UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Euclides Francisco António Amaro

Os Algoritmos C4.5 e *Hoeffding Tree* Aplicados a Mineração de Dados Educacionais Referente ao Exame Nacional de Desempenho de Estudante (ENADE) em Ciência da Computação

Criciúma
2019

Euclides Francisco António Amaro

Os Algoritmos C4.5 e *Hoeffding Tree* Aplicados a Mineração de Dados Educacionais Referente Ao Exame Nacional de Desempenho de Estudante (ENADE) em Ciência da Computação

Trabalho de Conclusão de Curso, apresentado para obtenção do grau de Bacharel no curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientadora: Profa. Dra. Merisandra Côrtes de Mattos Garcia

Criciúma

2019

EUCLIDES FRANCISCO ANTÓNIO AMARO

OS ALGORITMOS C4.5 E *HOEFFDING TREE* APLICADOS A MINERAÇÃO DE DADOS EDUCACIONAL REFERENTE AO EXAME NACIONAL DE DESEMPENHO DE ESTUDANTE (ENADE) EM CIÊNCIA DA COMPUTAÇÃO

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Inteligência Artificial.

Criciúma, 10 de dezembro de 2019

BANCA EXAMINADORA

Profa. Merisandra Côrtes de Mattos - Doutora - (UNESC) - Orientador

Profa. Ana Claudia Garcia Barbosa - Mestra - (UNESC)

Prof. Kristian Madeira - Doutor - (UNESC)

Decido este trabalho, aos meus pais, Paulo Júnior e Mariana Amaro, meu irmão Eurico Amaro e aos meus amigos que sempre me apoiaram pelo apoio e pela inspiração.

AGRADECIMENTOS

A Deus por ter me dado saúde, força foco e fé para supera todas dificuldades encontradas ao decorrer curso.

Agradeço os meus pais e toda minha família e especialmente aqueles que depositaram a confiança em mim.

Agradecer a minha orientadora, Professora Doutora Merisandra Cortês de Mattos Garcia e todos amigos que me deram força e coragem nesta fase de luta.

“Educar é crescer. E crescer é viver. Educação é, assim, vida no sentido mais autêntico da palavra.”

Anísio Teixeira

Resumo

Ao decorrer da globalização e a alta demanda de informações, surgiu-se a necessidade de armazenamento das mesmas. A partir deste contexto, observa-se o *data science,* que compreende as etapas pertinentes à limpeza, elaboração e análise de dados com mecanismos aplicados a fim de se extrair dados e obter intuições por meio de informações da base de dados. O Exame Nacional de Avaliação do Estudante (ENADE) tem como objetivo avaliar o grau dos conhecimentos dos estudantes, referentes aos conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos, a partir do desempenho individual destes no Exame. Nesta pesquisa realizou-se *mineração de dados* educacionais por meio da tarefa de classificação, a partir do método de indução de árvores de decisão, empregando-se os algoritmos C4,5 e *Hoeffding Tree.* Os dados estudados foram extraídos do ENADE do Curso de Ciência da Computação das três bases: Universidade do Extremo Sul Catarinense, Associação Catarinense das Fundações Educacionais e Santa Catarina. A base de dados analisada encontra-se disponível no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira e possui dados referentes às provas da área de Ciência da Computação dos anos de 2011, 2014 e 2017. Após a execução da mineração de dados dos modelos obtidos, estes foram analisados por meio das medidas de qualidade, como a acurácia, a fim de se identificar qual dos dois algoritmos gerou o melhor modelo. A base que se destaca com o classificador que tem o melhor resultado é a base da Universidade do Extremo Sul Catarinense com o algoritmo C4.5 e valor da acurácia 98,79.

Palavras chaves: ENADE; *Data Scienc*e; Mineração de dados; Agoritmo C4.5 e *Hoeffding Tree*.

Abstract

In the course of globalization and the high demand for information, the need for storage of information arose. From this context, one observes data science, which includes anything pertinent to data cleaning, elaboration and analysis with mechanisms applied in order to extract data and obtain intuitions through information from the database. The National Student Assessment Exam (ENADE) aims to assess the degree of students' knowledge regarding the syllabus provided in the curriculum guidelines of their respective courses, based on their individual performance in the Exam. In this research, educational data were mining through the classification task, using the decision tree induction method, using the C4.5 and Hoeffding Tree algorithms. The data studied were extracted from the ENADE of the Computer Science Course of the three bases: University of the Southern Santa Catarina, Santa Catarina Association of Educational Foundations and Santa Catarina. The database analyzed is available on the website of the Anísio Teixeira National Institute for Educational Studies and Research and has data related to the Computer Science tests of 2011, 2014 and 2017. After the data mining of the obtained models, these were analyzed through quality measures, such as accuracy, in order to identify which of the two algorithms generated the best model. The base that stands out with the classifier that has the best result is the base of the University of Far South Catarinense with the C4.5 algorithm and accuracy value 98.79.

Keywords: ENADE; Data Science; Data mining; C4.5 Agorithm and Hoeffding Tree.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACAFE	Associação Catarinense das Fundações Educacionais
DCBD	Descoberta de Conhecimentos em Bases de Dados
DM	Mineração de dados
ENADE	Exame Nacional de Desempenho de Estudante
INEP	Instituto Nacional de Pesquisa Educacional Anísio Teixeira
KDD	*Knowledge Discovery in Databases*
FN	Falso Negativo
FP	Falso Positivo
PDE	Plano de Desenvolvimento da Educação
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Com vastas quantidades de dados disponíveis, as empresas em quase todos os setores estão focadas na exploração de dados. O volume e a variedade de dados ultrapassou em muito a capacidade da análise manual e, em alguns casos excederam a capacidade dos bancos de dados convencionais. Ao mesmo tempo, os computadores se tornaram muito mais poderosos, a rede é onipresente e foram desenvolvidos algoritmos que podem conectar conjuntos de dados para permitir uma ampliação das análises do que anteriormente era possível. A convergência desses fenômenos deram origem a cada vez mais difundida aplicação comercial da *data science* (PROVOST; FAWCETT, 2013). Os avanços da tecnologia de informação proporcionaram as empresas o armazenamento de uma ampla quantidade de dados (TAN; STEINBACH; KUMAR, 2009), com isso inviabilizou-se a análise destes dados por meio dos métodos e das técnicas tradicionais (WITTEN; FRANK, 2005, tradução nossa). Em meados da década de 90, especialmente em 1996, originou-se a Descoberta de Conhecimento em Bases de Dados composta pelo conceito de mineração de dados, que se refere a aplicação de métodos específicos para identificar padrões em um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

A mineração de dados reúne diferentes áreas do conhecimento como a estatística, aprendizado de máquina, reconhecimento de padrões, computação evolutiva, teoria da informação, processamento de sinais, visualização e recuperação de dados (TAN; STEINBACH; KUMAR, 2009). A mineração de dados pode ser aplicada em diferentes áreas do conhecimento, como por exemplo, no que se refere a dados educacionais.

A mineração de dados educacionais é um campo de pesquisa interdisciplinar que envolve a evolução de métodos na exploração de dados originais de uma instituição educacional (SILVA; SILVA, 2014), empenhando-se nas aplicações das ferramentas e técnicas de mineração de dados (PATEL, 2017). A sua aplicação é essencial para a Educação, no que se refere ao futuro da aprendizagem e aperfeiçoamento nesta área (NETO; CASTRO, 2017), a fim de se explorar estes dados em busca de novos conhecimentos (AYUB et al., 2017 tradução nossa). Assim, pode-se analisar se os métodos educacionais empregados têm gerado bons resultados (BAKER; ISOTANI; CARVALHO, 2011).

No Brasil, o Instituto Nacional de Pesquisa Educacional Anísio Teixeira (INEP), vinculado ao Ministério da Educação, tem como objetivo auxiliar na produção de políticas educacionais nos diferentes níveis de ensino, em busca de uma educação de qualidade com objetivo de desenvolver o país econômica e socialmente (INEP, 2018).

Com base no Exame Nacional de Desempenho de Estudante (ENADE) que é uma prova para avaliar os estudantes concluintes de um curso de graduação, com base nos conteúdos previstos nas diretrizes curriculares e apreendidos na formação acadêmica, avaliando a evolução destes estudantes, o desenvolvimento das competências e habilidades em sua área de formação, bem como o conhecimento da realidade do Brasil e do mundo. Esta avaliação é um dos itens que compõe o Sistema Nacional de Avaliação da Educação Superior (SINAES) (ENADE, 2018).

O ENADE iniciou-se em 2004, sendo separado por áreas de conhecimento, as quais são avaliadas a cada três anos. No primeiro ano os cursos da área da saúde realizam as provas, no segundo ano as ciências exatas e licenciaturas, e o terceiro ano é integrado pelas áreas de ciências sociais aplicadas e ciências humanas (CRETTON; GOMES, 2016).

Alguns estudos têm sido realizados envolvendo a aplicação da mineração de dados em dados do ENADE, como por exemplo, a pesquisa de Cretton e Gomes (2016) que aplica técnicas de mineração de dados por meio de classificação na base de dados do ENADE com enfoque nos cursos de Medicina para adquirir conhecimento sobre as questões respondidas e o nível de dificuldade encontrado no componente específico da avaliação; Vista, Fiqueiró e Mozzaquatro (2017) aplicam mineração de dados por meio da tarefa de agrupamento e do algoritmo *K-Means* nos microdados do ENADE para examinar o desenvolvimento dos acadêmicos nos cursos de Ciência da Computação no Rio Grande do Sul.

A mineração dos dados ocorre por meio de diferentes tarefas, dentre elas: associação, regressão, previsão de séries temporais, detecção de desvios, agrupamento e classificação (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A tarefa de classificação tem a característica de reconhecer um determinado registro e dizer a qual classe ele pertence (CAMILO; SILVA, 2009). A classificação pode ser compreendida como o aprendizado de uma função *f* que mapeia cada conjunto de atributos *x* em um rótulo de classe *y* pré-definido (TAN;

STEINBACH; KUMAR, 2009). Existem vários métodos de classificação, dentre eles: redes neurais artificiais, bayesianos e indução de árvores de decisão.

A indução de árvores de decisão, conforme Tan, Steinbach e Kumar (2009), é um método de classificação bastante usado e de acordo com Dejaeger (2011) e empregado na exploração de conhecimentos em instituições acadêmicas. As árvores de decisão são geradas por nós, onde os nós não terminais executam uma determinada ação de escolha de decisão (MAIA; GOMES; CHAGAS, 2017). Dentre os algoritmos de indução de árvore de decisão tem-se o C4.5 e o *Hoeffding tree* (LEMOS; STEINER; NIEVOLA, 2005). O algoritmo C4.5 cria um modelo de decisão partindo do nó pai, de modo que cada um dos nós possa ser examinado individualmente para determinar a relevância de sua relação ou a existência dela (CRETTON; GOMES, 2016). Além disso, o algoritmo C4.5 sugere métodos baseados em conceitos de informações que dependem de hipóteses (VELHO et al., 2009).

O Algoritmo de *Hoeffding Tree*, também conhecido como *Very Fast Decision Tree* (VFDT), é um tipo de algoritmo que baseia a indução da árvore de decisão em um processo de aprendizagem, que recebendo sequencialmente cada instância de treino e construindo recursivamente a árvore por meio da substituição folhadas por nós de decisão. (MENEZES; ZAVERUCHA, 2011).

Tendo em vista os aspectos salientados, propõe-se nesta pesquisa realizar a mineração de dados educacionais por meio da tarefa de classificação pelo método de indução de árvores de decisão, empregando-se os algoritmos C4.5 e *Hoeffding Tree,* nos dados do ENADE do curso de Ciência da Computação da UNESC, ACAFE e Santa Catarina, os dados encontram-se disponíveis na base do INEP.

1.1 OBJETIVO GERAL

Disponibilizar um modelo de classificação dos dados do ENADE referente ao Curso de Ciência da Computação da UNESC, ACAFE e Santa Catarina.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta pesquisa constituem em:

a) descrever os conceitos de mineração de dados, mineração de dados educacionais, classificação, indução de árvores de decisão e os algoritmos C4.5 e *Hoeffding Tree*;
b) aplicar a tarefa de classificação pelo método de indução de árvores de decisão;
c) Utilizar os dados da prova do ENADE em Ciência da Computação do sul Catarinense.
d) empregar os algoritmos classificadores C4.5 e *Hoeffding Tree*;
e) comparar os modelos obtidos pelos algoritmos C4.5 e *Hoeffding Tree* por meio de medidas de qualidade em mineração de dados.

1.3 JUSTIFICATIVA

A mineração de dados distingue-se por absorver grandes quantidades de dados, com a finalidade de descobrir conhecimentos úteis e também de prever resultados futuros (TAN; STEINBACH; KUMAR, 2009). A mineração de dados educacionais propõe mais qualidade de busca de resposta sobre os dados na área da educação (BAKER; ISOTANI; CARVALHO, 2011). Para isso, procura analisar características do estudante para criar informações e avisos que sejam capazes de auxiliar os professores no processo de tomada de decisões, o que pode contribuir na atenuação dos problemas educacionais (FERREIRA, 2015).

Dentre os métodos de classificação o que emprega a indução de árvores de decisão, facilita a compreensão dos modelos obtidos, devido a sua configuração no formato de árvore. Constituindo-se em um método eficaz para examinar problemas de classificação (CRETTON; GOMES, 2016).

Este método destaca-se por ir além do conjunto que foi abordado nos treinamentos, gerando uma árvore com capacidade de analisar outro objeto. Emprega a estratégia de dividir–para-conquistar, transformando um problema complexo em problemas mais simples (CARVALHO, 2005). A árvore de decisão tem a habilidade de possibilitar ao usuário definir o objeto de saída, a partir de uma quantidade de

dados, identificando as causas mais influentes que estão ligadas ao objeto (PELEGRIN et al., 2006).

O algoritmo C4.5 toma conhecimentos importantes para a elaboração de uma árvore de decisão a partir de uma base de dados, sendo capaz de apresentar os conhecimentos para a tomada de decisão (LORENZETT; TELÖCKEN, 2016). O algoritmo *Hoeffding Tree* tem algumas características que melhoram o seu desempenho, sendo capaz de aprender com as classes desequilibradas e erro de classificação assimétrico, com base na sua natureza incremental o algoritmo *Hoeffding* não está adequado ou influencia a qualidade da árvore (DOMINGOS; ULTEN, 2011 nossa tradução).

O ENADE tem a capacidade de avaliar a evolução dos estudantes no que diz respeito aos conteúdos presentes nas Diretrizes Curriculares dos Cursos de Graduação, analisando os conhecimentos e as habilidades dos estudantes em termos de formação geral e profissional (SILVA, 2016).

A análise dos dados do ENADE em Ciência da Computação é importante, pois pode proporcionar no auxílio em qualidade nas instituições que oferecem este curso, bem como melhorar os pontos em função das dificuldades encontradas pelos alunos (CRETTON; GOMES, 2016).

1.4 ESTRUTURA DO TRABALHO

Este trabalho está dividido em uma ordenação de sete capítulos. Sendo que o primeiro capítulo relata sobre uma breve descrição do tema do trabalho, trazendo a introdução, o objetivo geral e os específicos e ainda a justificativa do mesmo trabalho.

Salienta-se que no segundo capítulo trata dos assuntos relacionados a descoberta do conhecimento, enquadrando os processos de KDD que empregam seleção de dados, pré-processamento dos dados, transformação dos dados, DM e interpretação ou avaliação do conhecimento.

No terceiro capítulo aborda sobre pesquisa teórica referente a classificação: conceitos, indução de árvore de decisão onde incorpora os assuntos sobre como construir uma árvore de decisão e tipo de árvore de decisão que empregam sobre os algoritmos C4.5 e *Hoeffding Tree*.

O quarto capítulo trata sobre as medidas de qualidade para classificadores que versam sobre acurácia, sensibilidade, especificidade, kappa.

E acrescentando no quinto capítulo os trabalhos correlatos ao estudo desenvolvido.

O sexto capítulo traz a estrutura do trabalho desenvolvido: metodologias aplicadas para o desenvolvimento.

No sétimo capítulo, por fim, é apresentada a conclusão do trabalho e sugestões para trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO DE BASE DADOS

No decorrer da globalização e da demanda de informações, teve-se a necessidade de armazenar as informações. Diante deste contexto, nasceu a necessidade de extrair novos conhecimentos. Já no período de 1980 a 1990 surgiu uma nova área da computação que é a Descoberta de Conhecimento em Bases de Dados (DCBD), do inglês *Knowledge Discovery in Databases* (KDD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

KDD compreende os processos que geram conhecimentos a partir de informações de dados narrados na sua principal qualidade de extração não-trivial de conhecimentos ocultamente armazenados em base de dados (SOARES JUNIOR; QUINTELLA, 2005). Constituindo-se em descobertas de padrões e habilidades por análises enormes de conjuntos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

A extração de conhecimento de uma determinada base de dados, através da aplicação do processo de KDD requer o entendimento das diferenças entre os dados, informação e conhecimento, conforme ilustra a figura 1.

Figura 1 - Hierarquia entre os dados, informação e conhecimentos

Fonte: (GOLDSCHMIDT, PASSOS, 2005).

Os dados contidos na base da pirâmide, podem ser entendidos, como itens importantes, captados e guardados por aplicativos de tecnologia de informação (GOLDSHIMIDT; PASSOS, 2005).

Para poder chegar em informação, é essencial distingui-la de dados (SETZER, 2014). A informação é uma ferramenta que se coloca ao dispor das organizações, para tornar-se um integrante essencial da compreensão dos agentes organizacionais que compartilham sobre uma referência, com a possibilidade de gerar novos requisitos internos e externos adequado ao sucesso das organizações (SANTOS; RAMOS, 2009).

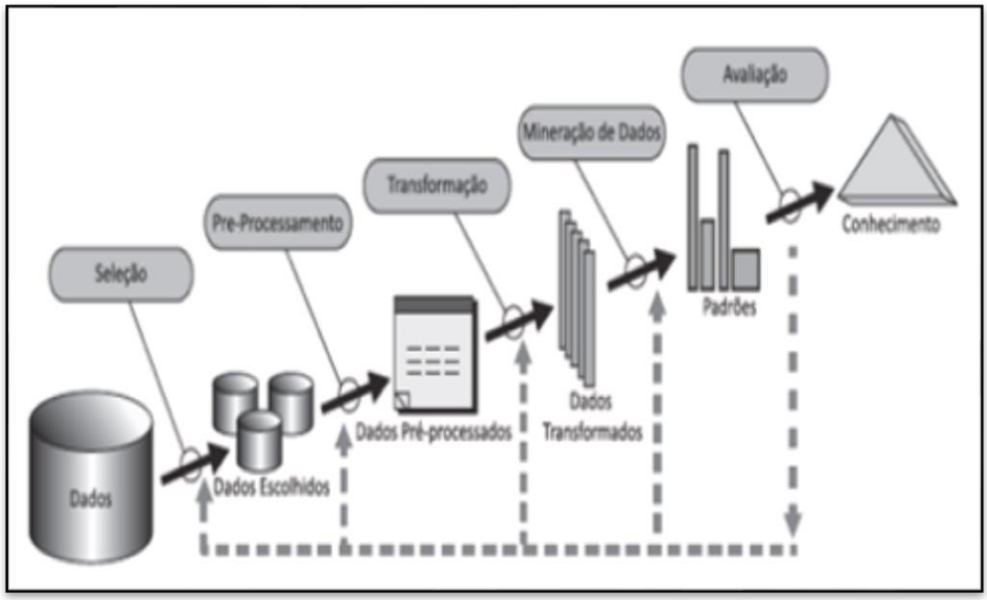
No topo da pirâmide é onde se encontra o conhecimento procurado em determinado padrão ou conjunto de padrões, devido ao relacionamento dados e informações (GOLDSHIMIDT; PASSOS, 2005).

A descoberta de conhecimento em base de dados tem como ligações interdisciplinares, envolvendo diversos campos, entre eles inteligência artificial, matemática, banco de dados, estatística e reconhecimento de padrões. Além disso, o processo de KDD combina técnicas, métodos e algoritmos de todas as áreas com o objetivo principal de extrair conhecimento, a partir de grandes bases de dados (RUBERT; MARIOTTO, 2013).

2.1 PROCESSOS KDD

O processo de KDD é o método que envolve descoberta de conhecimento em modelos de bases de dados com foco de extrair conhecimentos em conjuntos de dados. Também uma fase iterativa, quando o processo de KDD é executado e repetidas várias vezes, até que o resultado seja adequado e satisfatório (SASSI, 2006). O KDD compreende as seguintes etapas: seleção dos dados, pré-processamento dos dados, transformação dos dados, DM e interpretação ou avaliação do conhecimento (figura 2).

Figura 2 – Descoberta do conhecimento



Fonte: Fayyad; Piatetsky-Shapiro; Smyth, (1996).

2.1.1 Seleção de dados

A seleção de dados reporta uma etapa de organização dos dados oriundos de uma fonte de dados. Nem sempre uma fonte de dados é proveniente de um *Data Werehouse*[1] ou Sistema de Banco de dados, pode ser executável a partir de uma extração de dados reunidos em depósito ou repositório, como, biblioteca virtuais, tabelas construídas a partir de questionários, planilhas, pesquisas de internet (COELHO, 2006).

Para esta etapa entrar em ação, são necessários alguns conhecimentos em estrutura de dados e sobre os próprios dados para que os atributos possam ser selecionados de forma estratégica, e assim, aumentar as possiblidades de sucesso no resultado da aplicação da metodologia (ASSEISS, 2017).

2.1.2 Pré-processamento

Os dados a serem trabalhados são em incompletos, inseguros e com padrões ruidosos, o que tornaria o resultado imperfeito (XIONG; PANDEY;

[1] Data Werehouse é considerado como a melhor abordagem para a transformação das grandes quantidades de dados existentes nas organizações em informação útil e confiável, que atenda ao processo de tomada de decisão (MONTEIRO; PINTO; COSTA, 2004).

STEINBACH, 2006). Portanto, deve-se realizar o pré-processamentos que a remoção de ruídos ajustado, tem a capacidade de organizar os dados importantes para estruturar lidar com estratégias de ausência de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa). Além disso é responsável por aumentar a qualidade de dados, e consequentemente aumentam também as medidas de qualideas do modelo gerado (APPEL, 2010).

Um dos campos importantes no pré-processamento de dados é a redução de dados, limpeza de campos vazios, ela auxilia na redução do custo computacional do processo de KDD.

2.1.3 Transformação dos dados

A etapa de transformação dos dados é também conhecida como codificação de dados. Tem como objetivo transformar, simplificar para possibilitar o trabalho com informações relevantes, consolidando o conjunto de dados para as próximas etapas (ASSEISS, 2017). A padronização de dados pode ser normalizada consoante a um determinado intervalo de idade ou faixa de valores, por exemplo: idade {0... 18}. faixa 1; {19… 25}. faixa 2; {26… 30}. Faixa de valores exemplo: -1.0 a 1,0 ou de 0 a 1. (SASSI, 2006).

Essa etapa de transformação nos dados, oferece uma metodologia, para auxilia na descoberta final de conhecimento que é processo de avaliar dos resultados, uma vez que os atributos presentes foram inseridos e simplificados na etapa em que está a ser abordado (ASSEISS, 2017).

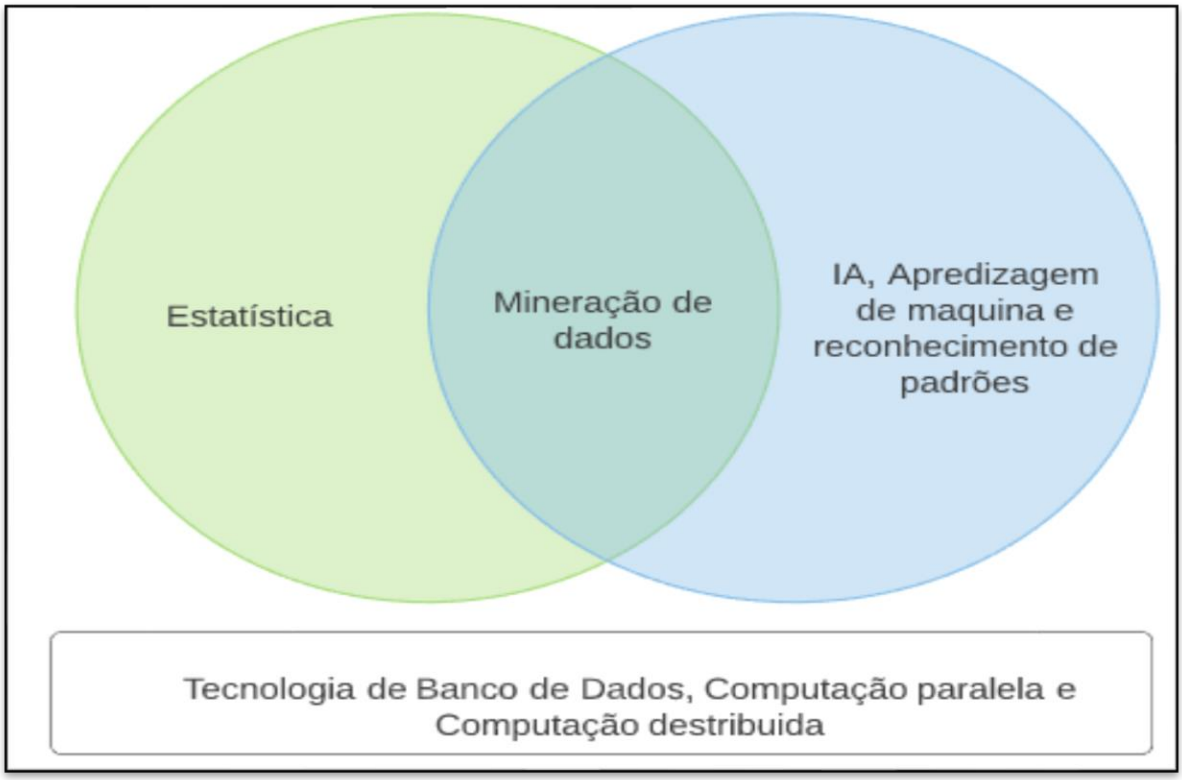
2.1.4 Mineração de dados

Mineração de dados (DM) também conhecida pelo termo em português Mineração de Dados, é um dos componentes principais do KDD, (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa). A mineração de dados emprega métodos tradicionais de análise de dados, bem como de aprendizado de máquina para processar uma vasta quantidade de volume de dados.

O DT tem a capacidade de selecionar padrões e encontrar ligações e tendências entre as informações. Assim as empresas podem antecipar aos concorrentes e surpreender os seus usuários (VELOSO et al., 2011).

O DM reúne vários saberes como algoritmos de busca, reconhecimento de padrões, aprendizagem de máquina, técnicas de modelagem e teorias de aprendizagem da inteligência artificial, amostragem, estimativa e teste de hipótese. Com o crescimento da aplicação do DM surgiu a necessidade de incorporar rapidamente outras áreas como recuperação de informações, processamentos de sinais, computação evolutiva, otimização e teoria de informações (figura 3).

Figura - 3: Mineração de dados como uma confluência.



Fonte: (Alessio, 2004).

2.1.4.1 Mineração de dados educacional

O mineração de dados Educacional (EDM) é um campo de pesquisa interdisciplinar que envolve a evolução de métodos na exploração de dados originais de uma instituição educacional (SILVA; SILVA, 2014), empenhando-se nas aplicações das ferramentas e técnicas de DM (PATEL, 2017). A sua aplicação é essencial para a Educação, no que se refere ao futuro da aprendizagem e ao perfeiçoamento nesta área (NETO; CASTRO, 2017), explorando-se estes dados em busca de novos conhecimentos (AYUB et al., 2017 tradução nossa).

O MDE é uma área que averigua algoritmos estatísticos e de aprendizado de máquina, aplicando sobre os diferentes tipos de dados educacionais, a fim de solucionar questões de pesquisa (ROMERO; VENTURA, 2010).

O MDE pode ser aplicado em qualquer uma das partes envolvidas dos sistemas educacionais, como alunos, educadores, administradores e pesquisadores. A interação de fornecimento de feedback podendo auxiliar nas recomendações de melhorias no processo de aprendizagem dos alunos, no desempenho do ensino, nas atividades propostas e nas tomadas de decisões *(BAKHSHINATEGH et al, 2017,* tradução nossa*).*

2.1.5 Interpretação ou avaliação do conhecimento

É nesta fase também conhecida como pós-processamento, que se realiza uma das maiores inquietações sobre questões de identificar os padrões descobertos na fase de mineração de dados, aqueles que são mais surpreendentes e interessantes ao usuário (CARVALHO et al, 2017), este elementos deve poder armazenar todos os dados relacionados aos elementos e seus resultados, tratá-los de forma adequada, e apresentar de maneira simples (ADAIME, 2005).

Considerando o conhecimento absorvido, podendo ser empregado para soluções do mundo real, seja por meio de um sistema inteligente ou por um ser humano fornecendo apoio em um processo de tomada de decisão (MELANDA, 2004).

3 CLASSIFICAÇÃO: CONCEITOS, INDUÇÃO DE ÁRVORE DE DECISÃO

A classificação pode ser compreendida como o aprendizado de uma função *f* que mapeia cada conjunto de atributos *X* em um rótulo de classe *Y* pré-definido (TAN; STEINBACH; KUMAR, 2009 tradução nossa). Esses modelos são chamados de classificadores que tendo-se rótulos de classe categóricos (discretos, não ordenados). (HAN; KAMBER; PEI, 2012, tradução nossa).

A classificação constitui-se em um métodos de aprendizado de maquina(HAN; KAMBER; PEI, 2012, tradução nossa), que já classificadas, as quais são base para entender um modelo adequado para classificar novas instâncias. É regularrmente referido como um aprendizado inspecionado, devido ao padrão de funcionamento, que quando o algoritmo entra em atuação é supervisionado por meio do precedente conhecimento do resultado da classificação para cada instância utilizada no treinamento (MACIEL et al., 2015).

Com base no algoritmo de classificação tem-se como objetivo encontrar uma relação entre um atributo e uma classe, podendo criar uma regra. Neste contexto a tarefas de classificação pode, aproveitar essa regra para indicar um novo registro para classe (CARVALHO et al., 2012)

As regras de ordenação apresentam uma lista de decisão, a ser executada em sequência. Umas da prioridade da lista é que a regra que aparece primeiro tem a maior preferência para prever a classe (CARVALHO et al., 2012)

Dentre os métodos de classificação o que emprega a indução de árvores de decisão, facilita a compreensão dos modelos obtidos, devido a sua configuração no formato de árvore. Constituindo-se em um método eficaz para examinar problemas de classificação (CRETTON; GOMES, 2016).

Existem vários métodos de aprendizado de máquina, para execução da classificação, dentre eles: redes neurais artificiais, bayesianos e indução de árvores de decisão. O método de indução de árvore de decisão é um dos mais empregados na classificação dando a sua facilidade de entendimento, velocidade e robutez a ruídos (BARROS et al., 2017 tradução nossa). Portanto, a classificação destaca-se por ir além do conjunto que foi abordado no treinamento, gerando uma árvore com capacidade de avaliar outro objeto, emprega a estratégia de dividir–para-conquistar, transformando um problema complexo em problemas mais simples (CARVALHO, 2005).

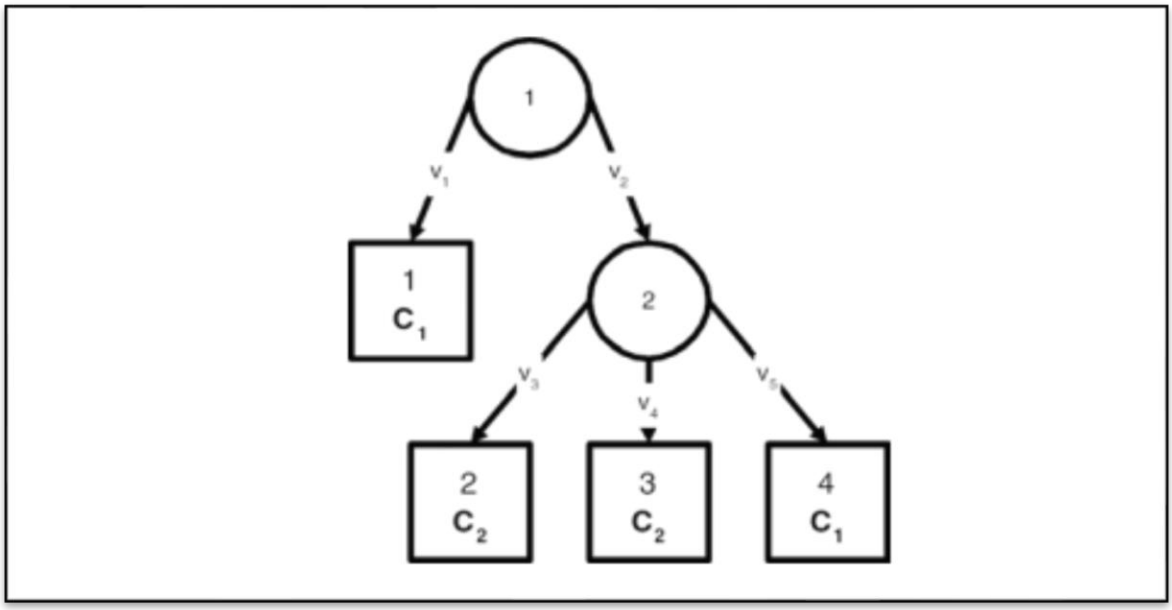
3.1 INDUÇÃO DE ARVORE DE DECISÃO

Árvore de decisão é um método que demostra uma organização de árvore do tipo fluxograma, que tem sido bastante utilizada na representação dos modelos de classificação, devido à sua natureza compreensível que se aproxima do raciocínio humano (BARROS et al., 2017 tradução nossa). É De acordo com Dejaeger (2011) este método é muito empregado na exploração de conhecimentos em instituições acadêmicas. As árvores de decisão são geradas por nós, em que os nós não terminais executam uma determinada ação de escolha de decisão conforme ilustra a figura 4 (MAIA; GOMES; CHAGAS, 2017). As árvores de decisão são exibições gráficas hierárquicas do conhecimento extraído de uma base usada para classificar os dados, relacionando os objetos. Este método e bastente utilizado, (TAN; STEINBACH & KUMAR, 2009, tradução nossa):

- em função da sua facilidade de compreensão, podendo facilmente serem convertidas em um conjunto de regras
- da presença de ruído.

Em geral, os classificadores e árvore de decisão têm boa precisão. No entanto, o uso bem sucedido pode depender dos dados disponíveis. Os algoritmo de Indução de árvore de decisão por exemplo, têm sido usados para classificação em muitas áreas de aplicação, como medicina, fabricação e produção, análise financeira, astronomia e biologia molecular (HAN; KAMBER; PEI, 2012, nossa tradução).

Figura 4 – Árvore de Decisão



Fonte: Alves (2002).

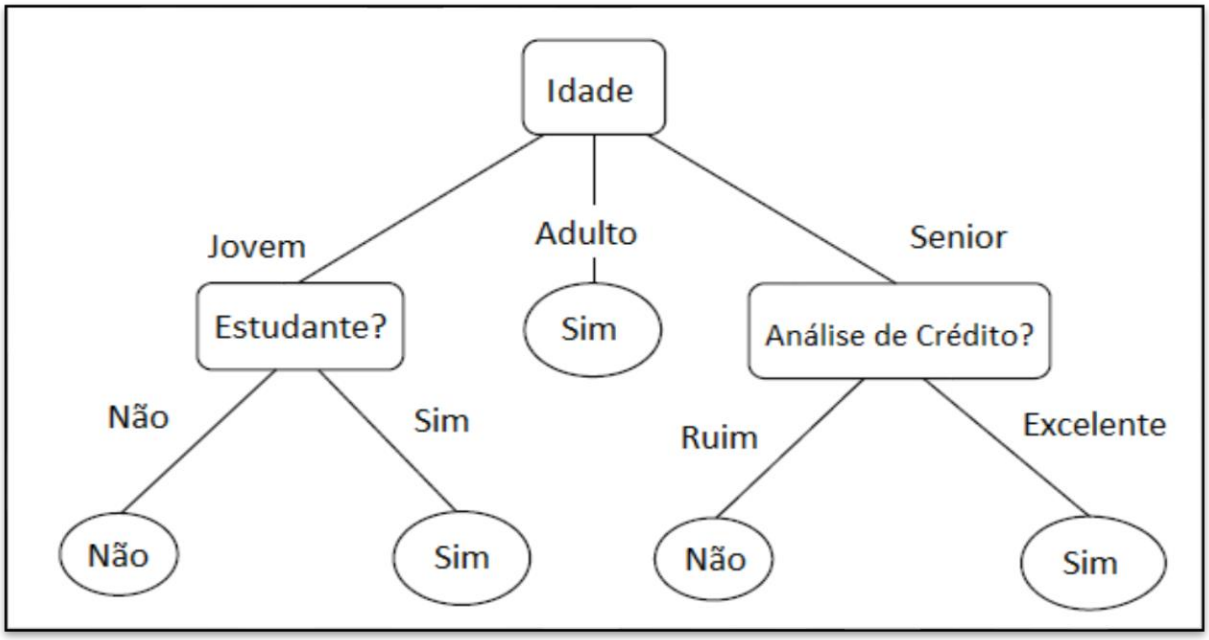
3.1.1 Como construir uma árvore de decisão

A construção da árvore de decisão consiste em uma estrutura apresentam os nós a serem divididos em conjuntos disjuntos de um atributo a ser recebido (PAULA, 2002).

Umas das questões mais genérica para indução de árvores de decisão é particionar, recursivamente, grupos de exemplos classificados até que uma regra de parada seja encontrada. Partição é uma conexão de teste que tem no minino um conjunto de saída, para poder gerar uma ligação de saída a cada valor novo de entrada. Ocorrendo para cada nó direcionado e tratando os nós de partição como um subproblema, para o qual um nó filho é gerado recursivamente. Um dos parâmentros comuns de parada é quando todos os nós filhos de ligação estão no mesmo nível ou na mesma classe (HALMENSCHLAGER, 2002 tradução nossa).

Na árvore de decisão cada aresta de uma árvore é um padrão definido, dependendo da regra ou a definição do nó, dividindo os valores de um atributo de uma maneira que pode discriminar as variável dependente. Para cada divisão, tem-se a possiblidade de poder gerar dois novos ramos, que depois podem se dividir sucessivamente. Assim, as árvores consistem em indicadores de particionamento recursivo conforme figura 5 (HAN, KAMBER, & PEI, 2009 tradução nossa).

Figura 5 – Construção de uma árvore de decisão



Fonte: Jorge Luis Cavalcanti Ramos (2016)

3.1.2 Tipos de árvore de decisão

As árvore de decisão são geradas por algoritmos que tomam umas série de decisões, localmente ótimas, sobre qual atributo usar para particionar os dados. Um deles é o algoritmo de hunt [2]que é a base do desenvolvimento de muitos algortimos para indução de árvore de decisão existentes, incluindo o ID3, CART, C4.5 e *Hoeffding Tree* (TAN; STEINBACH & KUMAR, 2009).

3.1.2.1 Algoritmo C4.5

Os algoritmos de aprendizado de máquina são projetados para aprender quais são os atributos mais adequados, para serem empregados e tomar suas decisões (WITTEN; FRANK, 2005, tradução nossa).

O algoritmo C4.5 procura um espaço de hipóteses expressivo, seu viés indutivo tém preferência por pequenas árvores do que grandes árvores (MITCHELL, 1997, tradução nossa).

[2] Hunt uma árvore de decisão cresce de uma forma recursiva pelo particionamento dos registros de treino em sucessivos subconjuntos mais puros. É aplicado recursivamente a cada filho do nodo raiz. (OLIVEIRA, 2010)

C4.5 possui como sua principal vantagem a simplicidade: de fazer combinanações com o método de indução de árvores de decisão de cima para baixo, com aprendizagem de regras e conquista, produz bons conjuntos de regras sem qualquer necessidade de otimização global (WITTEN; FRANK, 2005, tradução nossa).

Dentre as particularidades interessantes do algoritomo C4.5, tem-se a não obrigatóriedade em fazer divisão binária no seu particionamento, podendo fazer a geração de pequenas árvores e formação de agrupamentos de valores em conjuntos. Essa habilidade de gerar árvorer menor, faz com que ela seja facilmente entendida. E além disso, ela é consequentemente inclinada a precisão, visto que possibilita buscar a menor árvore possível (OLIVEIRA, 2013).

3.1.2.1.1 Funcionamento do algoritmo C4.5

O funcionamento do algoritmo C4.5 é realizado a partir da construção de um algoritmo guloso, que utiliza a técnica de dividir para conquistar. Esse algoritmo guloso tem como atividade eliminar problemas, de forma local. O algoritmo C4.5 tem a particularidade de calcular a árvore no geral para informar qual nó passa a ser o nó pai de uma árvore, para se ter a melhor decisão (CARVALHO, 2014).

O algoritmo C4.5 analisa qual atributo é aproveitado como teste em cada nó. Para se classificar esse teste nos subconjuntos que são gerados é necessário calcular o critério de seleção de dados, por meio da razão do ganho apesar disso é necessário se compreender o ganho de informação (CARVALHO, 2014).

A entropia da informação é um cálculo utilizado para selecionar as impurezas em um determinado subconjunto. Como isso, é possível reduzir a quantidade de informações, para se classificar uma instância (HAN, KAMBER, 2001). Com base nisso, pode-se calcular o ganho de informação.

$$E(T) = -\sum_{n=i}^{c} pilog2(pi)$$

$$\text{pi} = \frac{CiT}{T}$$

a) P_i → Probabilidade de uma instância aliatória i pertecente a uma classe Ci;

b) C_i,T → É a quantidade de instância de T que pertence a Ci;
c) Log_2 → é utilizada na medida de bits.

Após o calculo da entropia original, encontra-se uma diferença da operação descoberta e a entropia calculada após o particionamento do conjunto. Devido a isso, é importante calcular a entropia do conjunto do particionamento.

$$Ec(T) = -\sum_{i=1}^{V} \left(\frac{Ti}{T}.E(Ti)\right) Cv$$

a) V → é a quantidade de valores que uma caracteristica C pode assumir;
b) Ti → é o subconjunto de instância aonde C assume o valor Cv;
c) E(Ti) → é a entropia do subconjunto.

O ganho de informação é uma operação que determina uma caraterística e reparte o conjunto de avaliação de acordo com a sua classicação (MITCHELL, 1997). Também tem como função eliminar a quantidade de informações desnecessárias para se classificar uma instância.

$$Ganho(C) = E(T) - Ec(T)$$

A divisão da informação é uma operação que penaliza as árvores com maior quantidade de valores (MITCHELL, 1997).

$$DivInfoc(T) = -\sum_{i=1}^{V} \frac{Ti}{T}.log2\left(\frac{Ti}{T}\right)$$

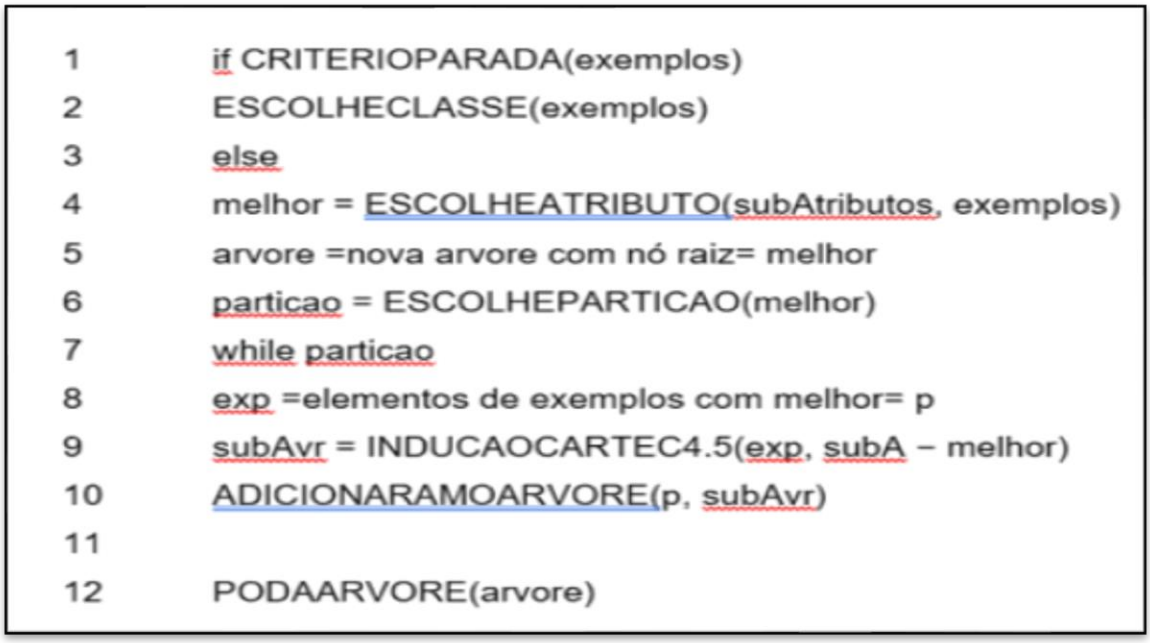
Com base na entropia da união dos dados de ganho de informação e de divisão do ganho tem-se a razão do ganho (CARVALHO, 2014).

$$RazãoGanho(T) = \frac{Ganho(C)}{DivInfoc(T)}$$

O algoritmo C4.5 passa pelo critério de classificação para o desenvolvimento da árvore de decisão, Figura 6: pseudocódigo do algoritmo C4.5, compreendido pela:

a) escolha das condições para dividir cada nó;
b) o critério que devem usar para dividir um nó pai em seus filhos;
c) decidir quando um nó se torna um nó terminal (para divisão);
d) atribuir uma classe a esse nó terminal.

Figura 6- Algoritmo C4.5



Fonte: (BARBOSA; TIAGO; ANDREA, 2010).

O método *EscolheAtributo*() funciona como uma pesquisa gulosa tendo como característica a redução da divisão de dados por técnicas de entropia. Já a função *EscolheParticao*() realiza a definição de escolha gerando em cada nó um ou mais ramos para direcionar ao próximo destino. O *CriterioParada*() elabora uma atividade de que o algoritmo deixa de dividir quando as folhas se encontram na mesma classe (MANTAS; ABELLÁN; CASTELLANO, 2016). O método *EscolhaClasse()* é a fase em que o algoritmo define o passo para a próxima decisão por parâmetro, onde passará o endereço da classe para o nó terminal. A eliminação da *PodaArvore*() é visto como um elemento significativo no desenvolvimento da árvore de decisão, que estabelece dimensões da árvore, removendo erros, para uma classificação mais apropriada gerando uma estrutura de árvore com melhor desempenho. Assim, o

reaproveitamento do próprio conjunto de treinamento para praticar a remoção de uma árvore (BARBOSA; CARNEIRO; TAVERES, 2012).

3.1.2.2 Algoritmo *Hoeffding Tree*

O Algoritmo *Hoeffding* Tree*,* também conhecido como *Very Fast Decision Tree* (VFDT), é baseado em árvore de decisão que aplica métodos que superam *trade-off,* podendo gerar uma árvore assintótica[3] que é criada por algoritmo tradicional (DOMINGOS; HULTEN, 2000, tradução nossa); (HANG; FONG, 2010, *tradução nossa*).

Trade-off é um método capaz de contabilizar e estruturar elementos que estão relacionados com o processo. Desde que a quantificação foi desempenhada a decisão de escolha terá mais facilidade na execução (MARTENSSON, 2005). No procedimento de admissão é considera-se a gama de alvos específicos que uma árvore pode desenvolver na medida que busca ferramentas para atingir objetivos mais amplos (NATIONAL RESEARCH COUNCIL, 1999).

Hoeffding analisa o conjunto de treinamento passo a passo e aperfeiçoam o modelo aprendido sucessivamente, com base na quantidade de exemplos que são averiguados, ocorrendo uma ação de estatística que é chamada de limitante de *Hoeffding*. O algoritmo tem autonomia de selecionar um novo valor para teste, podendo assim realizar a divisão nesta folha. Por ele ser um algoritmo incremental possibilita realizar a tarefa de classificação antes da etapa de aprendizado ser totalmente executada (MENEZES, 2011).

Quando uma amostra entra, percorre a árvore da raiz até uma folha, avaliando o atributo essencial em cada nó. Depois que o exemplo atinge uma folha, a estatística é atualizada. O nó da árvore de decisão contém o número de valores possíveis para o atributo escolhido sobre o teste de divisão instalado (HANG; FONG, 2010). Os principais elementos do *Hoeffding* são:

a) estado da árvore não pode conter mais de uma única folha raiz;
b) definição da função de avaliação heurística (detonado por G()).

---

[3] Árvore Assintótica são árvare balanceada em que o nível da classe não pode ter uma diferença de dois nó.

3.1.2.2.1 Limitante de Hoeffding

Seja r̄ uma variável aleatória real cujo domínio tem tamanho R. Considere n observações independentes desta variável e seja r sua média observada. O limitante de *Hoeffding* certifica que $P(\mu_r \geq \bar{r} - €) = 1 - \delta$, onde μ_r é a verdadeira média da variável r, € representa o limite *Hoeffding* e δ é um número muito pequeno.

$$€ = \sqrt{\frac{R2\ln\left(\frac{1}{\delta}\right)}{2n}}$$

3.1.2.2.2 Arvores de Hoeffding

A definição da função de avaliação heurística (detonado por G()), elabora uma árvore de decisão com o ganho informação, podendo aplicar estatística para a contagem de um valor discreto (HANG; FONG, 2010, tradução nossa).

O limite de *Hoeffding* pode ser utilizada associando r̄ com o ganho médio de informação de um valor selecionado. Isso fornece um meio em estatística para marcar um atributo correto, sem ter que prejudicar a divisão nos nós de decisão da árvore. Primeiramente reconhece os dois atributos com os maiores ganhos individuais de informação. Com um número suficiente, n instâncias como apresentadas pelo limite de *Hoeffding*, a diferença no ganho de informação GΔ entre os dois atributos (A1, A2), agora pode ser aplicada para identificar o atributo correto para dividir em (A1 tendo o maior ganho de informação). Sempre que GΔ > İ, o limite de *Hoeffding* pode afirmar que A1 é a escolha exata com uma probabilidade de 1 – δ (HOEGLINGER; PEARS, 2007, tradução nossa).

Existem caso em que o melhor atributo A1, não seja muito melhor que A2, porém, pode-se executar o particionamento empregando-se o A1, podendo não ser tão benéfico quanto A2, dependendo do que se procura considera-se um atributo extra, Aø, que na verdade reproduz o valor da heurística considerando apenas os exemplos na folha. Desta forma trata-se do caso de uma divisão no melhor atributo para não estimular uma diferenciação verdadeiramente boa (MENEZES, 2011), segue-se o *pseudo-código do Hoeffding* conforme ilustra a figura 7.

Entradas:

a) S é uma sequência de exemplos,
b) X é um conjunto de atributos discretos,
c) G(.) é uma função de avaliação de particionamento,
d) δ é um menos a probabilidade desejada de se escolher o atributo correto em um dado nó.
e) Saída:
f) HT é uma árvore de decisão

Segue, então, a explicação da figura 7 do pseudocódigo do algoritmo de *Hoeffding tree*. Na linha 1, tem-se um nó chamado L_1 que será indicado como a raiz, já na linha 2 é declarado um contador global de folha X, além disso tem como funcionalidade de não representar teste caso a árvore esteja vazia. Já na linha 3 tem-se uma variação de G para analisar classe mais preenchida, o atributo da heurística de $X_\emptyset$ é considerado também como uma entropia do conjunto. Na linha 4 realiza-se uma condição para reconhecer o nível da classe. Inicialmente, os contadores do nó, encontrado pela condição: i = atributo X_i que pertence a X, realizar uma ação na posição i ou j até uma determinada classe, para satisfazer a inserção do valor zero, com base na linha 5 e 6.

Após inicialização, os exemplos de S começam a serem inseridos pela árvore em desenvolvimento. Cada exemplo $(X, Y_k) \in S$, vão percorrendo até alcançar em uma folha L ajustada, (linhas 7 e 8).

Quando o exemplo (X, Y_k) é encaminhado para a folha L, cada folha incrementa o contador, conforme os valores dos atributos de X (linhas 9 e 10). A linha 11 apresenta atualização de L, para reproduzir a classe mais repetida até então.

Em questões de exemplos inseridos pertencerem à única classe, então L deve espera que mais exemplos cheguem, pois não será realizado nenhum particionamento. Para se executar, a heurística deve ser computada para cada atributo, a fim de escolher os dois melhores (linhas 12 a 15).

Depois de selecionar os dois melhores valores, aplica-se o limite de *Hoeffding*, para analisar a diferença entre ambos os atributos, devendo ser suficiente para se efetuar um particionamento adequado nos valores escolhidos (Linha 16 e 17).

Nas linhas 18 a 20, estatisticamente X_a recebe o melhor valor, L passará de uma simples folha a um nó interno, para que possa adicionar uma nova aresta e

apontar para o novo endereço de uma folha, essa sequência segue a cada valor de X_a.

Para cada novo nó filho L_m, analogamente separa-se um valor $X_\emptyset$. Os contadores dos filhos são sempre iniciados com valor 0 (linhas 21 a 23).

Após concluir a primeira inserção, retorna ao laço principal, descrito na linha 7. Desta forma, mais um exemplo será dirigido até uma folha, a qual consultará a adaptação de um particionamento. Ao finalizar os exemplos o processo termina.

Figura 7- Pseudocódigo do algoritmo de Hoeffding Tree.

Procedimento HoeffdingTree $(S, \mathbf{X}, G, \delta)$

1: Seja HT uma árvore com uma única folha l_1 (a raiz).
2: Seja $\mathbf{X}_1 = \mathbf{X} \cup \{X_\emptyset\}$.
3: Seja $\overline{G_1}(X_\emptyset)$ o valor de $\overline{G}$ obtido ao se predizer a classe mais frequente em S.
4: Para cada classe y_k
5: Para cada valor x_{ij} de cada atributo $X_i \in \mathbf{X}$
6: Faça $n_{ijk}(l_1) = 0$.
7: Para cada exemplo $(\mathbf{x}, y_k)$ em S
8: Desça $(\mathbf{x}, y)$ até uma folha l usando HT.
9: Para cada x_{ij} em $\mathbf{x}$ tal que $X_i \in \mathbf{X}_l$
10: Incremente $n_{ijk}(l)$.
11: Rotule a folha l com a classe majoritária dos exemplos que chegaram nela até agora.
12: Se os exemplos que chegaram até agora em l não forem todos da mesma classe, então
13: Calcule $\overline{G_l}(X_i)$ para cada atributo $X_i \in \mathbf{X}_l$ - $\{X_\emptyset\}$ usando os contadores $n_{ijk}(l)$.
14: Faça X_a ser o atributo com o maior valor de $\overline{G_l}$.
15: Faça X_b ser o atributo com o segundo maior valor de $\overline{G_l}$.
16: Calcule ϵ usando a equação 3.4.
17: Se $\overline{G_l}(X_a)$ - $\overline{G_l}(X_b) > \epsilon$ e $X_a \neq X_\emptyset$, então
18: Substitua l por um nó interno cujo teste é o atributo X_a (ou seja, faça o *particionamento* com o atributo X_a).
19: Para cada ramo do *particionamento*
20: Adicione uma nova folha l_m, e faça $\mathbf{X}_m = \mathbf{X}$ - $\{X_a\}$.
21: Faça $\overline{G_m}(X_\emptyset)$ ser o valor de $\overline{G}$ obtido ao se predizer a classe mais frequente em l_m.
22: Para cada classe y_k e cada valor x_{ij} de cada atributo $X_i \in \mathbf{X}_m$ - $\{X_\emptyset\}$
23: Faça $n_{ijk}(l_m) = 0$.
24: Retorne HT.

Fonte: Menezes (2011).

4 MEDIDAS DE QUALIDADE EM MINERAÇÃO DE DADOS PARA CLASSIFICADORES

A utilização dos classificadores vem da precisão de separação das informações a fim de proporcionar decisões, estando a escolha sua diretamente vinculada ao tipo de dado a ser empregado na aplicação (ALBERTO, 2012; SANTOS, 2017).

A ação da perfomance do comportamento de cada algoritmo classificador é mensurada, para isso é necessário calcular uma matriz de confusão (ALBERTO, 2012). As classificações corretas desta matriz são armazenadas nas células da diogonal principal e as incorretas são registradas nas demais celulas. O resultado é conhecido como Verdadeiro Positivo (VP), quando o algoritmo que classifica encontra casos de uniformidade positivos, por exemplo, alunos aprovados. O Verdadeiro Negativo (VN) é quando a classe recebe casos que também são de consônância negativa, por exemplo, alunos não aprovados. Resposta falso positivo (FP) é quando o algoritmo classifica como positivo, o é negativa. E falso negativo (FN) é a situação inversa do FP, conforme ilustra figura 8 (LOPEZ, 2014).

Figura 8: Matriz de Confusão

Fonte: Lopez (2014)

As medidas de classificadores mais utilizadas são as derivadas da chamda matriz de confusão (MEYNARD; QUINN, 2007), tendo-se a acurácia, sensibilidade, especificidade, precisão, kappa e roc (BARRETO, 2008).

4.1 Acurácia

Acurácia é calculada sobre os erros achados, por meio de uma equação, obtendo assim uma medida mais segura sobre a competência do modelo de demostrar o processo gerador dos dados (BRUM, 2016). Esta medida é amplamente utilizada, sendo adequada para problemas de classificação. (GUILLET; HAMILTON, 2007, tradução nossa).

$$Ac = \frac{VP + VN}{VP + FP + VN + FN}$$

4.2 Sensibilidade

Sensibilidade, que também é conhecida como Recall, calcula a quantidade de positivos corretos em relação ao total de positivos da amostra (GUILLET; HAMILTON, 2007 tradução nossa).

$$\text{Sensibilidade} = \frac{Verdadeiro\ positivos}{total\ de\ positivos\ de\ amostra} = \frac{VP}{VP + FN}$$

4.3 Índice Kappa

Índice Kappa é definido como uma medida que descreve e testa o grau de concordância entre a classe predita e a real, variando de 0 e 1, representada na tabela 1 (GUILLET; HAMILTON, 2010).

$$Cobertura = \frac{\text{total de acertos} - \text{proporção de acertos esperada}}{\text{total de amostras} - \text{proporção de acertos esperada}}$$

$$K = \frac{(VP + VN) - \{\frac{[\,(\text{VP} + \text{FN}) * (\text{VN} + \text{FP})\,] + [\,(\text{FP} + \text{VN}) * (\text{FN} + \text{VN})\,]}{(\text{VP} + \text{FN}) + (\text{VN} + \text{FP})}\}}{[\,(\text{VP} + \text{FN}) + (\text{VN} + \text{FP})\,] - \{\frac{[\,(\text{VP} + \text{FN}) * (\text{VN} + \text{FP})\,] + [\,(\text{FP} + \text{VN}) * (\text{FN} + \text{VN})\,]}{(\text{VP} + \text{FN}) + (\text{VN} + \text{FP})}\}}$$

Tabela 1- Interpretação dos valores do índice kappa

Valor de Kappa	**0 a 0,20**	**0,21 – 0,40**	**0,41 – 0,60**	**0,61 - 80**	**0,81 - 1**
Nível de concordância	Ruim	Fraca	Boa	Muito boa	Excelente

Fonte: (SILVA et al. 2011)

4.4 *F-measure*

F-measure é também denominado *F-score* é calculado por uma média ponderada entre a confiabilidade positiva e sensibilidade. O resultado obtido quanto mais próximo de um, melhor é a performance, e quanto mais próximo de zero (0) o resultado tem um desempenho ruim, para a classe ligada ao conceito de positivo (CASTRO; BRAGA 2017).

$$\mathrm{F-measure} = \frac{(1+\beta) * Precisão * sensibilidade}{\beta^2 * Precisão + sensibilidade} = \frac{(1+\beta) * \frac{VN}{VN+FP} * \frac{VP}{VP+FN}}{\beta^2 * \frac{VN}{VN+FP} + \frac{VP}{VP+FN}}$$

$$\beta = \frac{Sensibilidade}{Precisão}$$

5 TRABALHOS CORRELATOS

Diante da evolução da tecnologia, surgiram várias aplicações que auxiliam no desenvolvimento das pesquisas em diferentes áreas do conhecimento. As áreas são, por exemplo: saúde, transações comerciais e educação. Devido á quantidade de dados armazenados, possibilita que se desenvolvam várias pesquisas.

5.1 APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO ENADE COM ENFOQUE NOS CURSOS DE MEDICINA

A partir de um artigo publicado na resvista Acta Biomedica Brasilensia, aplicou-se a técnica com o objetivo de absorver conhecimento da base de dados do ENADE do curso de Medicina, bem como a resposta sobre o nível de dificuldade do componente específico da prova. Aplica-se o processo de KDD que compreende um agrupamento de técnicas capaz de examinar e explorar informações úteis das bases de dados por meio da identificações de padrões.

Empregou-se o algoritmo J48, verificando-se a influência da classe e das cacterísticas de instituições de ensino superior na criação dos perfis, sendo diretamente conectadas ao nível de perfomance dos estudantes e seu conceito acerca da prova. Foi possível constatar que os acadêmicos, tanto do Rio de Janeiro quanto de São Paulo, quando descendentes de universidades sem fins lucrativos, alcançaram na sua maioria, um resultado ruim, com nota menor que sessenta. Até então em São Paulo, os estudantes das instituições municipais, além de ganharem um rendimento negativo, também responderam como “fácil” o nível de dificuldade do componente específico do exame. (CRETTON; GOMES, 2016).

5.2 ANÁLISE DOS ALGORITMOS DE MINERAÇÃO J48 E APRIORI APLICADOS NA DETECÇÃO DE INDICADORES DA QUALIDADE DE VIDA E SAÚDE

Artigo publicado na Revista Interdisciplinar de Ensino, Pesquisa e Extensão. Este artigo abordou um estudo sobre as técnicas de mineração de dados, como associação e classificação, as quais aplicadas na identificação e classificação de indicadores da saúde, a fim de gerar um perfil de utilizador.

Teve como objetivo a implementação dos algoritmos J48 e Apriori, medindo e comparando o seu desempenho. Com a realização da pesquisa determinou-se que a técnica de classificação manisfestou o melhor desempenho no reconhecimento e geração de perfis de usuários da sua base de dados (LIBRELOTTO; MOZZAQUATRO, 2013).

5.3 MINERAÇÃO EM BASES DE DADOS DO INEP: UMA ANÁLISE EXPLORATÓRIA PARA NORTEAR MELHORIAS NO SISTEMA EDUCACIONAL BRASILEIRO

O Instituto Nacional de Estudo e Pesquisa Educacionais Anísio Teixeira (INEP) contêm uma base de dados que é nativa para investigações estatísticas e avaliativos em diversos níveis e particularidade de ensino, como por exemplo a Educação Básica, que é executada a partir do Sistema de Avaliação da Educação Básica (Saeb).

A pesquisa teve como objetivo identificar causas que associem o perfil de professores que instruem em matemática com a habilidade obtida por seus alunos.

Com finalidade de analisar o Ensino Fundamental público brasileiro, por intermédio das ferramentas de testes e questionários aos alunos, professores e diretores, torna possível a filtragem de informações importantes para identificação de alternativas tendo em vista à melhoria da qualidade do ensino (FONSECA; NAMEN, 2016).

5.4 INVESTIGAÇÃO ACERCA DOS FATORES DETERMINANTES PARA A CONCLUSÃO DO ENSINO FUNDAMENTAL UTILIZANDO MINERAÇÃO DE DADOS EDUCACIONAIS NO CENSO ESCOLAR DA EDUCAÇÃO BÁSICA DO INEP 2014

Este trabalho teve como objetivo a identificação de vários fatores relacionados à conclusão do Ensino Fundamental, empregando ferramentas de mineração de dados operadas aos Micro dados do Censo Escolar da Educação Básica do INEP de 2014. A importância desta abordagem é o manuseio dos dados públicos que geram informações sobre o acadêmico, turmas, escolas e docentes de todo o Brasil. Os experimentos produzidos a partir de árvores de decisão, por meio do

algoritmo J48, no primeiro instante, foram averiguados na sua totalidade de informações e, no segundo momento, as informações foram compartilhadas entre: dados dos alunos, dados das turmas e dados das escolas. Os experimentos executados com todas as informações, foram analisados 176 atributos da base de dados associadas forneceram precisão de média de 96,17% (FERREIRA, 2015).

5.5 ESTUDO DA EVASÃO EM ALUNOS DE GRADUAÇÃO POR MEIO DE MINERACÃO DE DADOS

Umas das melhores maneiras de se abordar a evasão, é predizer como tomar medidas para evitar que tal ação ocorra e para que essa antecipação ocorra por meio de mineração de dados. O objetivo desse trabalho é verificar em uma plataforma de dados de educação, a Plataforma Nilo Peçanha, o perfil dos estudantes que desistem de cursos graduação de dados e verificar quais algoritmos de classificação teria melhor eficácia. Por meio de classificação foi empregado 13 classificadores para avaliar o melhor desempenho. Verificou-se que o algoritmo C4.5 teve a melhor performance, levando em conta que também foi aplicado o classificador *hoeffding tree* (GOMES, 2019).

6 TRABALHO DESENVOLVIDO

Este trabalho teve como ponto de partida, à busca de pesquisas de trabalhos bibliográficos semelhantes ou próximos das ferramentas de estudo a serem aplicadas no decorrer do desenvolvimento do trabalho como KDD, MD, árvore de decisão, algoritmos C4,5 e *Hoeffding Tree*. O desenvolvimento deste projeto trata-se de uma pesquisa quantitativa e aplicada.

Este trabalho tem como objetivo empregar as tarefas de classificação por meio de indução de árvores de decisão, assim como analisar as saídas fornecidas pela ferramenta WEKA com a versão 3.8.3 e comparar os modelos obtidos pelos algoritmos C4,5 e *Hoeffding Tree*, por meio de medidas de qualidade em mineração de dados. Para a realização deste trabalho foi utilizada uma base de dados educacional aberta em formato CSV disponível no site do INEP.

6.1 METODOLOGIA

Este trabalho tem como finalidade a realização de ações com técnicas de Mineração de dados, por intermédio de base de dados Instituto Nacional Estudo e Pesquisa Anísio Teixeira (INEP). A base escolhida especificamente é a base de dados do ENADE que tem como significado Exame Nacional de Desempenho de Estudante. Segundo Francisco e Monteiro Constituído a partir da Lei No. 10.861, de 14 de abril de 2004, surge com o objetivo de estabelecer um instrumento para analisar o desempenho estudantil, de modo que seja pontual o controle das diretrizes curriculares no decorrer da formação. Nesse sentido, considerando a estrutura da avaliação que se deposita ao estudante, a intenção de constituir um instrumento com a capacidade de fortalecer o desenvolvimento de uma área específica de conhecimento, de modo a considerar um grande conjunto de análises em seus resultados (FRANCISCO; MONTEIRO, 2016).

É importante salientar que a prova do ENADE é executada todos os anos, porém cada curso só é avaliado de três em três anos. É assim que funciona em todas as áreas (área entendida como um conjunto de cursos), que foram divididas em três grupos. Em 2004, o exame foi aplicado ao grupo da Saúde e das Ciências Agrárias; em 2005, ao grupo das Engenharias (46 engenharias) e das Licenciaturas; em 2006, ao grupo das Ciências Sociais Aplicadas e aos demais cursos. Em 2007, volta o grupo

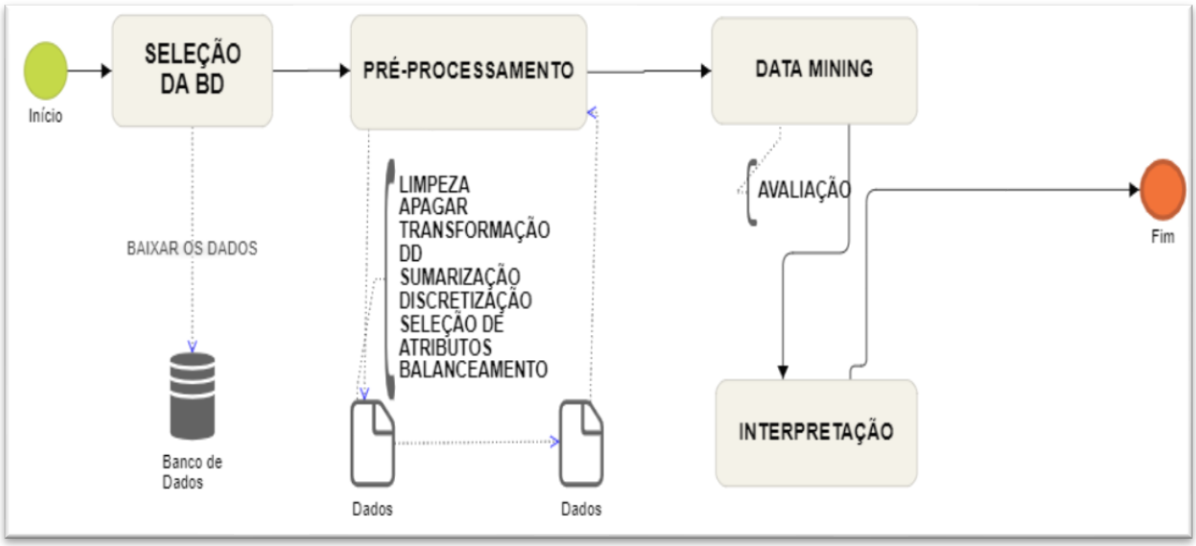
I; em 2008, o grupo II e assim sucessivamente. Todos os campos de graduação (RISTOFF; GIOLO, 2006).

O ENADE conta com o questionário do estudante e com o questionário do coordenador, que são dois instrumentos que auxiliam os órgãos de regulação na compreensão dos desafios que envolvem uma determinada área em avaliação (FRANCISCO; MONTEIRO, 2016). Com base nisso, procura-se avaliar o desempenho dos acadêmicos em relação aos conteúdos selecionados e submetidos nas diretrizes curriculares de seus pertencentes cursos e de modo que seja possível enxergar o perfil destes acadêmicos.

A área de conhecimento das ciências exatas, onde insere-se o curso de Ciência da Computação, realizou ENADE nos anos de 2008, 2011, 2014 e 2017. Os dados obtidos desta ação avaliativa, permitem conhecer o modo de funcionamento e as características do curso de ciência da computação (ENADE, 2015).

No ENADE em Ciência da Computação são avaliados os conteúdos aprendidos no decorrer da formação dos alunos concluintes do curso, tendo esta prova duração de quatro horas. Com base na avaliação, tem-se o desempenho geral dos estudantes no componente de formação geral e de no conhecimento específico da avaliação (VISTA et al., 2018). Na figura 10 tem-se as etapas de realização da pesquisa.

Figura 10 - Etapas e realização da pesquisa



Fonte: Do autor.

6.1.1 Seleção da Base de Dados

Os dados fornecidos para elaboração deste trabalho estão disponíveis no portal do INEP[4]. Onde foi selecionada a prova do ENADE que é um microdados, dos quatros últimos anos que ocorreu a prova do curso de Ciência da Computação (2008, 2011, 2014, e 2017), que procura avaliar o desenvolvimento dos estudantes de ciência da computação.

A primeira vez que se aplicou a prova do ENADE em ciência da computação, foi no de 2005 (BRASIL, 2019).

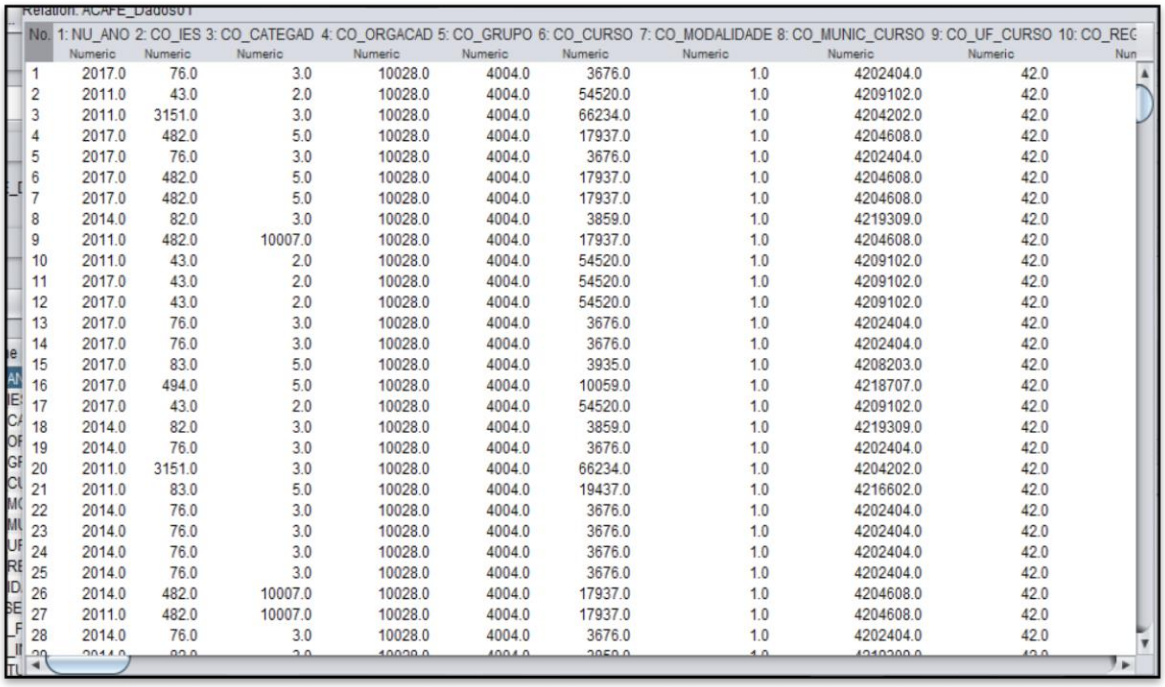
Após ter concluído o processo de execução de *download* dos dados. A extensão do arquivo encontrava-se no formato txt, imediatamente abriu-se o arquivo e copiou-se os dados informados que foram passados para a ferramenta Excel, visto que a prova do ENADE foi dividida em três grupos. Percebeu-se que a base utilizada é a do segundo grupo (Engenharias e das Licenciaturas). Constatou-se que a base está composta por nove partes que são: Informações das instituições de ensino superior e do curso; Informações dos estudantes; Números de itens da parte objetiva; Vetores; Tipos de situação das questões da parte discursiva; Notas na formação geral e componentes específicos; Questionário de percepção da prova e questionário dos estudantes.

A base de dados de 2005 e 2008 não foram empregadas devido às quantidades de atributos existentes diferentes dos últimos três anos, com isso eliminou-se da pesquisa. Esta base de dados que se encontra, carrega com 150 atributos, e em cada anos teve-se uma quantidade de instância diferentes. Em 2011 teve 376.181 instancias; em 2014, 481.721 instancias e em 2017, 537.437 instancias. Com objetivo de transformar em um único arquivo, quando se alcançar os dados concretos, na (figura 11).

.

[4] http://portal.inep.gov.br/web/guest/microdados

Figura 11 - Estado inicial da base de dados referente ao ENADE.



Relation: ACAPE_Dados01

No.	1: NU_ANO	2: CO_IES	3: CO_CATEGAD	4: CO_ORGACAD	5: CO_GRUPO	6: CO_CURSO	7: CO_MODALIDADE	8: CO_MUNIC_CURSO	9: CO_UF_CURSO	10: CO_REG
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nun
1	2017.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
2	2011.0	43.0	2.0	10028.0	4004.0	54520.0	1.0	4209102.0	42.0	
3	2011.0	3151.0	3.0	10028.0	4004.0	66234.0	1.0	4204202.0	42.0	
4	2017.0	482.0	5.0	10028.0	4004.0	17937.0	1.0	4204608.0	42.0	
5	2017.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
6	2017.0	482.0	5.0	10028.0	4004.0	17937.0	1.0	4204608.0	42.0	
7	2017.0	482.0	5.0	10028.0	4004.0	17937.0	1.0	4204608.0	42.0	
8	2014.0	82.0	3.0	10028.0	4004.0	3859.0	1.0	4219309.0	42.0	
9	2011.0	482.0	10007.0	10028.0	4004.0	17937.0	1.0	4204608.0	42.0	
10	2011.0	43.0	2.0	10028.0	4004.0	54520.0	1.0	4209102.0	42.0	
11	2017.0	43.0	2.0	10028.0	4004.0	54520.0	1.0	4209102.0	42.0	
12	2017.0	43.0	2.0	10028.0	4004.0	54520.0	1.0	4209102.0	42.0	
13	2017.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
14	2017.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
15	2017.0	83.0	5.0	10028.0	4004.0	3935.0	1.0	4208203.0	42.0	
16	2017.0	494.0	5.0	10028.0	4004.0	10059.0	1.0	4218707.0	42.0	
17	2017.0	43.0	2.0	10028.0	4004.0	54520.0	1.0	4209102.0	42.0	
18	2014.0	82.0	3.0	10028.0	4004.0	3859.0	1.0	4219309.0	42.0	
19	2014.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
20	2011.0	3151.0	3.0	10028.0	4004.0	66234.0	1.0	4204202.0	42.0	
21	2011.0	83.0	5.0	10028.0	4004.0	19437.0	1.0	4216602.0	42.0	
22	2014.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
23	2014.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
24	2014.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
25	2014.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	
26	2014.0	482.0	10007.0	10028.0	4004.0	17937.0	1.0	4204608.0	42.0	
27	2011.0	482.0	10007.0	10028.0	4004.0	17937.0	1.0	4204608.0	42.0	
28	2014.0	76.0	3.0	10028.0	4004.0	3676.0	1.0	4202404.0	42.0	

Fonte: Do autor (2019).

Nesta base, foi possível obter dados relacionados aos estudantes que se encontravam no final do curso que prestaram os respetivos exames nos anos de 2011, 2014 e 2017, tais estudantes, se tornaram adequados para fazer a prova, podendo assim avaliar o curso de ciência da computação.

A base de dados também estava acompanhada de um arquivo denominado como Dicionário de variáveis do micro dados do ENADE, onde é explicado cada significado dos atributos como nome, tipo, tamanho, descrição e categoria, na (figura 12).

Figura 12- Dicionário de variável.

DICIONÁRIO DE VARIÁVEIS – ENADE 2017					
Nº	Nome	Tipo	Tamanho	Descrição	Categoria
1	NU_ANO	Numérica	4	Ano de realização do exame	2017
PARTE 1 – INFORMAÇÕES DA INSTITUIÇÃO DE ENSINO SUPERIOR E DO CURSO					
2	CO_IES	Numérica	5	Código da IES (e-Mec)	Entre 1 e 19739 (Identificação da IES conforme e-MEC)
3	CO_CATEGA	Numérica	1	Código da categoria administrativa da IES	1 = Pública Federal 2 = Pública Estadual 3 = Pública Municipal 4 = Privada com fins lucrativos 5 = Privada sem fins lucrativos 7 = Especial
4	CO_ORGAC	Numérica	5	Código da organização acadêmica da IES	10019 = Centro Federal de Educação Tecnológica 10020 = Centro Universitário 10022 = Faculdade 10026 = Instituto Federal de Educação, Ciência e Tecnologia 10028 = Universidade
					21 = Arquitetura e Urbanismo 72 = Tecnologia em Análise e Desenvolvimento de Sistemas 76 = Tecnologia em Gestão da Produção Industrial 79 = Tecnologia em Redes de Computadores 701 = Matemática (Bacharelado) 702 = Matemática (Licenciatura) 903 = Letras-Português (Bacharelado) 904 = Letras-Português (Licenciatura) 905 = Letras-Português e Inglês (Licenciatura) 906 = Letras-Português e Espanhol (Licenciatura) 1401 = Física (Bacharelado) 1402 = Física (Licenciatura) 1501 = Química (Bacharelado) 1502 = Química (Licenciatura) 1601 = Ciências Biológicas (Bacharelado) 1602 = Ciências Biológicas (Licenciatura) 2001 = Pedagogia (Licenciatura) 2401 = História (Bacharelado) 2402 = História (Licenciatura) 2501 = Artes Visuais (Licenciatura) 3001 = Geografia (Bacharelado)

Fonte: Do autor (2019).

6.1.2 Pré-processamento

A fase de pré-processamento contempla a etapa de seleção de dados. Para esta fase de análise, utilizou-se a ferramenta *Excel* com a e versão 18.1910.1283.0, onde foi requerida a limpeza de campos vazios, retirada de palavras com acentuação ou caráter especial.

Analisou-se também a quantidade de atributos existentes em seus respetivos anos e a diferença entre o posicionamento de variável em anos diferentes. Organizou-se a posição dos atributos para poder transformar em um único arquivo dos três respectivos anos.

Foram apagados alguns campos como ID_status, Amostra e TP_Semestre por não se encontrarem nos dados de 2017 e também foram apagadas linhas no *Excel* que contém campos vazios.

6.1.2.1 Etapa de transformação de dados

É nesta fase em que os dados sofrem alterações devido ao valor que não são reconhecidos na ferramenta *Weka*. Com isso alguns valores foram alterados para possibilitar o trabalho com informações relevantes, consolidando o conjunto de dados para transformá-los em uma maneira mais apropriada para a mineração, sabendo que, WEKA não reconhece valores com caráter especiais, elaborado na tabela 2.

Tabela 2 - Tabela de alteração de valores com caráteres especiais.

Nº	Nome	Tipo	Descrição	Categoria e Troca
30	DS_VT_ESC_OFG	Caractere	Vetor que representa a escolha de resposta da parte objetiva da formação geral	1 letra por item Intervalo de A a E) '.' = em branco => M '*' = múltiplo => N
61	CO_RS_I1	Caractere	1 - Qual o grau de dificuldade desta prova na parte de Formação Geral?	* = Resposta anulada => AB . = Não respondeu => AC
62	CO_RS_I2	Caractere	2 - Qual o grau de dificuldade desta prova na parte do Componente Específico?	* = Resposta anulada => AB . = Não respondeu => AC
63	CO_RS_I3	Caractere	3 - Considerando a extensão da prova, em relação ao tempo total, você considera que a prova foi:	* = Resposta anulada => AB . = Não respondeu => AC
64	CO_RS_I4	Caractere	4 - Os enunciados das questões da prova na parte de Formação Geral estavam claros e objetivos?	* = Resposta anulada => AB . = Não respondeu => AC
65	CO_RS_I5	Caractere	5 - Os enunciados das questões na parte do Componente Específico estavam claros e objetivos?	* = Resposta anulada => AB . = Não respondeu => AC
66	CO_RS_I6	Caractere	6 – As informações /instruções fornecidas para a resolução das questões foram suficientes para resolvê-las?	* = Resposta anulada => AB . = Não respondeu => AC
67	CO_RS_I7	Caractere	7 - Você se deparou com alguma dificuldade ao responder à prova. Qual?	* = Resposta anulada => AB . = Não respondeu => AC
68	CO_RS_I8	Caractere	8 - Considerando apenas as questões objetivas da prova, você percebeu que:	* = Resposta anulada => AB . = Não respondeu => AC

69	CO_RS_I9	Caractere	9 - Qual foi o tempo gasto por você para concluir a prova?	* = Resposta anulada => AB . = Não respondeu => AC

Fonte: Do autor.

Após os dados estarem corretamente organizados, geraram-se três arquivos com os dados do ENADE em Ciência da Computação, que representam os dados da UNESC; de Santa Catarina e das universidades que fazem parte da Associação Catarinense das Fundações Educacionais (ACAFE). A ACAFE é formada por dezesseis universidades, sendo que nove possuem o curso de ciência da computação. Para se fazer a filtragem do campo Código da IES (e-Mec) (CO_IES), teve-se a necessidade de acessar a plataforma do e-Mec para abstrair os códigos das nove universidades do sistema ACAFE.

Os dados sofreram transformações com intuito de aprimorá-los, para que os testes fossem feitos apenas com informações relevantes. Outro ponto importante é a execução dessa etapa que possibilita otimizar o tempo de processamento do algoritmo a ser utilizado na fase de mineração de dados.

6.1.2.2 Sumarização

A sumarização tem como um dos propósitos, o agrupamento dos registros de várias bases de dados com valores idênticos. Por isso a tabela de sumarização se torna primordial quando é necessário englobar o conjunto de cada base selecionada, para uma única base. Como foram baixadas as três últimas provas do ENADE no departamento de ciência da computação as três bases possuem características de atributos idênticos teve a necessidade de unificar todas elas.

6.1.2.3 Discretização

Os algoritmos de classificação, necessitam que os dados estejam no formato de atributos categorizados. Assim, como muitas vezes é necessário transformar um atributo contínuo em um categorizado (discretização), e tanto as variáveis discretas como contínuas, podem precisar de alteração em um ou mais

atributos. Adicionalmente se um ou mais atributos categorizados possuir um número grande de valores (categorias), ou se algum valor ocorra raramente, então pode beneficiar para determinar tarefas de mineração de dados reduzir o número de categorias combinando em alguns valores. Esse método foi aplicado nas classes separando as notas dos estudantes em duas categorias, utilizando a ferramenta do Excel e as suas fórmulas.

a) Nota acima de 54.9 que é representada pela letra A;

b) E nota igual ou menor que 54.9, que é representado pela letra R.

É importante destacar que não há reprovação nas provas aplicadas do ENADE, logo, os números escolhidos para a pesquisa não correspondem a aprovações e reprovações.

Visto que no conjunto original foi reforçado ou aplicado a discretização em todas as bases de experimentos realizados, levando em conta que a classe maioritária é a classe de estudantes com a nota abaixo ou igual de 54,9 e a minoritária têm a nota maior que 54,9. Dentre os 138 atributos encontrados na base. A variável que foi selecionada foi a de nota geral.

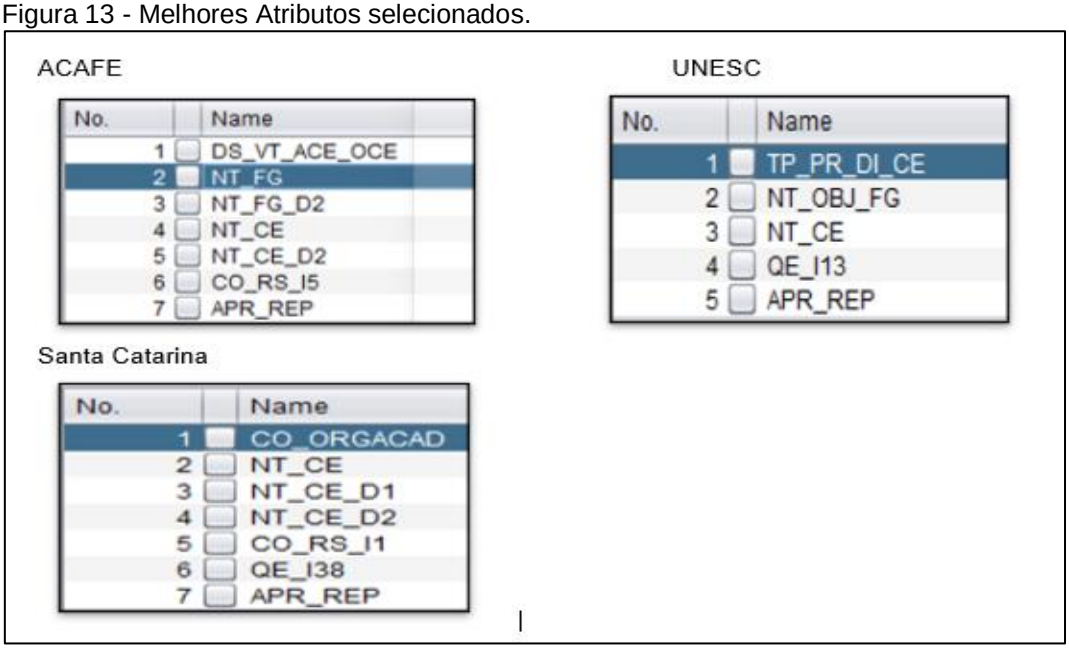
6.1.2.4 Seleção de Atributos

Com vasta quantidade de atributos encontrados na base de dados optou-se por aplicar a escolha de atributos. A seleção de atributos permite a ordenação das variáveis segundo a importância da classe escolhida, com isso causa a redução da dimensionalidade de espaço de busca de atributos e a remoção de dados contendo ruídos (LEE, 2005). Teve-se como opção para resolver este problema, de escolher 138 melhores atributos selecionados dos três arquivos que contém a mesma quantidade de atributos, usou-se o modelo de filtro que, segundo Wang (2013), resulta das particularidades dos dados para executar a avaliação e selecionar o subgrupo com a melhor caraterística, sem o envolvimento de nenhum algoritmo de mineração de dados. O método seleciona os dados previamente e subsequentemente realiza o processo de classificação, contudo, não pondera as interações entre os atributos (CHUANG et al., 2011).

O filtro (*AtributeSelection)* não necessita de um algoritmo de aprendizado de máquina para executar a seleção de características, utilizando-se somente das próprias habilidades para poder avaliar os subconjuntos, geralmente necessitando de

um menor poder computacional para ser utilizado. O método de seleção de atributos empregado foi o *Correlation-based Feature Subset Selection* (CfSubsetEval) que avalia o valor de um subconjunto de atributos, considerando a sua capacidade preditiva, optando-se pelos que são altamente correlacionados com a classe (figura 13 e tabela 3).

Figura 13 - Melhores Atributos selecionados.



Fonte: Do autor (2019).

Tabela 3 - Tabela de dicionário de melhores atributos selecionados

Atributos	**Descrição**
CO_TURNO_GRADUACAO	Código do turno de graduação
CO_ORGACAD	Código da categoria administrativa da IES Código da categoria administrativa da IES
DS_VT_ACE_OCE	Vetor que representa os acertos da parte objetiva do componente específico
NT_FG	Nota bruta na formação geral - Média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral. (valor de 0 a 100)
NT_OBJ_FG	Nota bruta na parte objetiva da formação geral. (valor de 0 a 100)
NT_CE	Nota bruta no componente específico - Média ponderada da parte objetiva (85%) e discursiva (15%) no componente específico. (valor 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. (valor de 0 a 100)
NT_CE_D2	Nota da questão 2 da parte discursiva do componente específico. (valor de 0 a 100)
CO_RS_I5	Os enunciados das questões na parte do Componente Específico estavam claros e objetivos?
CO_RS_I1	Qual o grau de dificuldade desta prova na parte de Formação Geral?

Fonte: Do autor.

6.1.2.5 Balanceamento de classes

No conjunto de dados desbalanceados, obteve uma grande disparidade de quantidade de dados em cada classe. Os conjuntos de dados de cada classe que se deseja modelar devem ser equivalentes de forma que possam ser assimiladas as características de cada classe envolvida, podendo fazer com que o classificador possa atingir um nível de definição homogêneo (CECHINEL; CAMARGO, 2016). O desbalanceamento de uma classe tem a capacidade de influenciar nas ações dos desempenhos de um modelo de classificação, portanto buscam classificar corretamente as classes maioritárias, definido por Chawla et al. (2002).

O método de balanceamento explorado foi o *Synthetic Minority Oversampling Technique* (SMOTE) que é um algoritmo de dados artificiais com base na semelhança existente entre a classe minoritária (SCHIAVONI, 2010). Tem a capacidade de gerar artificialmente e poder duplicar ou multiplicar dependo dos exemplos da classe minoritária arrastando os vizinhos mais próximo do seu radar, levando a um conjunto de dados mais balanceado (CHAWLA 2002).

A técnica SMOTE, foi executada neste trabalho com o objetivo de refazer o ajuste de frequência relativa de classes majoritárias e minoritárias, introduzindo sinteticamente instâncias de classes minoritárias (WITTEN; FRANK; HALL, 2011). Empregou-se essa técnica SMOTE, disponível na Weka, como um filtro supervisionado, onde foi possível fazer alterações nas definições dos parâmetros ao aplicar esse método na ferramenta, com o percentual de sobre amostragem que foram

usados 700% para base da UNESC, 200% ACAFE, e SANTA CATARINA 200% e o número de vizinhos, utilizando o valor 5, sugerido pela ferramenta Weka. Segue-se na tabela 3 as informações de quantidade de classes minoritária e majoritária, e a quantidade de instância anterior e atual.

Tabela 4 – Tabela dos dados balanceados.

Base de dados	Classes (maior - minoritária)	Instância (antes)	Percentagem	Instância (atual)	Classes (maior - minoritária)
UNESC	**159 - 6**	**165**	**700%**	**207**	**159 - 48**
ACAFE	**667 – 114**	**781**	**200%**	**974**	**667- 307**
SANTA CATARINA	**898 - 195**	**1093**	**200%**	**1483**	**898 - 585**

Fonte: Do autor.

6.1.3 Etapa de mineração de dados

É nesta fase da metodologia que se explora a etapa de mineração de dados para a descoberta de conhecimentos em dados acadêmicos do ENADE. Ainda nessa etapa empregou-se as técnicas de algoritmo definidos. Esta é a fase em que os algoritmos entram em ação para mineração de dados. Em outras palavras, é onde a informação é descodificada para que se possa obter um novo conhecimento.

A ferramenta utilizada para utilização do processo de descoberta de conhecimento, é o conhecido software Waikato Environment for Knowledge Analysis (WEKA) é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados (WAIKATO, 2019). Ele contém ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização. O WEKA foi desenvolvido pela Universidade de Waikato em Nova Zelândia é um software *open source*. (WAIKATO, 2019). Para se pôr em prática foi necessário um estudo antecipado, de como usar e configurar os algoritmos.

Na utilização de consumir técnicas de *mineração de dados* com algoritmos de classificação é necessário que a base de dados seja repartida em dois conjuntos: treinamento e testes, os modelos são alcançados por intermédio de um conjunto de treinamento e logo são empregados para classificar as instâncias separadas no conjunto de teste. Para estratificação dos conjuntos foi utilizado o método chamado *Cross-Validation* ou validação cruzada (WITTEN; FRANK; HALL, 2011).

A validação cruzada serve para definir um conjunto de dados para testar o modelo durante o período de treinamento. Considerando que os dados não particionado em subconjuntos, uns para treinamento e outros para teste. No próximo passo, haverá troca de dados, a cada iteração o subconjunto passara a ser conjunto de teste e outro passarão a ser pertences ao conjunto de treinamento e vice-versa. O erro total +e dado pela quantidade de execuções de dados (TAN; STEINBACH; KUMAR, 2009).

A aplicação de validação cruzadas de *k* generaliza está abordagem decompondo os dados em *k*.Este processo é repetido K *k* vezes de forma que cada partição seja utilizada para teste (TAN; STEINBACH; KUMAR, 2009).

Ela inclui a partição do conjunto de dados de treinamento em subconjuntos, onde um subconjunto é mantido para testar o desempenho do modelo. é utilizada quando o conjunto de dados que possuímos é limitado (CECHINEL; CAMARGO, 2016). A quantidade de *folds* utilizada é de 10 que é normalmente o padrão aplicado em validação cruzada no WEKA.

A repartição de dados em 10 partições (10-*folds*) tem se tornado uma estratégia padrão, visto que, testes em vários bancos de dados e com diferentes técnicas de mineração têm mostrado que 10 seria um número adequado para obtenção de uma boa estimativa de erro (*WITTEN; FRANK; HALL*, 2011).

Após ter passado por esses todos os passos, passamos para fase de classificação. Onde foi selecionado o modelo de classificação que é a árvore de decisão, para poder fazer a escolhas dos classificadores que são o C4.5 e *Hoeffding Tree*.

Com o intuito de exibir padrões presentes relacionados dentro da base de dados., optou-se por utilizar os algoritmos de aprendizado da árvore de decisão. É um dos mais utilizados e intuitivos, pois nele os padrões encontrados são modelados e exibidos em formato de árvore, com isso torna fácil o compreendimento na observação de ligações de um padrão de árvore (PANG-NING; STEINBACH; KUMAR, 2009). Para tal foram selecionados os algoritmos C4.5 E Hoeffding tree.

Estes algoritmos foram escolhidos através do vasto uso em tarefas de mineração de dados e sendo capaz de apresentar os conhecimentos para a tomada de decisão. Hoeffding Tree tem algumas características que melhoram o seu desempenho, sendo capaz de aprender com as classes desequilibradas e erro de classificação assimétrico, com base na sua natureza incremental.

C4.5 propõem uma aprendizagem em uma árvore de decisão, a procurar um espaço de hipóteses completamente expressivo e assim evitar as dificuldades de espaços de possibilidades restritas.

Hoeffding tree elabora uma árvore de decisão com a informação do ganho, podendo aplicar estatística suficiente para a contagem de um valor discreto.

O algoritmo pode ser modificado, selecionando "*choose*" na primeira opção que aparece dentro "*classify*" que é *Classifier*, apresentada na Figura 19. Com isso, uma tela de pesquisa irá de abrir, onde terá a possiblidade de escolher a pasta "*trees*", onde se encontrará todos os algoritmos de árvore de decisão, na qual tivemos a opção de selecionar dois algoritmos.

6.2 Experimento Realizado

Com os todos os passos prosseguidos e escritos nos subcapítulos anteriores foram desenvolvidos 6 experimentos descritos na tabela 5, com intuito de serem destinado analisar o comportamento dos classificadores que foram apurados em relação aos conjuntos de bases dados originais. Tendo em vista as considerações causadas pelo pré-processamento e o balanceamento de classes, que influenciariam no impacto de analises de medidas de qualidade e no desempenho dos algoritmos escolhidos para a classificação.

Tabela 5 – Descrição dos experimentos realizado.

Experimentos	Descrição
1	Base de dados da Unesc (discretizadas)
2	Base de dados da Unesc (discretizadas e balanceada)
3	Base de dados da Acafe (discretizadas)
4	Base de dados da Acafe (discretizadas e balanceada)
5	Base de dados de Santa Catarina (discretizadas)
6	Base de dados de Santa Catarina (discretizadas e balanceada)

Fonte: Do autor.

6.3 ANÁLISE RESULTADO OBTIDOS

Com a pesquisa efetuada neste trabalho e a execução dos fundamentos alcançados por meio de pesquisas bibliográficas, foram organizadas 3 bases

diferentes de uma determinada região catarinense em seus respetivos anos de 2011, 2014 e 2017 no curso de ciência da computação. Foi definido o algoritmo para a *data maning*, teve várias combinações possíveis para execução dos classificadores na ferramenta Weka. Tendo em vista alguns pontos principais de medidas para avaliação dos modelos mostrou-se os primordiais que são curva Roc,acurácia, taxade verdadeiros positivos ou TP-Rate, matriz de confusão e coeficiente Kappa.

Com bases aos autores Tan, Steinbach e Kumar, as medidas adequadas para avalição de performance ou desempenho geral são: medidas como acurácia e kappa. Tendo em vista que algumas medidas são mais eficazes para avaliação dos modelos balanceados, mostrou-se os primordiais para análise em questão aos modelos ao ser operados, que são: acurácia, taxa de verdadeiros positivos ou TP-Rate e coeficiente Kappa

A eficiência das medidas de qualidade empregadas e descritas nas etapas anteriores, gerando consequência nos resultados para análise e avalição das medidas, podendo assim, ser possível identificar o modelo final mais apropriado para as 3 bases de dados geradas.

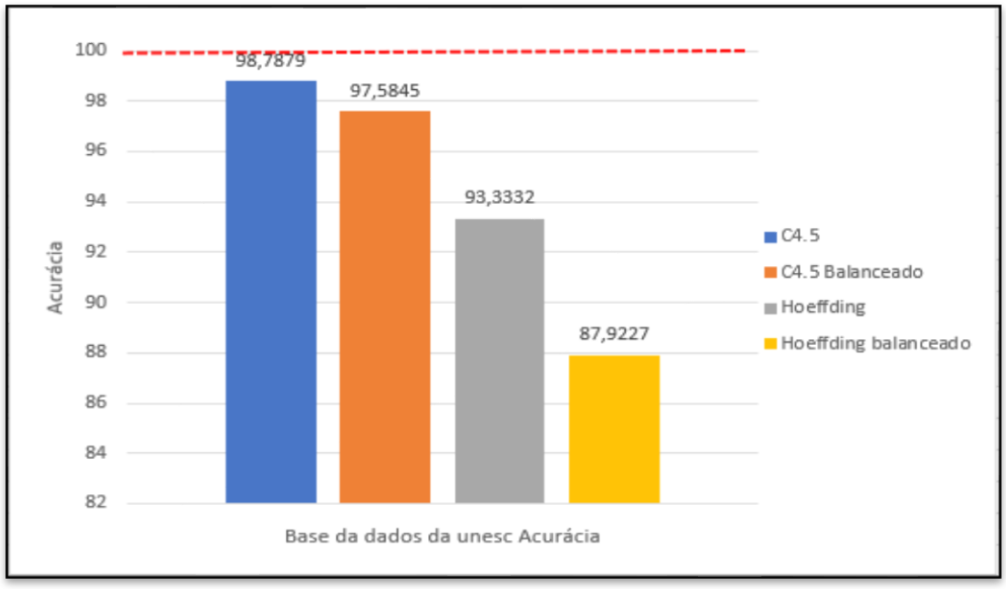
6.4.1 Resultados gerados pela base da UNESC com os algoritmos C4.5 e *Hoffding tree*

Com os resultados alcançados pela base da UNESC (discretizada) foram executados dois classificadores, assimilando a classe escolhida. Foram discutidos por meio de medidas de classificação como acurácia, coeficiente *kappa*, percentuais de verdadeiros positivos e *F-measure*.

Com base no teste feito no modelo da base da UNESC, a figura 14 representa os resultados alcançados com percentual da acurácia, com os dois algoritmos citados, para a realização da execução das bases de dados geradas. Com a observação no gráfico pode-se notar uma linha vermelha tracejada denominada como meta, ela é encontrada em todos os gráficos, quanto mais à taxa de acurácia se avizinha-se da linha de meta, maior é o percentual de acertos do modelo da base gerada, no entanto, deve-se cuidar do *overfitting*[5].

[5] *Overfitting* ocorre no momento em que o classificador está demasiadamente adaptado ao conjunto de treinamento.

Figura 14 – Base da Unesc.



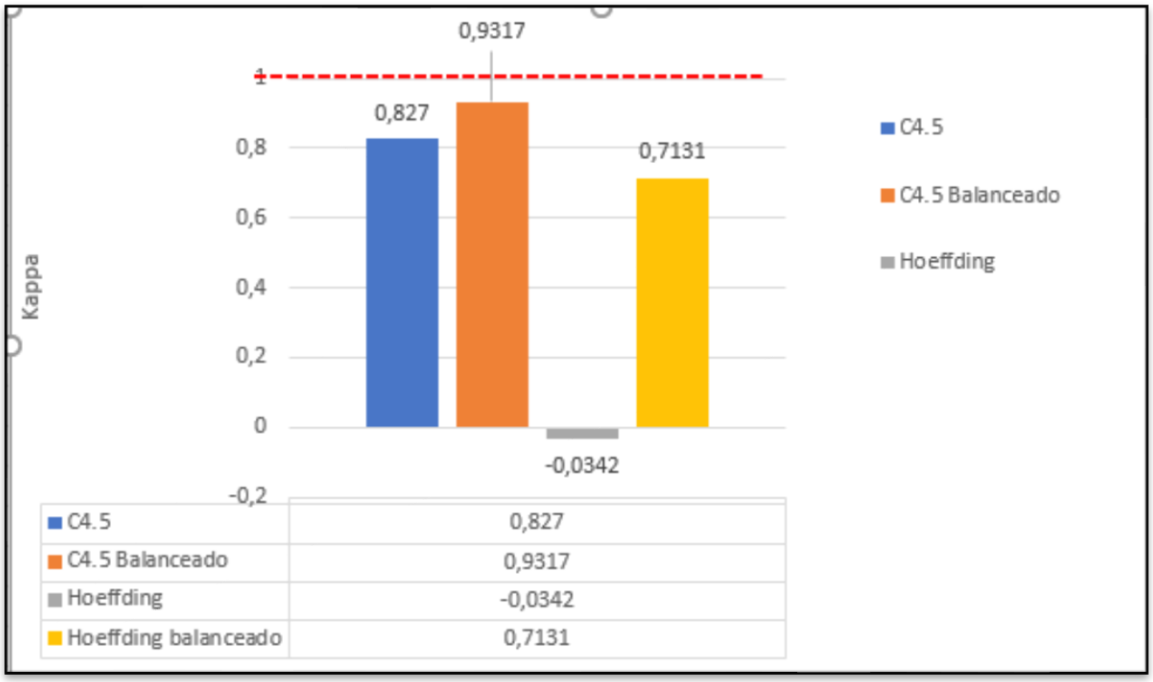
Fonte: Do autor.

Constata-se que os classificadores empregados alcançaram resultado com base o pré-processamento, a classe escolhida e o balanceamento aplicado, com a percentagem mais altas e o algoritmo C4.5 98,79% com a base de dados dados da UNESC (discretizadas), e segundo ponto mais alto também alcançado com o algoritmo C4.5 mas com alteração na base de dados da unesc (discretizadas e balanceada) com o valor 97,58. Analisando a linha de meta tracejada de no gráfico e com o comportamento dos algoritmos executado, todos alcançaram as taxas mais próximas da linha meta.

Designa-se que a base de dados da unesc (discretizadas e balanceada), aplicando o algoritmo *hoeffding tree,* que com isso atingiu a taxa de percentual mais baixo relacionando ao desempenho de outros testes na base. Com base a (figura 14), o resultado obtido no percentual é 87,92%.

A partir das medidas de qualidades foi avaliado o resultado do coeficiente kappa gerado pela base de dados da UNESC (discretizada) e (discretizadas e balanceada). Do mesmo jeito que na acurácia foi apresentada a linha de meta, os valores mais adjacentes a 1 são os modelos que alcançaram o melhor desempenho nesta medida (figura 15).

Figura 15 - Coeficiente Kappa na base da Unesc.



Fonte: Do autor.

A amostra mais alta alcançada ocorreu com o classificador C4.5 que teve mudança no balanceamento de instância a atingiu o valor de 0,93. Pode-se observar que os valores mais próximos da linha de meta.

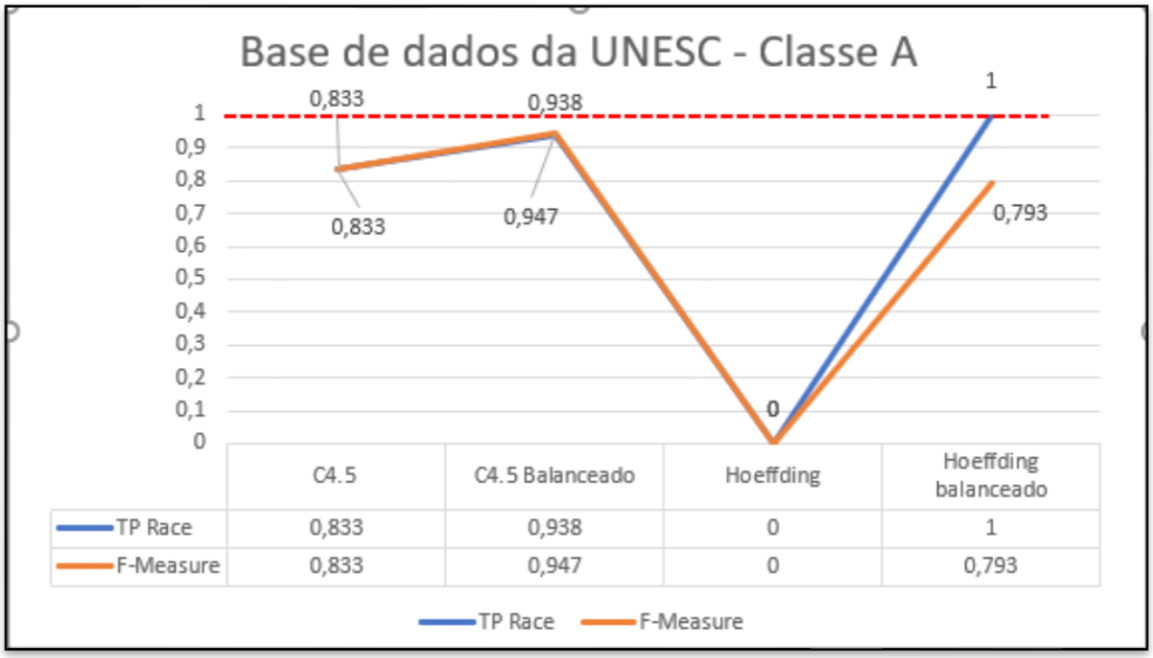
Em relação ao comportamento da execução geral no modelo da base da UNESC tendo em conta acurácia quanto ao coeficiente *Kappa*, os melhores resultados foram encontrados no desempenho do algoritmo C4.5, onde acurácia teve mais alto na UNESC (discretizadas) e o coeficiente *Kappa* na base da UNESC (discretizadas e balanceadas) com aplicação da técnica *SMOTE* com percentual de 700%.

Em seguida analisou-se a performance do modelo gerado nas classes “A” que representa nota igual ou de maior que 54.9 e classe “R” que representam as notas menor que 54.9, verificou-se a taxa de verdadeiros positivos e a medida *F-Measure.* delas.

Nas taxas de verdadeiros positivos para a classe “A” que chegaram mais próximo da vizinhança da linha de meta na (figura 16), mostra que a classificador *hoffding tree* alcançou a alinha de meta com o valor 1 tendo em conta que foi aplicado o balanceamento da técnica *SMOTE* com a percentagem 700%, e o classificador C4.5 aproximou-se da linha de meta com o resultado 0,94 também teve a mesma

quantidade de alteração no balanceamento. O gráfico também expressa o classificador com valor mais distante da linha de meta que é o *hoeffding tree*.

Figura 16 - *F-measure* e *TP-Rate* da classe "A".

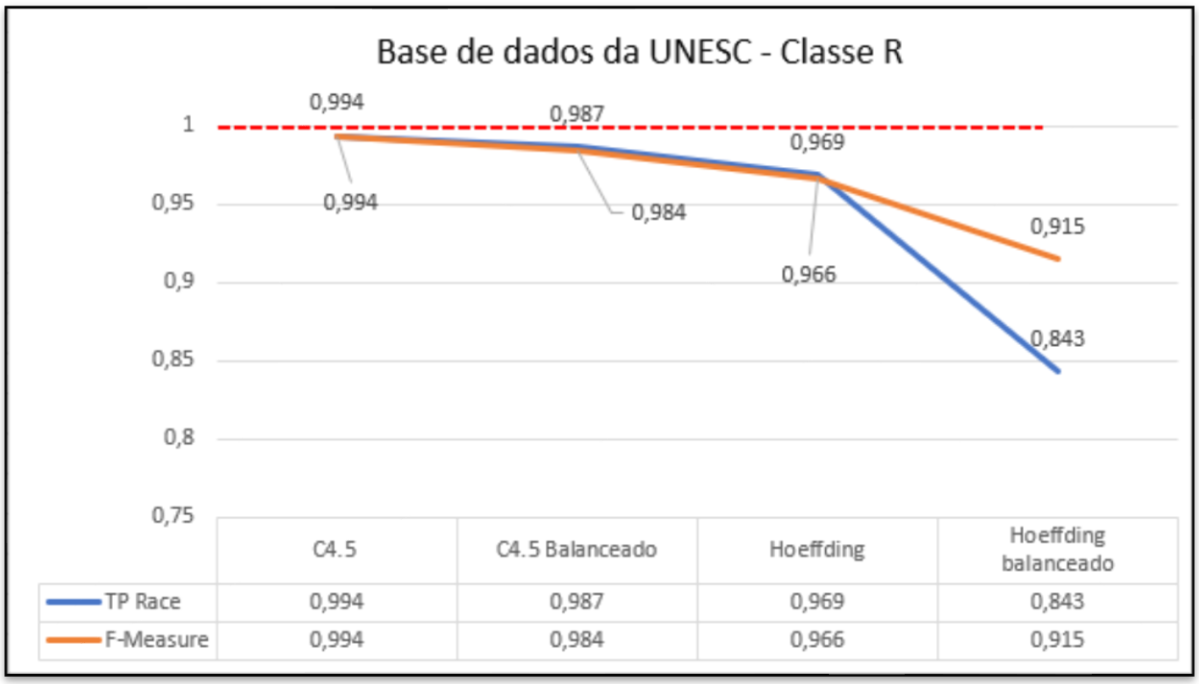


Fonte: Do autor.

Partindo para analisar a medida *F-Measure* para a classe “A”, pode-se enxergar no gráfico na (figura 16), que o classificador com o maior resultado foi o C4.5 balanceado que teve o valor 0,95. O gráfico também expressa o classificador com valor mais distante da linha de meta que é o *hoeffding tree* que se encontra com o valor zero.

Também analisou-se a taxas de verdadeiros positivos para a classe “R” que chegaram mais próximo da vizinhança da linha de meta (figura 17), mostra que a classificador C4.5 alcançou a alinha de aproximação com o valor 0,99 tendo em conta que foi na base original, e com a base balanceada se aproximou da linha de meta com o resultado 0,99. O gráfico também expressa o classificador com que mais se distância da linha de meta que é o *hoeffding tree* com o resultado 0,84.

Figura 17 - *F-measure* e *TP-Rate* da classe "R".



Fonte: do autor.

Partindo para analisar a medida *F-Measure* para a classe “R”, pode-se perceber que no gráfico (figura 17), que o classificador com o maior resultado foi o C4.5 balanceado que teve o valor 0,99. O gráfico também agrega informação sobre o classificador com valor mais distante da linha de meta que é o *hoeffding tree* que se encontra com o valor 0,91.

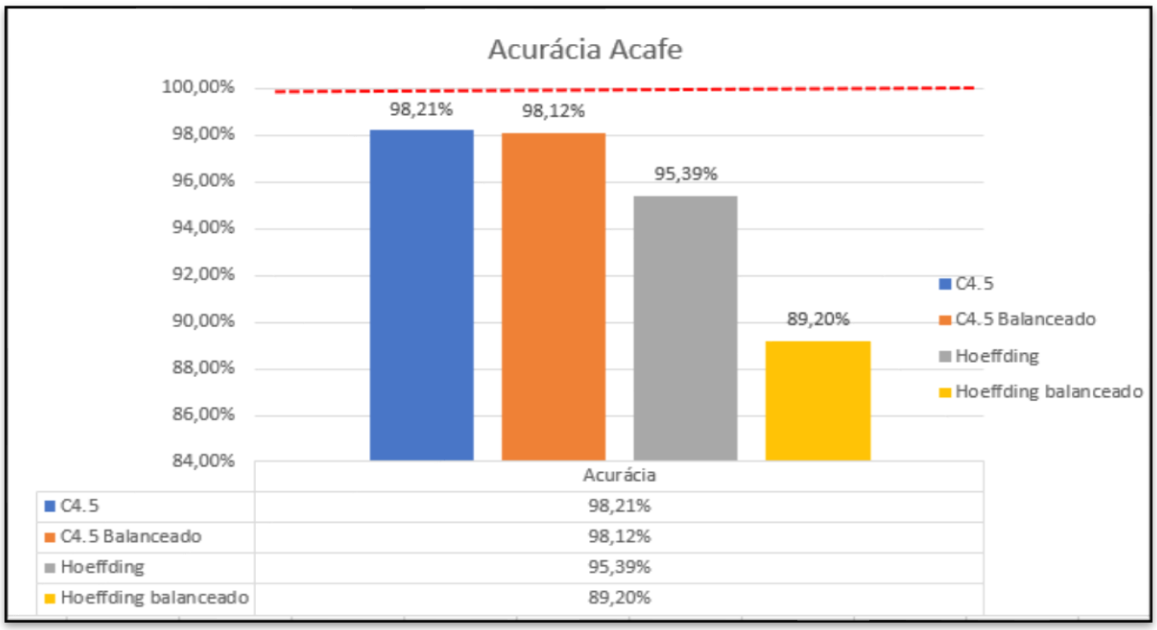
6.4.2 Resultados gerados pela base da Acafe com os algoritmos C4.5 e *Hoffding tree*

Conforme os resultados obtidos na base da Acafe (Discretizada) e Acafe (Discretizada Balanceado) foram empregados dois diferentes classificadores, consoante à classe selecionada (APR_REP). Foram debatidos por meio de medidas de classificação como acurácia, coeficiente *Kappa*, percentuais de verdadeiros positivos e *F-measure*.

Diante da execução dos classificadores empregados alcançaram as taxas de percentuais mais altos, o algoritmo C4.5 98,21% com a base de dados da ACAFE (discretizadas), e segundo ponto mais alto também alcançado com o algoritmo C4.5, mas com alteração na base de dados da ACAFE (discretizadas e balanceada) com o valor 98,12%. Diante a linha de meta tracejada no gráfico e com o comportamento dos

algoritmos executado, todos alcançaram as taxas mais próximas da linha meta (figura 18).

Figura 18 – Acurácia da base da Acafe.

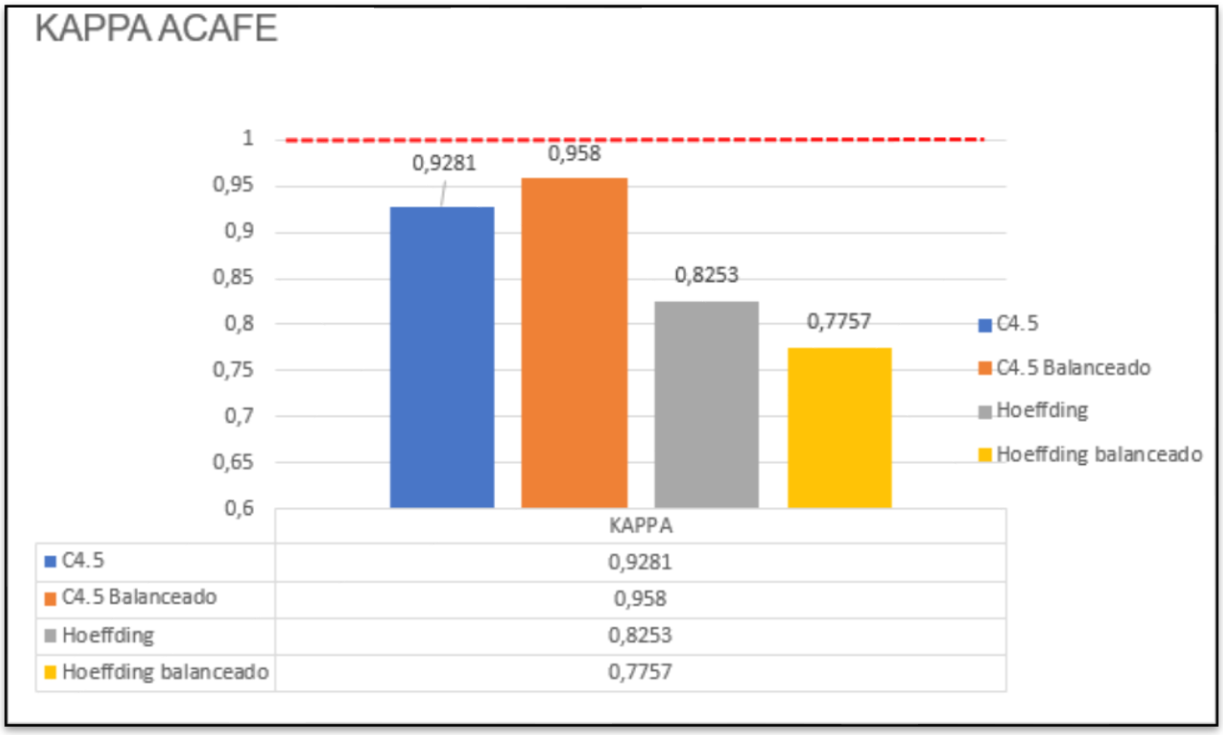


Fonte: Do autor.

Os percentuais de acurácia que mais se distanciaram da vizinhança da linha de meta e o *hoeffding tree* com 89,20% que teve alteração na base original com a aplicação da técnica SMOTE (figura 18).

Conforme a medida de qualidade foi avaliada o resultado do coeficiente *Kappa* gerado pela a base de dados da Acafe (discretizada) e (discretizadas e balanceada)., os valores mais adjacentes a linha de meta é 0,96 do algoritmo C4.5 que teve como consequência a técnica de balanceamento. Considerando também que o *Hoeffding Tree* teve o menor valor atingido sobre o efeito de aplicação de balanceamento alcançado com o valor de desempenho de 0,78 (figura 19).

Figura 19 - Base da Acafe.

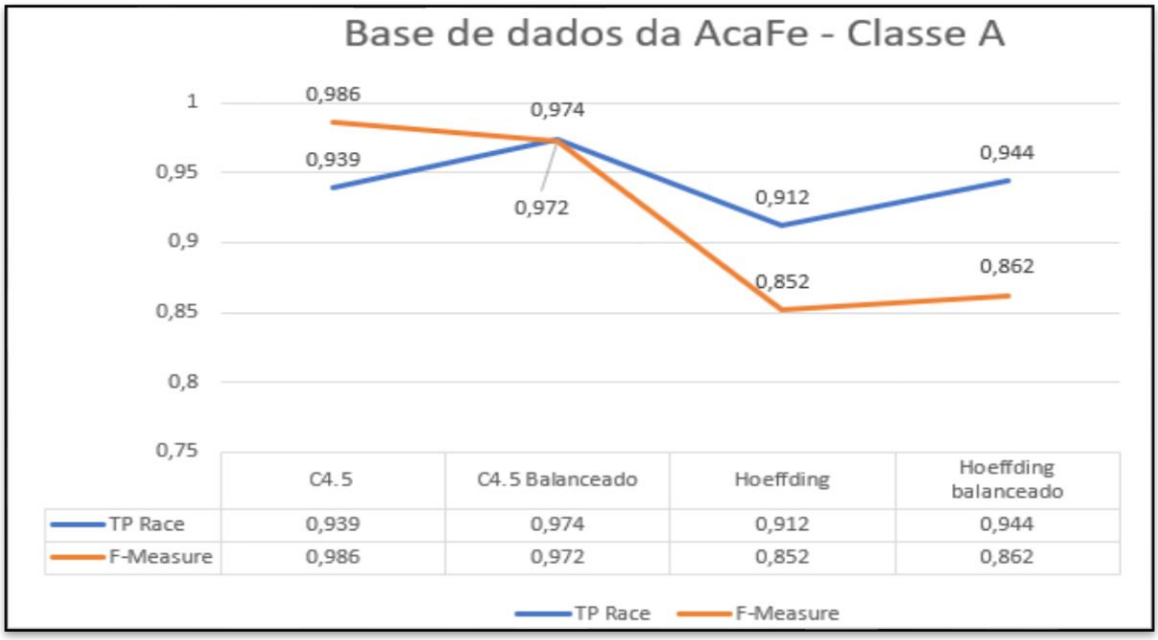


Fonte: Do autor.

No decorrer da análise das taxas de verdadeiros positivos da classe “A” (figura 20), verifica-se que o valor mais alto é obtido pelo algoritmo *C4.5* que empregou a técnica de balanceamento, chegando mais próximo da linha meta no gráfico com o valor de 0,97. Tendo em conta que o valor que mais se afastou da linha de meta não tem um valor tão distanciado dos outros apresentados no gráfico, O resultado alcançado foi com o algoritmo hoeffding tree de 0,91.

Procedendo com um dos elementos de análise de medida *F-Measure* para a classe “A”, pode-se enxergar no gráfico (figura 20), que o classificador com o maior resultado foi o C4.5 que tem o valor 0,99. O gráfico também expressa o classificador com valor mais distante da linha de meta que é o *hoeffding tree* que se encontra com o valor 0,85 na base original da ACAFE.

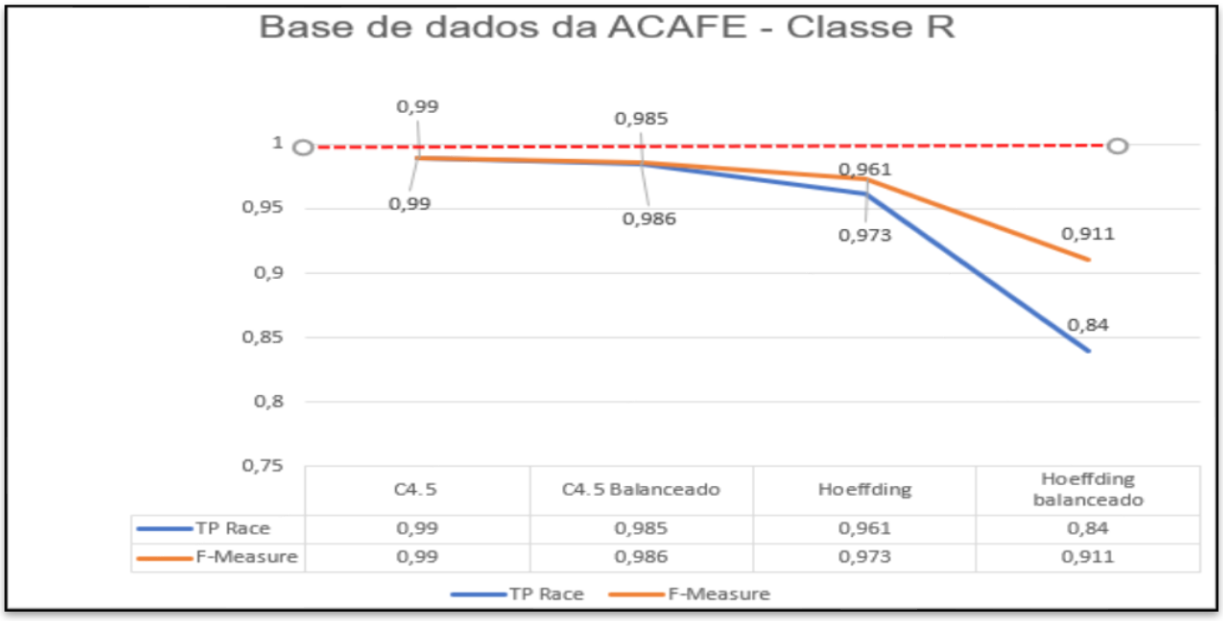
Figura 20 – ACAFE classe "A".



Fonte: Do autor.

Tratando-se da classe “R”, por sua vez, dispõem-se resultados com valores mais chegado da vizinhança da linha meta, tendo em conta a taxa de verdadeiros positivos, ao utilizar C4.5, chegando a valor 0,99 na base original da Acafe, conforme (figura 21). De acordo com a ilustração do gráfico encontra-se a taxas mais distantes da linha meta em relação aos outros testes, possuindo assim o valor 0,84.

Figura 21 - *F-Measure* e *TP-Rate* da classe "R".



Fonte: Do autor.

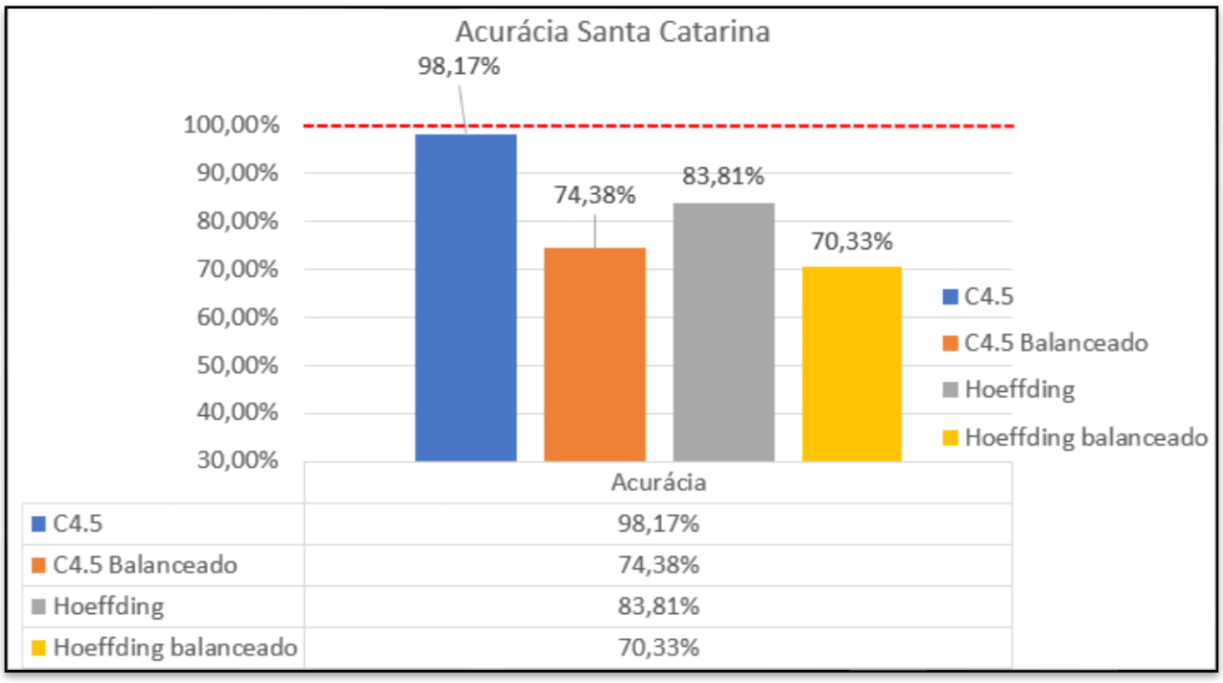
Com o destaque na análise de medida *F-Measure* para a classe “R”, podendo enxergar no gráfico (figura 21), que o classificador que está mais perto da linha de meta e o C4.5 que tem o valor 0,99. O gráfico também informa o classificador que mais se distância da linha de meta com o resultado encontrado no algoritmo *hoeffding tree* tendo valor de 0,91 na base original da ACAFE com o efeito do balanceamento.

6.4.3 Resultados gerados pela base da Santa Catarina com os algoritmos C4.5 e Hoffding tree

Diante do protótipo obtido da base dados de Santa Catarina foram debatidos por meio de medidas de classificação como acurácia, coeficiente *Kappa,* percentuais de verdadeiros positivos e *F-Measure*.

Conforme a execução dos algoritmos utilizados alcançou os percentuais de acurácia, podendo analisar na (figura 22), informando o valor do algoritmo C4.5 98,17% com a base de dados Base de dados da Unesc (discretizadas balanceada).

Figura 22 – Base de dados de Santa Catarina.

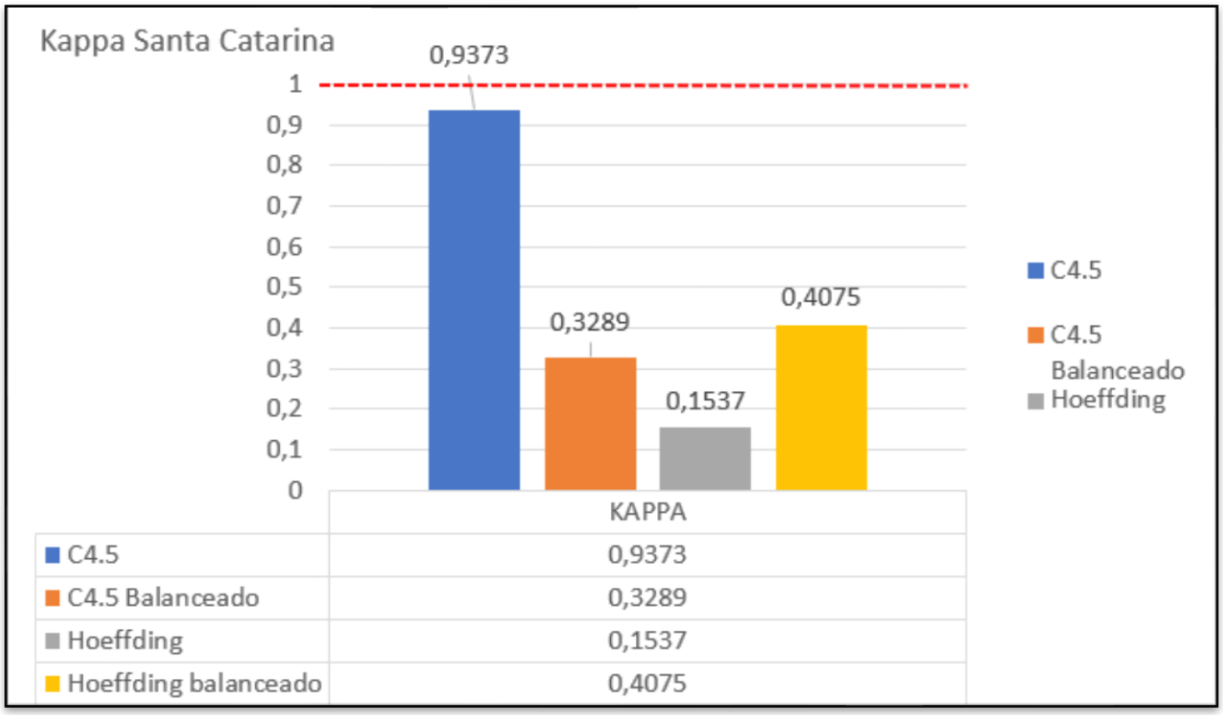


Fonte: Do autor.

Os percentuais de acurácia que mais se distanciaram da vizinhança da linha de meta e o *hoeffding tree* com 70,33% que teve alteração na base original com a aplicação da técnica SMOTE.

O resultado a seguir foi o avaliado por coeficiente *Kappa*, os valores encontrados mais adjacentes a linha de meta é 0,94 do algoritmo C4.5 na base da de Santa Catarina. Considerando também que o *hoeffdin tree* que teve o menor resultado atingido na base original da ACAFE alcançado com o valor de desempenho de 0,15 (figura 23).

Figura 23 - Base da Santa Catarina.

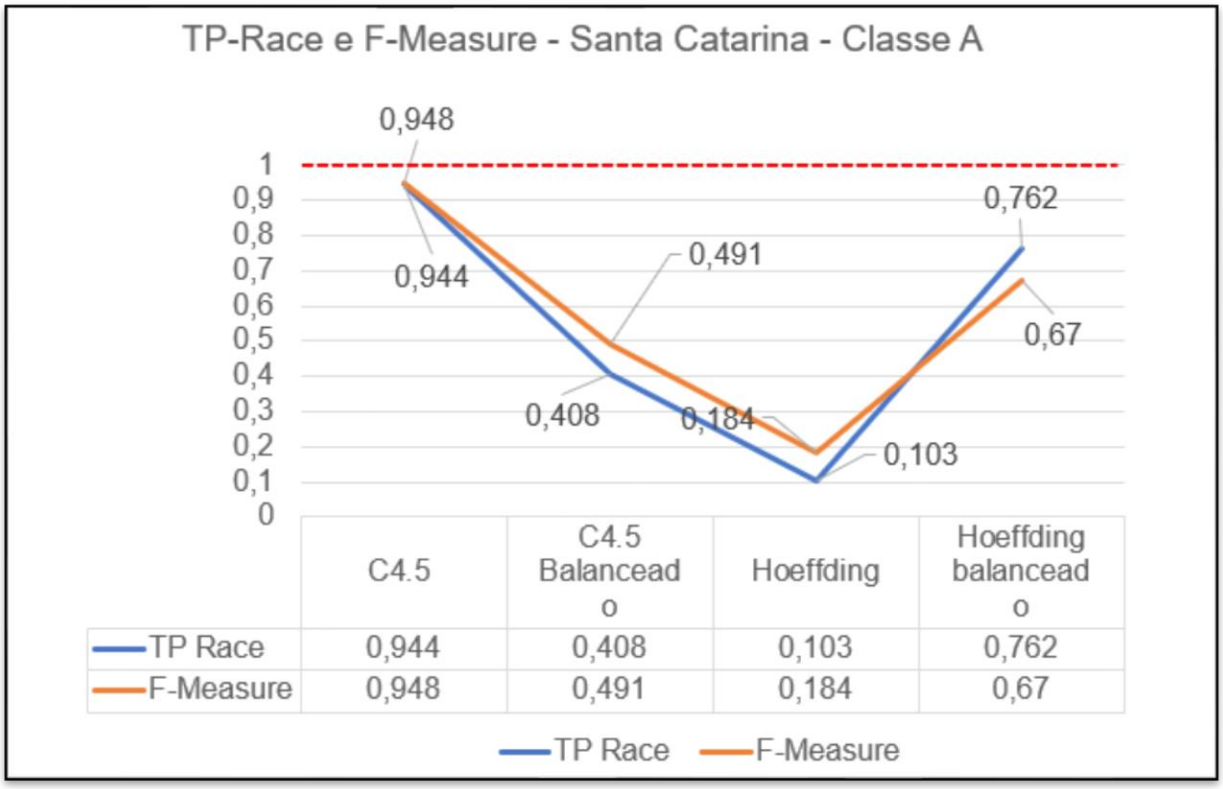


Fonte: Do autor.

No de correr da análise das taxas de verdadeiros positivos da classe “A” na (figura 24), verifica-se que o valor mais alto é obtido pelo algoritmo C4.5 chegando próximo da linha meta no gráfico com o valor de 0,94. Tendo em conta que o valor que mais se afastou da linha de meta apresentado no gráfico alcançou o valor foi com o algoritmo *hoeffding tree* 0,10.

Observando a medida *F-Measure* para a classe “A”, pode-se afirmar que na figura 24, o classificador com o maior resultado foi o C4.5 com o valor 0,94. O gráfico também expressa o classificador com valor que se distância da linha de meta que é o *hoeffding tree* que se encontra com o valor 0,18.

Figura 24 - *F-Measure* e *TP-Rate* da classe "A".

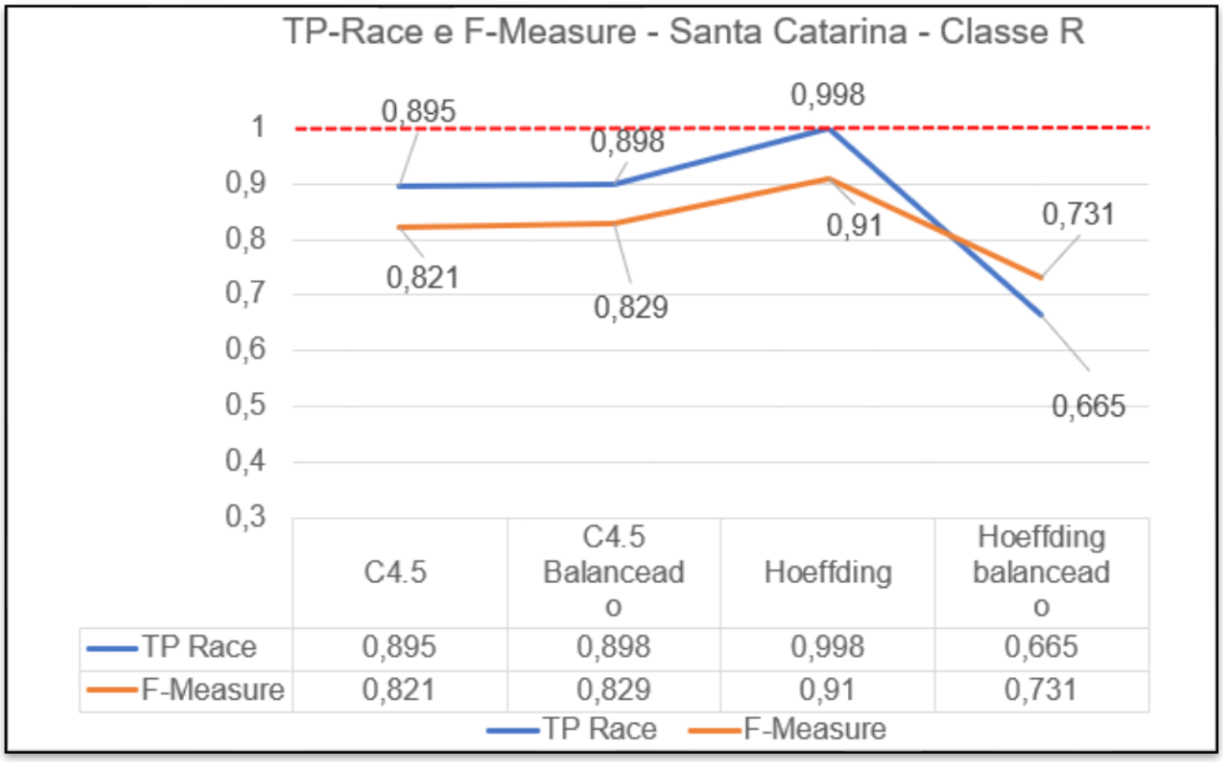


Fonte: Do autor.

A classe “R”, por sua vez, apresenta resultados com valores mais próximo da vizinhança da linha meta, tendo em conta a taxa de verdadeiros positivos, ao utilizar *hoffding tree* tendo o valor mais alto 0,99 na base original, conforme (figura 25). De acordo com a ilustração do gráfico encontra-se a taxas mais distantes da linha meta em relação aos outros testes, possuindo assim o valor 0,66 na base balanceada.

Com o destaque na análise de medida *F-Measure* para a classe “R”, podendo notar na (figura 25), o classificador que está mais perto da linha de meta e o *hoeffding tree* na base original com valor 0,91. O gráfico informa o classificador que mais se distância da linha de meta com o resultado encontrado no algoritmo *hoeffding tree* tendo valor de 0,73 com alteração no balanceamento.

Figura 25 - F-measure e TP-Rate da classe "R".



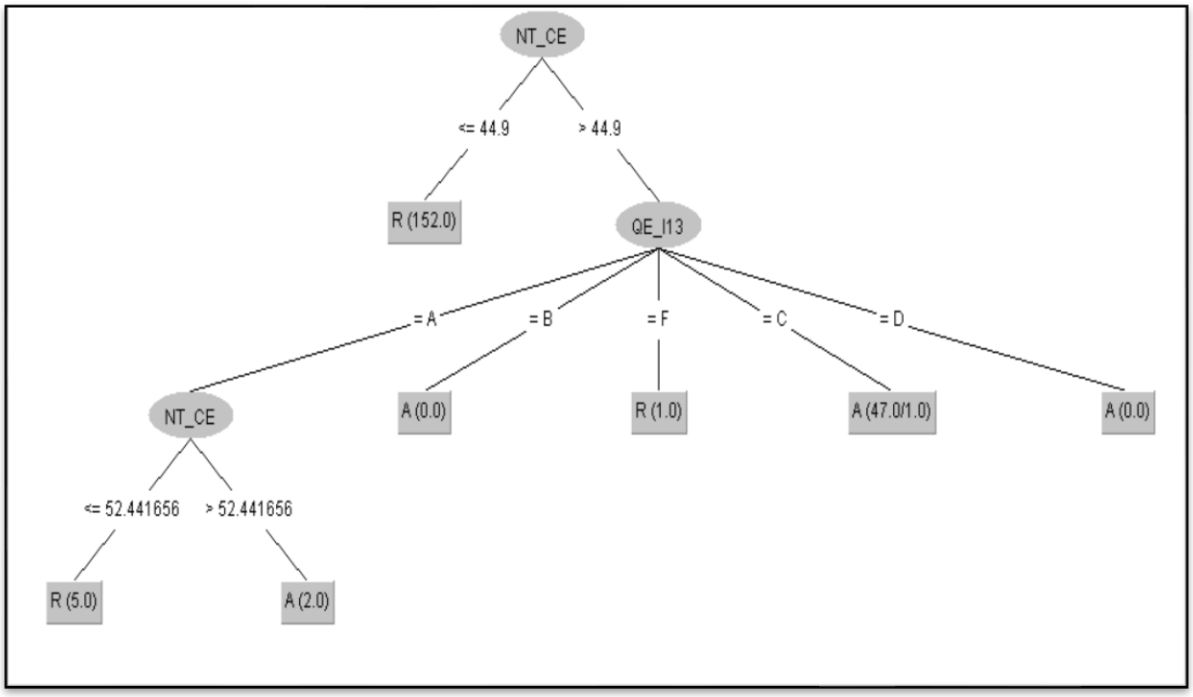
Fonte: Do autor

6.4.4 Gerando árvore de conhecimento nas três bases geradas

Árvore gerada com a base de Unesc balanceada aplicando o algoritmo C4.5, nos dados de ciência da computação

Observando a Figura 26, é possível analisar as principais amostras referentes às notas brutas no componente específico (NT_CE). Nota-se que cento e cinquenta e dois estudantes tiveram uma nota específica menor ou igual que 44,9; nota-se também que os estudantes que tiveram nota específica superior a 44,9 se possuíam bolsas de estudo (QE_I13). Quarenta e oito estudantes responderam que possuíam bolsa de extensão e um estudante, que possuía outro tipo de bolsa de estudos acadêmica. Cinco estudantes que tiraram a NT_CE inferior a 52,44 não possuíam nenhuma bolsa e dois estudantes que também tiveram nota superior a 52,44 responderam que não.

Figura 26 - Árvore de decisão referente a base da UNESC



Fonte: Do autor

Árvore gerada com a base de ACAFE aplicando o algoritmo C4.5, nos dados de ciência da computação, na (figura 27) são demonstrados os resultados mais significativos pertinentes ao Tipo de situação da questão 1 da parte discursiva do componente específico (TP_SCE_D1).

A figura 27 apresenta a árvore de decisão gerada a partir das principais regras derivadas das Tais árvores possibilita uma visualização mais intuitiva dos padrões encontrados, facilitando assim, a obtenção de conhecimento.

Destaca-se Tipo de situação da questão 1 da parte discursiva do componente específico (TP_SCE_D1) menor ou igual 333 foram destacados como 380 alunos ausente e presente que deixaram questão 1 da parte discursiva do componente específico em branco. E estudantes que selecionar opção maior que 333 apresentam uma situação de carga hora de trabalho (QE_I10), em que 318 estudantes de ciência da computação responderam que tem uma atividade diária de trabalho de 40 horas semanais e 50 deles tiveram uma nota superior a 54,9, 14 estudante que tiveram a nota menor que 54,9 responderam que tem 20 horas semanais, e outros 38 estudantes que semanalmente tem uma carga horária de 21h a 39h 6 deste estudante tiveram uma nota superior que 54,9 e 32 ficaram com a nota menor.

Continuando na figura 27: dezenove estudantes responderam que trabalham eventualmente, estes também ficaram na condição de situação financeira incluindo bolsa (QE_I09), dois destes estudantes tem a condição financeira de não ter renda e gastos financiados por programas governamentais e tem a nota superior que 54,9; oito dos alunos ficaram com a nota menor que 54,9 e assumiram ter renda e receber ajuda da família ou de outras pessoas para financiar gastos; nove estudantes responderam não ter renda e possuir gastos financiados pela família ou por outras pessoas; sete deles alcançaram uma nota superior que 54,9. Noventa e dois estudantes declararam que não possuem trabalho.

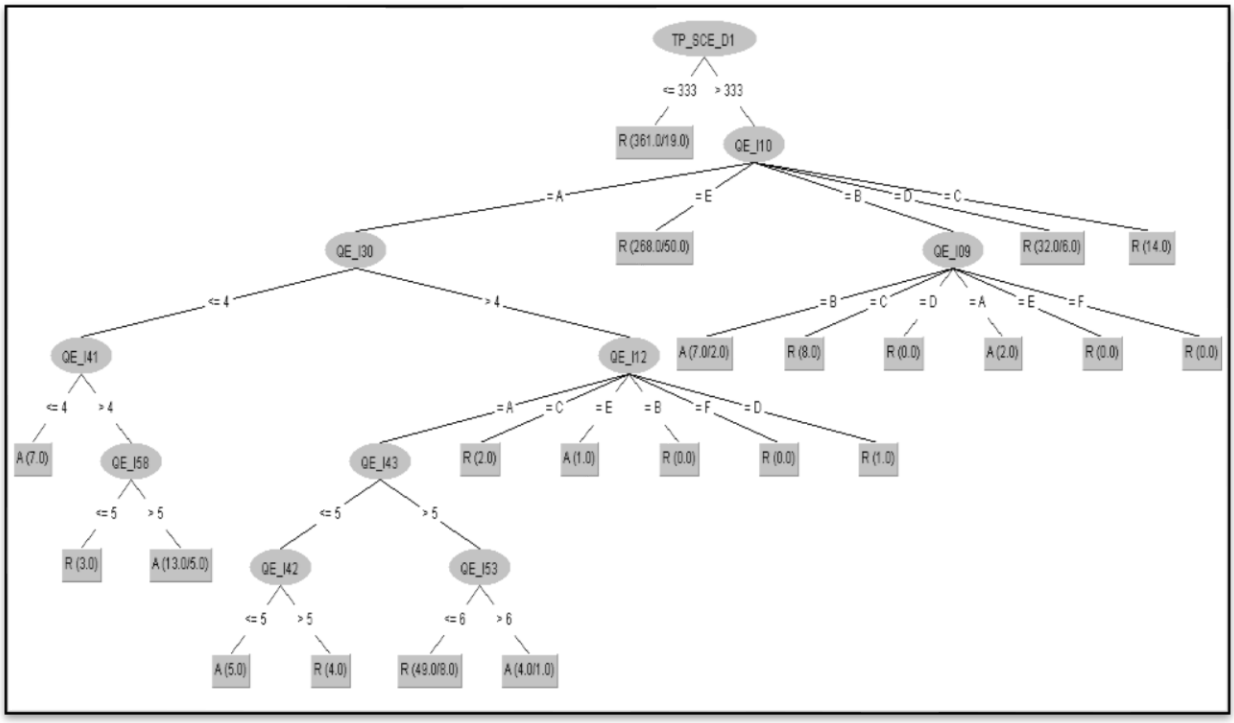
Observou-se que o curso propiciou experiências de aprendizagem inovadoras (QE_I30): vinte e um estudantes selecionaram valores menores ou iguais a quatro, e discordam das experiências de aprendizagem inovadoras; esteve disponível para orientação acadêmica dos estudantes (QE_41) alcançaram uma nota superior a 54,9 e entre vinte e um estudantes, alguns responderam que concordam totalmente, outros que não sabem responder e outros, não se aplicam, onde três estudantes ficaram com uma nota menor ou igual que 54,9; disseram que não recebem/discordam dos professores utilizaram tecnologias de informação e comunicação (TIC's), onde alguns dos dezoito estudantes declararam que concordam totalmente, não sei responder e não se aplica, sendo assim treze deles alcançaram uma nota superior a 54.

Os estudante que tiveram experiências de aprendizagem inovadoras (QE_I30), quatro estudantes ganharam bolsa ao longo da sua trajetória acadêmica; dois estudantes ganharam bolsa extensão e ficaram com a nota menor que 54,9; Um estudante que teve bolsa de monitoria/tutoria teve a nota menor que 54,9 e um estudante ganhou bolsa PET[6] e teve uma nota superior 54,9.

Notou-se também que para os alunos que não ganharam nenhuma bolsa, foram oferecidas oportunidades para que participassem de programas, projetos ou atividades de extensão universitária, sendo que nove estudantes discordam totalmente quatro deles ficaram com a nota menor ou igual que 54,9 e cinco alcançaram uma nota superior que 54,9.

[6]O Programa de Educação Tutorial (PET) insere estudantes de graduação em projetos de educação tutorial com o objetivo de aplicar seus conhecimentos e ampliar sua formação Brasil (2018).

Figura 27 – Árvore de decisão referente a base da ACAFE.



Fonte: Do autor.

A base de Santa Catarina gera uma árvore de conhecimentos em dados de Ciência da Computação, com o algoritmo C4.5 apresentado na figura 28 pertinentes ao tipo de bolsa de estudos ou financiamento do curso (QE_I11).

Na figura 28 observa-se conhecimento em que tipo de bolsa de estudos ou financiamento do curso você recebeu para custear todas ou a maior parte das mensalidades QE_I11. Dois estudantes que não alcançaram a nota superior a 54,9 possuíam bolsa de financiamento oferecida pela própria instituição; seis estudantes que faziam parte da bolsa financiamento bancário; cinco ficaram com a nota inferior a 54,9 e um estudante teve a nota superior que 54,9. A figura também apresenta que quatro estudantes fazem parte da bolsa ProUni Parcial e FIES. Observa-se que sessenta estudantes possuíam bolsa oferecida pela própria instituição e nove estudantes alcançaram uma nota superior 54,9; sessenta estudantes obtiveram uma nota inferior ou igual a 54,9.

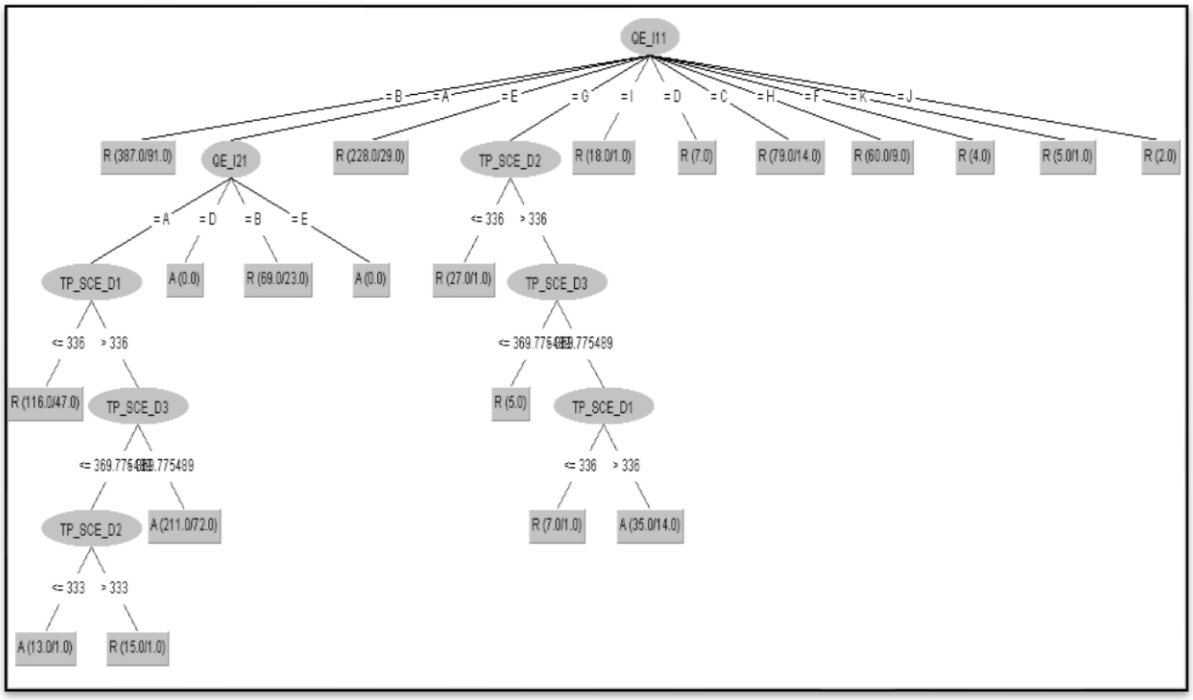
A figura 28 mostra que noventa e três estudantes com bolsa ProUni integral se destacaram: catorze estudantes com valores superiores que 54,9 e setenta e nove acadêmicos com valores menores ou iguais a 54,9; apenas sete estudantes fazem parte do ProUni parcial; também se constata que dezenove estudantes possuem bolsa oferecida por outra entidade (empresa, ONG, etc) e dezoito estudantes tiveram uma nota inferior ou igual a 54,9; um acadêmico teve o valor superior a 54,9.

Analisou-se também que trezentos e oitenta e sete acadêmicos tiveram uma nota menor ou igual que 54,9 e noventa e um acadêmicos tiraram valores maiores que 54,9. Nenhum desses estudantes possuía uma bolsa, embora seu curso não seja gratuito.

Dando continuidade na figura 28, observa-se que oitenta e nove acadêmicos faziam parte da bolsa oferecida pelo Governo Estadual, Distrital ou Municipal. Onde vinte e oito bolsistas do Governo nos tipos de situações da questão 2 da parte discursiva do componente específico TP_SCE_D2 optaram por questão em branco, questão zerada por motivo de resposta nula e questão zerada por motivo de resposta divergente com a temática, vinte e oito desses bolsistas alcançaram uma nota inferior que 54,9 e um bolsista teve a nota superior que 54,9.

A árvore de decisão também informa os estudandes que não têm nenhuma bolsa, pois o curso é gratuito, nota-se que noventa e dois acadêmicos não possuem alguém em sua família que tenha concluído um curso superior (QE_21); sessenta e nove estudantes tiveram o valor menor ou igual a 54,9 e vinte e três estudantes com a nota superior a 54,9.

Figura 28 - Árvore de decisão referente a base da Santa Catarina.



Fonte: Do autor.

6.3 DISCUSSÃO DOS RESULTADOS

Segundo Cretton (2016) realizaram quatro experimentos em dados educacionais da prova do ENADE, podendo assim analisar o desempenho da acurácia onde empregaram o algoritmo J48 que é o mesmo que o C4.5. Cada uma das quatro bases executadas obteve um nível de confiança distinto, com o valor da primeira base de 80,86%, na segunda 80,73%, na terceira 83,82% e na quarta base 84,05% com o valor mais alto.

Nessa sequência, notou-se que os resultados alcançados dos trabalhos do Cretton podem ser comparados com o desenvolvimento do trabalho atual. Observou-se que os melhores resultados alcançados nas três bases criadas, com aplicação do algoritmo C4.5 com os valores na base da Acafe 98,21%, Acafe sem UNESC 98,56%, UNESC 98,79% e Santa Catarina 98,79% superam os resultados do trabalho do Cretton.

Além disso, pode-se comparar o percentual de acurácia obtido por algoritmo C4.5 na base que tem resultado mais próximo da linha de meta com o valor de 98,79% com a base original da UNESC tendo em conta que é o melhor resultado obtido nas quatro bases geradas, comparando com o estudo feito por Rubert e Mariotto (2013), que aplicou a implementação dos classificadores de classificação e associação: J48 e A priori; Geração de perfis de usuários aplicando as técnicas de classificação e associação na identificação e classificação de indicadores de saúde. Alcançado assim o número de instâncias classificadas corretamente 85,61, tendo um desempenho muito menor que os trabalhos desenvolvidos na base da UNESC e outras bases do projeto.

Segundo o trabalho de Gomes, foram analisados 13 algoritmos, dos dois concedem com os meus classificadores usados, os resultados apresentados na pesquisa do Gomes foram *hoeffding tree* com o valor de 84,4956% e o C4.5 alcançou o resultado de 85,5433%. Esses são os resultados apresentados, que não foram mais eficientes que os da pesquisa atual onde o C4.5 teve como valor 98,79 e o *hoeffding tree* teve 95,39.

7 CONCLUSÃO

Levando-se em consideração a evolução da globalização e a alta demanda de informações, foi possível perceber que auxiliar na busca de respostas simplifica a compressão de diversos problemas pedagógicos. Este trabalho apresentou uma fundamentação sobre umas das ferramentas adequadas para se fazer a exploração dos possíveis conhecimentos.

Ao decorrer da introdução deste trabalho proposto, esta pesquisa realizou a mineração de dados educacionais por meio da tarefa de classificação nos dados do ENADE do Curso de Ciência da Computação. Dividindo a base em três modelos que são UNESC, ACAFE e Santa Catarina - os dados encontram-se disponíveis no portal do INEP. Diante disso foi exposto um referencial teórico sobre os temas tratados, assim como características e conceitos que foram selecionados para ajudar no processo de alcance dos objetivos propostos.

A mineração de dados traz um diferencial interessante para gerações de alertas importantes, podendo aplicar em qualquer uma das partes envolvidas de sistemas educacionais, como alunos, educadores, administradores e pesquisadores.

A intenção de fornecer *feedbacks* é a de ajudar nas orientações e melhorar o processo de descoberta de aprendizagem dos alunos. Ao fazer descobertas dos sistemas educacionais dos alunos, ajuda a melhorar o desempenho do ensino, as atividades propostas e a tomada de decisões.

Com o intuito de alcançar dados mais precisos, foram utilizados classificadores como C4.5 e *Hoeffding Tree*. Para a execução dos algoritmos foi necessário investigar como os dados estão organizados e aplicar o pré-processamento para avaliar qual atributo poderia ser definido como classe.

A ferramenta *Weka* teve uma grande importância para a realização da etapa de mineração de dados. A importação dos conjuntos de dados a serem explorados, implementa e executa os algoritmos de mineração de dados utilizados a tarefa de classificação de forma simples e flexível, mostrando-se uma ferramenta adequada para a metodologia de descoberta de conhecimento proposta.

Depois de ocorrerem diferentes análises comparativas entre as três bases geradas utilizadas na pesquisa, o conjunto de dados convenientes para classificação dos estudantes que tiveram uma nota superior a 54,9 e inferior ou igual que 54,9. A

base de dados da UNESC (discretizada) como o classificador C4.5 alcançou o valor de 98,79% nos dados originais, com apenas as classes discretizadas.

Durante o processo de desenvolvimento do trabalho, algumas dificuldades foram encontradas, umas foram compreendidas com o embasamento teórico adquirido no levantamento bibliográfico realizado. Uns dos maiores problemas encontrados foi a organização e a compreensão dos dados.

Recomendações para trabalhos futuros:

a) os algoritmos LMT e *randomforest* aplicados à mineração de dados educacionais do Exame Nacional de Desempenho de Estudante (ENADE) no curso de Pedagogia;
b) aplicação de regras de associação, voltados para a mineração de dados educacionais do Exame Nacional de Desempenho de Estudante (ENADE) em Ciência da Computação;
c) mineração de dados aplicada na base do ENADE com enfoque na criação de perfis dos estudantes que prestaram o exame no curso de Engenharia empregando classificadores.

REFERÊNCIAS

ABBAGNANO, Nícola. Dicionário de Filosofia. 3.ed. São Paulo: Mestre, 1982. 1.v.

ADAIME, Leonardo MÜller. **APLICAÇÃO DO VISUALIZATION TOOLKIT PARA PÓSPROCESSAMENTO DE ANÁLISES PELO MÉTODO DOS ELEMENTOS FINITOS.** 2005. 13 f. Dissertação (Mestrado) - Curso de Mecânica Computacional, Setores de Tecnologia e de Ciências Exata, Universidade Federal do Paraná, Curitiba, 2005. Disponível em: <https://www.acervodigital.ufpr.br/bitstream/handle/1884/3529/LeonardoAdaime-MSc.pdf?sequence=1>. Acesso em: 17 nov. 2019.

ALBERTO, B., Abordagens de Pré-processamento de dados em problemas de classificação com classes desbalanceadas., Master's Thesis, Centro Federal de Educação Tecnológica de Minas Gerais (Mestrado em Modelagem Matemática e Computacional), aug 2012.

ASSEISS, Maraísa da Silva Guerra. **APLICAÇÃO DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS ACADÊMICO UTILIZANDO AS TAREFAS DE AGRUPAMENTO E CLASSIFICAÇÃO.** 2017. 51 f. Dissertação (Mestrado) - Curso de Engenharia Elétrica., Automação, Unesp, Ilha Solteira, 2017. Cap. 4. <http://hdl.handle.net/11449/151251>. Acesso em: 19 set. 2018

AYUB, Mewati et al. Modelling online assessment in management subjects through educational mineração de dados. **2017 International Conference On Data And Software Engineering (icodse)**, [s.l.], v. 1, n. 6, p.1-6, nov. 2017. IEEE. http://dx.doi.org/10.1109/icodse.2017.8285881. Disponível em: <https://ieeexplore.ieee.org/document/8285881>. Acesso em: 02 out. 2018.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, [s.l.], v. 19, n.02, p.1-12, 31 ago. 2011. Comissao Especial de Informatica na Educacao. http://dx.doi.org/10.5753/rbie.2011.19.02.03. <http://www.upenn.edu/learninganalytics/ryanbaker/BD-RBIE-pt-v22.pdf>. Acesso em 15 set. 2018.

Barros, R. C.; Basgalupp, M. P.; De Carvalho, A. C.; Freitas, A. "A survey of evolutionary algorithms for decision-tree induction", IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 42–3, 2012, pp. 291–312. < https://ieeexplore.ieee.org/document/5928432 > Acesso em 25 set. 2018.

BAKHSHINATEGH, Behdad et al. Educational mineração de dados applications and tasks: A survey of the last 10 years. **Education And Information Technologies**, [s.l.], v. 23, n. 1, p.537-553, 3 jul. 2017. Springer Nature. http://dx.doi.org/10.1007/s10639-017-9616-z.

BARBOSA, Juliana Moreira; DE SENNA CARNEIRO, Tiago Garcia; TAVARES, Andrea Iabrudi. Métodos de Classificação por Árvores de Decisão Disciplina de Projeto e Análise de Algoritmos, 2012. < http://www.decom.ufop.br/menotti/paa111/files/PCC104-111-ars-11.1-JulianaMoreiraBarbosa.pdf > Acesso em 10 out. 2018.

BARRETO, Francisco Candido Cardoso. **Modelagem De Distribuição Potencial De Espécies Como Ferramenta Para Conservação: Seleção E Avaliação de Algoritmos E Aplicção Com Heliconuis Nattereri Felder 1865 (NYMPHALIDAE: HELICONIINAE.** 2008. 9 f. Tese (Doutorado) - Curso de Ciência da Computação, Universidade Federal de Viçosa, Viçosa, Mina Gerias, 2008.

BLAGOJEVIć, Marija; MICIć, Živadin. A web-based intelligent report e-learning system using mineração de dados techniques. **Computers & Electrical Engineering**, [s.l.], v. 39, n. 2, p.465-474, fev. 2013. Elsevier BV. http://dx.doi.org/10.1016/j.compeleceng.2012.09.011.

BRASIL. INEP. (Org.). **ENADE.** 2019. Disponível em: <http://portal.inep.gov.br/web/guest/microdados>. Acesso em: 10 maio 2019.

BRASIL. FNDE. (Org.). **Educação Tutorial.** 2018. Disponível em: <https://www.fnde.gov.br/index.php/programas/bolsas-e-auxilios/eixos-de-atuacao/educacao-tutorial>. Acesso em: 19 nov. 2019.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Estudo exploratório sobre o professor brasileiro com base nos resultados do Censo Escolar da Educação Básica 2007. Brasília, 2009a. Disponível em: <http://portal.mec.gov.br/dmdocuments/estudoprofessor.pdf>. Acesso em 25 Set. 2018.

BRASIL. UESB. (Org.). **Provas ENADE.** Disponível em: <http://www2.uesb.br/computacao/documentos/provas-enade/>. Acesso em: 18 nov. 2019.

BRUM, Bruna Conde Perez. **Estimação da Taxa de Churn para Clientes de uma Seguradora baseado em Técnicas de Reconhecimento de Padrões.** 2016. 24 f. Monografia (Especialização) - Curso de Business Intelligence, Engenharia Elétrica da Puc/rio, Rio, 2016. Disponível em: <http://www.ica.ele.puc-rio.br/Arquivos/monografias/TCC%20-%20BRUNA%20C.%20P.%20BRUM%20Estima%C3%A7%C3%A3o%20da%20Taxa%20de%20Churn%20para%20Clientes%20de%20uma%20Seguradora%20baseado%20em%20T%C3%A9cnicas%20de%20Reconhecimento%20de%20Padr%C3%B5es.pdf>. Acesso em: 19 out. 2018.

CARVALHO, Deborah Ribeiro et al. Mineração de Dados aplicada à fisioterapia. **Fisioterapia em Movimento**, [s.l.], v. 25, n. 3, p.595-605, set. 2012. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0103-51502012000300015.

CASTRO, Cristiano Leite de; BRAGA, Antônio Pádua. APRENDIZADO SUPERVISIONADO COM CONJUNTOS DE DADOS DESBALANCEADOS. **Scielo**

(revista Controle & Automação), Minas Gerais, n. 448, p.441-466, 5 set. 2011. Disponível em: <http://www.scielo.br/scielo.php?pid=S0103-17592011000500002&script=sci_abstract>. Acesso em: 20 out. 2018.

CECHINEL, Cristian; CAMARGO, Sandro da Silva. **Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa de Pesquisa:** Mineração de dados educacionais: avaliação e interpretação de modelos de classificação. 2. ed.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16:321-357.*

CRETTON, Nícollas Nogueira; GOMES, Georgia Rodrigues. APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO ENADE COM ENFOQUE NOS CURSOS DE MEDICINA. Acta Biomédica Brasiliensia, [s.l.], v. 7, n. 1, p.74-89, 20 jun. 2016. Universidade Iguacu - Campus V. http://dx.doi.org/10.18571/acbm.100.

CARVALHO, Hialo Muniz. **Aprendizado de Máquina voltado para Mineração de Dados: Árvores de Decisão.** 2014. 47 f. Monografia (Especialização) - Curso de Engenharia de Software), Universidade de Brasília, Brasília, 2014. Cap. 4.

CARVALHO, Luís Alfredo Vidal de. *Mineração de dados*: A Mineração de dados no *marketing*, medicina, economia, Engenharia e administração. Rio de Janeiro: Ciência Moderna, 2005. 19 f. Curso de Mestre em Engenheira Elétrica, Engenharia Elétric, Pee/ Coppe / Ufrj, Rio de Janeiro, 2014.

CARVALHO, Deborah Ribeiro et al. Ferramenta de pré e pós-processamento para mineração de dados. **SEMINÁRIO DE COMPUTAÇÃOSEMINCO**, v. 12, p. 131-139, 2003.

CHUANG, L.T.; YANG, C.H.; WU, K.C.; YANG, C.H. A hybrid feature selection method for DNA microarray data. Computers in Biology and Medicine, v. 41, p. 228-237, 2011.

CRETTON, Nícollas Nogueira; GOMES, Georgia Rodrigues. APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO ENADE COM ENFOQUE NOS CURSOS DE MEDICINA. Acta Biomédica Brasiliensia, [s.l.], v. 7, n. 1, p.74-89, 20 jun. 2016. Universidade Iguacu - Campus V. http://dx.doi.org/10.18571/acbm.100.

D., Michie; D.J., Spiegelhalter; C.C., Taylor. **Machine Learning, Neural and Statistical Classification.** São Francsico: Citeseer X, 1994. 120 p.

DOMINGOS, Pedro; HULTEN, Geoff. Mining High-Speed Data Streams. **Citeseerx**, [st], p.1-10, 2000. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.4248>. Acesso em: 11 nov. 2018.

Educação Básica do INEP 2014. Anais dos *Workshops* do Iv Congresso Brasileiro de Informática na Educação (cbie 2015), Porto Alegre, v. 1, n. 1, p.1034-1043, 2015.

ENADE. O que é o Enade? Disponível em: <http://portal.inep.gov.br/enade>. Acesso em: 22.abr. 2018.

MELANDA, Edson Augusto. **Pós-Processamento de Regras de Associação.** 2004. 22 f. Tese (Doutorado) - Curso de Ciência da Computação e Matemática Computacional, Sp São Carlos, 2004. Disponível em: <https://teses.usp.br/teses/disponiveis/55/55134/tde-13012006-171753/publico/TeseEdsonMelandaVersaoCorrigida.pdf>. Acesso em: 17 nov. 2019.

FRANCISCO, Thiago Henrique Almino; MONTEIRO, Erika Cristina Mendonça de Sousa. UMA REFLEXÃO SOBRE O ENADE: AS AÇÕES PARA A GESTÃO DE UM IMPORTANTE ELEMENTO DA AVALIAÇÃO. **2º Simpósio Avaliação da Educação Superior**, Porto Alegre, p.1-15, 02 set. 2016.

FERREIRA, Gisele. Investigação acerca dos fatores determinantes para a conclusão do Ensino Fundamental utilizando Mineração de Dados Educacionais no Censo Escolar da Educação Básica do INEP 2014. **Anais dos Workshops do Iv Congresso Brasileiro de Informática na Educação (cbie 2015)**, [s.l.], p.1-10, 26 out. 2015. Sociedade Brasileira de Computação - SBC. http://dx.doi.org/10.5753/cbie.wcbie.2015.1034.

FONSECA, Stella Oggioni da; NAMEN, Anderson Amendoeira. MINERAÇÃO EM BASES DE DADOS DO INEP: UMA ANÁLISE EXPLORATÓRIA PARA NORTEAR MELHORIAS NO SISTEMA EDUCACIONAL BRASILEIRO. **Educação em Revista**, [s.l.], v. 32, n. 1, p.133-157, mar. 2016. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/0102-4698140742.

GOMES, Gusttavo Nunes. **ESTUDO DA EVASÃO EM ALUNOS DE GRADUAÇÃO POR MEIO DE MINERAÇÃO DE DADOS.** 2019. 32 f. TCC (Graduação) - Curso de Sistema de Informaç]ao, Instituto Federal Goiano – Campus Ceres, Ceres - Go, 2019. Disponível em: <https://repositorio.ifgoiano.edu.br/handle/prefix/635>. Acesso em: 20 nov. 2019.

HALMENSCHLAGER, Carine. **Um Algoritmo de Indução de Árvore e regras de Decisão.** 2002. 34 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul. Centro de Estudos Interdisciplinares em Novas Tecnologias da Educação. Programa de Pós-graduação em Informática na Educação, Porto Alegre, 2002. Cap. 34.

GUILLET, Fabrice; HAMILTON, Howard J.. **Quality Measures in Mineração de dados.** Si: In-chief, 2010.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Mineração de dados Concepts and Techniques.** Waltham, Ma 02451, Usa: Elsevier Inc, 2012. 327 p.

HANG, Yang; FONG, Simon. Investigating the Impact of Bursty Traffic on Hoeffding

Tree Algorithm in Stream Mining over Internet. **2010 2nd International Conference On Evolving Internet**, [s.l.], p.1-6, set. 2010. IEEE. http://dx.doi.org/10.1109/internet.2010.33.

HOEGLINGER, Stefan; PEARS, Russel. Use of Hoeffding trees in concept based data stream mining. **2007 Third International Conference On Information And Automation For Sustainability**, [s.l.], p.57-62, dez. 2007. IEEE. http://dx.doi.org/10.1109/iciafs.2007.4544780.

INAP. Conheça o Inep. 2018. Disponível em: <http://portal.inep.gov.br/conheca-o-inep>. Acesso em: 22 jun. 2018.

KEEDWELL, Ed. An analysis of the area under the ROC curve and its use as a metric for comparing clinical scorecards. **2014 Ieee International Conference On Bioinformatics And Biomedicine (bibm)**, [s.l.], p.24-29, nov. 2014. IEEE. http://dx.doi.org/10.1109/bibm.2014.6999263.

LEE, Huei Diana. **Seleção de Atributos Importantes para a Extração de Conhecimento de Base de Dados.** 2005. 29 f. Tese (Doutorado) - Curso de Ciência da Computação, Ciências Matemática e Computação, Instituto de Ciências Matemática e Computação, São Carlos, 2005. Disponível em: <https://teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/publico/tese_huei.pdf>. Acesso em: 10 out. 2019.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, Patricia Mariotto. ANÁLISE DOS ALGORITMOS DE MINERAÇÃO J48 E APRIORI APLICADOS NA DETECÇÃO DE INDICADORES DA QUALIDADE DE VIDA E SAÚDE. **Revista Interdisciplinar de Ensino, Pesquisa e ExtensÃo**, Brasil, v. 1, p.1-12, 2013.

LOPEZ, Martín Esteban Andreoni. **Uma Arquitetura de Detecção e Prevenção de Intrusãopara Redes Definidas por Software.** 2014. 11 f. Dissertação (Mestrado) -

M, Anoopkumar; RAHMAN, A. M. J. Md. Zubair. A Review on Mineração de dados techniques and factors used in Educational Mineração de dados to predict student amelioration. **2016 International Conference On Mineração de dados And Advanced Computing (sapience)**, [s.l.], p.1-12, mar. 2016. IEEE. http://dx.doi.org/10.1109/sapience.2016.7684113.

MACIEL, Thales Vaz et al. Mineração de dados em triagem de risco de saúde. **Revista Brasileira de Computação Aplicada**, [s.l.], v. 7, n. 2, p.26-40, 13 maio 2015. UPF Editora. http://dx.doi.org/10.5335/rbca.2015.4651.

MANTAS, Carlos J.; ABELLÁN, Joaquín; CASTELLANO, Javier G.. Analysis of Credal-C4.5 for classification in noisy domains. **Expert Systems With Applications**, [s.l.], v. 61, p.314-326, nov. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2016.05.035.

MARTENSSON, F. Trade-off examples inside software engineering and computer science. **Blekinge Institute Of Technology, Karlskrona**, Si, v. 1, n. 1, p.1-6, 2005. Disponível em:

<https://pdfs.semanticscholar.org/d671/e1fce79502df40d424b94790444300f0d291.pdf>. Acesso em: 23 nov. 2018.

MENEZES, Glauber Marcius Cardoso. **HTILDE-RT: UM ALGORITMO DE APRENDIZADO DE ARVORES DE REGRESS~AO DE LOGICA DE PRIMEIRA ORDEM PARA FLUXOS DE DADOS RELACIONAIS.** 2011. 42 f. Dissertação (Mestrado) - Curso de Engenharia de Sistemas e Computac~ao, Universidade Federal do Rio de Janeiro, Rj, 2011. Cap. 03.

MEYNARD, Christine N.; QUINN, James F.. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. **Journal Of Biogeography**, [s.l.], v. 34, n. 8, p.1455-1469, 16 maio 2007. Wiley. http://dx.doi.org/10.1111/j.1365-2699.2007.01720.x.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. **Editora Manole Ltda**, São Paulo, p.93-56, 2003. Disponível em: <http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>. Acesso em: 21 out. 2018.

MONTEIRO, André Vinicius Gouvêa; PINTO, Marcos Paulo Oliveira; COSTA, Rosa Maria E. Moreira da. Uma aplicação de Data Warehouse para apoiar negócios. **Cadernos do Ime - SÉrie InformÁtica**, Rj, v. 16, p.1-11, 2004. Universidade de Estado do Rio de Janeiro. http://dx.doi.org/10.12957/cadinf.

National Research Council. 1999. *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions*. Washington, DC: The National Academies Press. doi: 10.17226/9632.

NETO, Francisco A. de ALMEIDA; CASTRO, ALBERTO. A *reference architecture for educational mineração de dados*. 2017 *Ieee Frontiers In Education Conference* (fie), [s.l.], p.1-8, out. 2017. IEEE. http://dx.doi.org/10.1109/fie.2017.8190728.

NAMEN, Anderson Amendoeira et al. Indicadores de qualidade do ensino fundamental: o uso das tecnologias de mineração de dados e de visões multidimensionais para apoio à análise e definição de políticas públicas. **Revista Brasileira de Estudos Pedagógicos**, [s.l.], v. 94, n. 238, p.677-700, dez. 2013. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s2176-66812013000300003.

OLIVEIRA, Werbeston Douglas de. **COMPARAÇÃO DO ALGORITMO C4.5 E MLP USADOS NA AVALIAÇÃO DA SEGUNRANÇA DINÂMICA E NO AUXILIO AO CONTROLE PREVENTIVO NO CONTEXTO FDA ESTABILIDADE TRANSITÓRIA DE SISTEMAS DE POTÊNCIA.** 2013. 26 f. Dissertação (Mestrado) - Curso de Engenharia ElÉtrica, Sistema de Energia, Campus UniversitÁrio do GuamÁ, BelÉm, 2013. Cap. 3.

PRATI, Rc; BATISTA, G. E. A. P. A.; MONARD, M. C.. Curvas ROC para avaliação de classificadores. **Revista Ieee América Latina**, Centro - São Carlos - Sp, n. 1, p.1-8, 2008. Disponível em:

<http://conteudo.icmc.usp.br/pessoas/gbatista/files/ieee_la2008.pdf>. Acesso em: 18 out. 2018.

PAULA, Maurício Braga de. **Indução Automática de árvore de decisão.** 2002. 48 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Sistema de Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2002. Cap. 5.

PATEL, Priti S. *Various Mineração de dados Techniques used to Study Student's Academic Performance.* ***International Journal Of Computer Science And Mobile Applications****, North Lalaguda,* v. 3, n. 55, p.55-58, 6 jun. 2017. Disponível em: <https://pdfs.semanticscholar.org/eb6f/6231139773d0de950c6758af180957dee4e8.pdf>. Acesso em: 21 abr. 2018.

PROVOST, Foster; FAWCETT, Tom. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, [s.l.], v. 1, n. 1, p.51-59, mar. 2013. Mary Ann Liebert Inc. http://dx.doi.org/10.1089/big.2013.1508. Disponível em: <https://www.liebertpub.com/doi/pdf/10.1089/big.2013.1508>. Acesso em: 18 nov. 2019.

QUINLAN, J. Ross; PUBLISHERS, Morgan Kaufmann. C4.5: Programs for Machine Learning by. **Springer Link**, Boston, v. 16, n. 1, p.236-240, set. 1994. Disponível em: <https://link-springer-com.ez318.periodicos.capes.gov.br/article/10.1007/BF00993309>. Acesso em: 25 out. 2018.

RAPHAELMELONI. **Experimento.** 2009. Disponível em: <https://www.maxwell.vrac.puc-rio.br/31439/31439_6.PDF>. Acesso em: 25 out. 2018.

RISTOFF, Dilvo; GIOLO, Jaime. O Sinaes como Sistema. **R B P G, Brasília**, [si], v. 3, n. 228, p.193-213, dez. 2006.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational Mineração de dados: A Review of the State of the Art. **Ieee Transactions On Systems, Man, And Cybernetics, Part C (applications And Reviews)**, [s.l.], v. 40, n. 6, p.601-618, nov. 2010. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tsmcc.2010.2053532.

SANTOS, Cinara de Jesus. **Avaliação do uso de classificadores para verificação de atendimento a critérios de seleção em programas sociais.** 2017. 18 f. Dissertação (Mestrado) - Curso de Aculdade de Engenharia, Faculdade de Engenharia, Universidade Federal de Juiz de Fora, Juiz de Fora, 2017. Cap. 18. Disponível em: <http://www.ufjf.br/pgmc/files/2009/08/Dissertacao_Cinara_070317.pdf>. Acesso em: 20. out. 2018.

SANTOS, Luciano Drosda M. dos et al. Procedimentos de Validação Cruzada em Mineração de dados para ambiente de Computação Paralela. **Departamento Acadêmico de Informática Universidade Tecnológica Federal do Paraná**, Curitiba, n. 235, p.233-236, 17 mar. 2009. Disponível em:

<http://www.lbd.dcc.ufmg.br/colecoes/erad/2009/047.pdf>. Acesso em: 07 jan. 20018.

SANTOS, Maribel Yasmina; RAMOS, Isabel. **Business Intelligence:** Tecnologia De Informação Na Gestão do Conhecimento. 2. ed. Lisboa, Portugal: Fca, 2009. 8 p.

SCHIAVONI, AndrÉ Spinelli. **UM ESTUDO COMPARATIVO DE MÉTODOS PARA BALANCEAMENTO DO CONJUNTO DE TREINAMENTO EM APRENDIZADO DE REDES NEURAIS ARTIFICIAIS.** 2010. 32 f. Monografia (Especialização) - Curso de Ciência da Computação, Ciência da Computação, Universidade Federal de Lavras, Lavras, 2010. Cap. 3. Disponível em: <http://repositorio.ufla.br/bitstream/1/5223/1/MONOGRAFIA_Um_estudo_comparativo_de_metodos_para_balanceamento_do_cnjunto_de_treinamento_em_aprendizado_de_redes_neurais_artificiais.pdf>. Acesso em: 3 jun. 2019.

SETZER, Valdemar W.. Dado, Informação, Conhecimento e Competência. **Depto. de Ciência da Computação, Universidade de São Paulo**: Depto. de Ciência da Computação, Universidade de São Paulo, São Paulo, n. 2, p.1-14, 12 set. 2014. Disponível em: <https://s3.amazonaws.com/academia.edu.documents/44270487/ART_2_GEST.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1539968154&Signature=nMieQ8NQrIbGzZ2k4ii4nhu2Sqs%3D&response-content-disposition=inline%3B%20filename%3DDado_Informacao_Conhecimento_e_Competenc.pdf>. Acesso em: 19 out. 2018.

SFERRA, Heloisa Helena; CORRÊA, Ângela M. C. Jorge. Conceitos e Aplicações de Mineração de dados. **Revista de Ciência & Tecnologia**, Piracicaba, São Paulo, v. 11, n. 22, p.19-34, 2013.

SILVA, Ivan T. Costa e et al. Variabilidade interobservadores no diagnóstico de lesões precursoras do câncer anal: estudo do cenário habitual. **Revista do Colégio Brasileiro de Cirurgiões**, [s.l.], v. 38, n. 6, p.372-380, dez. 2011. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0100-69912011000600002.

SILVA, Leandro A.; SILVA, Luciano. Fundamentos de Mineração de dados Educacionais. **Anais dos *Workshops* do Iii Congresso Brasileiro de Informática na Educação**

SOUZA, César. **Análise de Poder Discriminativo Através de Curvas ROC.** 2009. Disponível em: <http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/>. Acesso em: 25 out. 2018.

(cbie 2014), [s.l.], n. 568, p.568-581, 3 nov. 2014. Sociedade Brasileira de Computação - SBC. http://dx.doi.org/10.5753/cbie.wcbie.2014.568.

TAM, Pang-ning; STEINBACH, Michael; KUMAR, Cipin. INTRODUÇÃO AO DATA MINING. Rio de Janeiro: Ciência Moderna, 2009. 171 p.

VELOSO, Flávio Henrique da Silva et al. MINERAÇÃO DE DADOS, SEUS BENEFICIOS, UTILIZAÇÕES, METODOLOGIA, CAMPO DE ATUAÇÃO DENTRO DE GRANDES E PEQUENAS EMPRESAS. **Revista Eletrônica de Sistema de Informação e Gestão Tecnologica**, São Paulo, Franca, v. 1, n. 46, p.45-53, jan. 2011.

VISTA, Nicolas P. B. et al. Análise de Agrupamento Hierárquico aplicada aos microdados do ENADE do curso de graduação em Ciência da Computação. Revista Eletrônica Argentina-brasil de Tecnologias da Informação e Comunicação, Rs, v. 1, n. 5, p.1-12, 12 abr. 2018. Disponível em: <https://revistas.setrem.com.br/index.php/reabtic/article/view/267/122>. Acesso em: 29 abr.2018.

WAIKATO, The University Of. **Software de Aprendizado de Máquina em Java.** Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 30 maio 2019.

WANG, G.; SONG, Q.; XU, B.; ZHOU, Y. Selecting feature subset for high dimensional data via the propositional foil rules. Pattern Recognition, v. 46, p.199- 214, 2013a. Disponível em: . Acesso em: 30 abr. 2016.

WITTEEN, Ian H.; FRANK, Eibe. ***MINERAÇÃO DE DADOS:*** *Pratical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers Is An Imprint Of Elsevier., 2005.*

WITTEN, I. H.; FRANK, E.; HALL, M. A. Mineração de dados: Practical Machine Learning Tools and Techniques. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 9780123748560.

YANG, Zhiyong et al. Optimizing area under the ROC curve via extreme learning machines. **Knowledge-based Systems**, [s.l.], v. 130, p.74-89, ago. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.knosys.2017.05.013.

APÊNDICE – ARTIGO

Os Algoritmos C4.5 e *Hoeffding Tree* Aplicados a Mineração de Dados Educacionais Referente Ao Exame Nacional de Desempenho de Estudante (ENADE) em Ciência da Computação

Euclides Francisco A. Amaro[1], Merisandra C. de Mattos Garcia[2]

[1]Curso de Ciência da Computação Universidade do Extremo Sul Catarinense (UNESC) – Criciúma, SC – Brazil

{euclidesamaro, mem}@unesc.net

Abstract. In this research, educational data were mining through the classification task, using the decision tree induction method, using the C4,5 and Hoeffding Tree algorithms. The data studied were extracted from the ENADE of the Computer Science Course of the three bases: University of the Southern Santa Catarina, Santa Catarina Association of Educational Foundations and Santa Catarina. After the data mining of the obtained models, these were analyzed through quality measures, such as accuracy, in order to identify which of the two algorithms generated the best model.

Resumo. Nesta pesquisa realizou-se *mineração de dados* educacionais por meio da tarefa de classificação, a partir do método de indução de árvores de decisão, empregando-se os algoritmos C4.5 e *Hoeffding Tree.* Os dados estudados foram extraídos do ENADE do Curso de Ciência da Computação das três bases: Universidade do Extremo Sul Catarinense, Associação Catarinense das Fundações Educacionais e Santa Catarina. Após a execução da mineração de dados dos modelos obtidos, estes foram analisados por meio das medidas de qualidade, como a acurácia, a fim de se identificar qual dos dois algoritmos gerou o melhor modelo.

1. Introdução

Com vastas quantidades de dados disponíveis, as empresas em quase todos os setores estão focadas na exploração de dados. Ao mesmo tempo, os computadores se tornaram muito mais poderosos, a rede é onipresente e foram desenvolvidos algoritmos que podem conectar conjuntos de dados para permitir uma ampliação das análises do que anteriormente era possível. A convergência desses fenômenos deram origem a cada vez mais difundida aplicação comercial da *data science* (PROVOST; FAWCETT, 2013). Mineração de dados, que se refere a aplicação de métodos específicos para identificar padrões em um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

A mineração de dados educacionais é um campo de pesquisa interdisciplinar que envolve a evolução de métodos na exploração de dados originais de uma instituição educacional (SILVA; SILVA, 2014). empenhando-se nas aplicações das ferramentas e técnicas de mineração de dados (PATEL, 2017). A sua aplicação é essencial para a Educação, no que se refere ao futuro da aprendizagem e aperfeiçoamento nesta área (NETO; CASTRO, 2017).

No Brasil, o Instituto Nacional de Pesquisa Educacional Anísio Teixeira (INEP), vinculado ao Ministério da Educação, tem como objetivo auxiliar na produção de políticas educacionais nos diferentes níveis de ensino, em busca de uma educação de qualidade com objetivo de desenvolver o país econômica e socialmente (INEP, 2018).

Com base no Exame Nacional de Desempenho de Estudante (ENADE) que é uma prova para avaliar os estudantes concluintes de um curso de graduação, com base nos conteúdos previstos nas diretrizes curriculares e apreendidos na formação acadêmica, avaliando a evolução destes estudantes, o desenvolvimento das competências e habilidades em sua área de formação.

A mineração dos dados ocorre por meio de diferentes tarefas, dentre elas: associação, regressão, previsão de séries temporais, detecção de desvios, agrupamento e classificação (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). A tarefa de classificação tem a característica de reconhecer um determinado registro e dizer a qual classe ele pertence (CAMILO; SILVA, 2009). Existem vários métodos de classificação, dentre eles: redes neurais artificiais, bayesianos e indução de árvores de decisão.

A indução de árvores de decisão e empregado na exploração de conhecimentos em instituições acadêmicas (MAIA; GOMES; CHAGAS, 2017). Dentre os algoritmos de indução de árvore de decisão tem-se o C4.5 e o *Hoeffding tree* (LEMOS; STEINER; NIEVOLA, 2005).

O algoritmo C4.5 cria um modelo de decisão partindo do nó pai, de modo que cada um dos nós possa ser examinado individualmente para determinar a relevância de sua relação ou a existência dela (CRETTON; GOMES, 2016).

O Algoritmo de *Hoeffding Tree*,, é um tipo de algoritmo que baseia a indução da árvore de decisão em um processo de aprendizagem (MENEZES; ZAVERUCHA, 2011).

Tendo em vista os aspectos salientados, propõe-se nesta pesquisa realizar a mineração de dados educacionais por meio da tarefa de classificação pelo método de indução de árvores de decisão, empregando-se os algoritmos C4.5 e *Hoeffding Tree,* nos dados do ENADE do curso de Ciência da Computação da UNESC, ACAFE e Santa Catarina, os dados encontram-se disponíveis na base do INEP.

2. Métodos e Materiais

Para isso foi aplicado os algoritmos C4.5 e Hoeffding Tree e analisado por meio das medidas de qualidades escolhidas para identificar o modelo com melhores resultados. O software escolhido para a implementação de mineração dados foi Weka por ser uma ferramenta que possui diversos métodos de mineração de dados, bem como medidas de qualidade em classificação e por ser um software *free*.

2.1. Conjunto de dados

Os dados fornecidos para elaboração deste trabalho estão disponíveis no portal do INEP7. Onde foi selecionada a prova do ENADE que é uma microdados, dos quatros últimos anos que ocorreu a prova do curso de Ciência da Computação (2011, 2014, e 2017), que procura avaliar o desenvolvimento dos estudantes de ciência da computação.

Nesta base, foi possível obter dados relacionados aos estudantes que se encontravam no final do curso que prestaram os respetivos exames nos anos de 2011, 2014 e 2017, tais estudantes, se tornaram adequados para fazer a prova, podendo assim avaliar o curso de ciência da computação.

2.2. Pré-processamento

A fase de pré-processamento contempla a etapa de seleção de dados. Para esta fase de análise, utilizou-se a ferramenta Excel com a e versão 18.1910.1283.0, onde foi requerida a limpeza de campos vazios, retirada de palavras com acentuação ou caráter especial. Analisou-se também a quantidade de atributos existentes em seus respetivos anos e a diferença entre o posicionamento de variável em anos diferentes. Organizou-se a posição dos atributos para poder transformar

[7] http://portal.inep.gov.br/web/guest/microdados

em um único arquivo dos três respectivos anos. Foram apagados alguns campos como ID_status, Amostra e TP_Semestre por não se encontrarem nos dados de 2017 e também foram apagadas linhas no Excel que contém campos vazios.

2.3. Etapa de transformação de dados

É nesta fase em que os dados sofrem alterações devido ao valor que não são reconhecidos na ferramenta Weka. Com isso alguns valores foram alterados para possibilitar o trabalho com informações relevantes, consolidando o conjunto de dados para transformá-los em uma maneira mais apropriada para a mineração, sabendo que, WEKA não reconhece valores com caráter especiais. Os dados sofreram transformações com intuito de aprimorá-los, para que os testes fossem feitos apenas com informações relevantes.

2.4. Sumarização

A sumarização tem como um dos propósitos, o agrupamento dos registros de várias bases de dados com valores idênticos. Por isso a tabela de sumarização se torna primordial quando é necessário englobar o conjunto de cada base selecionada, para uma única base. Como foram baixadas as três últimas provas do ENADE no departamento de ciência da computação as três bases possuem características de atributos idênticos teve a necessidade de unificar todas elas.

2.5Discretização

Os algoritmos de classificação, necessitam que os dados estejam no formato de atributos categorizados. Assim, como muitas vezes é necessário transformar um atributo contínuo em um categorizado (discretização), e tanto as variáveis discretas como contínuas, podem precisar de alteração em um ou mais atributos. Adicionalmente se um ou mais atributos categorizados possuir um número grande de valores (categorias), ou se algum valor ocorra raramente, então pode beneficiar para determinar tarefas de mineração de dados reduzir o número de categorias combinando em alguns valores. Esse método foi aplicado nas classes separando as notas dos estudantes em duas categorias, utilizando a ferramenta do Excel e as suas fórmulas.

a) Nota acima de 54.9 que é representada pela letra A;
b) E nota igual ou menor que 54.9, que é representado pela letra R.

É importante destacar que não há reprovação nas provas aplicadas do ENADE, logo, os números escolhidos para a pesquisa não correspondem a aprovações e reprovações.

Visto que no conjunto original foi reforçado ou aplicado a discretização em todas as bases de experimentos realizados, levando em conta que a classe maioritária é a classe de estudantes com a nota abaixo ou igual de 54,9 e a minoritária têm a nota maior que 54,9. Dentre os 138 atributos encontrados na base. A variável que foi selecionada foi a de nota geral.

2.5. Seleção de Atributos

Com vasta quantidade de atributos encontrados na base de dados optou-se por aplicar a escolha de atributos. A seleção de atributos permite a ordenação das variáveis segundo a importância da classe escolhida, com isso causa a redução da dimensionalidade de espaço de busca de atributos e a remoção de dados contendo ruídos (LEE, 2005).

Teve-se como opção para resolver este problema, selecionar o filtro (AtributeSelection) não necessita de um algoritmo de aprendizado de máquina para executar a seleção de características, utilizando-se somente das próprias habilidades para poder avaliar os subconjuntos, geralmente necessitando de um menor poder computacional para ser utilizado. O método de seleção de atributos empregado foi o Correlation-based Feature Subset Selection (CfSubsetEval) que avalia o valor de um subconjunto de atributos, considerando a sua capacidade preditiva, optando-se pelos que são altamente correlacionados com a classe.

2.6 Balanceamento de classes

No conjunto de dados desbalanceados, obteve uma grande disparidade de quantidade de dados em cada classe. Os conjuntos de dados de cada classe que se deseja modelar devem ser equivalentes de forma que possam ser assimiladas as características de cada classe envolvida, podendo fazer com que o classificador possa atingir um nível de definição homogêneo (CECHINEL; CAMARGO, 2016). O desbalanceamento de uma classe tem a capacidade de influenciar nas ações dos desempenhos de um modelo de classificação, portanto buscam classificar corretamente as classes maioritárias, definido por Chawla et al. (2002).

A técnica SMOTE, foi executada neste trabalho com o objetivo de refazer o ajuste de frequência relativa de classes majoritárias e minoritárias, introduzindo sinteticamente instâncias de classes minoritárias (WITTEN; FRANK; HALL, 2011).

Empregou-se essa técnica SMOTE, disponível na Weka, como um filtro supervisionado, onde foi possível fazer alterações nas definições dos parâmetros ao aplicar esse método na ferramenta, com o percentual de sobre amostragem que foram usados 700% para base da UNESC, 200% ACAFE, e SANTA CATARINA 200% e o número de vizinhos, utilizando o valor 5, sugerido pela ferramenta Weka. Segue-se na tabela 3 as informações de quantidade de classes minoritária e majoritária, e a quantidade de instância anterior e atual.

Base de dados	Classes (maior - minoritária)	Instância (antes)	Percentagem	Instância (atual)	Classes (maior - minoritária)
UNESC	**159 - 6**	**165**	**700%**	**207**	**159 - 48**
ACAFE	**667 – 114**	**781**	**200%**	**974**	**667- 307**
SC	**898 - 195**	**1093**	**200%**	**1483**	**898 - 585**

Tabela dos dados balanceados.

2.7 Etapa de mineração de dados

É nesta fase da metodologia que se explora a etapa de mineração de dados para a descoberta de conhecimentos em dados acadêmicos do ENADE. Ainda nessa etapa empregou-se as técnicas de algoritmo definidos.

A ferramenta utilizada para utilização do processo de descoberta de conhecimento, é o conhecido software Waikato Environment for Knowledge Analysis (WEKA) é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados (WAIKATO, 2019). O WEKA foi desenvolvido pela Universidade de Waikato em Nova Zelândia é um software open source. (WAIKATO, 2019).

Na utilização de consumir técnicas de mineração de dados com algoritmos de classificação é necessário que a base de dados seja repartida em dois conjuntos: treinamento e testes. Para estratificação dos conjuntos foi utilizado o método chamado Cross-Validation ou validação cruzada (WITTEN; FRANK; HALL, 2011).

Foram desenvolvidos 6 experimentos descritos na tabela a seguir, com intuito de serem destinado analisar o comportamento dos classificadores C4.5 e Hoeffding tree.

Experimentos	Descrição
1	Base de dados da Unesc (discretizadas)
2	Base de dados da Unesc (discretizadas e balanceada)
3	Base de dados da Acafe (discretizadas)
4	Base de dados da Acafe (discretizadas e balanceada)
5	Base de dados de Santa Catarina (discretizadas)
6	Base de dados de Santa Catarina (discretizadas e balanceada)

Descrição dos experimentos realizado.

3. Resultados

Com os resultados alcançados pela base da UNESC (discretizada) foram executados dois classificadores, assimilando a classe escolhida. Foram discutidos por meio de medidas de classificação como acurácia, coeficiente kappa, percentuais de verdadeiros positivos e F-measure.

Constata-se que os classificadores empregados alcançaram resultado com base o pré-processamento, a classe escolhida e o balanceamento aplicado, com a percentagem mais altas e o algoritmo C4.5 98,79% com a base de dados dados da UNESC (discretizadas), e segundo ponto mais alto também alcançado com o algoritmo C4.5 mas com alteração na base de dados da unesc (discretizadas e balanceada) com o valor 97,58. Analisando a linha de meta tracejada de no gráfico e com o comportamento dos algoritmos executado, todos alcançaram as taxas mais próximas da linha meta.

Designa-se que a base de dados da unesc (discretizadas e balanceada), aplicando o algoritmo hoeffding tree, que com isso atingiu a taxa de percentual mais baixo relacionando ao desempenho de outros testes na base. O resultado obtido no percentual é 87,92%.

A partir das medidas de qualidades foi avaliado o resultado do coeficiente kappa gerado pela a base de dados da UNESC (discretizada) e (discretizadas e balanceada). Da mesmo jeito que na acurácia foi apresentada a linha de meta, os valores mais adjacentes a 1 são os modelos que alcançaram o melhor desempenho nesta medida.

A amostra mais alta alcançada ocorreu com o classificador C4.5 que teve mudança no balanceamento de instância a atingiu o valor de 0,93. Pode-se observar que os valores mais próximos da linha de meta.

Em relação ao comportamento da execução geral no modelo da base da UNESC tendo em conta acurácia quanto ao coeficiente Kappa, os melhores resultados foram encontrados no desempenho do algoritmo C4.5, onde acurácia teve mais alto na UNESC (discretizadas) e o coeficiente Kappa na base da UNESC (discretizadas e balanceadas) com aplicação da técnica SMOTE com percentual de 700%.

Em seguida analisou-se a performance do modelo gerado nas classes “A” que representa nota igual ou de maior que 54.9 e classe “R” que representam as notas menor que 54.9, verificou-se a taxa de verdadeiros positivos e a medida F-Measure. delas.

Nas taxas de verdadeiros positivos para a classe “A” que chegaram mais próximo da vizinhança da linha de meta, mostra que a classificador hoffding tree alcançou a alinha de meta com o valor 1 tendo em conta que foi aplicado o balanceamento da técnica SMOTE com a percentagem 700%, e o classificador C4.5 aproximou-se da linha de meta com o resultado 0,94 também teve a mesma quantidade de alteração no balanceamento.

Partindo para analisar a medida F-Measure para a classe “A”, pode-se enxergar que o classificador com o maior resultado foi o C4.5 balanceado que teve o valor 0,95. O gráfico também expressa o classificador com valor mais distante da linha de meta que é o hoeffding tree que se encontra com o valor zero.

Também analisou-se a taxas de verdadeiros positivos para a classe “R” que chegaram mais próximo da vizinhança da linha de meta, mostra que a classificador C4.5 alcançou a alinha de aproximação com o valor 0,99 tendo em conta que foi na base original, e

com a base balanceada se aproximou da linha de meta com o resultado 0,99. O gráfico também expressa o classificador com que mais se distância da linha de meta que é o hoeffding tree com o resultado 0,84.

Partindo para analisar a medida F-Measure para a classe “R”, pode-se perceber que no gráfico (figura 17), que o classificador com o maior resultado foi o C4.5 balanceado que teve o valor 0,99. O gráfico também agrega informação sobre o classificador com valor mais distante da linha de meta que é o hoeffding tree que se encontra com o valor 0,91.

3.1. Gerando árvore de conhecimento nas três bases geradas

Árvore gerada com a base de Unesc balanceada aplicando o algoritmo C4.5, nos dados de ciência da computação.

Observando a Figura 1, é possível analisar as principais amostras referentes às notas brutas no componente específico (NT_CE). Nota-se que cento e cinquenta e dois estudantes tiveram uma nota específica menor ou igual que 44,9; nota-se também que os estudantes que tiveram nota específica superior a 44,9 se possuíam bolsas de estudo (QE_I13). Quarenta e oito estudantes responderam que possuíam bolsa de extensão e um estudante, que possuía outro tipo de bolsa de estudos acadêmica. Cinco estudantes que tiraram a NT_CE inferior a 52,44 não possuíam nenhuma bolsa e dois estudantes que também tiveram nota superior a 52,44 responderam que não.

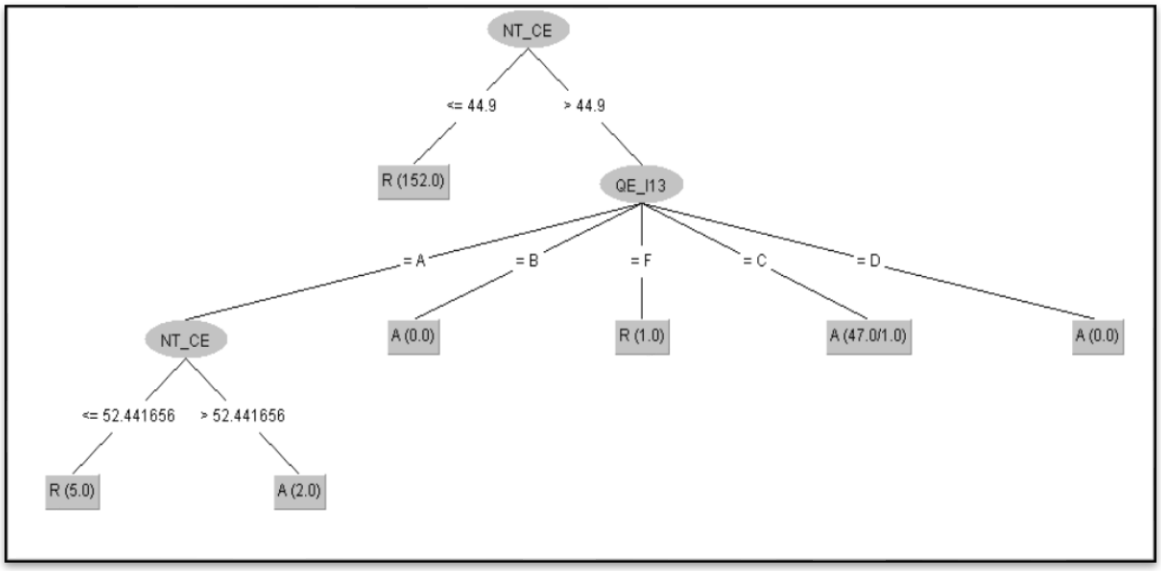


Figura 1 - Árvore de decisão referente a base da UNESC

3.2. Discussão dos resultados

Segundo Cretton (2016) realizaram quatro experimentos em dados educacionais da prova do ENADE, podendo assim analisar o desempenho da acurácia onde empregaram o algoritmo J48 que é o mesmo que o C4.5. Cada uma das quatro bases executadas obteve um nível de confiança distinto, com o valor da primeira base de 80,86%, na segunda 80,73%, na terceira 83,82% e na quarta base 84,05% com o valor mais alto.

Nessa sequência, notou-se que os resultados alcançados dos trabalhos do Cretton podem ser comparados com o desenvolvimento do trabalho atual. Observou-se que os melhores resultados alcançados nas três bases criadas, com aplicação do algoritmo C4.5 com os valores na base da Acafe 98,21%, Acafe sem UNESC 98,56%, UNESC 98,79% e Santa Catarina 98,79% superam os resultados do trabalho do Cretton.

Além disso, pode-se comparar o percentual de acurácia obtido por algoritmo C4.5 na base que tem resultado mais próximo da linha de meta com o valor de 98,79% com a base original da UNESC tendo em conta que é o melhor resultado obtido nas quatro bases geradas, comparando com o estudo feito por Rubert e Mariotto (2013), que aplicou a implementação dos classificadores de classificação e associação: J48 e A priori; Geração de perfis de usuários aplicando as técnicas de classificação e associação na identificação e classificação de indicadores de saúde. Alcançado assim o número de instâncias classificadas corretamente 85,61, tendo um desempenho muito menor que os trabalhos desenvolvidos na base da UNESC e outras bases do projeto.

Segundo o trabalho de Gomes, foram analisados 13 algoritmos, dos dois concedem com os meus classificadores usados, os resultados apresentados na pesquisa do Gomes foram hoeffding tree com o valor de 84,4956% e o C4.5 alcançou o resultado de 85,5433%. Esses são os resultados apresentados, que não foram mais eficientes que os da pesquisa atual onde o C4.5 teve como valor 98,79 e o hoeffding tree teve 95,39.

4. CONCLUSÃO

Levando-se em consideração a evolução da globalização e a alta demanda de informações, foi possível perceber que auxiliar na busca de respostas simplifica a compressão de diversos problemas pedagógicos. Este trabalho apresentou uma fundamentação sobre umas das ferramentas adequadas para se fazer a exploração dos possíveis conhecimentos.

Ao decorrer da introdução deste trabalho proposto, esta pesquisa realizou a mineração de dados educacionais por meio da tarefa de classificação nos dados do ENADE do Curso de Ciência da Computação. Dividindo a base em três modelos que são UNESC, ACAFE e Santa Catarina - os dados encontram-se disponíveis no portal do INEP. Diante disso foi exposto um referencial teórico sobre os temas tratados, assim como características e conceitos que foram selecionados para ajudar no processo de alcance dos objetivos propostos.

A mineração de dados traz um diferencial interessante para gerações de alertas importantes, podendo aplicar em qualquer uma das partes envolvidas de sistemas educacionais, como alunos, educadores, administradores e pesquisadores.

A intenção de fornecer feedbacks é a de ajudar nas orientações e melhorar o processo de descoberta de aprendizagem dos alunos. Ao fazer descobertas dos sistemas educacionais dos alunos, ajuda a melhorar o desempenho do ensino, as atividades propostas e a tomada de decisões.

Com o intuito de alcançar dados mais precisos, foram utilizados classificadores como C4.5 e Hoeffding Tree. Para a execução dos algoritmos foi necessário investigar como os dados estão organizados e aplicar o pré-processamento para avaliar qual atributo poderia ser definido como classe.

A ferramenta Weka teve uma grande importância para a realização da etapa de mineração de dados. A importação dos conjuntos de dados a serem explorados, implementa e executa os algoritmos de mineração de dados utilizados a tarefa de classificação de forma simples e flexível, mostrando-se uma ferramenta adequada para a metodologia de descoberta de conhecimento proposta.

Depois de ocorrerem diferentes análises comparativas entre as três bases geradas utilizadas na pesquisa, o conjunto de dados convenientes para classificação dos estudantes que tiveram uma nota superior a 54,9 e inferior ou igual que 54,9. A base de dados da UNESC (discretizada) como o classificador C4.5 alcançou o valor de 98,79% nos dados originais, com apenas as classes discretizadas.

REFERÊNCIAS

PROVOST, Foster; FAWCETT, Tom. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, [s.l.], v. 1, n. 1, p.51-59, mar. 2013. Mary Ann Liebert Inc. http://dx.doi.org/10.1089/big.2013.1508. Disponível em: <https://www.liebertpub.com/doi/pdf/10.1089/big.2013.1508>. Acesso em: 18 nov. 2019.

FAYYAD , Usama M.; *PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI Magazine, Providence* , v.17, n. 3, p. 37-54, Autumn 1996.

SILVA, Leandro A.; SILVA, Luciano. Fundamentos de Mineração de dados Educacionais. **Anais dos *Workshops* do Iii Congresso Brasileiro de Informática na Educação**

SOUZA, CÉSAR. ANÁLISE DE PODER DISCRIMINATIVO ATRAVÉS DE CURVAS ROC. **2009. DISPONÍVEL EM: <HTTP://CRSOUZA.COM/2009/07/13/ANALISE-DE-PODER-DISCRIMINATIVO-ATRAVES-DE-CURVAS-ROC/>. ACESSO EM: 25 OUT. 2018**

PATEL, Priti S. *Various Mineração de dados Techniques used to Study Student's Academic Performance.* ***International Journal Of Computer Science And Mobile Applications****, North Lalaguda,* v. 3, n. 55, p.55-58, 6 jun. 2017. Disponível em: <https://pdfs.semanticscholar.org/eb6f/6231139773d0de950c6758af180957dee4e8.pdf>. Acesso em: 21 abr. 2018.

NETO, Francisco A. de ALMEIDA; CASTRO, ALBERTO. A *reference architecture for educational mineração de dados*. 2017 *Ieee Frontiers In Education Conference* (fie), [s.l.], p.1-8, out. 2017. IEEE. http://dx.doi.org/10.1109/fie.2017.8190728.

INAP. Conheça o Inep. 2018. Disponível em: <http://portal.inep.gov.br/conheca-o-inep>. Acesso em: 22 jun. 2018.

CRETTON, Nícollas Nogueira; GOMES, Georgia Rodrigues. APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO ENADE COM ENFOQUE NOS CURSOS DE MEDICINA. Acta Biomédica Brasiliensia, [s.l.], v. 7, n. 1, p.74-89, 20 jun. 2016. Universidade Iguacu - Campus V. http://dx.doi.org/10.18571/acbm.100.

MENEZES, Glauber Marcius Cardoso. **HTILDE-RT: UM ALGORITMO DE APRENDIZADO DE ARVORES DE REGRESS~AO DE LOGICA DE PRIMEIRA ORDEM PARA FLUXOS DE DADOS RELACIONAIS.** 2011. 42 f. Dissertação (Mestrado) - Curso de Engenharia de Sistemas e Computac~ao, Universidade Federal do Rio de Janeiro, Rj, 2011. Cap. 03.