto de Gráficos 11:

Gráfico 11 - Curva KS dos classificadores A e B




Fonte: próprio autor (2018)

Neles pode se observar que o **Classificador A** obteve o KS máximo de 41,5%, e o **B** 23,9%. Os testes realizados mostram que o classificador **A** se mostra sempre mais confiável, uma vez que tende a cometer menos erros e demonstra menos variância enquanto à divisão realizada durante o treinamento, além de distinguir melhor entre as classes como é observado a partir do teste KS. Finalizada a validação dos algoritmos, realizando os testes propostos nos objetivos da pesquisa, torna-se possível revisitar o resultado dos classificadores, relacionando-a ao objetivo geral.

## 4.6 DISCUSSÃO DOS CLASSIFICADORES

Primeiramente, analisando a decisão de desenvolver dois classificadores, que surgiu da seguinte questão: como transformar o desempenho dos participantes em uma categoria? Como foi descrito anteriormente, foram tomadas várias abordagens, até o ponto em que foi adquirido conhecimento suficiente acerca do desenvolvimento prático do trabalho, momento este em que foi dada a característica-alvo a partir da separação pelo Conceito ENADE. Unindo a experiência tomada à importância da nota geral como uma medida de desempenho, foi decidido apresentar o desenvolvimento dos dois casos.

Tendo dois classificadores desenvolvidos e validados, foi possível observar em abordagens diferentes que as seguintes características socioeconômicas geram impacto no desempenho do estudante: a renda total familiar, a UF que o participante cursou o ensino médio, a idade, o tempo em curso, o tempo de egressão, o turno, a escolaridade paterna e materna. O impacto de algumas

dessas características na nota geral foi apresentado desde a analise descritiva como UF a que o participante cursou o ensino médio e a renda total familiar, e com o auxílio do **Apêndice E**, que descreve a correlação entre as características utilizadas pelos classificadores, junto à nota geral. Dentre as características socioeconômicas dos participantes que não foram abordadas anteriormente, pode se observar que quando a idade, o tempo em curso ou o tempo de egressão, aumentam, o desempenho é diminuído. Já o grau de escolaridade dos pais, e outras características como o tempo dedicado semanalmente aos estudos se mostram diretamente proporcionais à nota. A importância das características para a decisão da classificação pôde ser observada nos Gráficos 8 e 9.

Por fim, houveram também características acerca do curso que se mostraram influentes no desempenho, como a UF e a área de ensino em que se classifica, que foram abordadas durante a análise descritiva. As características identificadas, ao serem unidas poderão auxiliar as IES nas futuras edições do exame.

# 5 CONCLUSÃO

Os resultados mostraram que é possível predizer o desempenho dos estudantes participantes do ENADE a partir de informações socioeconômicas, considerando o desempenho como um limitador das notas ou interpretando-o com base no desempenho do curso a partir do CE. Tanto fazendo uso do modelo quanto ao se usufruir das regras geradas pelo algoritmo de árvore de decisão, pode-se classificar os futuros participantes das seguintes provas, que pertençam ao mesmo grupo do que foi selecionado em 2015. Ao verificar que um aluno se encaixa na hipótese de obter baixo desempenho, os cursos ou as instituições de ensino podem intensificar a orientação deste.

Atingindo esta etapa, foram realizados os objetivos propostos, o processo de limpeza, pré-processamento, tratamento e seleção dos dados, criação de modelos de aprendizagem e realização de testes, finalizando um ciclo do KDD (como apresentado anteriormente na Figura 3).

Com este resultado foi possível adquirir conhecimento não só sobre a metodologia que tem por objetivo obter conhecimento a partir de dados, mas também aprofundar o entendimento no problema, e exigiu dedicação máxima para entender o que deve ser feito. A resposta-problema da pesquisa foi respondida ao identificar as características socioeconômicas que mais influenciam no desempenho dos participantes do ENADE no ano de 2015.

Ao desenvolver deste trabalho, as dificuldades se resumiram em algumas principais: a primeira delas surgiu devido à falta de conhecimento acerca do assunto, uma vez que este trabalho se iniciou como um desafio a aprender sobre um novo tema. Aprendizagem de Máquina foi o escolhido, porém, inicialmente sem ter uma abordagem ou problema condizente com o tema, até ser decidido utilizar os dados públicos fornecidos pelo INEP.

A área estudada dispõe da maioria de seus materiais e pesquisas em inglês, que não foi um problema ao aprofundar no assunto, porém pela falta de entendimento prévio, a introdução ao assunto teria sido facilitada caso tivesse maior disponibilidade de materiais em português. Além disso o baixo aprofundamento do curso nos mais diversos estudos da estatística aqui abordados, unido à falta de abordagem prática da aplicação de Inteligência Artificial em sistemas computacionais deram a dificuldade de entendimento no assunto e na elaboração do trabalho.

Dito isto, a elaboração da pesquisa fez uso das mais diversas áreas abordadas durante a carreira acadêmica. Da divisão de tarefas em tarefas menores, para ter melhor controle do projeto, como pregam as metodologias de desenvolvimento ágil apresentadas na engenharia de software. Do controle, documentação e coleta de requisitos para tornar possível o acompanhamento e planejamento do trabalho, como apresentado nas disciplinas de análise de projetos. Dos algoritmos,

das linguagens de programação, as estruturas de dados e a matemática, que, assim como outros diversos assuntos da ciência da computação, aqui sempre se mostraram necessários.

Por fim, a pesquisa proporcionou a oportunidade de estudar e praticar um novo tema, que vem se mostrando cada vez mais importante, como foi apresentado. Foi possível aprender e se integrar com as ferramentas de análise de dados e de aprendizagem de máquina, e a experiência poderá criar outras oportunidades, além de ter desenvolvido um senso crítico acerca de como os dados são tratados, organizados e apresentados.

## 5.1 TRABALHOS FUTUROS

Os resultados apresentados são sempre passíveis de melhorias, uma vez que a metodologia utilizada para o desenvolvimento da pesquisa, o KDD, trata de ciclos em que cada iteração pode resultar no aperfeiçoamento de um resultado final. Tais melhorias podem ser alcançadas futuramente ao conhecer melhor o funcionamento da metodologia e das nuances de cada uma de suas etapas. Desde a etapa inicial do processo é passível de melhorias, na seleção de dados poderia ser buscada a integração de dados de anos anteriores, estudando a possibilidade de trabalhar com todos os grupos em que o INEP separa para a aplicação anual do exame.

Na etapa de mineração de dados, podem surgir outras propostas, como a de normalizar os dados, verificando de forma mais ampla a distribuição dos dados. Ou ainda, poderiam ser analisados os cursos como a entidade central, abordando não só as características socioeconômicas do grupo com auxílio de métricas estatísticas, acrescentando características como a quantidade de alunos inscritos, ou de faltantes.

Por fim, poderia ser continuada a pesquisa ao estudar o impacto de utilização dos resultados fornecidos nesta nos eventuais exames seguintes, validando se o proposto realmente auxiliaria as instituições de ensino, uma vez que é trabalhado com hipóteses estatísticas.

# REFERÊNCIAS

ADEODATO, P. J. L.; MELO, S. B. **Equivalência entre a Área sob a Curva Kolmogorov Smirnov e o Índice de Gini na Avaliação de Desempenho de Decisões Binárias.** Centro de Informática – Universidade Federal de Pernambuco. 2016. Disponível em: <https://pdfs.semanticscholar.org/3fb0/ba0f0eab9c9379f0bb4a0121184ae2ed77cd.pdf>. Acesso em: 14 abr. 2018.

ALBARDEIRO, L; GAMA, C; PEREIRA, M. F.; CHICHORRO, M. **Utilização do Teste Kolmogorov-Smirnov para estudos de proveniência sedimentar**. Comunicações Geológicas, Laboratório Nacional de Geologia e Energia IP, 2014. Disponível em: <http://www.lneg.pt/download/9786/66_2943_ART_CG14_ESPECIAL_III.pdf>. Acesso em: 14 abr. 2018.

AMO, S de. Técnicas de mineração de dados. **Jornada de Atualização em Informática**, 2004. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 30 set. 2017.

CANDÃO, J. de P.; REAL, E. M. Uma análise do perfil e desempenho de estudantes de Computação através da Mineração de Dados. **Anais do *Computer on the Beach***, p. 529-530, 2017. Disponível em: <https://siaiap32.univali.br/seer/index.php/acotb/article/view/10515>. Acesso em: 05 out. 2017.

CASTRO, L. N. de; FERRARI, D. G. **Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações.** São Paulo: Saraiva, 2016.

COSTA, E.; BAKER, R. J. D.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1-29, 2013. Disponível em: <http://www.br-ie.org/pub/index.php/pie/article/view/2341>. Acesso em: 06 out. 2017.

CRETTON, N. N.; GOMES, G. R. R. Aplicação de técnicas de Mineração e Dados na base de dados do ENADE com enfoque nos cursos de medicina. **Acta Biomedica Brasiliensia**, v. 7, n. 1, p.

74-89, 2016. Disponível em: <http://www.actabiomedica.com.br/index.php/acta/article/view/130>. Acesso em: 15 set. 2017.

FAWCETT, T. **ROC Graphs: Notes and Practical Considerations for Researchers**. Kluwer Academics Publishers, 2004. Disponível em: <http://binf.gmu.edu/mmasso/ROC101.pdf>. Acesso em: 05 abr. 2018.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de Pesquisa**. Editora da UFRGS – PLAGEDER, 2009.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas SA, 2008.

GOTTI, F. J. A.; COSTA, I.; SHIGUEMORI, E. H. ***Computational Intelligence applied to student's performance evaluation in Higher Education***. 2012. Disponível em <http://www.abepro.org.br/biblioteca/icieom2012_submission_173.pdf>. Acesso em: 12 set. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **NOTA TÉCNICA Nº 2/2017/CGCQES/DAES**. Disponível em <http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2015/nota_tecnica_daes_n22017_calculo_do_conceito_enade2015.pdf>. Acesso em: 08 out. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Manual do Enade 2015**. Disponível em: <http://portal.inep.gov.br/manuais>. Acesso em: 06 out. 2017.

MÜLLER, A. C; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists.** California: O'Reilly Media, 2016.

NOGUEIRA, E. D. A.; TSUNODA, D. F. **Mineração de Dados para análise da relação entre as características socioeconômicas de concluintes do ensino superior e o desempenho desses estudantes no ENADE 2012**. Revista Percurso, v. 15, n. 1, p. 245-268, 2015. Disponível em <http://revista.unicuritiba.edu.br/index.php/percurso/article/view/1102>. Acesso em: 07 set. 2017.

PEREIRA, G.; ORTIGÃO, Mª. I. R. **Pesquisa Quantitativa em educação: Algumas considerações**. Periferia – Publicações UERJ, v. 8, n. 1, p. 66-79, 2017.

PROVOST, F.; FAWCETT, T. **Data Science for Business**. California: O'Reilly Media, 2013.

SOUZA, H. V. L.; NEIVA, D. H.; CAVALCANTI, R. P.; RODRIGUES, R.L.; GOMES, A. S.; ADEODATO, P. J. L. Uma Análise preditiva de desempenho dos cursos no ENADE com base no perfil socioeconômico e desempenho no ENEM dos alunos. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. VI Congresso Brasileiro de Informática na Educação, 2017. Disponível em: <http://www.br-ie.org/pub/index.php/wcbie/article/view/7454/5250>. Acesso em: 10 mai. 2018.

SCUDINO, P. A. **A Utilização de Alguns Testes Estatísticos para Análise da Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro**. Trabalho de Conclusão de Curso, Universidade Federal Rural do Rio de Janeiro. 2008. Disponível em: <http://www.ufrrj.br/abelhanatureza/paginas/docs_estado/Estudomercado_mel.pdf>. Acesso em: 14 abr. 2018.

WIRTH, R. HIPP, J. ***CRISP-DM: Towards a Standard Process Model for Data Mining***. *In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, p. 29-39, 2000.

# APÊNDICE A – *SCRIPTS* SQL

**Importação da tabela**

```
LOAD DATA LOCAL INFILE '/home/pedro/Documentos/Microdados Enade
2015/2.DADOS/5.csv'
   INTO TABLE microdados_enade_2015 FIELDS TERMINATED by ';'
   OPTIONALLY ENCLOSED BY '"' LINES TERMINATED BY '\r\n';
```

**Conversão dos dados em formato VARCHAR para FLOAT**

```
UPDATE microdados_enade_2015 SET nt_obj_fg = REPLACE(nt_obj_fg,',','.');
UPDATE microdados_enade_2015 SET nt_dis_fg = REPLACE(nt_dis_fg,',','.');
UPDATE microdados_enade_2015 SET nt_fg     = REPLACE(nt_fg,    ',','.');
UPDATE microdados_enade_2015 SET nt_obj_ce = REPLACE(nt_obj_ce,',','.');
UPDATE microdados_enade_2015 SET nt_dis_ce = REPLACE(nt_dis_ce,',','.');
UPDATE microdados_enade_2015 SET nt_ce     = REPLACE(nt_ce,    ',','.');
UPDATE microdados_enade_2015 SET nt_ger    = REPLACE(nt_ger,   ',','.');
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_obj_fg FLOAT;
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_dis_fg FLOAT;
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_fg     FLOAT;
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_obj_ce FLOAT;
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_ce     FLOAT;
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_dis_ce FLOAT;
ALTER TABLE microdados_enade_2015 MODIFY COLUMN nt_ger    FLOAT;
```

**Criação auxiliar de tabela dos estados**

```
CREATE TABLE enade_2015_estados (codigo int, uf varchar(2), nome varchar(50));
```

**Estados com código, sigla e nome**

```
INSERT INTO enade_2015_estados VALUES (11, 'RO','Rondônia');
INSERT INTO enade_2015_estados VALUES (12, 'AC','Acre');
INSERT INTO enade_2015_estados VALUES (13, 'AM','Amazonas');
INSERT INTO enade_2015_estados VALUES (14, 'RR','Roraima');
INSERT INTO enade_2015_estados VALUES (15, 'PA','Pará');
```

```
INSERT INTO enade_2015_estados VALUES (16, 'AP','Amapa');
INSERT INTO enade_2015_estados VALUES (17, 'TO','Tocantins');
INSERT INTO enade_2015_estados VALUES (21, 'MA','Maranhão');
INSERT INTO enade_2015_estados VALUES (22, 'PI','Piauí');
INSERT INTO enade_2015_estados VALUES (23, 'CE','Ceará');
INSERT INTO enade_2015_estados VALUES (24, 'RN','Rio Grande do Norte');
INSERT INTO enade_2015_estados VALUES (25, 'PB','Paraíba');
INSERT INTO enade_2015_estados VALUES (26, 'PE','Pernambuco');
INSERT INTO enade_2015_estados VALUES (27, 'AL','Alagoas');
INSERT INTO enade_2015_estados VALUES (28, 'SE','Sergipe');
INSERT INTO enade_2015_estados VALUES (29, 'BA','Bahia');
INSERT INTO enade_2015_estados VALUES (31, 'MG','Minas Gerais');
INSERT INTO enade_2015_estados VALUES (32, 'ES','Espírito Santo');
INSERT INTO enade_2015_estados VALUES (33, 'RJ','Rio de Janeiro');
INSERT INTO enade_2015_estados VALUES (35, 'SP','São Paulo');
INSERT INTO enade_2015_estados VALUES (41, 'PR','Paraná');
INSERT INTO enade_2015_estados VALUES (42, 'SC','Santa Catarina');
INSERT INTO enade_2015_estados VALUES (43, 'RS','Rio Grande do Sul');
INSERT INTO enade_2015_estados VALUES (50, 'MS','Mato Grosso do Sul');
INSERT INTO enade_2015_estados VALUES (51, 'MT','Mato Grosso');
INSERT INTO enade_2015_estados VALUES (52, 'GO','Goiás');
INSERT INTO enade_2015_estados VALUES (53, 'DF','Distrito Federal');
INSERT INTO enade_2015_estados VALUES (99, 'N','Não se aplica');
```

**Busca por alunos presentes**

TOTAL DE ALUNOS: 549.487

TOTAL DE ALUNOS PRESENTES: 447.056

PORCENTAGEM PRESENTES: 81,36%

```
SELECT COUNT(*) AS QTD_TOTAL,
   (SELECT COUNT(*)
      FROM microdados_enade_2015
     WHERE tp_pres = 555) AS QTD_PRESENTES,
   ((SELECT COUNT(*)
```

```
    FROM microdados_enade_2015
   WHERE tp_pres = 555) * 100 / COUNT(*)) AS PORC_PRESENTES
FROM microdados_enade_2015;
```

**Busca por alunos que tiveram presença em todas as etapas**
Essa query acabou excluindo muitos alunos, então serão desconsideradas as colunas denominadas 'tp_sfg' e 'tp_sce', no qual tratam das questões dissertativas individualmente.

TOTAL DE ALUNOS: 549.487
TOTAL DE ALUNOS VALIDOS: 297.266
PORCENTAGEM VALIDOS: 54,10%

```
SELECT COUNT(*) AS QTD_TOTAL,
  (SELECT COUNT(*) FROM microdados_enade_2015
    WHERE tp_pres = 555
    AND tp_pr_ger = 555
    AND tp_pr_ob_fg = 555
    AND tp_pr_di_fg = 555
    AND tp_pr_ob_ce = 555
    AND tp_pr_di_ce = 555
    AND tp_sfg_d1 = 555
    AND tp_sfg_d2 = 555
    AND tp_sce_d1 = 555
    AND tp_sce_d2 = 555
    AND tp_sce_d3 = 555) AS QTD_VALIDOS,
  ((SELECT COUNT(*) FROM microdados_enade_2015
    WHERE tp_pres=555
      AND tp_pr_ger = 555
      AND tp_pr_ob_fg = 555
      AND tp_pr_di_fg = 555
      AND tp_pr_ob_ce = 555
      AND tp_pr_di_ce = 555
      AND tp_sfg_d1 = 555
      AND tp_sfg_d2 = 555
      AND tp_sce_d1 = 555
      AND tp_sce_d2 = 555
```

```
        AND tp_sce_d3 = 555) * 100 / COUNT(*)) AS PORC_VALIDOS
    FROM microdados_enade_2015;
```

**Busca por alunos presentes, desconsiderando as questões discussivas**

TOTAL DE ALUNOS: 549.487

TOTAL DE ALUNOS VALIDOS: 393.408

PORCENTAGEM VALIDOS: 71,60%

```
SELECT COUNT(*) AS QTD_TOTAL,
    (SELECT COUNT(*) FROM microdados_enade_2015
        WHERE tp_pres = 555
        AND tp_pr_ger = 555
        AND tp_pr_ob_fg = 555
        AND tp_pr_di_fg = 555
        AND tp_pr_ob_ce = 555
        AND tp_pr_di_ce = 555) AS QTD_VALIDOS,
    ((SELECT COUNT(*)
        FROM microdados_enade_2015
        WHERE tp_pres=555
            AND tp_pr_ger = 555
            AND tp_pr_ob_fg = 555
            AND tp_pr_di_fg = 555
            AND tp_pr_ob_ce = 555
            AND tp_pr_di_ce = 555) * 100 / COUNT(*)) AS PORC_VALIDOS
    FROM microdados_enade_2015;
```

**Busca por alunos que deixaram todas as provas em branco, mas participaram delas**

Resultado: 519

```
SELECT COUNT(*) AS PROVAS_EM_BRANCO
    FROM microdados_enade_2015
    WHERE tp_pres=555
        AND tp_pr_ger != 555
        AND tp_pr_ob_fg != 555
        AND tp_pr_di_fg != 555
```

```
    AND tp_pr_ob_ce != 555
    AND tp_pr_di_ce != 555
    AND tp_sfg_d1 != 555
    AND tp_sfg_d2 != 555
    AND tp_sce_d1 != 555
    AND tp_sce_d2 != 555
    AND tp_sce_d3 != 555;
```

**Porcentagem de alunos por estado**

```
SELECT ES.uf,
   (COUNT(MD.co_uf_curso) * 100 / (SELECT COUNT(*) FROM microdados_enade_2015) )
   AS porcentagem
   FROM microdados_enade_2015 MD
      INNER JOIN enade_2015_estados ES ON (MD.co_uf_curso = ES.codigo)
   group by ES.uf;
```

**Porcentagem de alunos por estado, com provas válidas**

```
SELECT ES.uf,
   (count(MD.co_uf_curso) * 100 / (SELECT COUNT(*) FROM microdados_enade_2015) )
   AS porcentagem
   FROM (SELECT * FROM microdados_enade_2015 WHERE tp_pres = 555
         AND tp_pr_ger = 555
         AND tp_pr_ob_fg = 555
         AND tp_pr_di_fg = 555
         AND tp_pr_ob_ce = 555
         AND tp_pr_di_ce = 555) MD
      INNER JOIN enade_2015_estados ES ON (MD.co_uf_curso = ES.codigo)
   GROUP BY ES.uf;
```

**Cálculo de egressão e de tempo em curso.**

```
SELECT ano_fim_2g,
   ano_in_grad,
```

```
    (ano_in_grad - ano_fim_2g) AS EGRESSAO,
    (2015 - ano_in_grad) AS TEMPO_CURSO
    FROM microdados_enade_2015;
```

**Criação de colunas com tempo de egressão, e de tempo em curso**

```
ALTER TABLE microdados_enade_2015
    ADD anos_egressao INT;
ALTER TABLE microdados_enade_2015
    ADD anos_em_curso INT;
UPDATE microdados_enade_2015 SET anos_egressao = (ano_in_grad - ano_fim_2g);
UPDATE microdados_enade_2015 SET anos_em_curso = (2015 – ano_in_grad);
```

**Criando tabela final**

```
CREATE TABLE microdados_final AS (
    SELECT qe_i4 AS escol_paterna,
        qe_i5 AS escol_materna,
        qe_i8 AS ren_total,
        qe_i10 AS sit_trabalho,
        qe_i12 AS aux_permanencia,
        qe_i16 AS uf_ens_medio,
        qe_i17 AS cat_esc_ens_medio,
        qe_i21 AS concl_es_familia,
        qe_i23 AS dedic_h_estudos,
        anos_egressao,
        anos_em_curso,
       nt_ger AS nota_geral
        FROM microdados_enade_2015
        WHERE tp_pres = 555
            AND tp_pr_ger = 555
            AND tp_pr_ob_fg = 555
            AND tp_pr_di_fg = 555
            AND tp_pr_ob_ce = 555
```

```
    AND tp_pr_di_ce = 555);
```

Tabela auxiliar para transformação das opções em números

```
CREATE TABLE aux_opcoes (numero INT, letra CHAR(1));
INSERT INTO aux_opcoes VALUES (1, 'A');
INSERT INTO aux_opcoes VALUES (2, 'B');
INSERT INTO aux_opcoes VALUES (3, 'C');
INSERT INTO aux_opcoes VALUES (4, 'D');
INSERT INTO aux_opcoes VALUES (5, 'E');
INSERT INTO aux_opcoes VALUES (6, 'F');
INSERT INTO aux_opcoes VALUES (7, 'G');
INSERT INTO aux_opcoes VALUES (8, 'H');
INSERT INTO aux_opcoes VALUES (9, 'I');
```

**Cópia da tabela final para realizar transformação nos dados**

```
CREATE TABLE microdados_final_v2 AS (SELECT * FROM microdados_final);
```

**Transformação de cada coluna em numérica**

```
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.escol_paterna
    SET microdados_final_v2.escol_paterna = aux_opcoes.numero;
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.escol_materna
    SET microdados_final_v2.escol_materna = aux_opcoes.numero;
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.ren_total
    SET microdados_final_v2.ren_total = aux_opcoes.numero;
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.sit_trabalho
    SET microdados_final_v2.sit_trabalho = aux_opcoes.numero;

UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.aux_permanencia
```

```
    SET microdados_final_v2.aux_permanencia = aux_opcoes.numero;
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.cat_esc_ens_medio
    SET microdados_final_v2.cat_esc_ens_medio = aux_opcoes.numero;
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.concl_es_familia
    SET microdados_final_v2.concl_es_familia = aux_opcoes.numero;
UPDATE microdados_final_v2
    JOIN aux_opcoes ON aux_opcoes.letra = microdados_final_v2.dedic_h_estudos
    SET microdados_final_v2.dedic_h_estudos = aux_opcoes.numero;
UPDATE microdados_final_v2
JOIN enade_2015_estados ON enade_2015_estados.uf=microdados_final_v2.uf_ens_medio
    SET microdados_final_v2.uf_ens_medio = enade_2015_estados.codigo;
```

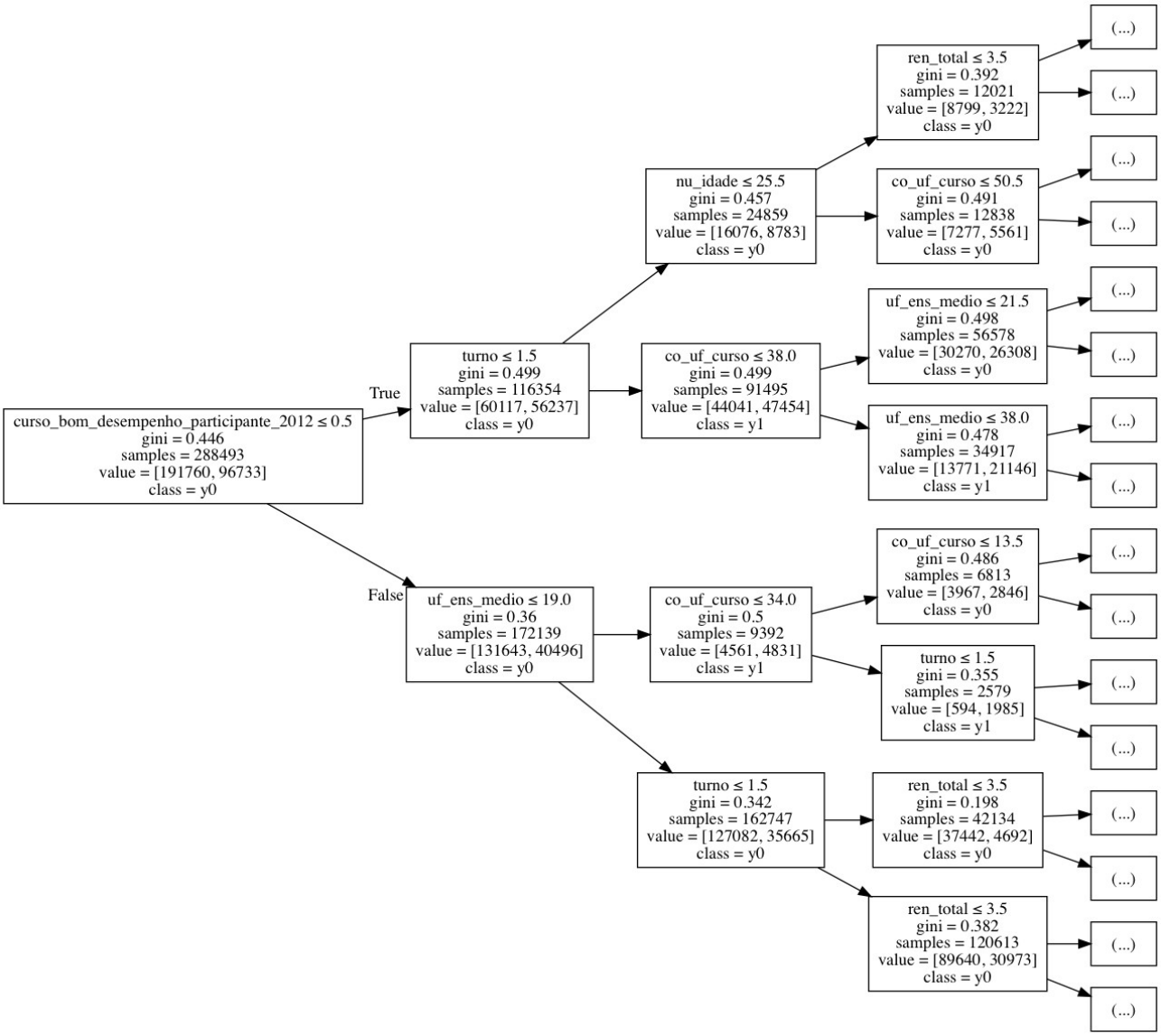
**Modificando o tipo das variáveis alteradas para INT**

```
ALTER TABLE microdados_final_v2 MODIFY COLUMN escol_paterna INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN escol_materna INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN ren_total INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN sit_trabalho INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN aux_permanencia INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN cat_esc_ens_medio INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN concl_es_familia INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN dedic_h_estudos INT;
ALTER TABLE microdados_final_v2 MODIFY COLUMN uf_ens_medio INT;
```

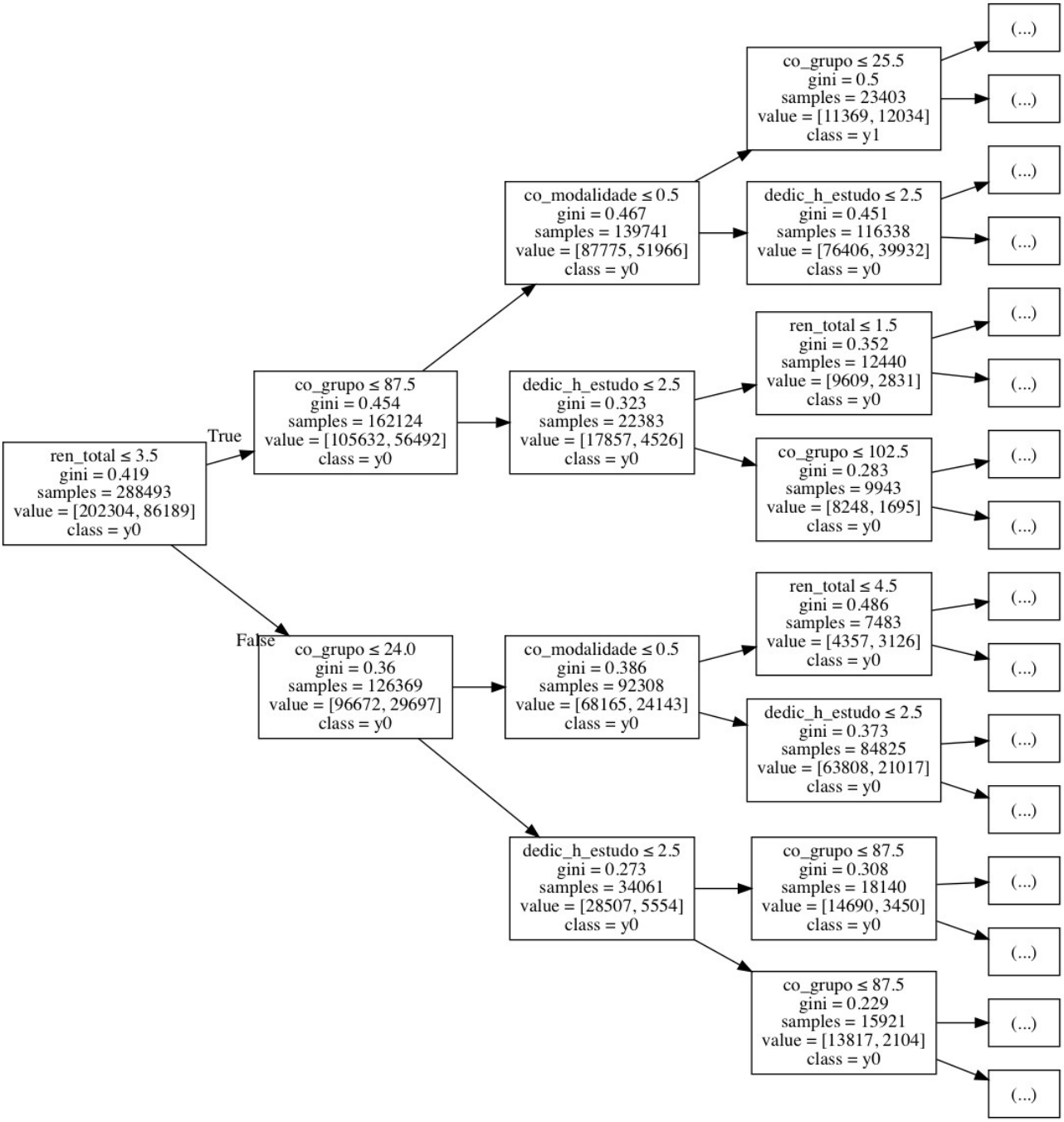
# APÊNDICE B – DICIONÁRIO DE DADOS RESULTANTE

| Nome da Variável | Descrição | Tipo |
|---|---|---|
| escol_paterna | Formação escolar do pai | Categórica |
| escol_materna | Formação escolar da mãe | Categórica |
| ren_total | Renda total familiar | Categórica |
| sit_trabalho | Situação de trabalho do estudante, por hora semanal | Categórica |
| aux_permanencia | Recebimento de auxílio permanência durante a trajetória acadêmica | Categórica |
| uf_ens_medio | Estado em que concluiu o ensino médio | Categórica |
| cat_esc_ens_medio | Tipo de escola que o estudante cursou no ensino médio | Categórica |
| concl_es_familia | Indicador de concluintes de ensino superior na família | Binária |
| dedic_h_estudo | Quantidade de horas dedicadas aos estudos semanalmente | Categórica |
| anos_egressao | Quantidade de anos que durou o período entre o fim do ensino médio e início do ensino superior | Numérica discreta |
| anos_em_curso | Quantidade de anos que durou o período entre o ano de início do ensino superior e o ano de realização do exame | Numérica discreta |
| nota_geral | Nota geral do estudante na prova do ENADE | Numérica continua |

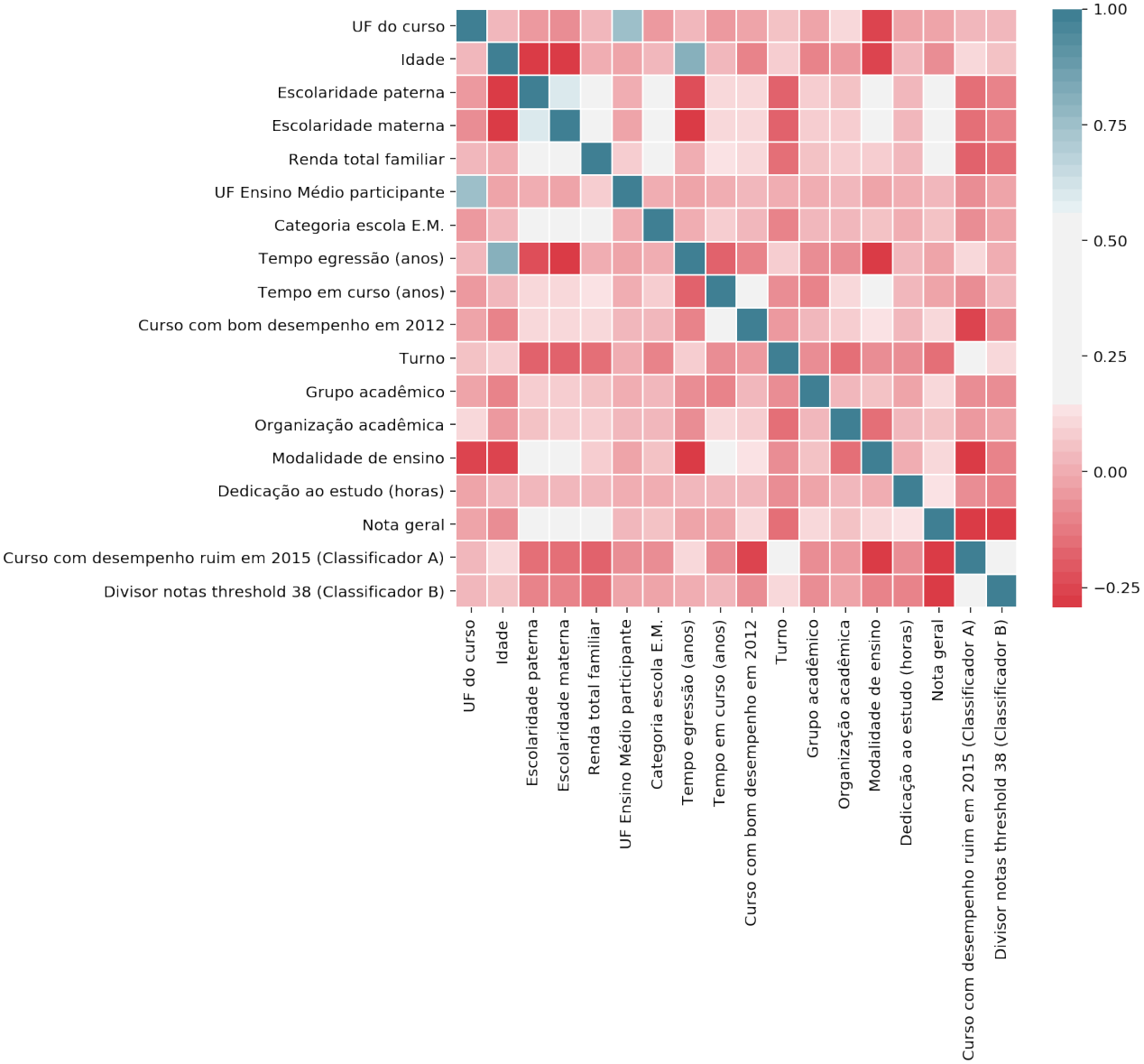
# APÊNDICE C – ÁRVORE DE GRAU 4 DO CLASSIFICADOR A



# APÊNDICE D – ÁRVORE DE GRAU 4 DO CLASSIFICADOR B


ren_total ≤ 3.5
gini = 0.419
samples = 288493
value = [202304, 86189]
class = y0
True
False
co_grupo ≤ 87.5
gini = 0.454
samples = 162124
value = [105632, 56492]
class = y0
co_grupo ≤ 24.0
gini = 0.36
samples = 126369
value = [96672, 29697]
class = y0
co_modalidade ≤ 0.5
gini = 0.467
samples = 139741
value = [87775, 51966]
class = y0
dedic_h_estudo ≤ 2.5
gini = 0.323
samples = 22383
value = [17857, 4526]
class = y0
co_modalidade ≤ 0.5
gini = 0.386
samples = 92308
value = [68165, 24143]
class = y0
dedic_h_estudo ≤ 2.5
gini = 0.273
samples = 34061
value = [28507, 5554]
class = y0
co_grupo ≤ 25.5
gini = 0.5
samples = 23403
value = [11369, 12034]
class = y1
dedic_h_estudo ≤ 2.5
gini = 0.451
samples = 116338
value = [76406, 39932]
class = y0
ren_total ≤ 1.5
gini = 0.352
samples = 12440
value = [9609, 2831]
class = y0
co_grupo ≤ 102.5
gini = 0.283
samples = 9943
value = [8248, 1695]
class = y0
ren_total ≤ 4.5
gini = 0.486
samples = 7483
value = [4357, 3126]
class = y0
dedic_h_estudo ≤ 2.5
gini = 0.373
samples = 84825
value = [63808, 21017]
class = y0
co_grupo ≤ 87.5
gini = 0.308
samples = 18140
value = [14690, 3450]
class = y0
co_grupo ≤ 87.5
gini = 0.229
samples = 15921
value = [13817, 2104]
class = y0
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

# APÊNDICE E – CORRELAÇÃO DAS CARACTERÍSTICAS UTILIZADAS NOS CLASSIFICADORES



# ANEXO A – DICIONÁRIO DE VARIÁVEIS

| Nome | Descrição |
|---|---|
| nu_ano | Ano de realização do exame |
| co_grupo | Código da Área de enquadramento do curso no Enade |
| co_ies | Código da IES (e-Mec) |
| co_catad | Código da categoria administrativa da IES |
| co_orgac | Código da organização acadêmica da IES |
| co_munic_curso | Código do município de funcionamento do curso |
| co_uf_curso | Código da UF de funcionamento do curso |
| co_regiao_curso | Código da região de funcionamento do curso |
| co_curso | Código do curso no Enade |
| co_modalidade | Modalidade de Ensino |
| nu_idade | Idade do inscrito em 22/11/2015 |
| tp_sexo | Sexo do inscrito |
| ano_fim_2g | Ano de conclusão do 2º grau |
| ano_in_grad | Ano de início da graduação |
| tp_semestre | Semestre de graduação |
| in_matutino | Indicador de turno matutino |
| in_vespertino | Indicador de turno vespertino |
| in_noturno | Indicador de turno noturno |
| id_status | Indicativo de inscrito regular ou irregular |
| amostra | Indicativo de estudante selecionado para a amostra |
| tp_inscricao | Indicador de concluinte / ingressante |
| tp_def_fis | Indicador de deficiência física |
| tp_def_vis | Indicador de deficiência visual |
| tp_def_aud | Indicador de deficiência auditiva |
| nu_item_ofg | Quantidade de itens da parte objetiva de Formação Geral |
| nu_item_ofg_z | Quantidade de itens da parte objetiva de Formação Geral que foram excluídos devido a anulação |
| nu_item_ofg_x | Quantidade de itens da parte objetiva de Formação Geral que foram excluídos devido ao coeficiente pontobisserial menor que 0,20 |
| nu_item_ofg_n | Quantidade de itens da parte objetiva de Formação Geral que não se aplicam ao grupo de curso |
| vt_gab_ofg_orig | Vetor que representa o gabarito original de Formação Geral |
| vt_gab_ofg_fin | Vetor que representa o gabarito final de Formação Geral |
| nu_item_oce | Quantidade de itens da parte objetiva de Componente Específico |
| nu_item_oce_z | Quantidade de itens da parte objetiva de Componente Específico que foram excluídos devido a anulação |
| nu_item_oce_x | Quantidade de itens da parte objetiva de Componente Específico que foram excluídos devido ao coeficiente pontobisserial menor que 0,20 |
| nu_item_oce_n | Quantidade de itens da parte objetiva de Componente Específico que não se aplicam ao grupo de curso |
| vt_gab_oce_orig | Vetor que representa o gabarito original de Componente Específico |

| | |
|---|---|
| vt_gab_oce_fin | Vetor que representa o gabarito final de Componente Específico |
| tp_pres | Tipo de presença |
| tp_pr_ger | Tipo de presença na prova |
| tp_pr_ob_fg | Tipo de presença na parte objetiva na formação geral |
| tp_pr_di_fg | Tipo de presença na parte discursiva na formação geral |
| tp_pr_ob_ce | Tipo de presença na parte objetiva no componente específico |
| tp_pr_di_ce | Tipo de presença na parte discursiva no componente específico |
| tp_sfg_d1 | Situação da questão 1 da parte discursiva da formação geral |
| tp_sfg_d2 | Situação da questão 2 da parte discursiva da formação geral |
| tp_sce_d1 | Situação da questão 1 da parte discursiva do componente específico |
| tp_sce_d2 | Situação da questão 2 da parte discursiva do componente específico |
| tp_sce_d3 | Situação da questão 3 da parte discursiva do componente específico |
| vt_esc_ofg | Vetor que representa a escolha da parte objetiva da formação geral - 1 letra por item, '.'=em branco, '*'=múltiplo |
| vt_ace_ofg | Vetor que representa os acertos da parte objetiva na formação geral |
| vt_esc_oce | Vetor que representa a escolha da parte objetiva do componente específico - 1 letra por item, '.'=em branco, '*'=múltiplo |
| vt_ace_oce | Vetor que representa os acertos da parte objetiva do componente específico |
| nt_obj_fg | Nota bruta na parte objetiva da formação geral - Convertida para escala de 0 a 100 |
| nt_fg_d1_pt | Nota de Língua Portuguesa da questão 1 da parte discursiva da formação geral - Convertida para escala de 0 a 100 |
| nt_fg_d1_ct | Nota de Conteúdo da questão 1 da parte discursiva da formação geral - Convertida para escala de 0 a 100 |
| nt_fg_d1 | Nota da questão 1 da parte discursiva da formação geral - Média ponderada da parte de Língua Portuguesa (20%) e Conteúdo (80%) da Questão 1 da parte discursiva (0 a 100) |
| nt_fg_d2_pt | Nota de Língua Portuguesa da questão 2 da parte discursiva da formação geral - Convertida para escala de 0 a 100 |
| nt_fg_d2_ct | Nota de Conteúdo da questão 2 da parte discursiva da formação geral - Convertida para escala de 0 a 100 |
| nt_fg_d2 | Nota da questão 2 da parte discursiva na formação geral - Média ponderada da parte de Língua Portuguesa (20%) e Conteúdo (80%) da Questão 2 da parte discursiva (0 a 100) |
| nt_dis_fg | Nota bruta na parte discursiva da formação geral - Convertida para escala de 0 a 100 |
| nt_fg | Nota bruta na formação geral - Média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral (0 a 100) |
| nt_obj_ce | Nota bruta na parte objetiva do componente específico - Convertida para escala de 0 a 100 |
| nt_ce_d1 | Nota da questão 1 da parte discursiva do componente específico - Convertida para escala de 0 a 100 |
| nt_ce_d2 | Nota da questão 2 da parte discursiva do componente específico - Convertida para escala de 0 a 100 |
| nt_ce_d3 | Nota da questão 3 da parte discursiva do componente específico - Convertida para escala de 0 a 100 |
| nt_dis_ce | Nota bruta na parte discursiva do componente específico - Convertida para escala de 0 a 100 |
| nt_ce | Nota bruta no componente específico - Média ponderada da parte objetiva (85%) e discursiva |

| | |
|---|---|
| | (15%) no componente específico (0 a 100) |
| nt_ger | Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%) (0 a 100) |
| qp_i1 | 1 - Qual o grau de dificuldade desta prova na parte de Formação Geral? |
| qp_i2 | 2 - Qual o grau de dificuldade desta prova na parte do Componente Específico? |
| qp_i3 | 3 - Considerando a extensão da prova, em relação ao tempo total, você considera que a prova foi: |
| qp_i4 | 4 - Os enunciados das questões da prova na parte de Formação Geral estavam claros e objetivos? |
| qp_i5 | 5 - Os enunciados das questões na parte do Componente Específico estavam claros e objetivos? |
| qp_i6 | 6 - As informações/instruções fornecidas para a resolução das questões foram suficientes para resolvê-las? |
| qp_i7 | 7 - Você se deparou com alguma dificuldade ao responder à prova. Qual? |
| qp_i8 | 8 - Considerando apenas as questões objetivas da prova, você percebeu que: |
| qp_i9 | 9 - Qual foi o tempo gasto por você para concluir a prova? |
| qe_i1 | Item 1 do Questionário do Estudante |
| qe_i2 | Item 2 do Questionário do Estudante |
| qe_i3 | Item 3 do Questionário do Estudante |
| qe_i4 | Item 4 do Questionário do Estudante |
| qe_i5 | Item 5 do Questionário do Estudante |
| qe_i6 | Item 6 do Questionário do Estudante |
| qe_i7 | Item 7 do Questionário do Estudante |
| qe_i8 | Item 8 do Questionário do Estudante |
| qe_i9 | Item 9 do Questionário do Estudante |
| qe_i10 | Item 10 do Questionário do Estudante |
| qe_i11 | Item 11 do Questionário do Estudante |
| qe_i12 | Item 12 do Questionário do Estudante |
| qe_i13 | Item 13 do Questionário do Estudante |
| qe_i14 | Item 14 do Questionário do Estudante |
| qe_i15 | Item 15 do Questionário do Estudante |
| qe_i16 | Item 16 do Questionário do Estudante |
| qe_i17 | Item 17 do Questionário do Estudante |
| qe_i18 | Item 18 do Questionário do Estudante |
| qe_i19 | Item 19 do Questionário do Estudante |
| qe_i20 | Item 20 do Questionário do Estudante |
| qe_i21 | Item 21 do Questionário do Estudante |
| qe_i22 | Item 22 do Questionário do Estudante |
| qe_i23 | Item 23 do Questionário do Estudante |
| qe_i24 | Item 24 do Questionário do Estudante |
| qe_i25 | Item 25 do Questionário do Estudante |
| qe_i26 | Item 26 do Questionário do Estudante |
| qe_i27 | Item 27 do Questionário do Estudante |

| | |
|---|---|
| qe_i28 | Item 28 do Questionário do Estudante |
| qe_i29 | Item 29 do Questionário do Estudante |
| qe_i30 | Item 30 do Questionário do Estudante |
| qe_i31 | Item 31 do Questionário do Estudante |
| qe_i32 | Item 32 do Questionário do Estudante |
| qe_i33 | Item 33 do Questionário do Estudante |
| qe_i34 | Item 34 do Questionário do Estudante |
| qe_i35 | Item 35 do Questionário do Estudante |
| qe_i36 | Item 36 do Questionário do Estudante |
| qe_i37 | Item 37 do Questionário do Estudante |
| qe_i38 | Item 38 do Questionário do Estudante |
| qe_i39 | Item 39 do Questionário do Estudante |
| qe_i40 | Item 40 do Questionário do Estudante |
| qe_i41 | Item 41 do Questionário do Estudante |
| qe_i42 | Item 42 do Questionário do Estudante |
| qe_i43 | Item 43 do Questionário do Estudante |
| qe_i44 | Item 44 do Questionário do Estudante |
| qe_i45 | Item 45 do Questionário do Estudante |
| qe_i46 | Item 46 do Questionário do Estudante |
| qe_i47 | Item 47 do Questionário do Estudante |
| qe_i48 | Item 48 do Questionário do Estudante |
| qe_i49 | Item 49 do Questionário do Estudante |
| qe_i50 | Item 50 do Questionário do Estudante |
| qe_i51 | Item 51 do Questionário do Estudante |
| qe_i52 | Item 52 do Questionário do Estudante |
| qe_i53 | Item 53 do Questionário do Estudante |
| qe_i54 | Item 54 do Questionário do Estudante |
| qe_i55 | Item 55 do Questionário do Estudante |
| qe_i56 | Item 56 do Questionário do Estudante |
| qe_i57 | Item 57 do Questionário do Estudante |
| qe_i58 | Item 58 do Questionário do Estudante |
| qe_i59 | Item 59 do Questionário do Estudante |
| qe_i60 | Item 60 do Questionário do Estudante |
| qe_i61 | Item 61 do Questionário do Estudante |
| qe_i62 | Item 62 do Questionário do Estudante |
| qe_i63 | Item 63 do Questionário do Estudante |
| qe_i64 | Item 64 do Questionário do Estudante |
| qe_i65 | Item 65 do Questionário do Estudante |
| qe_i66 | Item 66 do Questionário do Estudante |
| qe_i67 | Item 67 do Questionário do Estudante |
| qe_i68 | Item 68 do Questionário do Estudante |

# ANEXO B – QUESTIONÁRIO DO ESTUDANTE

1. Qual o seu estado civil?

A ( ) Solteiro(a).

B ( ) Casado(a).

C ( ) Separado(a) judicialmente/divorciado(a).

D ( ) Viúvo(a).

E ( ) Outro.

2. Como você se considera?

A ( ) Branco(a).

B ( ) Negro(a).

C ( ) Pardo(a)/mulato(a).

D ( ) Amarelo(a) (de origem oriental).

E ( ) Indígena ou de origem indígena.

3. Qual a sua nacionalidade?

A ( ) Brasileira.

B ( ) Brasileira naturalizada.

C ( ) Estrangeira.

4. Até que etapa de escolarização seu pai concluiu?

A ( ) Nenhuma.

B ( ) Ensino Fundamental: 1o ao 5o ano (1a a 4a série).

C ( ) Ensino Fundamental: 6o ao 9o ano (5a a 8a série).

D ( ) Ensino Médio.

E ( ) Ensino Superior - Graduação.

F ( ) Pós-graduação.

5. Até que etapa de escolarização sua mãe concluiu?

A ( ) Nenhuma.

B ( ) Ensino fundamental: 1o ao 5o ano (1a a 4a série).

C ( ) Ensino fundamental: 6o ao 9o ano (5a a 8a série).

D ( ) Ensino médio.

E ( ) Ensino Superior - Graduação.

F ( ) Pós-graduação.

6. Onde e com quem você mora atualmente?

A ( ) Em casa ou apartamento, sozinho.

B ( ) Em casa ou apartamento, com pais e/ou parentes.

C ( ) Em casa ou apartamento, com cônjuge e/ou filhos.

D ( ) Em casa ou apartamento, com outras pessoas (incluindo república).

E ( ) Em alojamento universitário da própria instituição.

F ( ) Em outros tipos de habitação individual ou coletiva (hotel, hospedaria, pensão ou outro).

7. Quantas pessoas da sua família moram com você? Considere seus pais, irmãos, cônjuge, filhos e outros

parentes que moram na mesma casa com você.

A ( ) Nenhuma.

B ( ) Uma.

C ( ) Duas.

D ( ) Três.

E ( ) Quatro.

F ( ) Cinco.

G ( ) Seis.

H ( ) Sete ou mais.

8. Qual a renda total de sua família, incluindo seus rendimentos?

A ( ) Até 1,5 salário mínimo (até R$ 1.086,00).

B ( ) De 1,5 a 3 salários mínimos (R$ 1.086,01 a R$ 2.172,00).

C ( ) De 3 a 4,5 salários mínimos (R$ 2.172,01 a R$ 3.258,00).

D ( ) De 4,5 a 6 salários mínimos (R$ 3.258,01 a R$ 4.344,00).

E ( ) De 6 a 10 salários mínimos (R$ 4.344,01 a R$ 7.240,00).

F ( ) De 10 a 30 salários mínimos (R$ 7.240,01 a R$ 21.720,00).

G ( ) Acima de 30 salários mínimos (mais de R$ 21.720,01).

9. Qual alternativa a seguir melhor descreve sua situação financeira (incluindo bolsas)?

A ( ) Não tenho renda e meus gastos são financiados por programas governamentais.

B ( ) Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas.

C ( ) Tenho renda, mas recebo ajuda da família ou de outras pessoas para financiar meus gastos.

D ( ) Tenho renda e não preciso de ajuda para financiar meus gastos.

E ( ) Tenho renda e contribuo com o sustento da família.

F ( ) Sou o principal responsável pelo sustento da família.

10. Qual alternativa a seguir melhor descreve sua situação de trabalho (exceto estágio ou bolsas)?

A ( ) Não estou trabalhando.

B ( ) Trabalho eventualmente.

C ( ) Trabalho até 20 horas semanais.

D ( ) Trabalho de 21 a 39 horas semanais.

E ( ) Trabalho 40 horas semanais ou mais.

11. Que tipo de bolsa de estudos ou financiamento do curso você recebeu para custear todas ou a maior parte das mensalidades? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.

A ( ) Nenhum, pois meu curso é gratuito.

B ( ) Nenhum, embora meu curso não seja gratuito.

C ( ) ProUni integral.

D ( ) ProUni parcial, apenas.

E ( ) FIES, apenas.

F ( ) ProUni Parcial e FIES.

G ( ) Bolsa oferecida por governo estadual, distrital ou municipal.

H ( ) Bolsa oferecida pela própria instituição.

I ( ) Bolsa oferecida por outra entidade (empresa, ONG, outra).

J ( ) Financiamento oferecido pela própria instituição.

K ( ) Financiamento bancário.

12. Ao longo da sua trajetória acadêmica, você recebeu algum tipo de auxílio permanência? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.

A ( ) Nenhum.

B ( ) Auxílio moradia.

C ( ) Auxílio alimentação.

D ( ) Auxílio moradia e alimentação.

E ( ) Auxílio Permanência.

F ( ) Outro tipo de auxílio.

13. Ao longo da sua trajetória acadêmica, você recebeu algum tipo de bolsa acadêmica? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.

A ( ) Nenhum.

B ( ) Bolsa de iniciação científica.

C ( ) Bolsa de extensão.

D ( ) Bolsa de monitoria/tutoria.

E ( ) Bolsa PET.

F ( ) Outro tipo de bolsa acadêmica.

14. Durante o curso de graduação você participou de programas e/ou atividades curriculares no exterior?

A ( ) Não participei.

B ( ) Sim, Programa Ciência sem Fronteiras.

C ( ) Sim, programa de intercâmbio financiado pelo Governo Federal (Marca; Brafitec; PLI; outro).

D ( ) Sim, programa de intercâmbio financiado pelo Governo Estadual.

E ( ) Sim, programa de intercâmbio da minha instituição.

F ( ) Sim, outro intercâmbio não institucional.

15. Seu ingresso no curso de graduação se deu por meio de políticas de ação afirmativa ou inclusão social?

A ( ) Não.

B ( ) Sim, por critério étnico-racial.

C ( ) Sim, por critério de renda.

D ( ) Sim, por ter estudado em escola pública ou particular com bolsa de estudos.

E ( ) Sim, por sistema que combina dois ou mais critérios anteriores.

F ( ) Sim, por sistema diferente dos anteriores.

16. Em que unidade da Federação você concluiu o ensino médio?

( ) AC ( ) DF ( ) MT ( ) RJ ( ) SE

( ) AL ( ) ES ( ) PA ( ) RN ( ) SP

( ) AM ( ) GO ( ) PB ( ) RO ( ) TO

( ) AP ( ) MA ( ) PE ( ) RR ( ) Não se aplicada

( ) BA ( ) MG ( ) PI ( ) RS

( ) CE ( ) MS ( ) PR ( ) SC

17. Em que tipo de escola você cursou o ensino médio?

A ( ) Todo em escola pública.

B ( ) Todo em escola privada (particular).

C ( ) Todo no exterior.

D ( ) A maior parte em escola pública.

E ( ) A maior parte em escola privada (particular).

F ( ) Parte no Brasil e parte no exterior.

18. Qual modalidade de ensino médio você concluiu?

A ( ) Ensino médio tradicional.

B ( ) Profissionalizante técnico (eletrônica, contabilidade, agrícola, outro).

C ( ) Profissionalizante magistério (Curso Normal).

D ( ) Educação de Jovens e Adultos (EJA) e/ou Supletivo.

E ( ) Outra modalidade.

19. Quem lhe deu maior incentivo para cursar a graduação?

A ( ) Ninguém.

B ( ) Pais.

C ( ) Outros membros da família que não os pais.

D ( ) Professores.

E ( ) Líder ou representante religioso.

F ( ) Colegas/Amigos.

G ( ) Outras pessoas.

20. Algum dos grupos abaixo foi determinante para você enfrentar dificuldades durante seu curso superior e

concluí-lo?

A ( ) Não tive dificuldade.

B ( ) Não recebi apoio para enfrentar dificuldades.

C ( ) Pais.

D ( ) Avós.

E ( ) Irmãos, primos ou tios.

F ( ) Líder ou representante religioso.

G ( ) Colegas de curso ou amigos.

H ( ) Professores do curso.

I ( ) Profissionais do serviço de apoio ao estudante da IES.

J ( ) Colegas de trabalho.

K ( ) Outro grupo.

21. Alguém em sua família concluiu um curso superior?

A ( ) Sim.

B ( ) Não.

22. Excetuando-se os livros indicados na bibliografia do seu curso, quantos livros você leu neste ano?

A ( ) Nenhum.

B ( ) Um ou dois.

C ( ) De três a cinco.

D ( ) De seis a oito.

E ( ) Mais de oito.

23. Quantas horas por semana, aproximadamente, você dedicou aos estudos, excetuando as horas de aula?

A ( ) Nenhuma, apenas assisto às aulas.

B ( ) De uma a três.

C ( ) De quatro a sete.

D ( ) De oito a doze.

E ( ) Mais de doze.

24. Você teve oportunidade de aprendizado de idioma estrangeiro na Instituição?

A ( ) Sim, somente na modalidade presencial.

B ( ) Sim, somente na modalidade semipresencial.

C ( ) Sim, parte na modalidade presencial e parte na modalidade semipresencial.

D ( ) Sim, na modalidade a distância.

E ( ) Não.

25. Qual o principal motivo para você ter escolhido este curso?

A ( ) Inserção no mercado de trabalho.

B ( ) Influência familiar.

C ( ) Valorização profissional.

D ( ) Prestígio Social.

E ( ) Vocação

F ( ) Oferecido na modalidade a distância.

G ( ) Baixa concorrência para ingresso.

H ( ) Outro motivo.

26. Qual a principal razão para você ter escolhido a sua instituição de educação superior?

A ( ) Gratuidade.

B ( ) Preço da mensalidade.

C ( ) Proximidade da minha residência.

D ( ) Proximidade do meu trabalho.

E ( ) Facilidade de acesso.

F ( ) Qualidade/reputação.

G ( ) Foi a única onde tive aprovação.

H ( ) Possibilidade de ter bolsa de estudo.

I ( ) Outro motivo.

27. As disciplinas cursadas contribuíram para sua formação integral, como cidadão e profissional.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

28. Os conteúdos abordados nas disciplinas do curso favoreceram sua atuação em estágios ou em atividades de iniciação profissional.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

29. As metodologias de ensino utilizadas no curso desafiaram você a aprofundar conhecimentos e desenvolver competências reflexivas e críticas.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

30. O curso propiciou experiências de aprendizagem inovadoras.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

31. O curso contribuiu para o desenvolvimento da sua consciência ética para o exercício profissional.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

32. No curso você teve oportunidade de aprender a trabalhar em equipe.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

33. O curso possibilitou aumentar sua capacidade de reflexão e argumentação.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

34. O curso promoveu o desenvolvimento da sua capacidade de pensar criticamente, analisar e refletir sobre soluções para problemas da sociedade.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

35. O curso contribuiu para você ampliar sua capacidade de comunicação nas formas oral e escrita.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

36. O curso contribuiu para o desenvolvimento da sua capacidade de aprender e atualizar-se permanentemente.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

37. As relações professor-aluno ao longo do curso estimularam você a estudar e aprender.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

38. Os planos de ensino apresentados pelos professores contribuíram para o desenvolvimento das atividades acadêmicas e para seus estudos.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

39. As referências bibliográficas indicadas pelos professores nos planos de ensino contribuíram para seus estudos e aprendizagens.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

40. Foram oferecidas oportunidades para os estudantes superarem dificuldades relacionadas ao processo de formação.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

41. A coordenação do curso esteve disponível para orientação acadêmica dos estudantes.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

42. O curso exigiu de você organização e dedicação frequente aos estudos.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

43. Foram oferecidas oportunidades para os estudantes participarem de programas, projetos ou atividades de extensão universitária.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

44. Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

45. O curso ofereceu condições para os estudantes participarem de eventos internos e/ou externos à instituição.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

46. A instituição ofereceu oportunidades para os estudantes atuarem como representantes em órgãos colegiados.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

47. O curso favoreceu a articulação do conhecimento teórico com atividades práticas.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

48. As atividades práticas foram suficientes para relacionar os conteúdos do curso com a prática, contribuindo para sua formação profissional.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

49. O curso propiciou acesso a conhecimentos atualizados e/ou contemporâneos em sua área de formação.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

50. O estágio supervisionado proporcionou experiências diversificadas para a sua formação.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

51. As atividades realizadas durante seu trabalho de conclusão de curso contribuíram para qualificar sua formação profissional.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

52. Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios no país.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

53. Foram oferecidas oportunidades para os estudantes realizarem intercâmbios e/ou estágios fora do país.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

54. Os estudantes participaram de avaliações periódicas do curso (disciplinas, atuação dos professores, infraestrutura).

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

55. As avaliações da aprendizagem realizadas durante o curso foram compatíveis com os conteúdos ou temas trabalhados pelos professores.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

56. Os professores apresentaram disponibilidade para atender os estudantes fora do horário das aulas.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

57. Os professores demonstraram domínio dos conteúdos abordados nas disciplinas.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

58. Os professores utilizaram tecnologias da informação e comunicação (TICs) como estratégia de ensino (projetor multimídia, laboratório de informática, ambiente virtual de aprendizagem).

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

59. A instituição dispôs de quantidade suficiente de funcionários para o apoio administrativo e acadêmico.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

60. O curso disponibilizou monitores ou tutores para auxiliar os estudantes.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

61. As condições de infraestrutura das salas de aula foram adequadas.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

62. Os equipamentos e materiais disponíveis para as aulas práticas foram adequados para a quantidade de estudantes.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

63. Os ambientes e equipamentos destinados às aulas práticas foram adequados ao curso.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

64. A biblioteca dispôs das referências bibliográficas que os estudantes necessitaram.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

65. A instituição contou com biblioteca virtual ou conferiu acesso a obras disponíveis em acervos virtuais.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

66. As atividades acadêmicas desenvolvidas dentro e fora da sala de aula possibilitaram reflexão, convivência e respeito à diversidade.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

67. A instituição promoveu atividades de cultura, de lazer e de interação social.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica

68. A instituição dispôs de refeitório, cantina e banheiros em condições adequadas que atenderam as necessidades dos seus usuários.

1 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ 2 ◯ ( ) Não sei responder ( ) Não se aplica