FACULDADE PROFESSOR MIGUEL ÂNGELO DA SILVA SANTOS – FeMASS

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

A UTILIZAÇÃO DA TÉCNICA DE ÁRVORE DE DECISÃO PARA ENTENDER PADRÕES DA EVASÃO DOS ALUNOS DO CURSO DE SISTEMAS DE INFORMAÇÃO DA FEMASS

GUSTAVO RIBEIRO SACRAMENTO STRAUSS

MACAÉ

2022

FACULDADE PROFESSOR MIGUEL ÂNGELO DA SILVA SANTOS – FeMASS
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

GUSTAVO RIBEIRO SACRAMENTO STRAUSS

A UTILIZAÇÃO DA TÉCNICA DE ÁRVORE DE DECISÃO PARA ENTENDER PADRÕES DA EVASÃO DOS ALUNOS DO CURSO DE SISTEMAS DE INFORMAÇÃO DA FEMASS

Trabalho Final apresentado ao curso de graduação em Sistemas de Informação, da Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS), para obtenção do grau de BACHAREL em Sistemas de Informação.

Orientador: Prof. Dr. Alan Carvalho Galante

MACAÉ/RJ
2022

GUSTAVO RIBEIRO SACRAMENTO STRAUSS

A UTILIZAÇÃO DA TÉCNICA DE ÁRVORE DE DECISÃO PARA ENTENDER PADRÕES DA EVASÃO DOS ALUNOS DO CURSO DE SISTEMAS DE INFORMAÇÃO DA FEMASS

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Sistemas de Informação, da Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS), para obtenção do grau de BACHAREL em Sistemas de Informação.

Aprovada em ____ de _____ de 20____

BANCA EXAMINADORA

_____

Prof. Dr. Alan Carvalho Galante

Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS)

1º Examinador

_____

Prof. Dr. Isac Mendes Lacerda

Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS)

2º Examinador

DEDICATÓRIA

Dedico este trabalho aos meus pais, que sempre se fizeram presentes em minha vida e moldaram os valores bases do meu ser. Principalmente, a minha mãe, que mesmo após o falecimento do meu pai, não deixou de me incentivar nos estudos e acreditar em mim. A minha namorada, pela paciência, compreensão e carinho nos momentos de hesitação e cansaço. A meu primo, Luiz Carlos, que trilhou pioneiramente os caminhos do ensino superior na família Ribeiro e nunca deixou de me incentivar nesta mesma trajetória.

AGRADECIMENTO

A minha mãe, por sempre se manter presente em minha vida, me incentivando a continuar e me apoiando em todas as decisões.

Ao meu orientador, Alan Galante, pelo suporte e paciência em todas as tentativas de finalizar esse projeto até a chegada desse momento.

A todos os professores, que fizeram parte da minha vida acadêmica. Em especial, a professora Liliane Valério, que me orientou na primeira etapa desse projeto.

Aos amigos, que fiz durante essa jornada acadêmica e que, direta ou indiretamente, me ajudaram a chegar até aqui.

A minha namorada Tainá Carvalho e a nossa amiga Juliana Campos, que me resgataram de um momento de desespero e desistência, me trazendo de volta a trajetória para finalizar esse trabalho.

Aos membros da família Ferreira Monnerat, por me acolherem em um momento de necessidade e, ao mesmo tempo, muito decisivo da minha vida acadêmica.

E, por fim, a FeMASS pela oportunidade de realizar um estudo com dados da instituição.

“I really think a champion is defined not by their wins but by how they can recover when they fall.”

Serena Williams

RESUMO

A evasão no ensino superior ainda é um dos desafios que assolam tanto as instituições públicas quanto privadas. Essa problemática vem sendo tema de estudos com intuito de identificar os fatores que levam a evasão dos alunos, buscando criar padrões de modelo que possam auxiliar, preventivamente, a evitar a saída do aluno antes da conclusão do curso. Dentro do processo de mapeamento desses fatores, tem-se como uma das áreas de pesquisa o *Machine Learning*. Este trabalho utilizou da ferramenta Weka com o objetivo de analisar a evasão no curso de Sistemas de Informação, no ambiente da Faculdade Municipal Professor Miguel Ângelo da Silva Santos - FeMASS. Para tal, foi utilizada a técnica de árvore de decisão, baseada no algoritmo J48, em dados do curso com base na grade antiga e na grade nova. Como resultados da pesquisa, puderam ser extraídas as principais regras que mapeiam a evasão com base nas árvores de decisão geradas pelo algoritmo. Desta forma, este trabalho pode servir como base de trabalhos futuros na análise de fatores que levam a evasão do aluno na instituição.

Palavras-chave: *Machine Learning*. Algoritmo de Classificação. Weka. Evasão.

ABSTRACT

Dropout in higher education is still one of the challenges that plague both public and private institutions. This problem has been the subject of studies aiming to identify the factors that lead to student dropout, seeking to create model patterns that can help preventively to avoid the student's departure before the completion of the course. Within the process of mapping these factors, Machine Learning appears as one of the research areas. This study used the Weka tool with the objective of analyzing the dropout in the Information Systems course at the Faculdade Municipal Professor Miguel Ângelo da Silva Santos - FeMASS. To do this, the decision tree technique, based in the J48 algorithm, was used on course data based on the old and new grids. As results of the research, the main rules that map evasion based on the decision trees generated by the algorithm could be extracted. Thus, this work can serve as a basis for future studies in the analysis of factors that lead to student dropout in the institution.

Key words: Machine Learning. Classification algorithm. Weka. Dropout.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ADM Administração

ARFF *Attribute Relation File Format*

CSV *Comma-separated values*

ENADE Exame Nacional de Desempenho dos Estudantes

ENG Engenharia de Produção

FeMASS Faculdade Municipal Professora Miguel Ângelo da Silva Santos

FN *False Negative*

FP *False Positive*

GPL *General Public License*

ID3 *Iterative Dichotomiser 3*

IES Instituição de Ensino Superior

INEP Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

JSON *JavaScript Object Notation*

KDD *Knowledge Discovery Databases*

MAT Matemática

MEC Ministério da Educação

ML *Machine Learning*

SI Sistemas de Informação

TN *True Negative*

TP *True Positive*

UnB Universidade de Brasília

WEKA *Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento de uma sociedade está atrelado diretamente ao tripé “educação X cultura X tecnologia”, sendo a educação uma dimensão que fornece a estrutura de sustentação para alavancar as transformações sociais e o desenvolvimento econômico pretendidos por qualquer nação.

No Brasil, os segmentos que compõe a Educação Nacional estão divididos em educação básica (educação infantil, ensino fundamental e ensino médio) e educação superior, os quais são organizados em cooperação entre as esferas federal, estadual e municipal com atribuições próprias, regulamentadas por lei específica.

A literatura científica aponta que todos os níveis educacionais se deparam com um problema em comum, a evasão de alunos (SOUZA; NÓBREGA; AMORIM, 2017), recaindo o recorte da presente pesquisa na evasão de uma Instituição de Ensino Superior (IES) localizada na cidade de Macaé no estado do Rio de Janeiro.

O tema de evasão no ensino superior vem sendo observado desde 1995, quando os dados do Censo da Educação Superior no Brasil começaram a ser publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Os estudos do INEP começaram a ser fomentados pelo Ministério da Educação (MEC), que buscava reunir dados sobre o desempenho das universidades públicas brasileiras referente aos índices de diplomação, retenção e evasão dos estudantes de seus cursos de graduação (RIGO; BARBOZA; CAMBRUZZI, 2014; MANHÃES, L. M. B. *et al*., 2011).

No Brasil, segundo FILHO (2017), nos últimos quinze anos, as taxas de evasão se mantiveram aproximadamente constantes em 22%, com pequenas variações anuais. Em uma análise feita com os dados de 2014/2015 foi identificado que, dentre os cursos do setor público, a área de computação apresenta o segundo maior índice de evasão, que é de vinte e dois por cento, só perdendo para a área de matemática, que apresenta trinte e um por cento de evasão.

Nessa direção, infere-se que a gestão de IES, não pode prescindir de boas estratégias acadêmicas e estruturantes para acompanhamento dos alunos ingressos nos cursos de Sistemas de Informação e, principalmente, de um sistema de informações que lhe permita captar e gerenciar dados que resultem no aumento da assertividade na tomada de decisão concernente à política de permanência no ensino superior. A busca pela criação de um sistema de informação funcional deve ser precedida pela coleta de dados de diversas naturezas, sejam objetivos ou subjetivos, de modo que sejam extraídos e, posteriormente convertidos para a tomada de decisão gerencial.

Com base nos conceitos supracitados, é extremamente oportuno que uma IES possa contar com ferramentas e sistemas de análise de dados educacionais, como o *Machine learning*, por exemplo, pois possibilita o desenvolvimento e adaptação de métodos que possam detectar comportamentos ligados a evasão no ensino, sendo assim possível uma melhor compreensão do fenômeno a ser estudado.

A Descoberta de Conhecimentos em Banco de Dados (do Inglês KDD, *Knowledge Discovery Databases*) tem como objetivo extrair, de uma grande base de dados, conhecimentos a partir do uso de diversas técnicas de diferentes áreas de conhecimento, tais como: estatística, matemática, bancos de dados, entre outras (CASTANHEIRA, 2008; BAKER, ISOTANI, CARVALHO, 2011).

Dentro do contexto de *Machine Learning*, o foco dessa área está no desenvolvimento de métodos para análises de dados nos ambientes educacionais (CASTANHEIRA, 2008). Assim, trata-se de uma ferramenta oportuna para explorar dados coletados durante anos e armazenados nas bases de dados das IES e definir perfis de alunos propensos a evasão, por exemplo.

Pensar nos fatores multidimensionais que podem explicar as razões para índices tão alarmantes numa profissão, que é um dos símbolos do paradigma das ciências modernas, levou o autor deste projeto a estabelecer contato com alunos e ex-alunos, que saíram do curso de Sistemas de Informação antes do término da graduação, na intenção de conhecer um pouco das razões de suas desistências. Outro fator motivante apareceu durante uma aula da disciplina de Ética no Contexto Empresarial, que abrange três cursos da FeMASS, quando durante uma dinâmica de grupo foi observado que dos dez alunos do curso de sistemas matriculados na disciplina, apenas dois tinham a intenção de prosseguir no curso até a sua conclusão, instigando, assim, a proposição investigativa aqui descrita.

Os dados abordados até este momento não deixam dúvidas quanto à necessidade de investigação do fenômeno da evasão em cursos superiores e, principalmente no curso de Sistemas de Informação da FeMASS.

Adicionam-se à presente argumentação o quantitativo de alunos formados no referido curso da IES a ser analisada entre os anos de 2017 e 2021, cuja média foi de 4,3 alunos a cada ano/semestre (conforme verificado em documentos institucionais da FeMASS), além da escassez de produção científica local sobre o tema, uma vez que na base de dados de monografias do curso de SI foi somente uma pesquisa voltada à investigação da evasão no referido curso, fatos que justificam a necessidade de mais estudos afetos ao tema.

O presente trabalho realizou uma análise utilizando técnicas de *machine learning* no banco de dados fornecido pela FeMASS para identificação de padrões associados a evasão de alunos no curso de Sistemas de Informação entre os anos de 2017 e 2021, corroborando para validação das seguintes hipóteses: “Usando técnicas de ML, será possível criar regras para prever a evasão no curso de SI da FeMASS?”, “A reprovação ou aprovação de uma ou mais disciplinas dos primeiros semestres podem influenciar no abandono do curso?” e “É possível identificar outras variáveis que são relevantes para a conclusão ou não do curso?”

Com base nos conceitos apresentados anteriormente, pode-se afirmar que o objetivo geral dessa pesquisa foi analisar a evasão no curso de Sistemas de Informação da FeMASS usando algoritmos de *Machine Learning* antes e depois da atualização da grade do curso.

Trata-se dos objetivos específicos deste trabalho:

- Realizar estudo sobre evasão e *Machine Learning;*
- Coletar dados dos alunos do banco de dados da IES;
- Realizar pré-processamento dos dados da FeMASS a partir das variáveis selecionadas;
- Aplicar algoritmo de técnica de classificação J48 nos dados selecionado com a ferramenta Weka;
- Analisar resultados obtidos pelo algoritmo J48 sobre evasão no curso de Sistemas de Informação.

Sob o ponto de vista da natureza da pesquisa, ela é categorizada como pesquisa básica, que tem o objetivo de gerar novos conhecimentos, que podem ser úteis para novas pesquisas sem uma aplicação prática prevista (PRODANOV, FREITAS, 2013, p. 51).

O método científico utilizado nesse trabalho foi do tipo dedutivo. Conforme define Prodanov e Freitas (2013, p. 27), “o método dedutivo, de acordo com o entendimento clássico, é o método que parte do geral e, a seguir, desce ao particular. A partir de princípios, leis ou teorias consideradas verdadeiras e indiscutíveis, prediz a ocorrência de casos particulares com base na lógica”. Ou seja, analisando a literatura do tema e olhando para a problemática da evasão na educação no Brasil, será possível realizar um estudo de caso da evasão no curso de Sistemas de Informação na FeMASS. Durante o período de pesquisa não havia nenhum estudo ou ferramenta de análise da evasão que auxiliasse na tomada de decisão da gerência da instituição para amenizar tal problema.

Quanto ao objetivo de estudo deste trabalho, foi do tipo descritivo e exploratório, pois visa descrever as características de uma população ou fenômeno estabelecendo relações entre variáveis que podem assumir a forma de levamento, além de buscar mais informações sobre o

assunto da investigação, para facilitar a delimitação do tema e formulação de hipóteses (PRODANOV, FREITAS, 2013, p. 51-52). Destinou-se a esta pesquisa estudar a evasão no curso de Sistemas de Informação da FeMASS, para um melhor entendimento do fenômeno e fornecimento de dados para apoio a tomadas de decisão.

Dos procedimentos técnicos, foram utilizados a pesquisa bibliográfica que remete ao estudo de outras literaturas a respeito do tema, que deu base para o desenvolvimento do estudo, conforme descrito por Prodanov e Freitas (2013, p. 54) e a pesquisa documental na consulta dos dados, do banco de dados da faculdade, entre os anos de 2017 e 2021, para identificar os alunos que se categorizam no perfil de evadidos.

Quanto a abordagem usada foi de natureza quantitativa que, segundo Prodanov e Freitas, "considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las" (2013, p. 69). O trabalho criou relações com os dados do banco e gerará dados estatísticos a partir dessas relações para identificar perfis de evasão.

O capítulo um deste trabalho contextualizou os temas a respeito de *Machine Learning* e evasão no ensino superior no Brasil, assim como também apresentou conceitos iniciais e descreveu as delimitações desta pesquisa, seus objetivos, justificativa e metodologia, resumindo ao final a estruturação de cada capítulo presente na pesquisa.

O capítulo dois aborda a revisão de conceitos fundamentais para o estudo e identifica as principais ferramentas de análise de dados aplicadas em estudos semelhantes, dando enfoque na técnica de *Machine Learning* e o algoritmo de classificação utilizado no estudo.

O capítulo três traz informações sobre o curso de Sistema de Informação e da IES estudada, além de informações sobre a ferramenta utilizada no estudo da evasão.

O capítulo quatro sintetiza as aplicações de *Machine Learning* no banco de dados fornecido pela IES, aprofundando nas metodologias utilizadas para classificação dos dados de evasão com o uso do algoritmo classificador J48 na ferramenta Weka.

O capítulo cinco mostra os resultados obtidos da ferramenta Weka com as árvores de decisão de cada experimento realizado sobre evasão, evidenciando os conhecimentos encontrados e a comparação com as hipóteses levantadas no primeiro capítulo.

No sexto e último capítulo serão apresentadas as considerações finais sobre a pesquisa, assim como as opiniões do autor sobre os resultados obtidos e suas limitações e sugestões de possíveis trabalhos que poderão ser realizados a partir desse estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será fornecida a fundamentação teórica com base nos objetivos da pesquisa junto aos seguintes temas: Conceito de Evasão; KDD; *Machine Learning*; e Algoritmos de *Machine Learning*, no intuito de sustentar a elaboração e a execução da pesquisa.

2.1 Evasão

O problema da evasão nos cursos superiores pode surgir de múltiplos fatores como: questões vocacionais, institucionais, socioeconômicas, pessoas, de saúde, familiares, entre outros. Forma de ingresso, políticas governamentais ligadas a educação e o tipo de instituição (pública ou privada), também são fatores elencados a esse problema.

A evasão escolar pode ser vista, simplesmente, como a saída do aluno do curso antes que ele obtenha o diploma. Porém, alguns autores diferenciam esse conceito de evasão e os resultados obtidos podem se tornar diferentes conforme o conceito de evasão escolhido para o estudo. Neste capítulo serão vistos os diferentes conceitos de evasão na literatura.

A evasão é um dos impasses que ainda inquietam as instituições de ensino como um todo e a busca de suas causas tem sido o instrumento de muitos trabalhos e pesquisas educacionais (OLIVEIRA *et al*., 2018, p. 6).

Conforme a Comissão Especial de Estudos sobre Evasão do MEC (1997), são especificadas três modalidades de evasão, como forma de gerar um consenso conceitual e possibilitar a comparação de resultados. As três dimensões da evasão são:

Evasão de curso - se dá quando o aluno se desliga por diversas razões, tais como: por abandono (não efetua rematrícula), por mobilidade acadêmica (mudança de curso na mesma instituição), por desistência oficializada na instituição, por transferência para outra instituição, e por exclusão de prerrogativa da instituição;

Evasão da instituição - se dá pelo desligamento do aluno da instituição na qual está matriculado;

Evasão do sistema – se dá quando o aluno abandona de forma definitiva ou temporária do ensino superior.

Já em um olhar mais abrangente, define-se que um ponto comum entre estas dimensões da evasão é a saída do estudante do curso, ou seja, o destino do evadido pode ser outro curso, outra IES ou abandonar o sistema de ensino superior. Portanto, a evasão pode ser medida a partir da saída do curso (LOBO 2012; HOFFMANN, 2016).

Branco; Conte e Habowski (2020) sumarizam alguns conceitos de evasão de outros autores com observações sobre cada conceito conforme a seguir no Quadro 1.

Quadro 1 - Perspectivas de evasão

Conceitos de Evasão		
Autor	**Evasão**	**Observação**
SANTOS (2008)	É a desistência definitiva do curso.	Independente da etapa.
GAIOSO (2005)	Pode ocorrer antes de concluir o curso.	Refere-se à perda do estudante.
KIRA (1998)	Refere-se ao fato de não concluir o curso.	Ressalta a perda do estudante.
BAGGI; LOPES (2011)	Ressalta a ideia de não concluir o curso.	Ressalta a perda do estudante em um curso.
BRASIL (1997)	É a saída do curso de origem.	Desconsidera a mobilidade acadêmica (troca de curso).
FAVERO (2006)	É a desistência do curso.	Considera os que realizaram e os que nunca participaram das atividades propostas.
VARGAS (2007)	Considera o período de início e a saída do curso.	Busca identificar o período que ocorreu a evasão.
POLYDORO (2000)	Considera o abandono no curso e no sistema universitário.	Considera dois marcos – abandono no curso e no sistema universitário - relacionados à desistência definitiva dos estudos e evasão.
CARDOSO (2008)	Mobilidade do estudante para outro curso e a desistência do curso.	A troca de curso já é considerada uma evasão.

Fonte: Habowski; Branco; Conte (2020, p.4)

Neste trabalho, foi considerado evadido todo aluno que encerrou o seu vínculo de matrícula com a instituição, mesmo que o aluno não tenha frequentado uma aula.

2.1.1 Políticas de permanência

É importante destacar a definição de políticas de permanência, que auxiliam como medidas de prevenção a evasão dos discentes no ensino superior. Conforme estudo realizado por Vargas e Heringer, primeiro devemos diferenciar “permanência” de “assistência estudantil”. Para os autores permanência pode ser descritas como “possuem maior abrangência, incluindo aspectos relacionados a diferentes formas de inserção plena na universidade, como por exemplo, programas de iniciação científica e à docência, monitoria, apoio à participação em eventos, entre outras atividades” (VARGAS, HERINGER, 2017, p. 14). Já as políticas de assistencial estudantil estão inclusas nas políticas de permanência e são voltadas para ações necessárias para viabilizar a frequência às aulas e demais atividades de alunos em situação de vulnerabilidade e em circunstâncias que possam comprometer sua permanência na IES (VARGAS, HERINGER, 2017, p. 14).

A seguir são apresentados os principais modelos de permanência que mais aparecem nas IES descritos por Vargas e Heringer (2017) e suas classificações em 5 grandes grupos:

- Bolsa auxílio ou permanência: auxílio financeiro a estudantes em situação de vulnerabilidade socioeconômica, estudantes indígenas e estudantes quilombolas matriculados.
- Moradia: apoio à moradia como um auxílio financeiro ou residência estudantil.
- Alimentação: fornecem alimentação gratuita ou subsidiada através do restaurante universitário ou transferem um valor monetário para que o aluno providencie sua própria alimentação.
- Transporte: auxílio financeiro destinado a transporte, bem como créditos, vale-transporte ou similar, para transporte local municipal ou intermunicipal.
- Outros benefícios: aqui foram agrupados todos os benefícios que não se enquadravam nas opções anteriores.

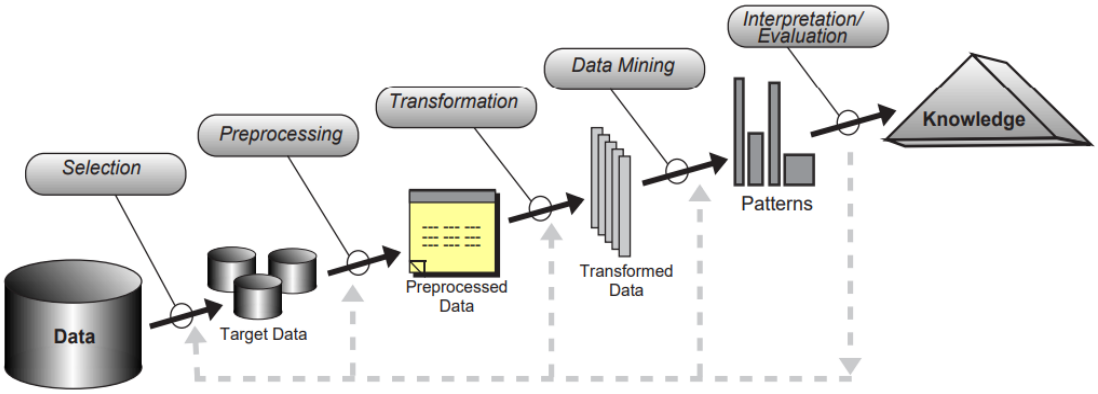
2.2 KDD

Knowledge Discovery in Databases (KDD) é definido por Fayyad *et. al.* (1996, p.3) como um processo não trivial de identificar padrões de dados válidos, novos, potencialmente úteis e compreensíveis dos dados.

Para padrões pode-se entender como a construção de modelos de dados ou a produção de qualquer descrição em alto nível desses modelos. Válidos, os padrões devem possuir certo grau de confiança para garantir que as descobertas desses padrões sejam aceitáveis. Os padrões descobertos devem ser desconhecidos pelo usuário para que sejam considerados novos. Potencialmente úteis são os padrões que podem trazer algum benefício. Por fim, os padrões devem ser compreensíveis de imediato ao usuário que está analisando ou após a realização de um pós-processamento.

Ainda para Fayyad *et. al.* (1996, p.3), o termo processo implica que o KDD é composto por algumas etapas, que envolvem preparação de dados, busca de padrões, avaliação de conhecimento e refinamento. Ou seja, todas repetidas em múltiplas iterações objetivando melhores resultados a cada próxima iteração, além de ser considerado também um processo interativo, pois envolve a participação de diferentes classes de usuários durante o processo de descoberta de conhecimento, sendo eles: especialistas de domínio, analistas e usuário final, que faz uso do conhecimento extraído ao longo do processo.

Figura 1 - Processo KDD



Fonte: Fayyad *et. al.* (1996, p.41)

Ainda seguindo a linha de raciocínio de Fayyad *et. al.* (1996), o processo de KDD é composta das etapas vistas na Figura 1 e explicitadas abaixo:

Seleção: Esta etapa compõe os conjuntos de dados e atributos selecionados aos quais o processo de descoberta vai operar. Apenas dados considerados relevantes para o objetivo são selecionados. É válido informar que os dados podem ser obtidos de diversos meios e formatos diferentes, como tabelas, banco de dados e arquivos de texto, o que dificulta esse processo inicial de obtenção dos dados.

Pré-processamento: Normalmente, as fontes de dados não são totalmente completas e nesta etapa são feitas as eliminações de dados inconsistentes (dados inválidos, incompletos, nulos ou repetidos), desta forma limpando a base de dados para que possa seguir para a próxima etapa.

Transformação: Com os dados pré-processados, inicia-se o processo de transformação, que tem o objetivo de atribuir um formato padrão para os dados. Nesse momento procura-se os atributos úteis ao estudo de acordo com o objetivo da análise. Também são utilizados métodos em busca de reduzir o número de variáveis.

Mineração de Dados: Nesta etapa é definida a estratégia mais apropriada para a mineração dos dados. Utilizando informações das etapas anteriores e com o objetivo do projeto estabelecido, nesta etapa são aplicados os algoritmos de *Machine Learning*.

Interpretação dos dados: Após passar por todas estas etapas, o conhecimento obtido deve ser interpretado e avaliado para verificação do alcance ou não do objetivo. A eficácia do processo KDD é determinada neste momento, pois só tem validade quando o conhecimento gerado é de fato aplicado para geração de valor (BALDASSO; CORTIMIGLA, 2019).

2.3 *Machine Learning*

Nas últimas duas décadas, o *Machine Learning* (ML) tornou-se um dos pilares da tecnologia da informação. Com a grande quantidade de dados gerados nos mais diversos meios, acredita-se que a análise inteligente de dados se torne ainda mais necessário para o progresso tecnológico (SMOLA e VISHWANATHAN, 2008).

Ainda sobre o ponto de vista de grandes quantidades de dados, destaca-se a variedades de origens desses dados, podendo dentre suas fontes serem: bancos de dados, *datawarehouses*, a web, outros repositórios de informações ou dados que são transmitidos para o sistema dinamicamente, dados de texto, dados de multimídia, entre outros (HAN, KAMBER, PEI, 2011, p. 8).

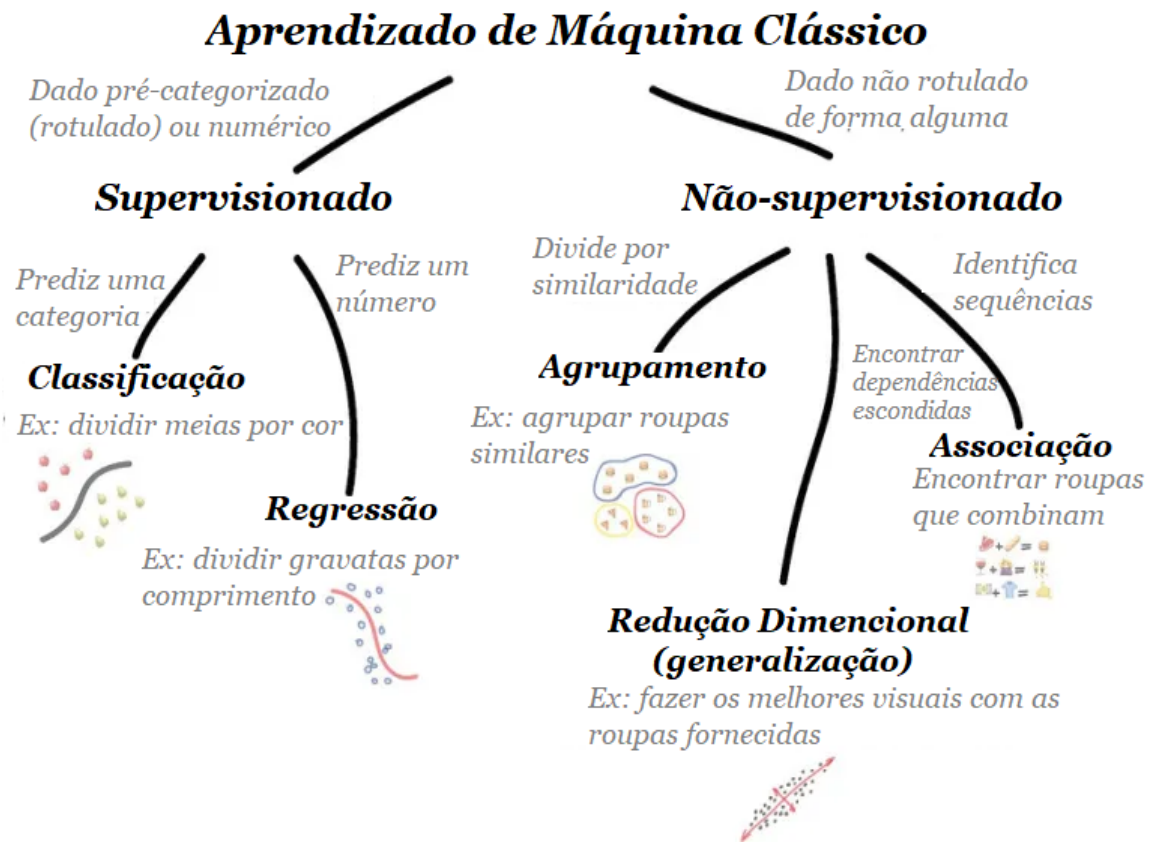
Para com Goldschmidt (2010) o "aprendizado de máquina é uma área da inteligência artificial que tem por objetivo desenvolver algoritmos e técnicas computacionais que permitam que o computador seja capaz de aprender", ou seja, o computador analisa séries históricas e a partir disso ele consegue adquirir conhecimento de forma automática, aperfeiçoando seu desempenho a determinada tarefa.

Ainda segundo Goldschmidt (2010), esse aprendizado só é possível devido aos sistemas de aprendizado. Estes sistemas são programas de computadores com a capacidade de tomar decisões com base em resoluções bem-sucedidas de problemas vistos anteriormente, ou seja, a partir da identificação de padrões históricos.

Smola e Vishwanathan (2008) afirmam que existem inúmeras aplicações do ML, os tipos de dados dos quais elas lidam e a formalização do problema de forma mais estilizada. Sendo este último fundamental para evitar reinventar a roda sempre que realizarmos uma nova análise. Assim, grande parte da ciência do ML é resolver esses problemas fornecendo garantias para as soluções.

Para entender melhor os algoritmos de ML, Goldschmidt (2010) discorre sobre o aprendizado indutivo, que é a base dos principais algoritmos usados no *Machine Learning*. É comum que os modelos de conhecimento sejam abstraídos a partir de sucessivas iterações sobre o conjunto de dados históricos disponíveis e podem ser divididos em duas grandes categorias: Aprendizado Supervisionado e Aprendizado não supervisionado.

Figura 2 - Tipos *Machine Learning*



Fonte: Grando (2022)

- Aprendizado Supervisionado – neste caso são disponibilizados ao algoritmo todas as entradas e as suas respectivas saídas. Em termos mais técnicos, Goldschmidt (2010) diz: “os exemplos históricos disponíveis devem conter qual a informação esperada a ser produzida pelo modelo de conhecimento que está sendo construído.

  O aprendizado supervisionado também é subdividido em 2 algoritmos:

  - Classificação (prediz uma categoria) – Exemplos: *Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbours, Support Vector Machine*.
  - Regressão (prediz um número) – Exemplos: Regressão Linear e Regressão Polinomial.

- Aprendizado não supervisionado – já para os não supervisionados são informadas somente as entradas dos dados sem informar as saídas. Dessa forma os algoritmos procuram agrupar exemplos históricos em função da similaridade que apresentam entre si (GOLDSCHIMIDT, 2010).

  Do aprendizado não supervisionado temos 3 subgrupos de algoritmos (PRIMÃO, 2022):

  - Agrupamento (divide por similaridade) – Exemplos: K-*Means*, *Clustering* Hierárquico [HCA], Maximização da Expectativa.

- Generalização (encontrar dependências escondidas) – Exemplos: Análise de Componentes Principais [PCA], Kernel PCA, *Locally*-Linear *Embedding* [LLE].
- Associação (identifica sequências) – Exemplos: *Apriori*, *Eclat*, FP-*growth*.

2.4 Algoritmos de *Machine Learning*

Nesta seção serão descritos dois exemplos de algoritmos de ML, sendo um de cada tipo de aprendizado (supervisionado e não supervisionado).

2.4.1 Classificação - Árvore de decisão

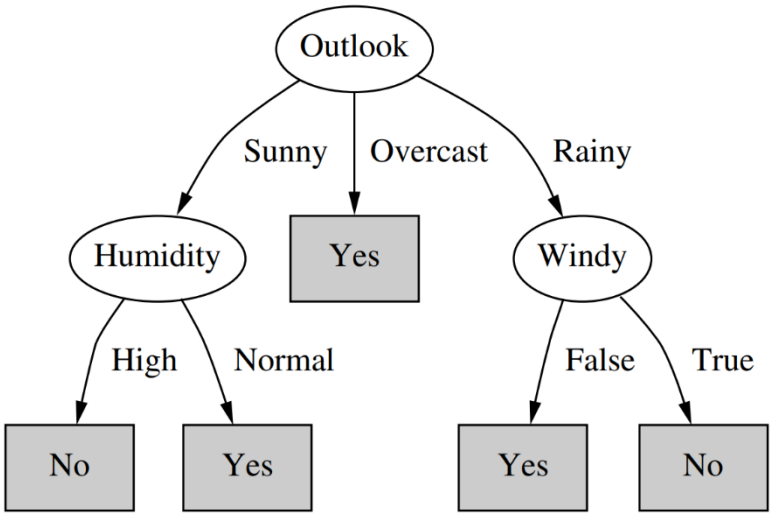
Árvores de decisão, um dos modelos mais difundidos dentro do ML, é representado em como um fluxograma em formato de árvore com a capacidade de classificar informações, numéricas ou simbólicas, baseando-se nos valores dos atributos de decisão (BALDASSO; CORTIMIGLA, 2019).

Segundo Librelotto e Mozzaquatro (2013), as árvores de decisão utilizam a abordagem dividir-para-conquistar, na qual um problema complexo é dividido em subproblemas mais simples, aplicando a mesma estratégia aos subproblemas até que sejam indivisíveis para representar uma classe.

Para Barbosa (2007), árvore de decisão é um método adequado quando o objetivo do estudo for classificação de dados ou predição de saídas. Também é conveniente utilizar este método quando o objetivo for gerar regras que podem ser facilmente entendidas, explicadas e traduzidas para uma linguagem mais natural.

A seguir na Figura 3 temos um exemplo de árvore de decisão, apresentada por Witten *et al*. (2017), para determinar se o tempo está propício ou não para uma partida de tênis.

Figura 3 - Exemplo árvore de decisão



Fonte: Witten *et al.* (2017, p. 109)

Na árvore apresentada na Figura 3 podemos extrair as seguintes regras como exemplo.

SE tempo = nublado ENTÃO jogar = sim

SE tempo = chuvoso E vento = verdadeiro ENTÃO jogar = não

SE tempo = ensolarado E umidade = normal ENTÃO jogar = sim

Dessa forma, podemos chegar na conclusão de que o jogo de tênis ocorrerá somente nas condições descritas nas regras 1 e 3.

Entre os principais algoritmos de árvore de decisão, podemos citar o ID3 (*Iterative Dichotomiser 3*) como precursor do algoritmo C4.5 e sua recodificação em Java o algoritmo J48. O ID3 foi desenvolvido por Quinlan (1979) como um algoritmo capaz de gerar árvores de decisão utilizando os conceitos de entropia e ganho de informação.

Entropia pode ser entendida como sendo o grau de pureza de determinado conjunto, ou seja, quanto menor a entropia, menos informação codificada em um ou mais atributos. Logo, quanto maior a entropia, maior a relevância desses atributos na descrição do conjunto de dados (GOLDSCHMIDT; PASSOS, 2005).

Já o ganho de informação calcula a redução da entropia e mede o quão bem um determinado recurso separa ou classifica as classes de destino. Ou seja, o recurso com maior ganho de informações é selecionado como o melhor atributo (WITTEN *et al*., 2017).

O algoritmo C4.5, proposto por Quinlan (1996), traz algumas melhorias em relação ao algoritmo ID3, listadas abaixo:

- Lidar com atributos contínuos e discretos;
- Lidar com dados de treinamento com atributos incompletos. Aceita atributos rotulados como “?” para casos de valores que não estejam presentes;
- Lidar com atributos com diferentes pesos;
- Poda de árvores após a criação. O algoritmo C4.5 permite retroceder na árvore após criada para tentar remover ramificações que não ajudam no processo de decisão.

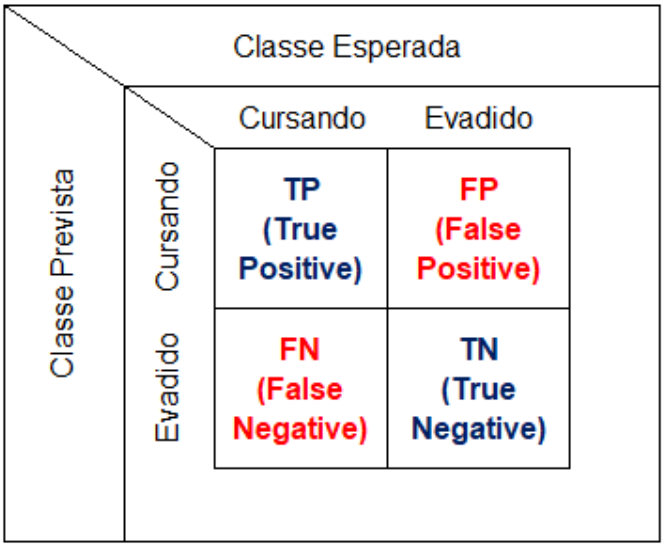
O algoritmo J48 surgiu da recodificação do C4.5 que, originalmente, foi escrito em linguagem C, para a linguagem JAVA (WITTEN *et al*., 2017). Desta forma, o J48 possui as mesmas características do algoritmo C4.5 descritas nos parágrafos anteriores.

Um dos motivos para grande utilização do algoritmo J48 pelos especialistas é que ele se adequa para os procedimentos envolvendo as variáveis tanto qualitativas contínuas quando discretas de uma base de dados. Além de ser considerado o que apresenta o melhor resultado em montagem de árvores de decisão (LIBRELOTTO; MOZZAQUATRO, 2013).

Outra característica dentro dos algoritmos de classificação, que ajuda na avaliação de desempenho do modelo, é o conceito de Matriz de Confusão, que traz informações de desempenho do modelo treinado através das classificações corretas em relação ao número de classificações indicadas pelo modelo (GOLDSCHMIDT; PASSOS, 2005).

A seguir, na Figura 4, apresenta-se uma matriz de confusão binária com base no modelo de matriz binária descrita por Witten *et al*., 2017; Primão, 2022.

Figura 4 - Matriz de confusão



Fonte: adaptado pelo autor (2022)

A matriz de confusão exemplificada na Figura 4 nos mostra 4 tipos de valores:

- TP (*True Positive*): Resultado esperado “cursando” e resultado previsto “cursando”;
- FP (*False Positive*): Resultado esperado “evadido” e resultado previsto “cursando”;
- FN (*False Negative*): Resultado esperado “cursando” e resultado previsto “evadido”;
- TN (*True Negative*): Resultado esperado “evadido” e resultado previsto “evadido”.

Com os valores descritos anteriormente podemos avaliar os conceitos de Acurácia e Precisão, que são utilizados também para avaliar os modelos de classificação.

Acurácia, segundo Primão (2022), é uma medida amplamento usada, pois mede a média geral do acerto do modelo ao classificar as classes seguindo a fórmula: Acurácia = (TP + TN) / (TP + TN + FP + FN).

Ainda segundo Primão (2022), precisão mostra quantas observações o modelo classificou como TP, ou seja, o percentual de predições corretas. Neste caso, pode-se achar o valor de precisão seguindo a fórmula: Precisão = TP / (TP + FP).

No caso da ferramenta WEKA, que foi selecionada para esse estudo, ela já realiza a geração da matriz de confusão e os cálculos de precisão e acurácia, conforme veremos no capítulo de resultados deste trabalho.

Ainda sobre o algoritmo J48, pode-se citar a capacidade de plotar a árvore de forma gráfica facilitando a visualização das decisões tomadas pelo modelo, além de permitir trabalhar com dados faltantes da base de dados, ignorando os campos de dados faltantes e trabalhando com os demais campos preenchidos para criação do modelo, algo que não é possível em outros algoritmos de árvore de decisão como o ID3, conforme visto na evolução do ID3 para o C4.5.

Com base nos parágrafos supracitados, o autor deste trabalho escolheu o algoritmo J48 como base para realizar os estudos sobre evasão na instituição.

2.4.2 Regras de associação – Apriori

Dentro do aprendizado não supervisionado temos alguns subgrupos de algoritmos de ML e dentre eles a Regra de Associação é uma das técnicas que tem como destaque o algoritmo Apriori sendo um dos principais algoritmos utilizados na literatura.

O algoritmo Apriori é um dos principais recursos para descoberta de regras de associação e é considerado um clássico para lidar com questões de extração de regras de associação em mineração de dados educacionais (FERNANDES, 2017).

O algoritmo tem por objetivo encontrar conjunto de itens que apareçam de forma simultânea e frequente em uma base de dados, como por exemplo produtos de uma loja que sejam vendidos de forma conjunta como um dos casos mais famosos da literatura sobre o *Walmart,* que descobriu uma regra de associação entre vendas de cervejas e fraldas (SILVA, 2004).

O uso de regras de associação é muito interessante na aplicação de dados de supermercado como visto no parágrafo anterior. Nesse sentido Hoed (2016) diz que o algoritmo pode nos mostrar se uma determinada transação de venda de um produto, um outro produto também é vendido pelo supermercado. Logo, o objetivo do algoritmo é encontrar todas as regras de associação relevantes entre os itens da base de dados, na lógica de: X(antecedente) => Y(consequente).

Ainda segundo Hoed (2016), as regras de associação têm uma utilidade muito evidente em transações comerciais e essa utilidade pode ser aplicada para outros campos de

estudo, como o caso da análise de dados do ensino. Pode-se, por exemplo, encontrar regras de associação entre evasão e o desempenho de alunos em determinadas disciplinas ou com a forma de ingresso, gênero, período de saída etc.

Apesar do algoritmo Apriori ser evidentemente muito utilizado na literatura e trazer bons resultados para o tipo de pesquisa realizado neste trabalho, o autor deste estudo encontrou diversas dificuldades em achar regras de associação relevantes durante a análise dos dados na ferramenta. Desta forma seguiu-se somente com a análise dos dados utilizando o algoritmo de classificação.

3 ESTUDO DE CASO DA FERRAMENTA WEKA

3.1 FeMASS

A Faculdade Municipal Professor Miguel Ângelo da Silva Santos – FeMASS é uma instituição educacional que tem a sociedade como princípio. A IES desenvolve suas atividades com o objetivo de garantir uma formação superior voltada a um ensino de qualidade, de acordo com exigências do Conselho Estadual de Educação, do mercado de trabalho e da sociedade. (PREFEITURA MUNICIPAL DE MACAÉ, 2022).

A FeMASS oferece, no ano de 2022, quatro cursos de graduação no período noturno. Suas atividades iniciaram em 2001 com o curso de Sistemas de Informação (SI) e expandiu em 2007 com a criação dos cursos de Engenharia de Produção (ENG) e Administração (ADM). O curso mais recente da instituição é o de Licenciatura de Matemática (MAT), que é o único curso com foco em formação pedagógica.

No presente momento da escrita deste trabalho todos os cursos da instituição possuem avaliação 4 ou 5 no ENADE (Exame Nacional de Desempenho dos Estudantes). O curso de ENG recebeu nota 5 no exame do ano de 2019, enquanto os demais cursos obtiveram nota 4 numa escala de 1 a 5.

Dentro desse contexto este trabalho propõe-se a utilizar uma ferramenta de ML com o objetivo de aplicar um algoritmo de classificação para análise de possíveis causas de evasão no curso de Sistemas de Informação da IES, estudada com base nos conceitos elucidados do capítulo 2 dessa pesquisa.

Espera-se com os resultados obtidos validar as hipóteses levantadas sobre o estudo de evasão no capítulo 1 deste trabalho.

3.2 Definição WEKA

O WEKA (*Waikato Environment for Knowledge Analysis*) é uma coleção de algoritmos de aprendizado de máquina e pré-processamento de dados que foi desenvolvido na Universidade de Waikato na Nova Zelândia. O WEKA foi desenvolvido em linguagem Java e é um *software* livre de código aberto emitido sob o domínio da licença GNU GPL (*General Public License*) (Witten *et al.*, 2017).

Ainda segundo Witten *et al.* (2017), a ferramenta fornece implementações de algoritmos de aprendizagem de máquina, que podem sem implementados facilmente sem necessidade de qualquer tipo de escrita de código de programação. A biblioteca inclui métodos para os principais problemas de mineração de dados: regressão, classificação, agrupamento,

regras de associação e seleção de atributos. Além disso, também fornece meios de visualização dos dados e ferramentas de pré-processamento.

O WEKA pode ser usado tanto via interface gráfica quanto por linha de código. Nessa pesquisa foi utilizada somente a interface gráfica do *software*.

Abaixo a Figura 5 apresenta a tela inicial da ferramenta na versão 3.8.6.

Figura 5 - Tela inicial Weka

Fonte: adaptado Weka (2022)

3.2.1 ARFF

Para abertura dos dados e aplicação das técnicas de ML, o Weka fornece algumas formas de obter esses dados: Abertura de arquivos, conexão com banco de dados, URL ou, até mesmo, gerar uma base de dados genérica.

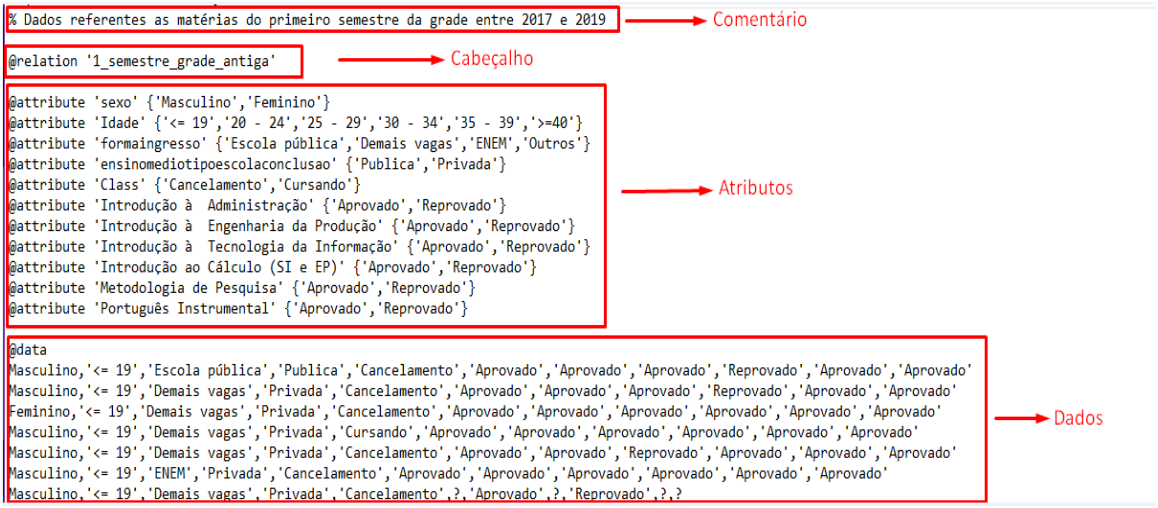
Nessa pesquisa foi utilizada uma base de dados CSV (*Comma-separated values*), posteriormente convertida para ARFF (*Attribute Relation File Format*), conforme passos descritos no próximo capítulo. Dentre os formatos de arquivos aceitos estão: CSV, JSON (*JavaScript Object Notation*) e ARFF. Este último sendo um arquivo próprio da ferramenta que tem a sua estrutura dividida em três partes: cabeçalho, declaração dos atributos e seção de dados.

O WEKA trabalha com os seguintes tipos de atributos:

- Numérico;
- Inteiros;
- Datas;
- *String*;
- Enumerados.

Foi utilizado nessa pesquisa atributos do tipo Enumerados para o tipo de análise realizada.

Figura 6 - Arquivo ARFF



Fonte: adaptado pelo autor (2022)

Na Figura 6, o autor mostra um exemplo de arquivo ARFF gerado para a pesquisa com as informações organizadas conforme a seguir:

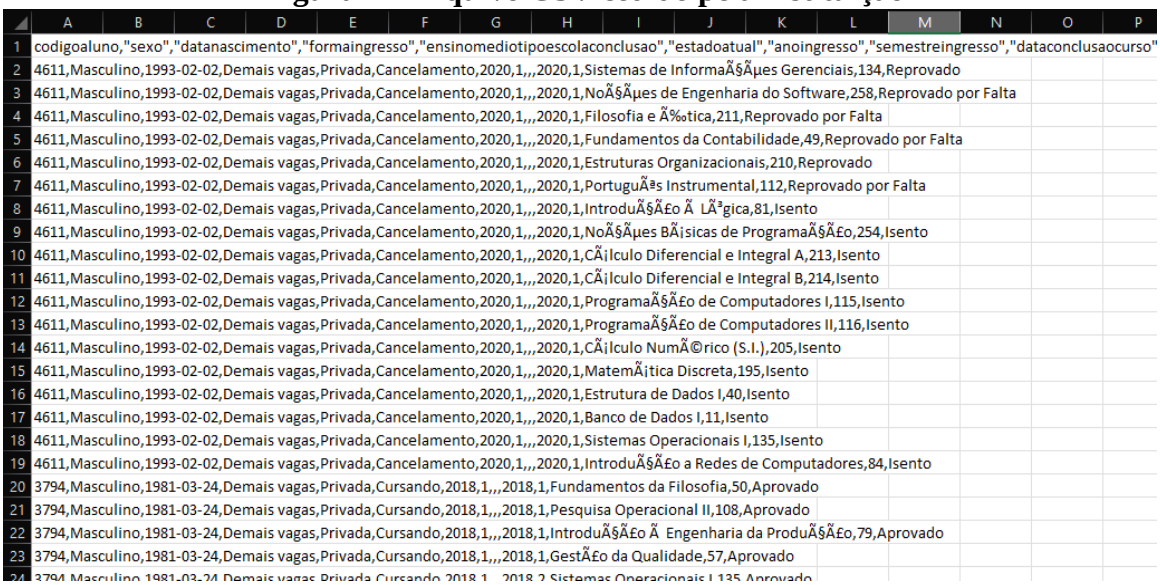
- ‘%’ corresponde a comentários
- ‘@relation’ equivale ao título do arquivo
- ‘@attribute’ declaração dos atributos
- ‘@data’ parâmetros de dados

Importante informar que na parte da declaração de atributos, os valores que serão preenchidos abaixo de ‘@data’ deve seguir a mesma sequência de declaração dos atributos. Para os atributos enumerados, os valores que podem ser lidos na base de dados são postos entre chaves ({}) e separados por vírgulas. Já a parte de dados abaixo do ‘@data’ também devem estar separados por vírgulas e os atributos que não possuírem valor definido devem ser representados pelo caractere ‘?’.

4 EXPERIMENTO

Foi realizado no início deste trabalho a compreensão do cenário sobre o tema evasão para definição dos dados que seriam necessários para posterior análise via WEKA. Dessa forma o primeiro passo do trabalho, após compreensão do cenário, foi a extração dos dados retirados do banco de dados da faculdade para um arquivo CSV conforme mostra a Figura 7 a seguir:

Figura 7 - Arquivo CSV cedido pela instituição



	A
1	codigoaluno,"sexo","datanascimento","formaingresso","ensinomediotipoescolaconclusao","estadoatual","anoingresso","semestreingresso","dataconclusaocurso"
2	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Sistemas de InformaÃ§Ãµes Gerenciais,134,Reprovado
3	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,NoÃ§Ãµes de Engenharia do Software,258,Reprovado por Falta
4	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Filosofia e Ã‰tica,211,Reprovado por Falta
5	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Fundamentos da Contabilidade,49,Reprovado por Falta
6	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Estruturas Organizacionais,210,Reprovado
7	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,PortuguÃªs Instrumental,112,Reprovado por Falta
8	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,IntroduÃ§Ã£o Ã LÃ³gica,81,Isento
9	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,NoÃ§Ãµes BÃ¡sicas de ProgramaÃ§Ã£o,254,Isento
10	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,CÃ¡lculo Diferencial e Integral A,213,Isento
11	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,CÃ¡lculo Diferencial e Integral B,214,Isento
12	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,ProgramaÃ§Ã£o de Computadores I,115,Isento
13	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,ProgramaÃ§Ã£o de Computadores II,116,Isento
14	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,CÃ¡lculo NumÃ©rico (S.I.),205,Isento
15	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,MatemÃ¡tica Discreta,195,Isento
16	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Estrutura de Dados I,40,Isento
17	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Banco de Dados I,11,Isento
18	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,Sistemas Operacionais I,135,Isento
19	4611,Masculino,1993-02-02,Demais vagas,Privada,Cancelamento,2020,1,,,2020,1,IntroduÃ§Ã£o a Redes de Computadores,84,Isento
20	3794,Masculino,1981-03-24,Demais vagas,Privada,Cursando,2018,1,,,2018,1,Fundamentos da Filosofia,50,Aprovado
21	3794,Masculino,1981-03-24,Demais vagas,Privada,Cursando,2018,1,,,2018,1,Pesquisa Operacional II,108,Aprovado
22	3794,Masculino,1981-03-24,Demais vagas,Privada,Cursando,2018,1,,,2018,1,IntroduÃ§Ã£o Ã Engenharia da ProduÃ§Ã£o,79,Aprovado
23	3794,Masculino,1981-03-24,Demais vagas,Privada,Cursando,2018,1,,,2018,1,GestÃ£o da Qualidade,57,Aprovado
24	3794,Masculino,1981-03-24,Demais vagas,Privada,Cursando,2018,1,,,2018,2,Sistemas Operacionais I,135,Aprovado

Fonte: adaptado pelo autor (2022)

A Faculdade forneceu os dados relacionados aos alunos referentes ao período de 2017 a 2021, compreendendo os 5 anos do recorte temporal deste estudo. A partir desses dados foram realizadas as etapas do processo KDD conforme serão descritos adiante.

4.1 Dados

Para realização do trabalho foram extraídos do banco de dados da FeMASS os seguintes dados com base na pesquisa bibliográfica realizada no capítulo de referencial teórico:

Quadro 2 - Variáveis selecionadas para o estudo

Variável	Descrição
codigoaluno	Código único gerado para identificação do aluno
sexo	Masculo ou Feminino
datanascimento	Data usada para calcular a idade do aluno
formaingresso	Cotas, ENEM, Demais vagas
ensinomediotipoescolaconclusao	Pública ou Privada
estadoatual	Cursando, Formado, Cancelamento
anoingresso	Ano de ingresso na faculdade
semestreingresso	Semestre ingressado
dataconclusaocurso	Ano de formação dos alunos formados
anocursoudisciplina	Ano que cursou determinada disciplina
semestrecursoudisciplina	Semestre que cursou a disciplina
nomedisciplina	Nome da disciplina
codigodisciplina	código único da disciplina
situacao	Aprovado, Reprovado ou Isento
codigogradecurricular	1 - Grade antiga / 10 - Grade nova

Fonte: adaptado pelo autor (2022)

Abaixo foram descritas com mais detalhes as variáveis do Quadro 2 utilizadas no estudo.

- Codigoaluno – variável de identificação única de cada aluno que estava matriculado ou evadido no período que tange essa pesquisa;
- Sexo – gênero informado pelo aluno no cadastro da instituição;
- Datanascimento – data de nascimento do aluno;
- Formaingresso – forma de ingresso do aluno na faculdade (ENEM, cotas, demais vagas, reingresso, transferência interna ou transferência externa);
- Ensinomediotipoescolaconclusao – qual tipo de rede o aluno frequentou o ensino médio (pública ou privada);
- Estadoatual – qual o estado da matrícula do aluno no ano de 2022 (cancelamento, formado, cursando, trancamento de matrícula);
- Anoingresso – essa variável representa o ano que o aluno se matriculou na instituição;
- Semestreingresso – essa variável distingue em qual dos dois semestres o aluno ingressou na faculdade;
- Anocursoudisciplina – ano que o aluno que cursou determinada disciplina;
- Semestrecursoudisciplina – semestre que um aluno cursou a disciplina no ano especificado;
- Nomedisciplina – nome por extenso das disciplinas ofertadas no curso de Sistemas de Informação;
- Codigodisciplina – código único identificador da disciplina;
- Situação – situação do aluno na disciplina;
- Codigogradecurricular – variável gerada após identificar que os anos selecionados para o estudo tinham diferenças entre a grade curricular nova e a grade curricular antiga da instituição.

Para chegar na seleção das variáveis usadas no estudo, o autor buscou na literatura as características mais usadas entre os estudos de trabalhos de pesquisa semelhantes. Abaixo foram extraídas duas tabelas como exemplo de outros autores em suas respectivas pesquisas. Quadro 3 - Primão (2022) faz um resumo de variáveis mais importantes, segundo a literatura usada em sua pesquisa e Quadro 4 - Hoed (2016) descreve quais variáveis foram extraídas para um estudo de evasão na UnB - Universidade de Brasília.

Quadro 3 - Lista de variáveis segundo a Literatura

Variáveis	Referências	Quantidade
Sexo	Casanova *et al.* (2021); Silva *et al.* (2020.2); Freitas *et al.* (2020); Silva *et al.* (2020.1); Suharjito (2019); Muñiz *et al.* (2019); Adejo e Connolly (2018); Costa *et al.* (2017).	8
Nota média disciplinas concluídas	Muñiz *et al.* (2019); Suharjito (2019); Beaulac e Rosenthal (2019); Costa *et al.* (2017); Adekitan e Salau (2019); Ezz e Elshenawy (2019); Zulfiker *et al.* (2020).	7
Idade	Costa *et al.* (2017); Adejo e Connolly (2018); Suharjito (2019); Casanova *et al.* (2021); Silva *et al.* (2020.2); Muñiz *et al.* (2019).	6
Renda familiar per capita	Costa *et al.* (2017); Adejo e Connolly (2018); Muñiz *et al.* (2019); Suharjito (2019); Silva *et al.* (2020.1); Freitas *et al.* (2020).	6
Estado civil	Costa *et al.* (2017); Muñiz *et al.* (2019); Suharjito (2019); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	5
Cidade do aluno	Costa *et al.* (2017); Adejo e Connolly (2018); Muñiz *et al.* (2019); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	5
Raça	Adejo e Connolly (2018); Silva *et al.* (2020.1); Freitas *et al.* (2020); Silva *et al.* (2020.2).	4
Campus/cidade campus	Costa *et al.* (2017); Adejo e Connolly (2018); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	4
Ano/semestre ingresso	Costa *et al.* (2017); Muñiz *et al.* (2019); Silva *et al.* (2020.2).	3
Tipo de ingresso	Adejo e Connolly (2018); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	3
Origem do ensino anterior	Freitas *et al.* (2020); Silva, Cabral e Pacheco (2020).	2
Turno curso	Silva, Almeida e Ramalho (2020); Costa *et al.* (2017).	2
Curso	Casanova *et al.* (2021); Beaulac e Rosenthal (2019).	2
Reprovações	Shirasu e Albuquerque (2016); Ezz e Elshenawy (2019)	2
Semestre	Costa *et al.* (2017); Beaulac e Rosenthal (2019).	2
Índice de desenvolvimento humano por município	Freitas *et al.* (2020).	1
Disciplinas concluídas	Muñiz *et al.* (2019).	1
Tipo de aprendizagem	Adejo e Connolly (2018).	1

Fonte: Primão (2022, p. 35-36)

Quadro 4 - Variáveis utilizadas no estudo da UnB

Variável	Descrição
numero_matricula	Número de matricula do aluno no curso
nome	Nome do aluno
evadido	Situação do aluno se evadido ou não
periodos_cursados	Número de períodos cursados
forma_ingresso	Forma de ingresso do aluno
sexo	Sexo do aluno
cotista	Informa se o aluno ingressou por cotas ou sistema universal
reprovado_calculo_1	Informa se o aluno reprovou na disciplina Cálculo 1
reprovado_algoritmo	Informa se o aluno reprovou em disciplinas de algoritmos
curso	Nome do curso do aluno
telefone	Número de telefone do aluno
email	Endereço de mail do aluno

Fonte: Hoed (2016, p. 48)

Desta forma o autor dessa pesquisa pode chegar as variáveis utilizadas nesse estudo que conforme mostrado no Quadro 2.

4.2 Limpeza e pré-processamento

Segundo Fayyad *et al.* (1996), estão incluídos no processo de limpeza e pré-processamento dos dados as operações de limpeza para redesenhar os dados. Por meio desse modelo foram definidas estratégias para remover inconsistências de dados ou até mesmo mudanças de sequência caso necessário.

Para dar início ao trabalho, foi necessário realizar a limpeza de dados da base enviada pela faculdade. No primeiro momento, a base enviada pela instituição apresentava 9.526 linhas referente a todos os alunos e disciplinas que estes cursaram até o ano de 2021. Um dos primeiros passos foi a transcrição de todas as disciplinas de linhas para colunas, para que assim fosse possível identificar cada aluno em uma única linha da tabela. Dessa forma, houve uma redução de 9.526 linhas para 506 linhas, sendo esta a quantidade de alunos nos parâmetros de consulta do banco de dados dos anos de 2017 a 2022.

Após essa primeira alteração foi realizada a redução de *status* de algumas variáveis dentro da base. Logo, os alunos que estavam com situação atual “aguardando x dia” para inscrição foram alterados para “Cancelamento”, pois esses alunos não realizaram matrícula até o início do segundo semestre de 2022. Ainda nesta mesma coluna de estado do aluno, foram removidos todos os alunos que constavam como “Formado”, reduzindo a base de dados para 495 linhas.

Também houve a remoção de um aluno da base, pois na sua data de nascimento constava o mesmo ano em que o aluno ingressou na faculdade, assim essa linha foi removida deixando a base com 494 linhas únicas.

Todos os alunos que estavam em *status* “trancamento de matrícula” foram alterados para “cursando”, pois apesar desses alunos não estarem no momento com alguma disciplina cadastrada, eles ainda mantem ligação com a instituição e nesse estudo foi considerado “evadido” todo o aluno que perdeu o vínculo com o curso de Sistemas de Informação.

Para os alunos que fizeram a mesma matéria em semestres ou anos diferentes foi considerada apenas a última tentativa em que o aluno realizou a matéria. Assim, caso o aluno tenha reprovado no primeiro semestre e no segundo semestre tenha sido aprovado na mesma matéria, foi considerado apenas o último *status* de aprovado na matéria.

Para realizar a separação da faixa etária dos alunos, o campo “datanascimento” foi transformado em idade, de acordo com o ano de ingresso do estudante. Posteriormente, o campo

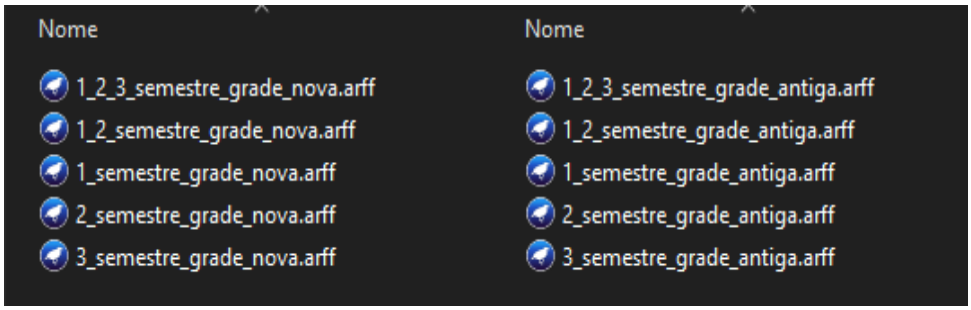
idade foi modificado para as faixas etárias “<=19”, “20-24”, “25-29”, “30-34”, “35-39” e “>=40”.

O campo “formaingresso” teve uma padronização dos *status* “Reingresso”, “Transferência externa” e “Transferência Interna”. Esses 3 possíveis valores foram sintetizados como “Outros”, pois havia baixa incidência deles na base de dados que pudesse gerar alguma análise relevante ao estudo.

Por último, todos os alunos que isentaram alguma matéria e tinham o *status* “Isento”, tiveram o dado modificado para “Aprovado”, pois a isenção da matéria é a garantia de aprovação dela para aquele aluno.

Após a limpeza e transformação da base de dados, foi realizada a conversão do arquivo CSV para ARRF e gerados 10 novos arquivos para análise referente aos 3 primeiros semestres do aluno, com o intuito de identificar quais matérias desses períodos podem influenciar na desistência ou não desse aluno, conforme Figura 8.

Figura 8 - Arquivos ARFF gerados para o estudo



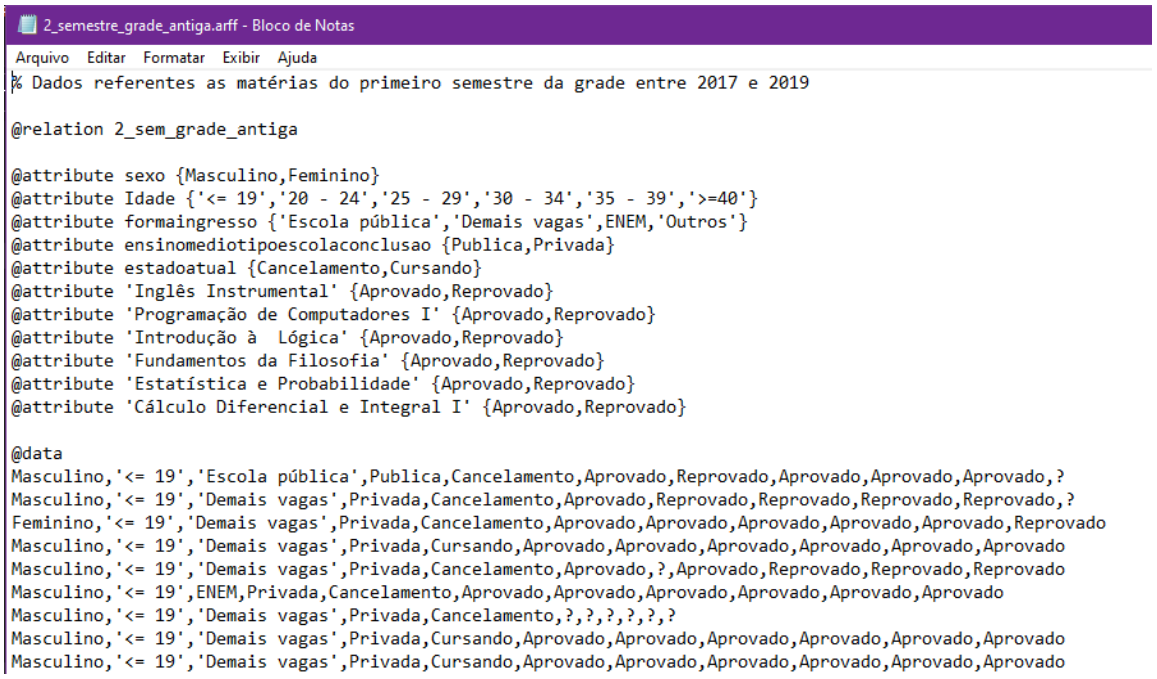
Fonte: adaptado pelo Autor (2022)

Os arquivos foram separados por semestre e grade curricular, visto que a partir do primeiro semestre de 2020 foi realizada a atualização da grade curricular do curso de sistemas de informação e as matérias correspondentes do primeiro ao terceiro semestre entre as duas grades não são totalmente equivalentes, levando assim a necessidade de realizar tal separação para que a evasão fosse analisada de forma mais precisa para cada versão da grade curricular. Para realizar tal separação foi utilizado o código de grade nova para separar todos os alunos que se encontravam associados a essa nova grade, pois alguns alunos da grade antiga tiveram a opção de migrar para grade nova.

4.3 *Machine Learning* no Weka

Na criação dos arquivos ARFF algumas colunas foram removidas, deixando somente as colunas que interessavam para o estudo, conforme mostrado a seguir na Figura 9 como exemplo dos dados do segundo semestre da grade antiga.

Figura 9 - Arquivo ARFF 2º semestre da grade antiga



2_semestre_grade_antiga.arff - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
% Dados referentes as matérias do primeiro semestre da grade entre 2017 e 2019

@relation 2_sem_grade_antiga

@attribute sexo {Masculino,Feminino}
@attribute Idade {'<= 19','20 - 24','25 - 29','30 - 34','35 - 39','>=40'}
@attribute formaingresso {'Escola pública','Demais vagas',ENEM,'Outros'}
@attribute ensinomediotipoescolaconclusao {Publica,Privada}
@attribute estadoatual {Cancelamento,Cursando}
@attribute 'Inglês Instrumental' {Aprovado,Reprovado}
@attribute 'Programação de Computadores I' {Aprovado,Reprovado}
@attribute 'Introdução à  Lógica' {Aprovado,Reprovado}
@attribute 'Fundamentos da Filosofia' {Aprovado,Reprovado}
@attribute 'Estatística e Probabilidade' {Aprovado,Reprovado}
@attribute 'Cálculo Diferencial e Integral I' {Aprovado,Reprovado}

@data
Masculino,'<= 19','Escola pública',Publica,Cancelamento,Aprovado,Reprovado,Aprovado,Aprovado,Aprovado,?
Masculino,'<= 19','Demais vagas',Privada,Cancelamento,Aprovado,Reprovado,Reprovado,Reprovado,Reprovado,?
Feminino,'<= 19','Demais vagas',Privada,Cancelamento,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado,Reprovado
Masculino,'<= 19','Demais vagas',Privada,Cursando,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado
Masculino,'<= 19','Demais vagas',Privada,Cancelamento,Aprovado,?,Aprovado,Reprovado,Reprovado,Reprovado
Masculino,'<= 19',ENEM,Privada,Cancelamento,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado
Masculino,'<= 19','Demais vagas',Privada,Cancelamento,?,?,?,?,?,?
Masculino,'<= 19','Demais vagas',Privada,Cursando,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado
Masculino,'<= 19','Demais vagas',Privada,Cursando,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado
```

Fonte: adaptador pelo Autor (2022)

Se o arquivo do 2° semestre for comparado ao arquivo do acumulado 1º e 2º semestres, temos apenas o incremento dos atributos de matérias do 1° semestre, conforme mostra Figura 10.

Figura 10 - Arquivo ARFF acumulado 1º e 2º semestre grade antiga

1_2_semestre_grade_antiga.arff - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
% Dados referentes as matérias do primeiro e segundo semestre da grade entre 2017 e 2019

@relation 1_2_sem_grade_antiga

@attribute sexo {Masculino,Feminino}
@attribute Idade {'<= 19','20 - 24','25 - 29','30 - 34','35 - 39','>=40'}
@attribute formaingresso {'Escola pública','Demais vagas',ENEM,'Outros'}
@attribute ensinomediotipoescolaconclusao {Publica,Privada}
@attribute estadoatual {Cancelamento,Cursando}
@attribute 'Introdução à  Administração' {Aprovado,Reprovado}
@attribute 'Inglês Instrumental' {Aprovado,Reprovado}
@attribute 'Programação de Computadores I' {Aprovado,Reprovado}
@attribute 'Introdução à  Engenharia da Produção' {Aprovado,Reprovado}
@attribute 'Introdução à  Tecnologia da Informação' {Aprovado,Reprovado}
@attribute 'Introdução ao Cálculo (SI e EP)' {Aprovado,Reprovado}
@attribute 'Metodologia de Pesquisa' {Aprovado,Reprovado}
@attribute 'Introdução à  Lógica' {Aprovado,Reprovado}
@attribute 'Fundamentos da Filosofia' {Aprovado,Reprovado}
@attribute 'Português Instrumental' {Aprovado,Reprovado}
@attribute 'Estatística e Probabilidade' {Aprovado,Reprovado}
@attribute 'Cálculo Diferencial e Integral I' {Aprovado,Reprovado}

@data
Masculino,'<= 19','Escola
pública',Publica,Cancelamento,Aprovado,Aprovado,Reprovado,Aprovado,Aprovado,Reprovado,Aprovado,Aprovado,Aprovado,Aprovado,Aprovado,?
Masculino,'<= 19','Demais
vagas',Privada,Cancelamento,Aprovado,Aprovado,Reprovado,Aprovado,Aprovado,Reprovado,Aprovado,Reprovado,Reprovado,Aprovado,Reprovado,?
Feminino,'<= 19','Demais
```

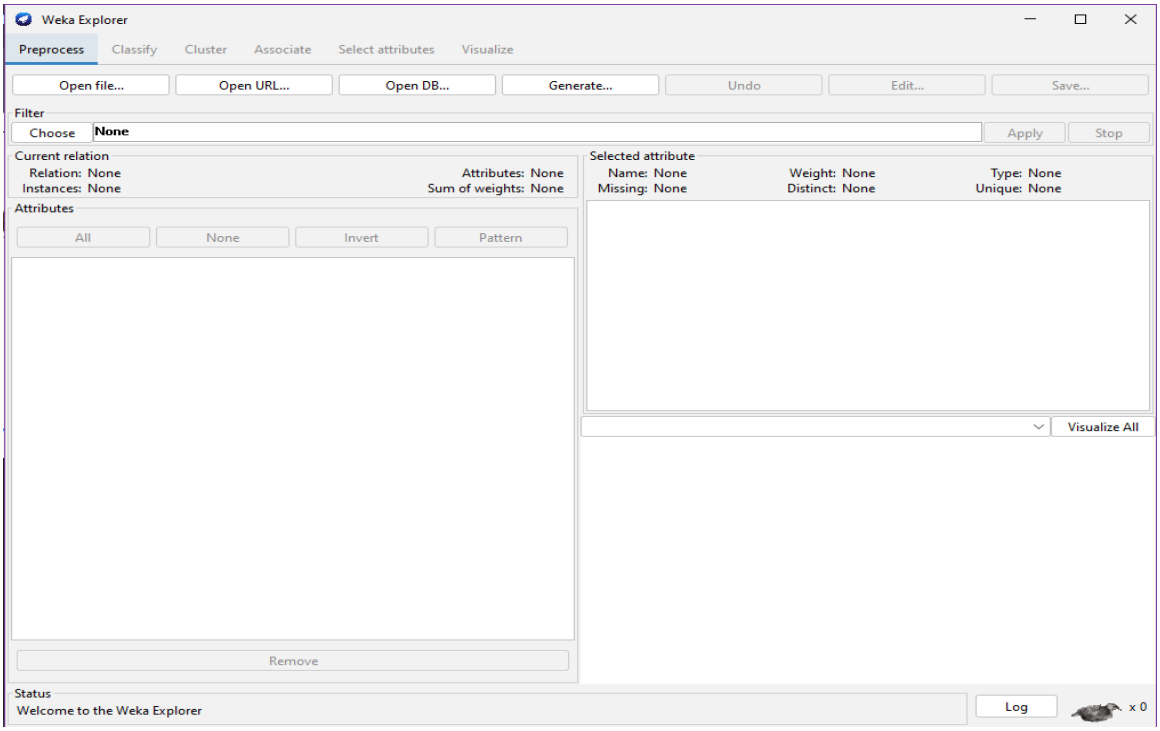
Fonte: adaptador pelo Autor (2022)

Após a padronização de todos os arquivos necessários, pôde-se iniciar a etapa de mineração dos dados na ferramenta WEKA, que será descrito no item adiante dessa pesquisa.

4.3.1 Utilizando dados no Weka

Conforme a figura 5 da tela inicial do weka apresentada na seção 3.2, foi utilizada a função "*Explorer*", que abre a tela apresentada abaixo na Figura 11:

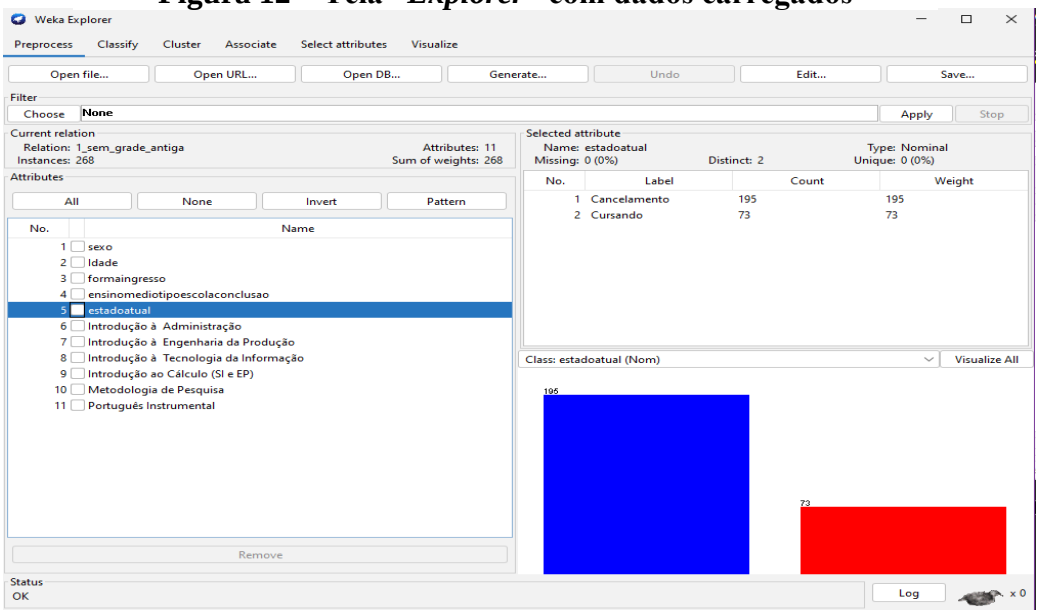
Figura 11 - Tela "*Explorer*" Weka



Fonte: adaptado Weka (2022)

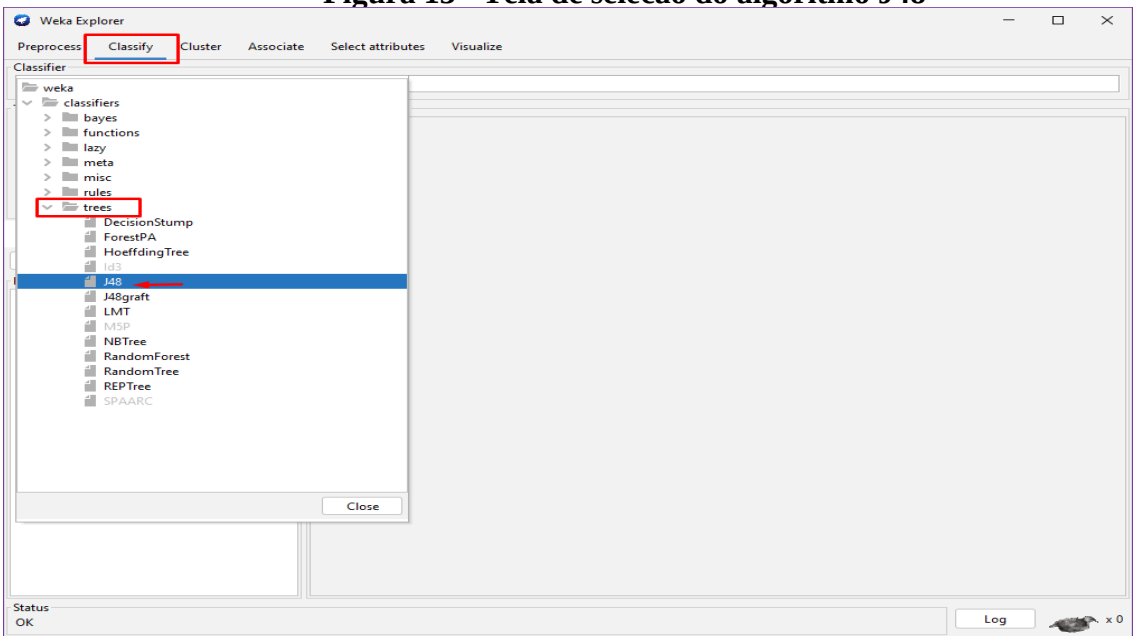
Nesta tela, foram abertos os arquivos através do botão "*Open file...*" que carregou os dados conforme mostra a Figura 12, exemplificando o arquivo ARFF do primeiro semestre da grade antiga. Em seguida foi selecionado o a opção *"Classify"* conforme mostra a Figura 13 para selecionar o algoritmo J48.

Figura 12 – Tela "*Explorer*" com dados carregados



Fonte: adaptado Weka (2022)

Figura 13 - Tela de selecão do algoritmo J48



Fonte: adaptado Weka (2022)

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), a etapa de mineração de dados é o momento para exploração e seleção do modelo, escolhendo então o algoritmo que será utilizado nos dados que foram tratados. Neste caso, utilizou-se o algoritmo de classificação J48 da técnica de ML.

A partir da hipótese sobre o aluno evadir do curso ao reprovar ou não em uma matéria, podemos seguir um modelo preditivo como base para chegar a tal conhecimento.

De acordo com Witten *et al.* (2017), o algoritmo J48 é um dos algoritmos mais robustos dos métodos de classificação e efetivo na geração de representatividade gráfica. Métodos de classificação nos ajudam a compreender a relação de “aprovação X cancelamento” de matrícula de um aluno, colaborando assim com o objetivo deste trabalho.

4.3.2 Experimentos realizados

Neste estudo foram realizados 5 experimentos para aplicação do algoritmo de ML J48 conforme descrito a seguir.

A – Classificação dos alunos evadidos e cursando do primeiro semestre de cada grade

B – Classificação dos alunos evadidos e cursando do segundo semestre de cada grade

C – Classificação dos alunos evadidos e cursando do terceiro semestre de cada grade

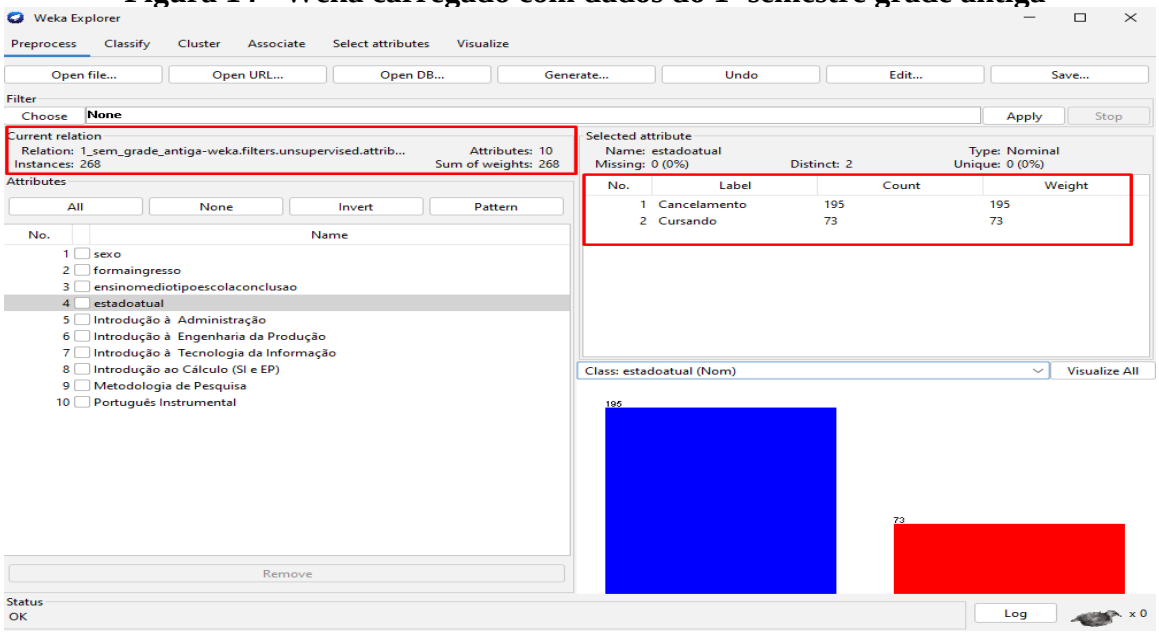
D – Classificação dos alunos evadidos e cursando do primeiro e segundo semestre de cada grade

E – Classificação dos alunos evadidos e cursando do primeiro, segundo e terceiro semestre de cada grade

4.3.2.1 Experimento A

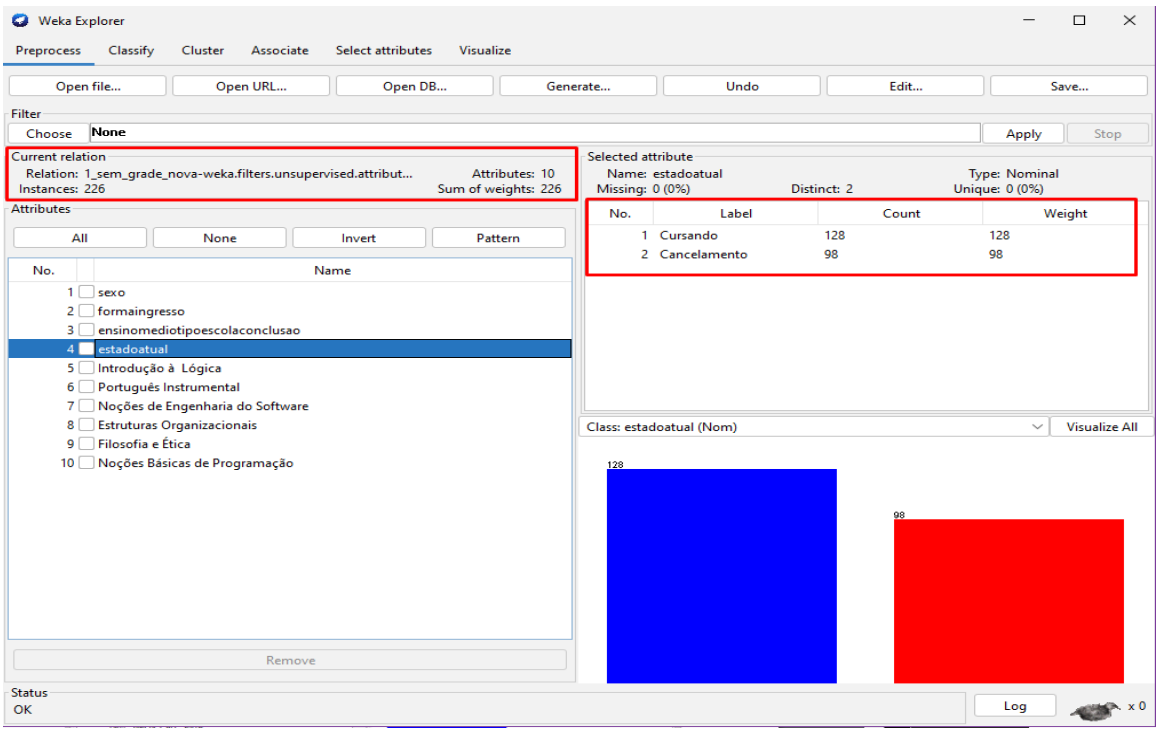
Dados alunos do primeiro semestre, grade antiga na Figura 14 e grade nova na Figura 15.

Figura 14 - Weka carregado com dados do 1º semestre grade antiga



Fonte: adaptado Weka (2022)

Figura 15 - Weka carregado com dados do 1º semestre grade nova



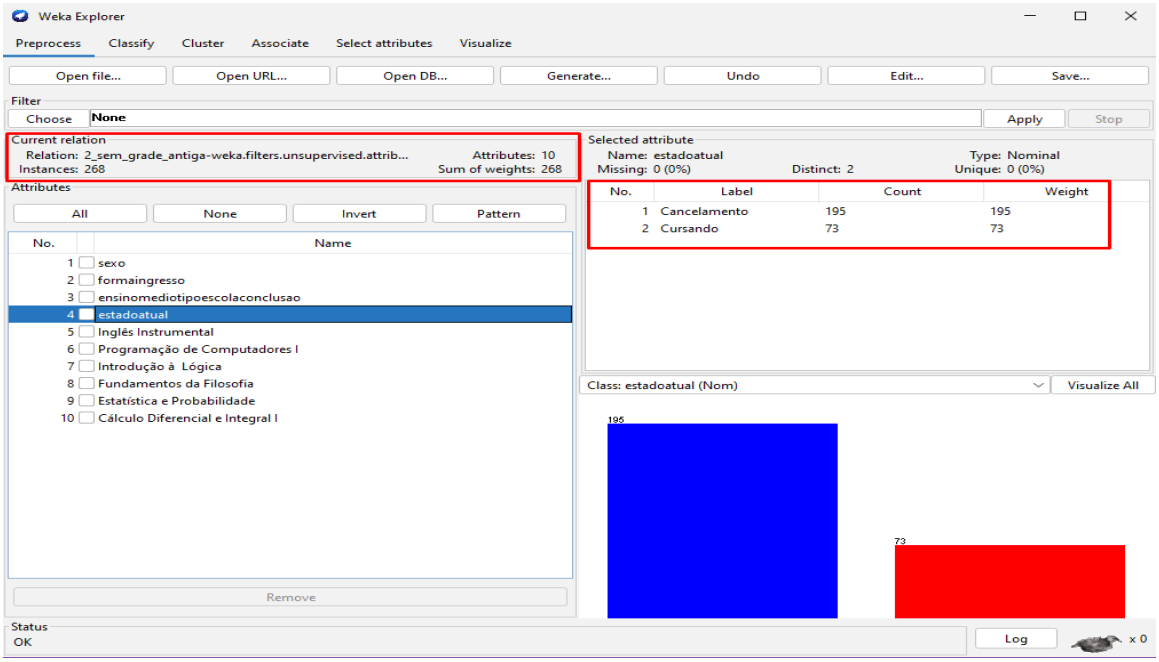
Fonte: adaptado Weka (2022)

Das 268 instâncias de alunos da grade antiga, podemos observar 195 em estado “Cancelamento” e 73 em estado “Cursando”. Já para as 226 instâncias da grade nova, podemos observar 128 alunos “Cursando” e 98 em “Cancelamento”.

4.3.2.2 Experimento B

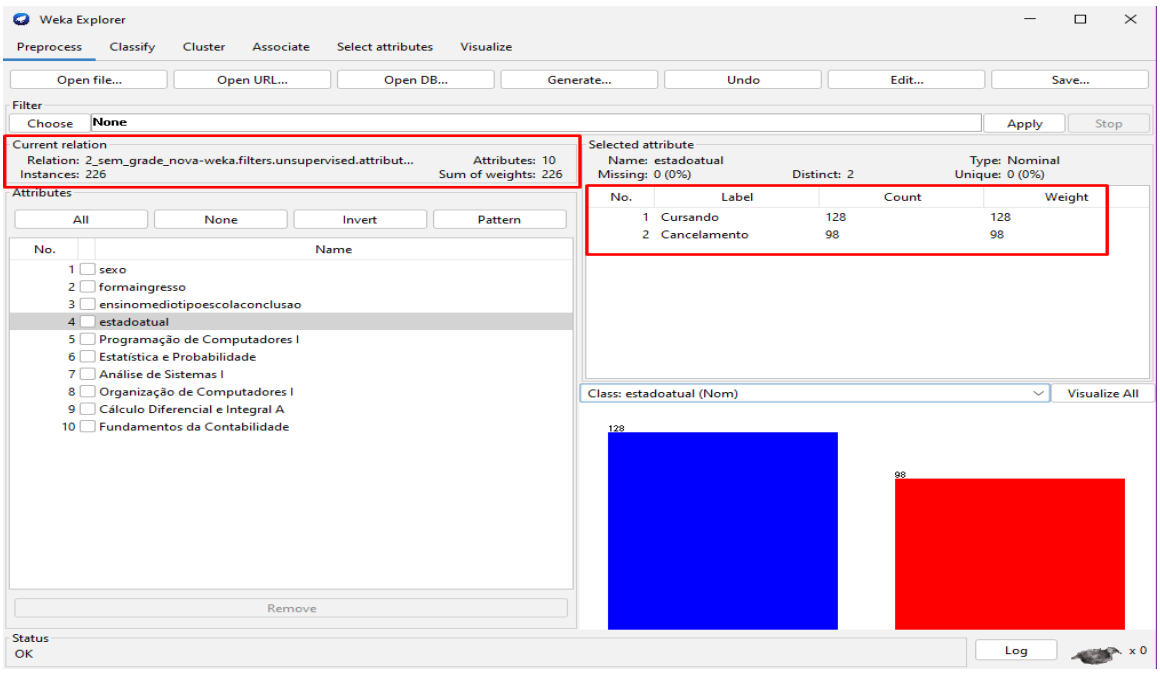
Dados alunos do segundo semestre, grade antiga na Figura 16 e grade nova na Figura 17.

Figura 16 - Weka carregado com dados do 2º semestre grade antiga



Fonte: adaptado Weka (2022)

Figura 17 - Weka carregado com dados do 2º semestre grade nova



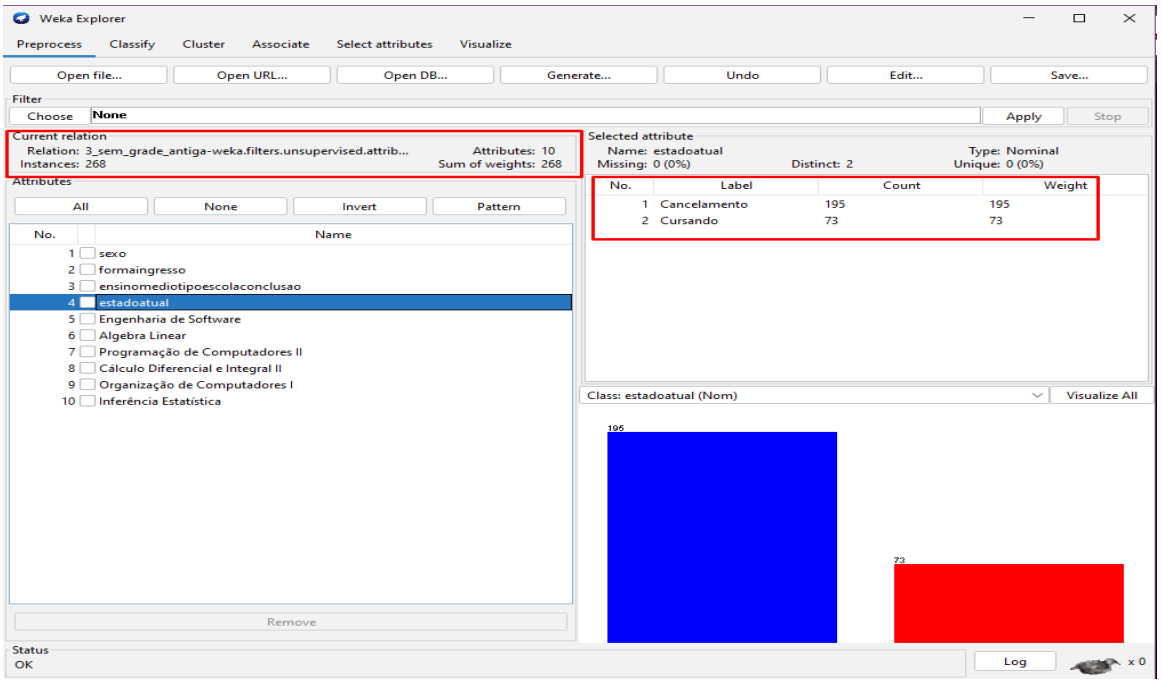
Fonte: adaptado Weka (2022)

Os números de alunos evadidos e cursando são os mesmos apresentados nas figuras 14 e 15 do experimento A. As mudanças significativas nesse experimento são as matérias que alteram de um semestre para o outro conforme pode-se observar nas imagens.

4.3.2.3 Experimento C

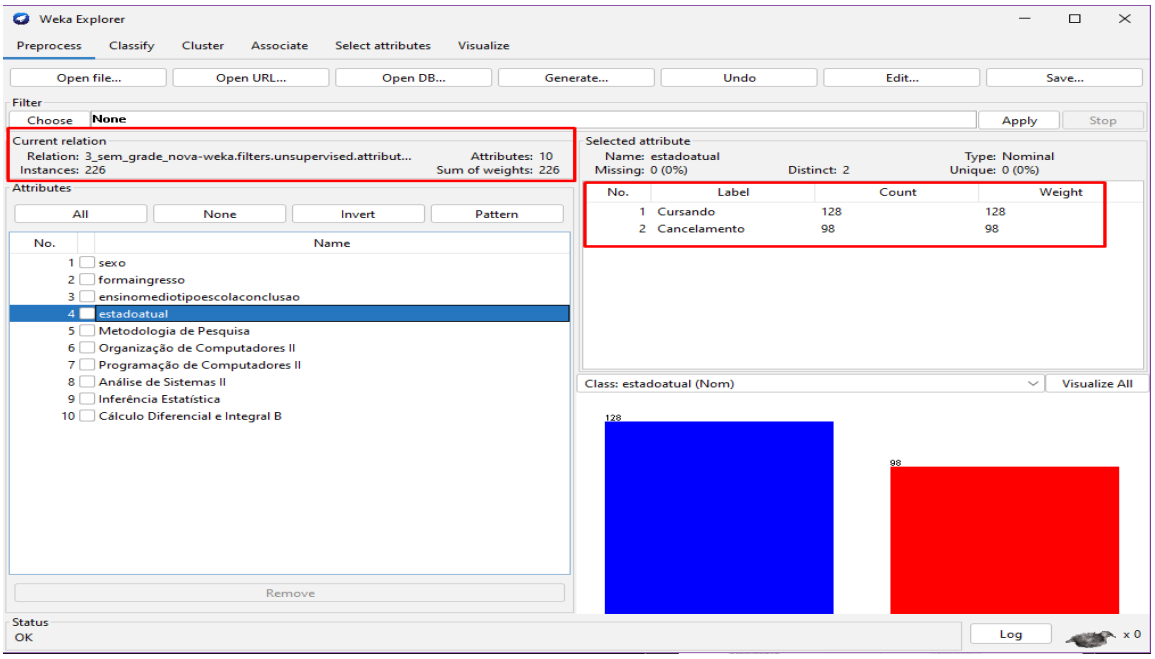
Dados alunos do terceiro semestre, grade antiga na Figura 18 e grade nova na Figura 19.

Figura 18 - Weka carregado com dados do 3º semestre grade antiga



Fonte: adaptado Weka (2022)

Figura 19 - Weka carregado com dados do 3º semestre grade nova



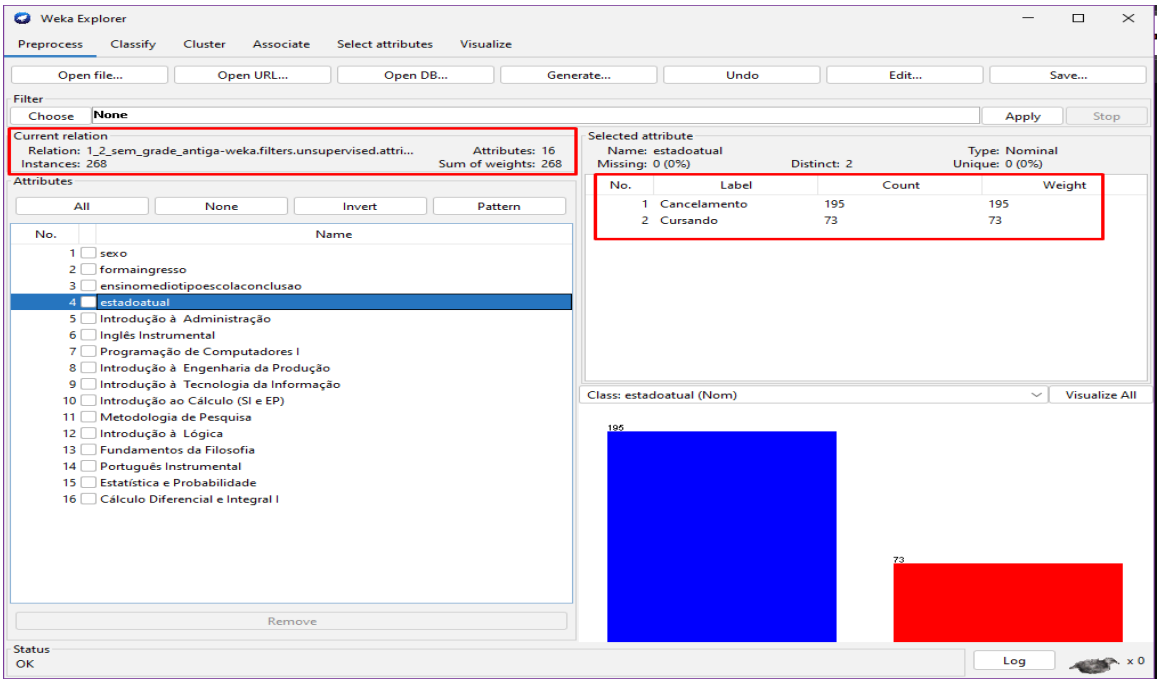
Fonte: adaptado Weka (2022)

Da mesma forma como ocorreu com o experimento B, neste experimento C os dados de alunos evadidos e cursando continuam os mesmos e a diferenciação entre eles se deu pelas matérias serem diferentes no terceiro semestre.

4.3.2.4 Experimento D

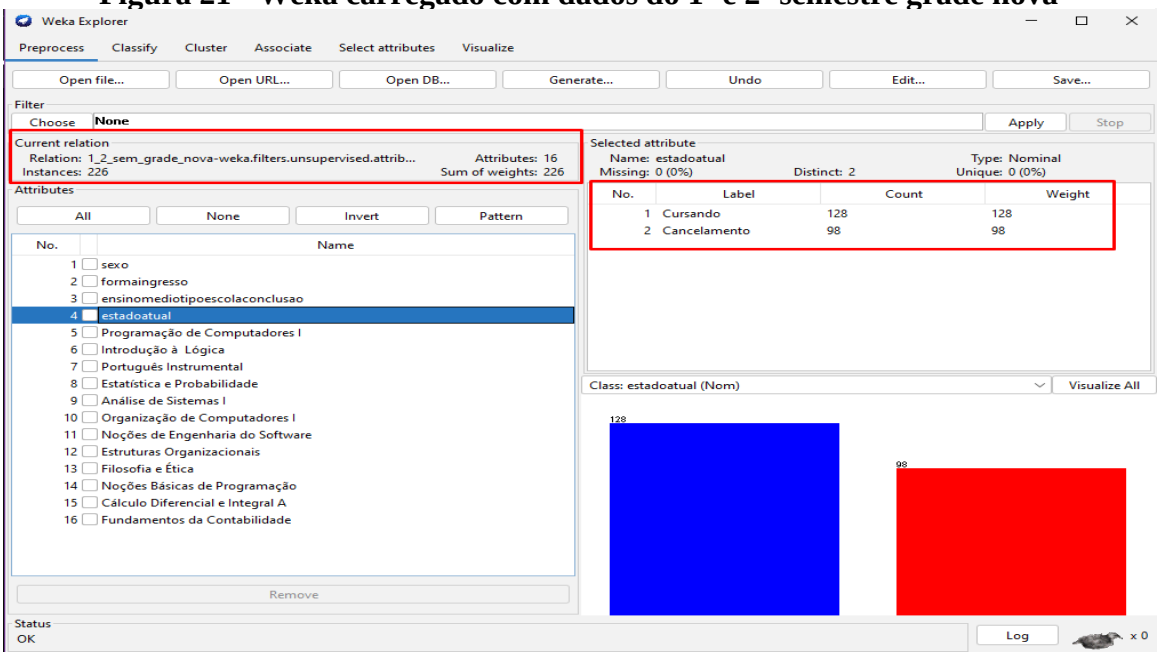
Dados alunos do primeiro e segundo semestre, grade antiga na Figura 20 e grade nova na Figura 21.

Figura 20 - Weka carregado com dados do 1º e 2º semestre grade antiga



Fonte: adaptado Weka (2022)

Figura 21 - Weka carregado com dados do 1º e 2º semestre grade nova



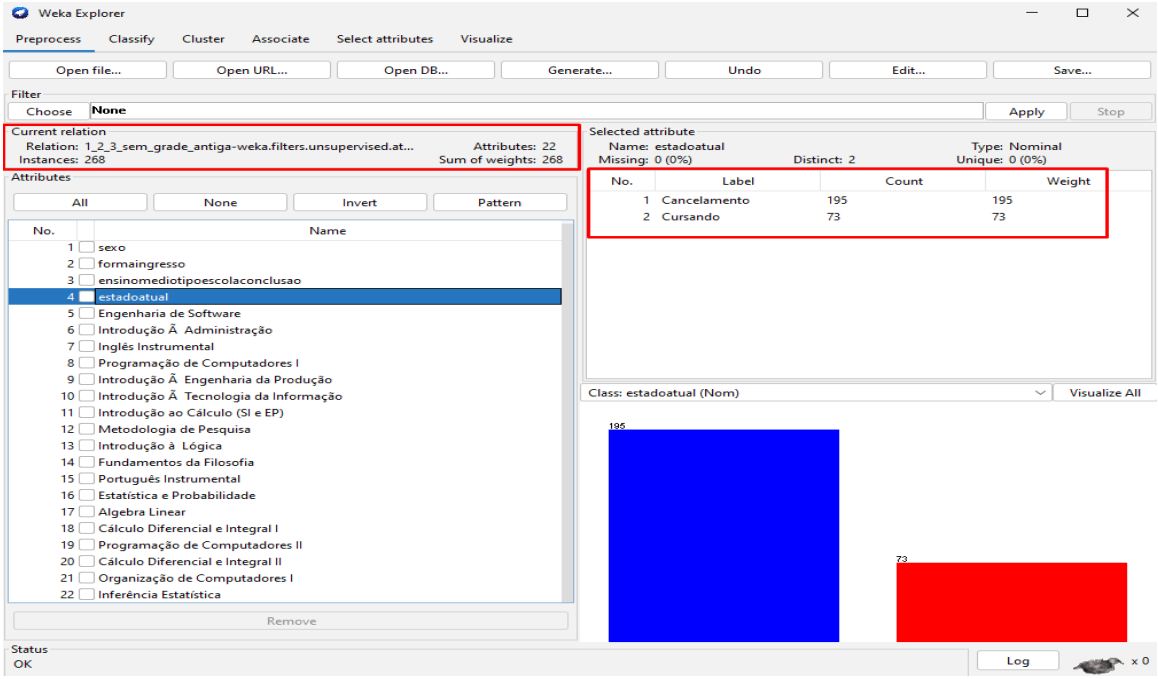
Fonte: adaptado Weka (2022)

No experimento D pode-se observar o acúmulo de matérias do primeiro e segundo semestre, mas com o mesmo quantitativo de alunos em situação de evasão e alunos em situação de continuidade do curso.

4.3.2.5 Experimento E

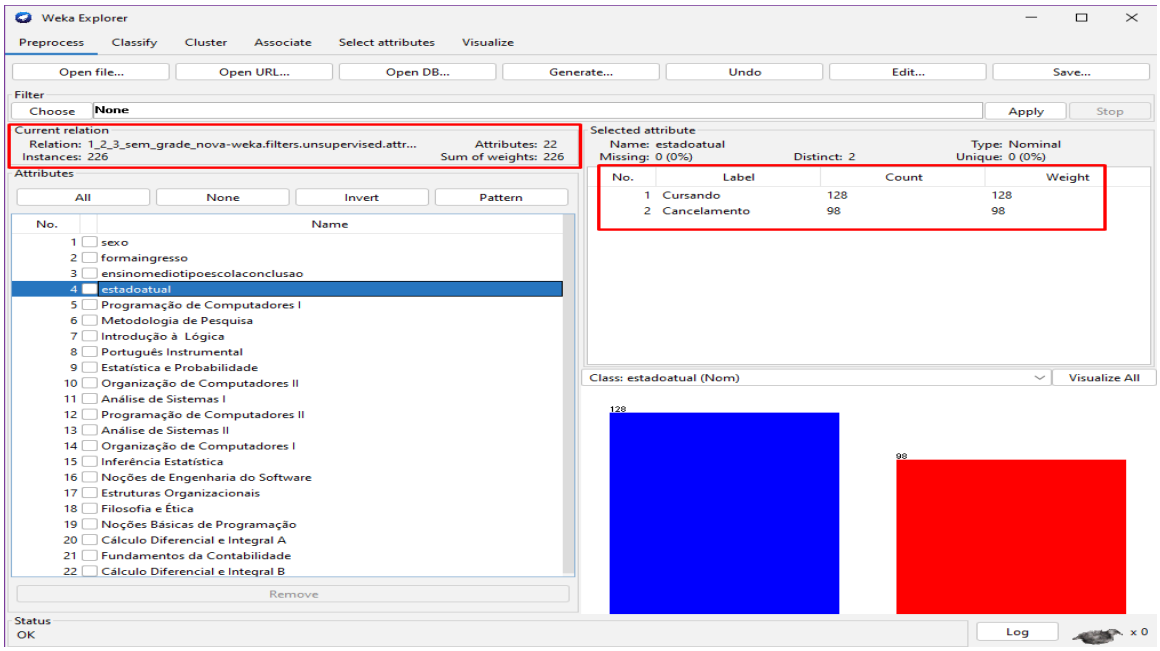
Dados alunos do primeiro, segundo e terceiro semestre, grade antiga na Figura 22 e grade nova na Figura 23.

Figura 22 - Weka carregado com dados do 1º, 2º e 3º semestre grade antiga



Fonte: adaptado Weka (2022)

Figura 23 - Weka carregado com dados do 1º, 2º e 3º semestre grade nova



Fonte: adaptado Weka (2022)

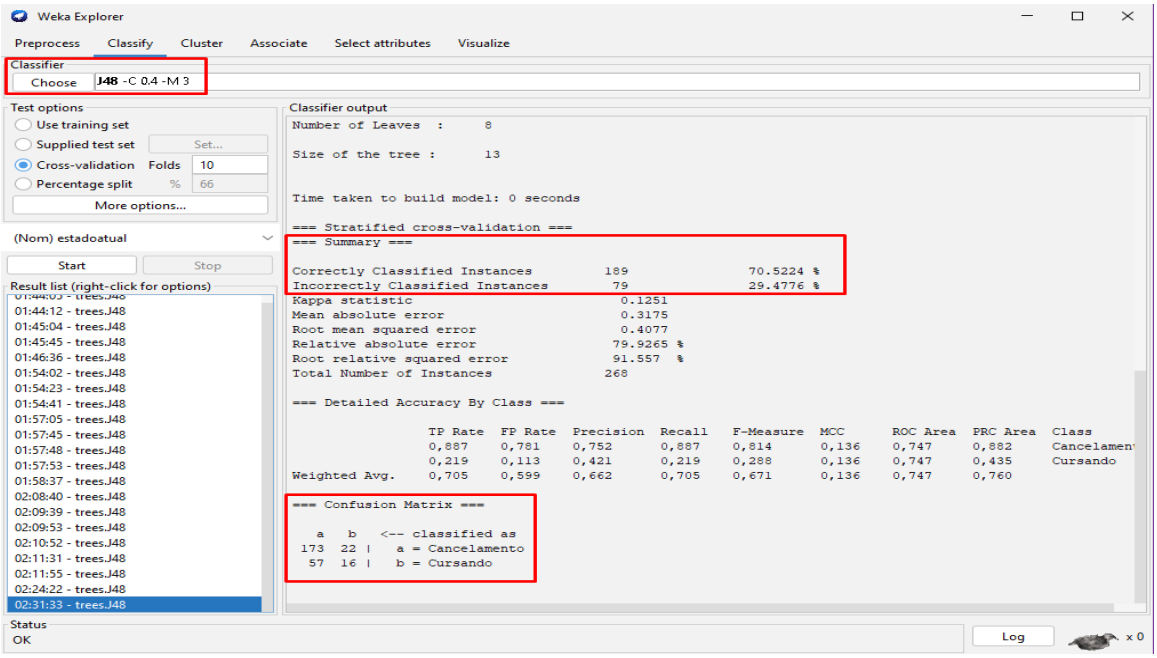
Por fim, o experimento E trouxe o cumulativo dos 3 primeiros semestres para os mesmos alunos evadidos e em curso da base de dados.

5 RESULTADOS

5.1 Resultados do experimento A

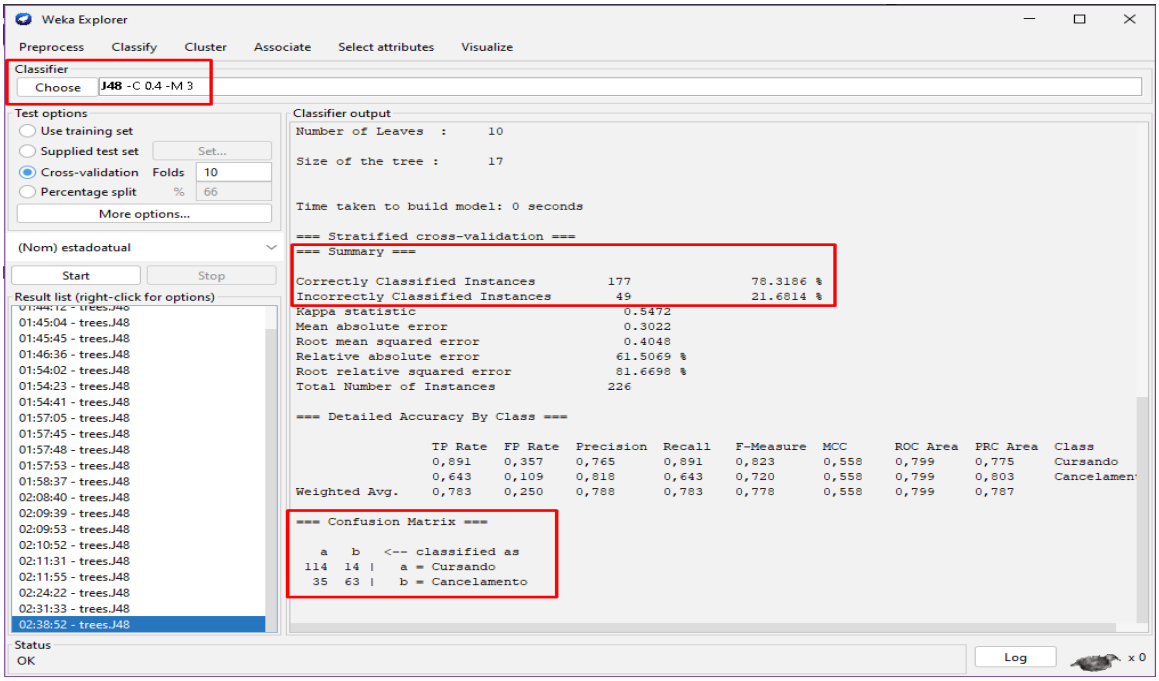
Classificação alunos do primeiro semestre, grade antiga na Figura 24 e grade nova na Figura 25.

Figura 24 - Saídas do algoritmo J48 do experimento A no Weka (grade antiga)



Fonte: adaptado Weka (2022)

Figura 25 - Saídas do algoritmo J48 do experimento A no Weka (grade nova)

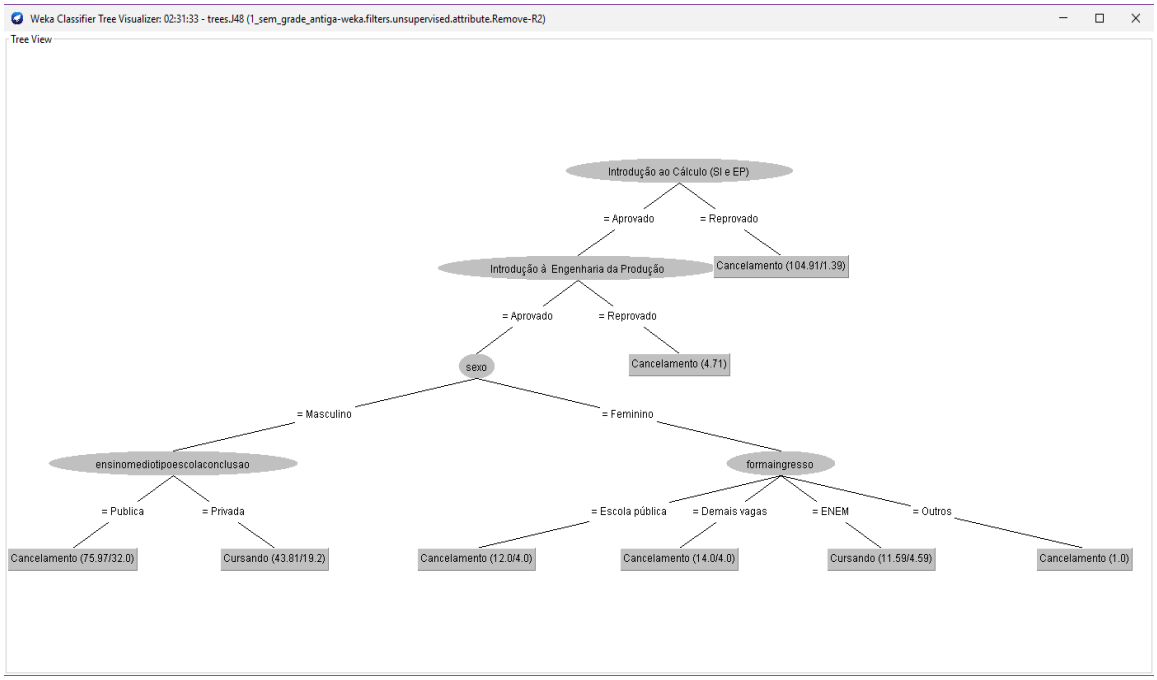


Fonte: adaptado Weka (2022)

Ao analisar os resultados do algoritmo de classificação J48 nos dados do primeiro semestre, pode-se observar ambos com mais de 70% de acerto na classificação das instâncias.

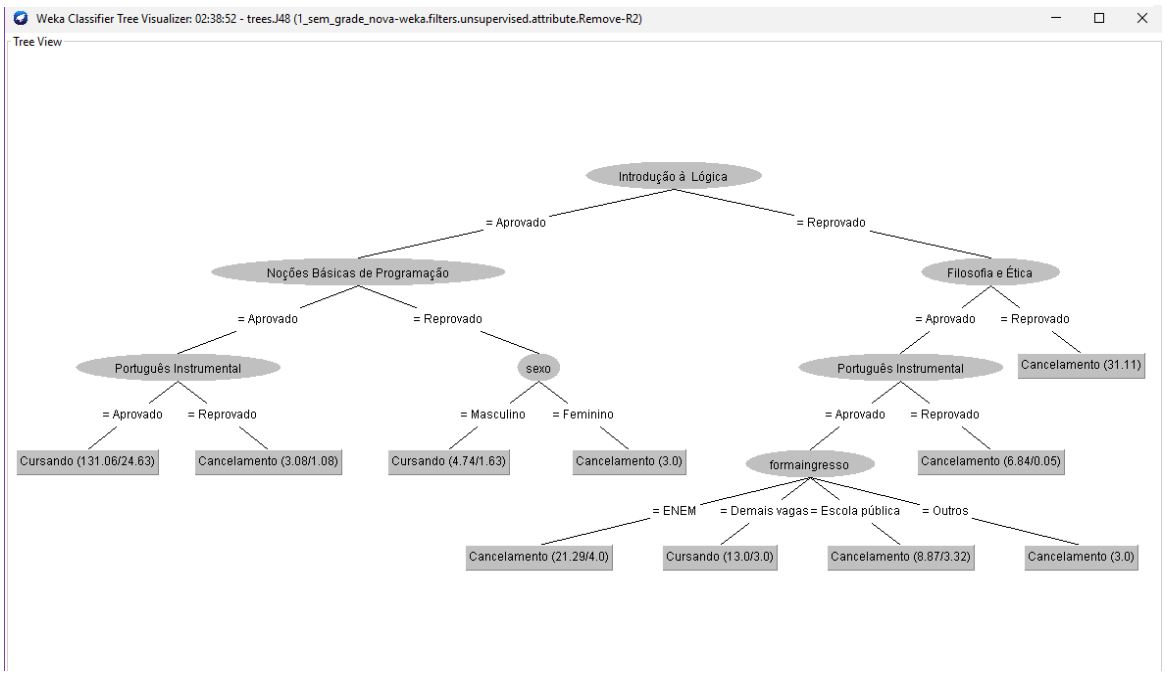
A seguir nas Figuras 26 e 27 serão apresentadas as árvores de decisão geradas pelo algoritmo na grade antiga e nova respectivamente.

Figura 26 - Árvore de decisão gerada no experimento A (grade antiga)



Fonte: adaptado Weka (2022)

Figura 27 - Árvore de decisão gerada no experimento A (grade nova)



Fonte: adaptado Weka (2022)

Já nas árvores de decisão, pode-se observar que a matéria de Introdução ao Cálculo foi o principal fator de cancelamento ou continuidade do aluno na graduação no primeiro semestre da grade antiga, enquanto na grade nova a matéria de Introdução à lógica se tornou a matéria mais relevante para essa análise.

Dessa forma pode-se extrair algumas regras mais relevantes:

- Grade antiga
    - SE Introdução ao Cálculo = reprovado ENTÃO estadoatual = cancelamento
    - SE Introdução ao Cálculo = aprovado E Introdução a Engenharia de Produção = reprovado ENTÃO estadoatual = cancelamento
    - SE Introdução ao Cálculo = aprovado E Introdução a Engenharia de Produção = aprovado E sexo = masculino E ensinomedio = publica ENTÃO estadoatual = cancelamento
    - SE Introdução ao Cálculo = aprovado E Introdução a Engenharia de Produção = aprovado E sexo = masculino E ensinomedio = privada ENTÃO estadoatual = cursando
- Grade Nova
    - SE Introdução a Lógica = reprovado E Filosofia e Ética = reprovado ENTÃO estadoatual = cancelamento
    - SE Introdução a Lógica = aprovado E Noções básicas de programação = aprovado E Português Instrumental = Aprovado ENTÃO estadoatual = cursando

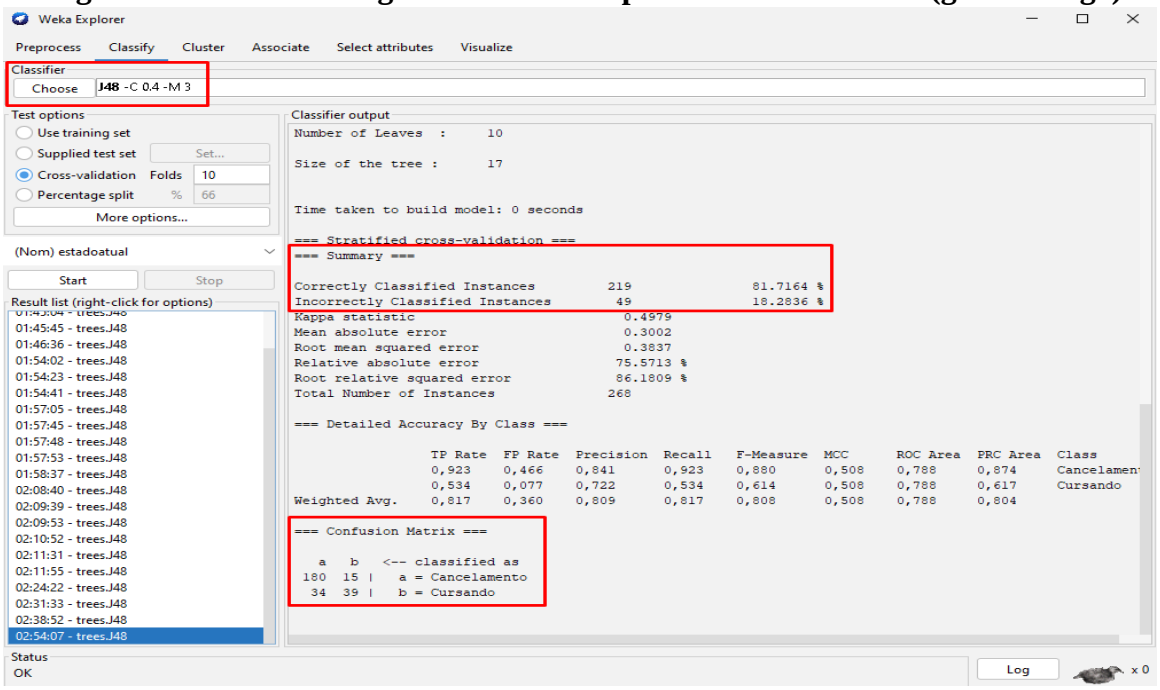
No que diz respeito a primeira árvore de classificação com alunos que fizeram matérias do primeiro semestre, observa-se que dos 101 alunos que reprovaram na matéria, 100 cancelaram a matrícula posteriormente, corroborando para a análise mostrada na árvore que se o aluno perde nessa matéria ele tem grandes chances de evadir do curso. Já na grade nova dos 83 alunos que perderam em lógica, 66 cancelaram a matrícula. A matéria Filosofia e Ética foi dada como o segundo fator determinante, visto que dos 30 alunos que perderam em ambas as matérias, todos evadiram do curso.

Desta forma tanto as regras descritas acima quanto as informações extraídas dos dados brutos corroboram para as hipóteses levantadas no capítulo 1. Validou-se tanto a extração de regras a partir da técnica de ML usando um algoritmo de classificação, quanto a identificação de matérias, que influenciam ou não na evasão, além de identificar que podem existir outros fatores como forma de tipo de escola cursada o ensino médio, sexo ou, até mesmo, forma de ingresso na instituição.

5.2 Resultados do experimento B

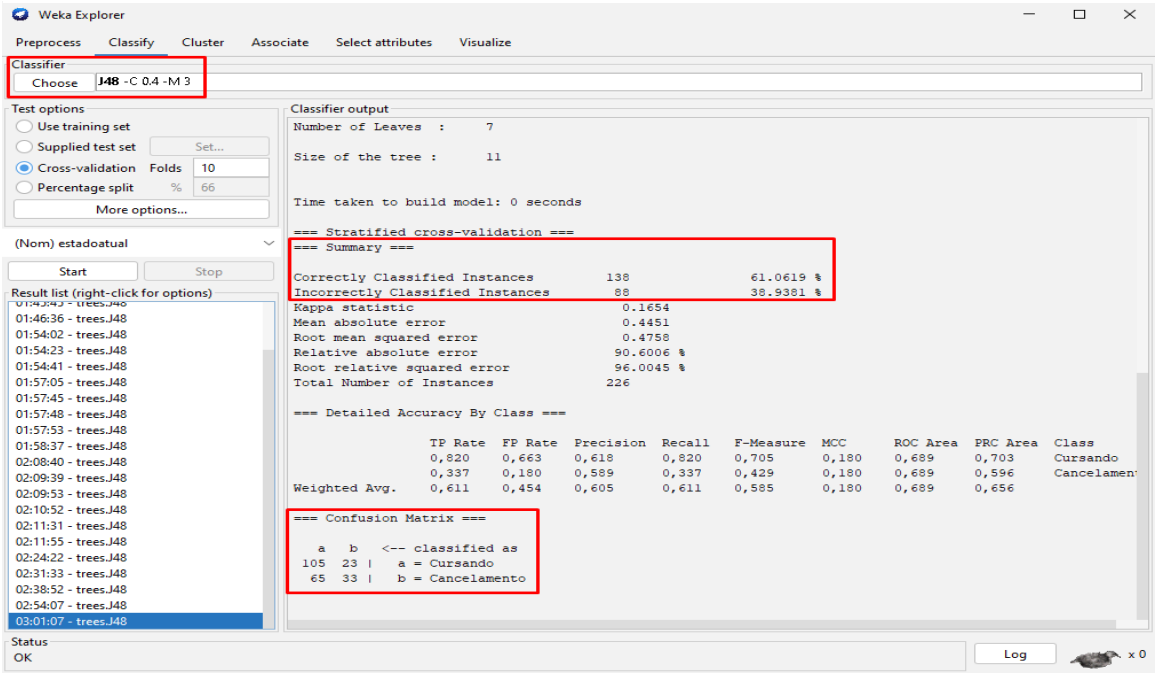
Classificação alunos do segundo semestre, grade antiga na Figura 28 e grade nova na Figura 29.

Figura 28 - Saídas do Algoritmo J48 no experimento B no Weka (grade antiga)



Fonte: adaptado Weka (2022)

Figura 29 - Saídas do Algoritmo J48 no experimento B no Weka (grade nova)



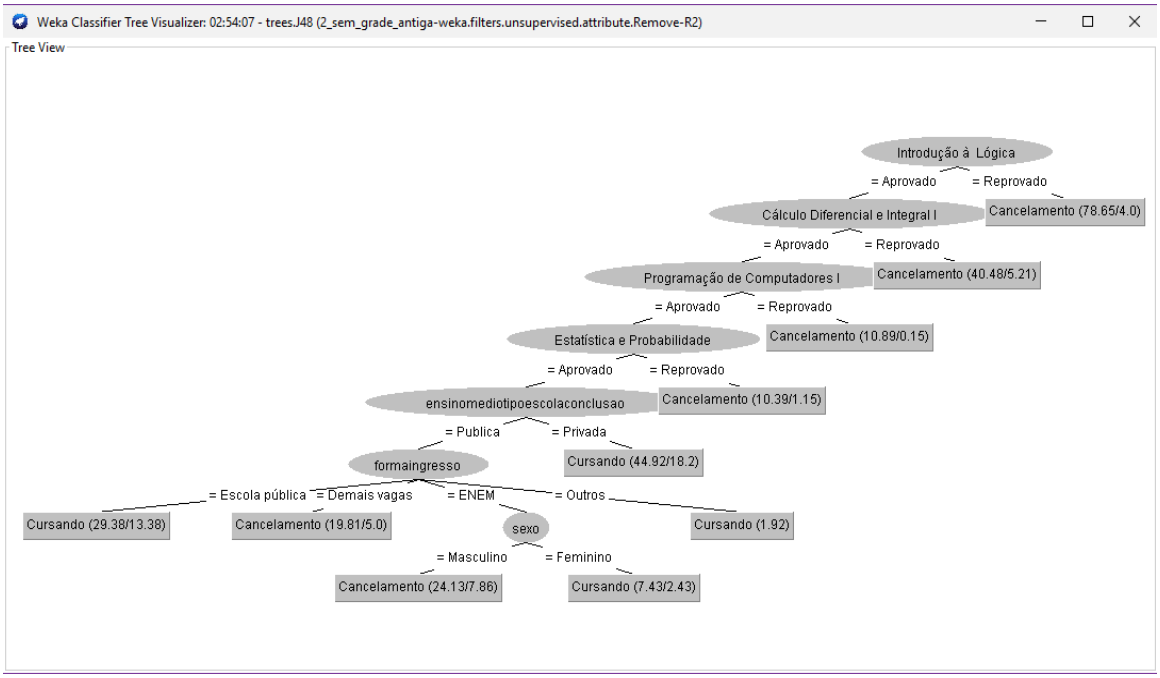
Fonte: adaptado Weka (2022)

Ao analisar os resultados do algoritmo de classificação J48 nos dados do segundo semestre, pode-se observar que as classificações da grade antiga atingem 81% de precisão,

enquanto há uma redução na previsão da classificação na grade nova, atingindo 61% de precisão.

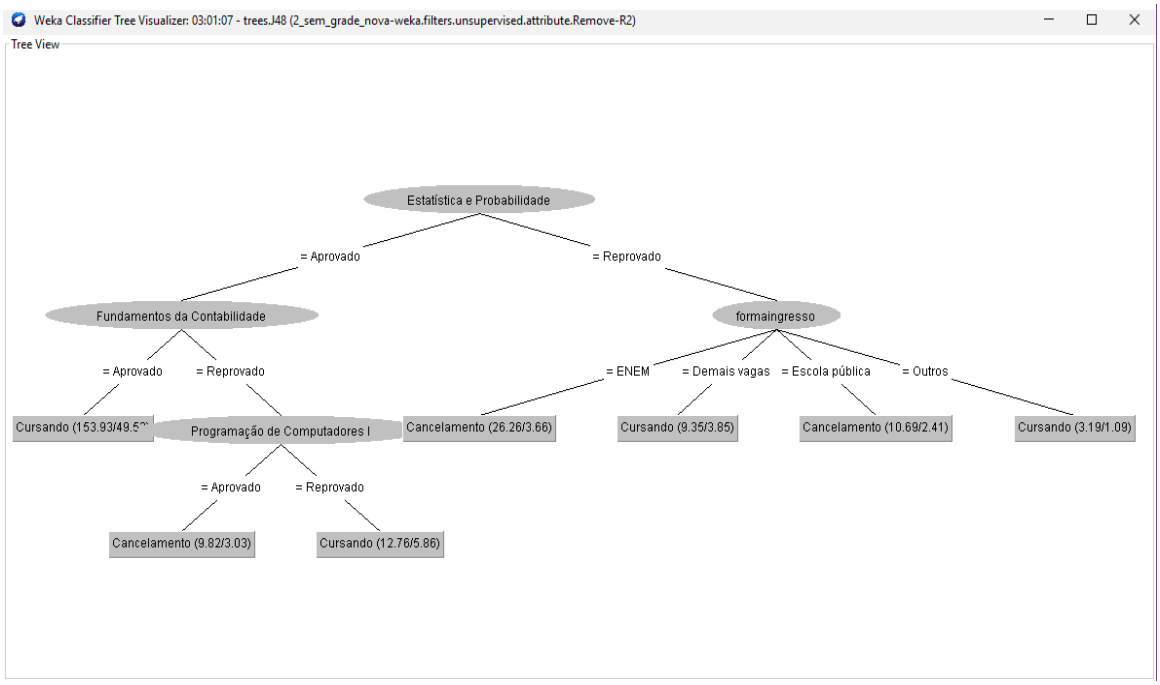
A seguir nas Figuras 30 e 31 serão apresentadas as árvores de decisão geradas pelo algoritmo na grade antiga e nova respectivamente.

Figura 30 - Árvore de decisão geradas no experimento B (grade antiga)



Fonte: adaptado Weka (2022)

Figura 31 - Árvore de decisão geradas no experimento B (grade nova)



Fonte: adaptado Weka (2022

Já nas árvores de decisão, pode-se observar que a matéria Introdução à lógica foi o principal fator de cancelamento ou continuidade do aluno na graduação no segundo semestre

da grade antiga, seguido de matérias como Cálculo Diferencial e Integral I, Programação de Computadores I e Estatística e Probabilidade. Enquanto na grade nova, a matéria Estatística de Probabilidade se tornou a matéria mais relevante para essa análise seguido de Fundamentos da Contabilidade e Programação de computadores.

Desta forma as seguintes regras mais relevantes foram extraídas das árvores de decisão:

- Grade Antiga
    - SE Introdução à Lógica = reprovado ENTÃO estadoatual = cancelamento
    - SE Introdução à Lógica = aprovado E Cálculo 1 = reprovado ENTÃO estadoatual = cancelamento
    - SE Introdução à Lógica = aprovado E Cálculo 1 = aprovado E Prog 1 = aprovado E Estatística e Probabilidade = aprovado E ensinomedio = privado ENTÃO estadoatual = cursando
- Grade Nova
    - SE Estatística e Probabilidade = reprovado E formaingresso = ENEM ENTÃO estadoatual = cancelamento
    - SE Estatística e Probabilidade = reprovado E formaingresso = Demais Vagas ENTÃO estadoatual = cursando
    - SE Estatística e Probabilidade = aprovado E Fundamentos da Contabilidade = aprovado ENTÃO estadoatual = cursando
    - SE Estatística e Probabilidade = aprovado E Fundamentos da Contabilidade = reprovado E Prog 1 = aprovado ENTÃO estadoatual = cancelamento
    - SE Estatística e Probabilidade = aprovado E Fundamentos da Contabilidade = reprovado E Prog 1 = reprovado ENTÃO estadoatual = cursando

Na árvore do experimento B da grade antiga temos como principal matéria Introdução à lógica seguida de Cálculo Diferencial Integral I. Dos 54 alunos que reprovaram no segundo semestre em Introdução a lógica, 50 tiveram a matrícula cancelada. Na grade nova a matéria Estatística de probabilidade entra como o nó mais importante da árvore, uma vez que dos 30 alunos reprovados, 24 cancelaram a matrícula.

Já olhando para a regras extraídas das árvores podemos perceber a importância de algumas matérias na continuidade ou não do curso, porém as duas últimas regras da grade nova

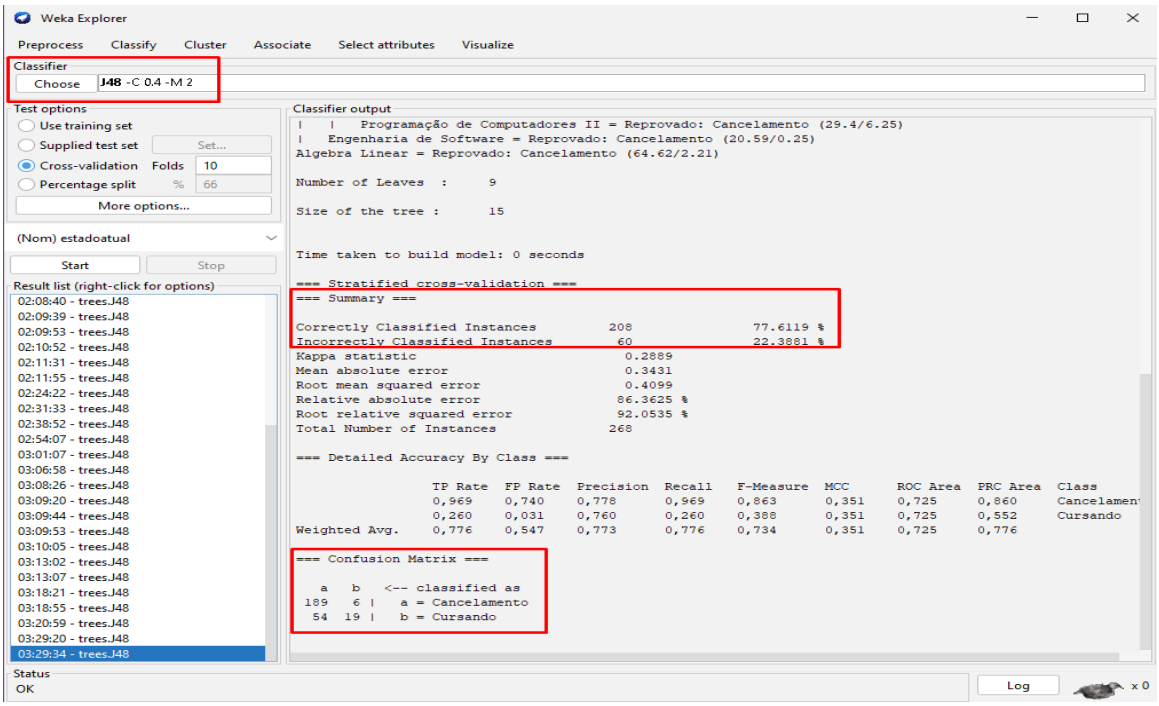
nos levam a um caminho intrigante. Se o aluno aprova em estatística, reprova em contabilidade e aprova em programação, ele evade do curso. Porém, se seguir o mesmo caminho e reprovar também em programação, o aluno continua cursando a faculdade.

Esse resultado oposto entre “aprovação-encerramento” e “reprovação-continuidade”, leva ao entendimento de que neste caso possa existir um outro fator não mapeado que influenciou na evasão/permanência do aluno no curso.

5.3 Resultados do experimento C

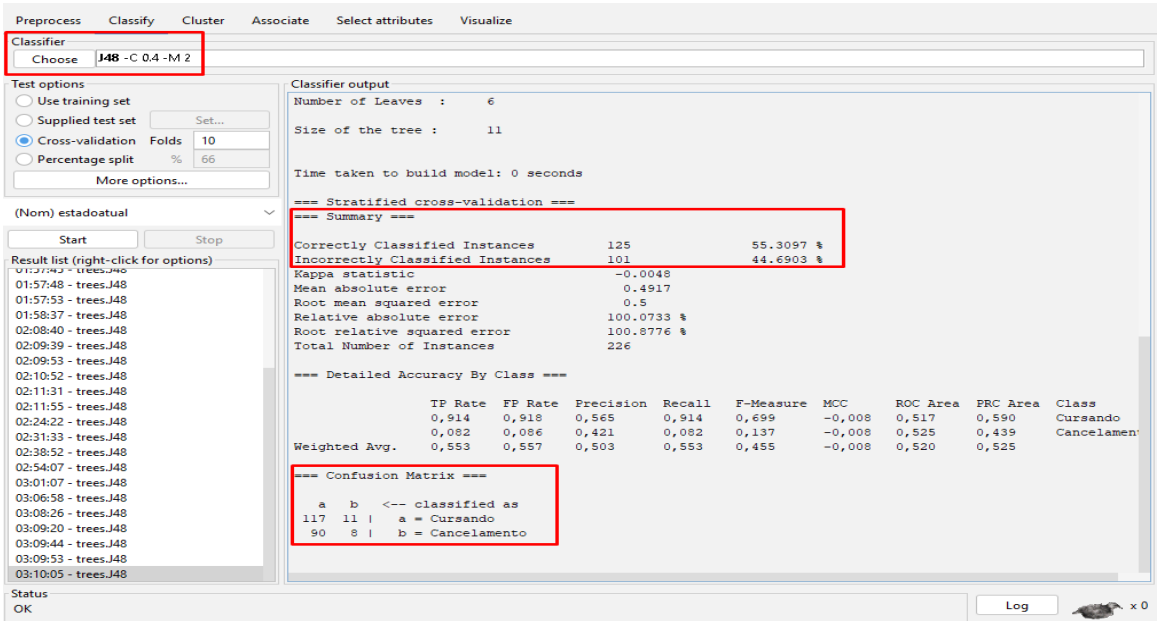
Classificação alunos do terceiro semestre, grade antiga na Figura 32 e grade nova na Figura 33.

Figura 32 - Saídas do Algoritmo J48 no experimento C no Weka (grande antiga)



Fonte: adaptado Weka (2022)

Figura 33 - Saídas do Algoritmo J48 no experimento C no Weka (grande nova)



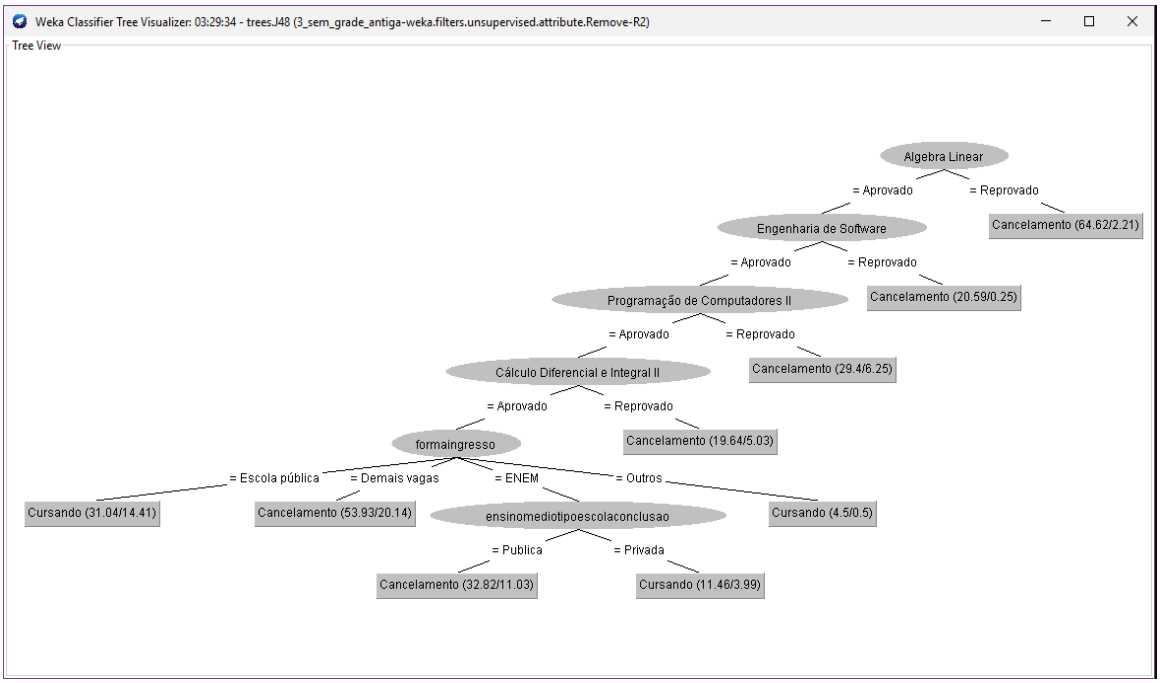
Fonte: adaptado Weka (2022)

Ao analisar os resultados do algoritmo de classificação J48 nos dados do terceiro semestre, pode-se observar que as classificações da grade antiga atingem 77% de precisão, enquanto há novamente uma redução na previsão da classificação na grade nova atingindo 55% de precisão.

Apenas para análise dos dados do terceiro semestre da grade antiga e nova, foi alterado um parâmetro do algoritmo para que fosse possível evidenciar na árvore de decisão as matérias referentes ao semestre em estudo. O parâmetro “minNumObj”, que é responsável por definir o número mínimo de observações permitidas por cada folha da árvore foi alterado de 3 para 2. Para os demais estudos não foi necessário realizar essa alteração.

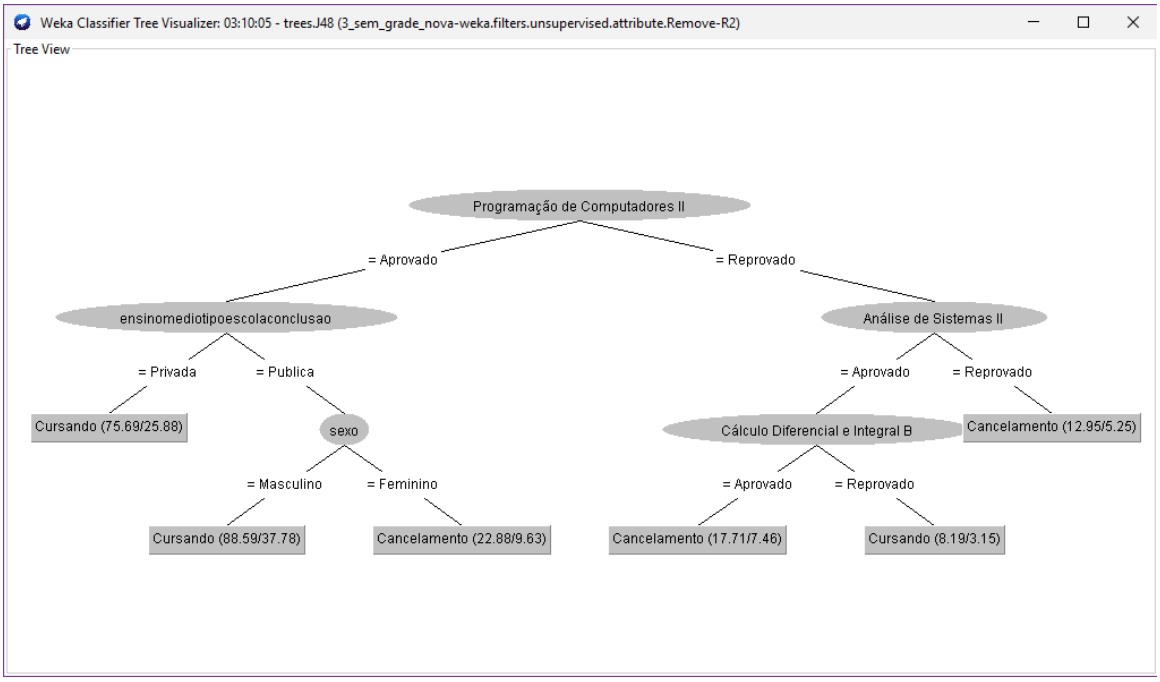
A seguir nas Figuras 34 e 35 serão apresentadas as árvores de decisão geradas pelo algoritmo na grade antiga e nova respectivamente.

Figura 34 - Árvore de decisão gerada no experimento C (grade antiga)



Fonte: adaptado Weka (2022)

Figura 35 - Árvore de decisão gerada no experimento C (grade nova)



Fonte: adaptado Weka (2022)

Já na árvore de decisão, pode-se observar a matéria Álgebra Linear como principal fator de cancelamento ou continuidade do aluno na graduação no terceiro semestre da grade antiga, seguida de matérias como Engenharia de *Software*, Programação de Computadores II e Cálculo Diferencial e Integral II. Enquanto na grade nova, a matéria Programação de Computadores II

se tornou a matéria mais relevante para essa análise, seguido de Análise de Sistemas II e Cálculo Diferencial e Integral B.

As seguintes regras mais relevantes puderam ser extraídas das árvores:

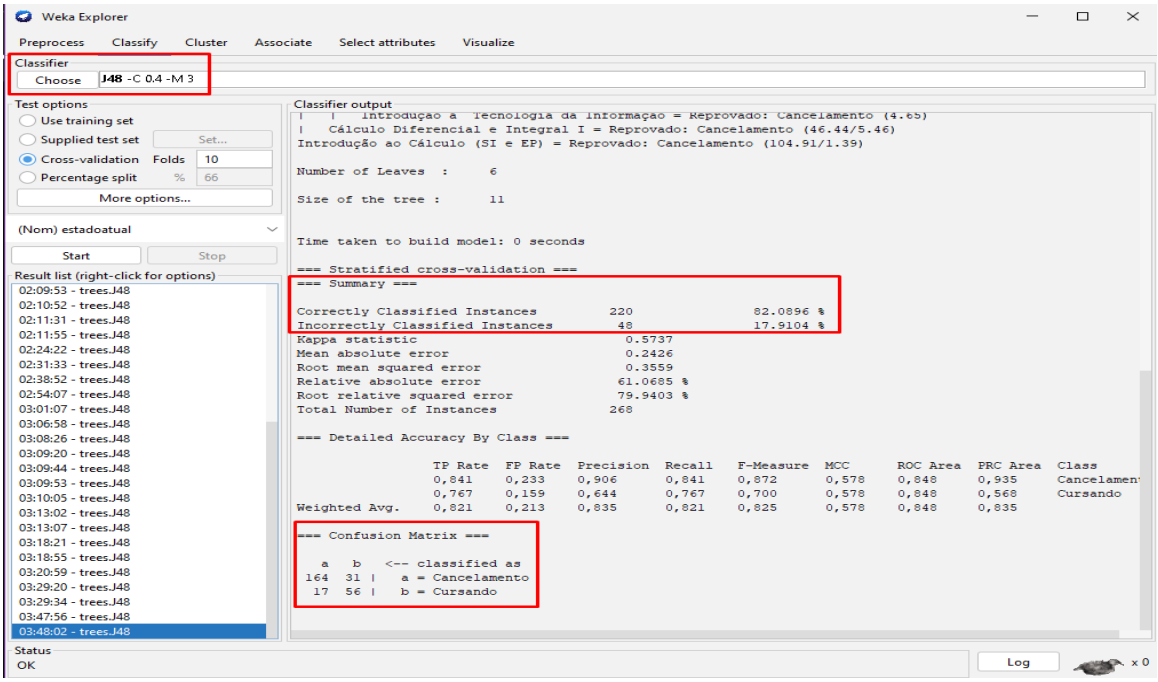
- Grade Antiga
    - SE Álgebra Linear = reprovado ENTÃO estadoatual = cancelamento
    - SE Álgebra Linear = aprovado E Engenharia de *Software* = reprovado ENTÃO estadoatual = cancelamento
    - SE Álgebra Linear = aprovado E Engenharia de *Software* = aprovado E Prog 2 = reprovado ENTÃO estadoatual = cancelamento
    - SE Álgebra Linear = aprovado E Engenharia de *Software* = aprovado E Prog 2 = aprovado E Cálculo 2 = aprovado E formaingresso = Escola Pública ENTÃO estadoatual = cursando
- Grade Nova
    - SE Prog 2 = aprovado E ensinomedio = privada ENTÃO estadoatual = cursando
    - SE Prog 2 = aprovado E ensinomedio = pública E sexo = masculino ENTÃO estadoatual = cursando
    - SE Prog 2 = aprovado E ensinomedio = pública E sexo = feminino ENTÃO estadoatual = cancelamento
    - SE Prog 2 = reprovado E Análise de Sistemas 2 = reprovado ENTÃO estadoatual = cancelamento

Nas árvores de decisão do experimento C pode-se observar novamente o aparecimento de outras variáveis como sexo e formaingresso aparecendo como determinantes no cancelamento ou continuidade do curso pelos alunos.

5.4 Resultados do experimento D

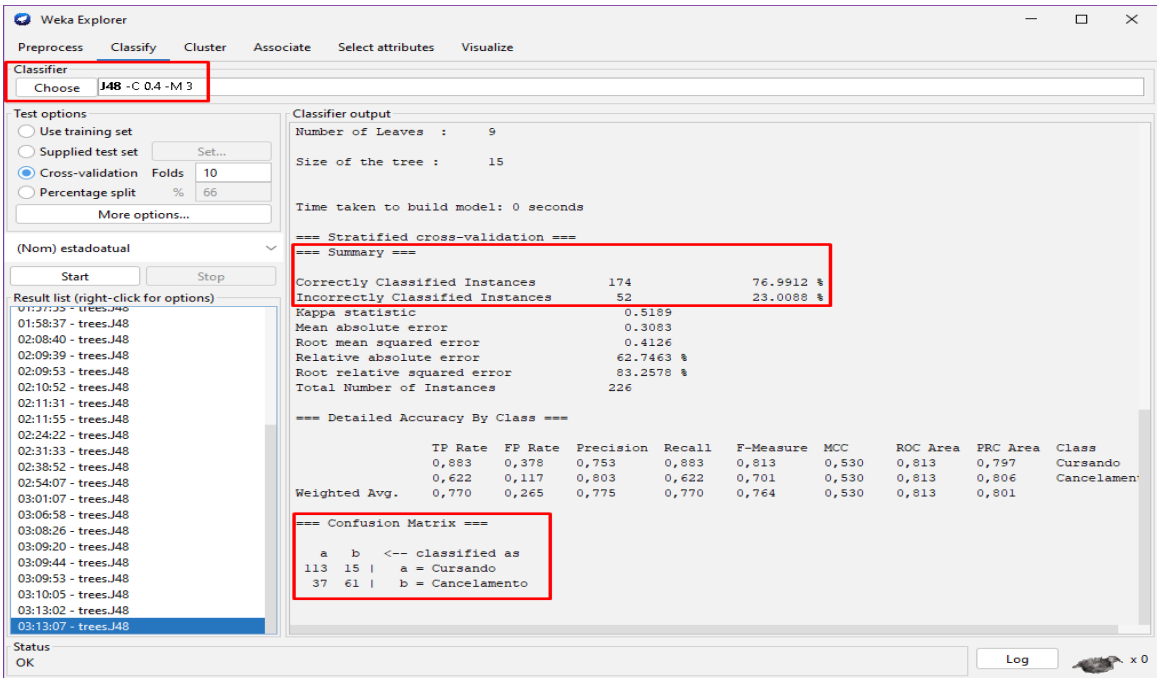
Classificação alunos do primeiro e segundo semestre, grade antiga na Figura 36 e grade nova na Figura 37.

Figura 36 - Saídas do Algoritmo J48 no experimento D no Weka (grade antiga)



Fonte: adaptado Weka (2022)

Figura 37 - Saídas do Algoritmo J48 no experimento D no Weka (grade nova)

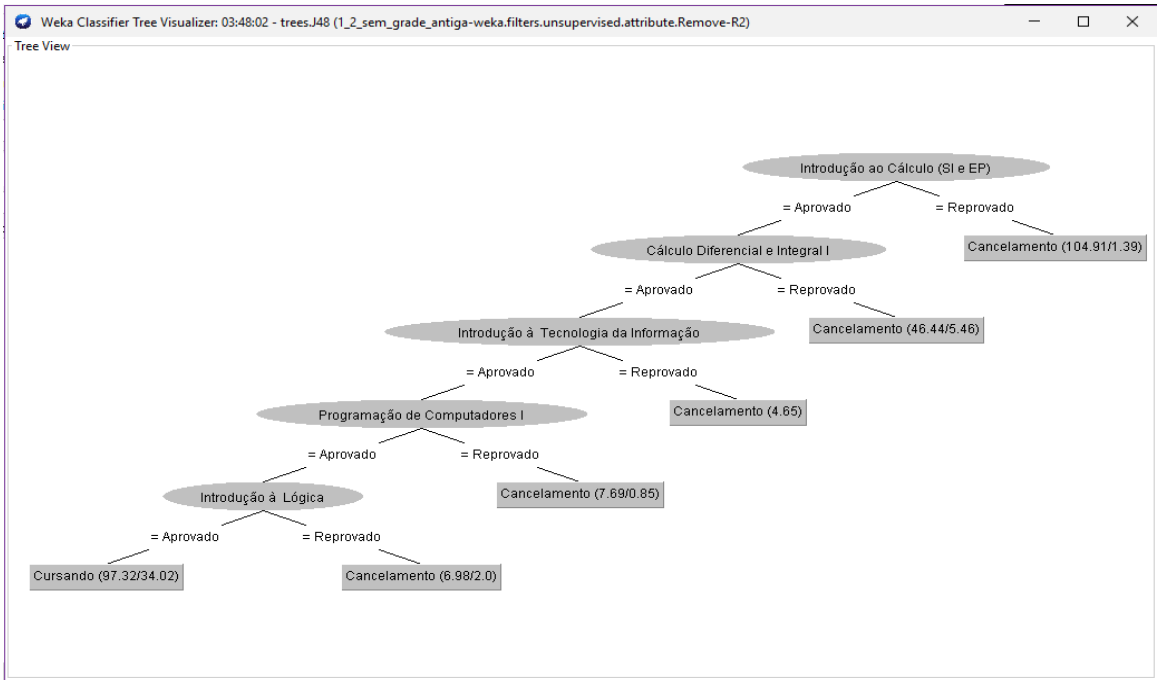


Fonte: adaptado Weka (2022)

Ao analisar os resultados do algoritmo de classificação J48 nos dados do primeiro e segundo semestre, pode-se observar que as classificações da grade antiga atingiram 82% de precisão, enquanto na previsão da classificação na grade nova atingiram 76% de precisão.

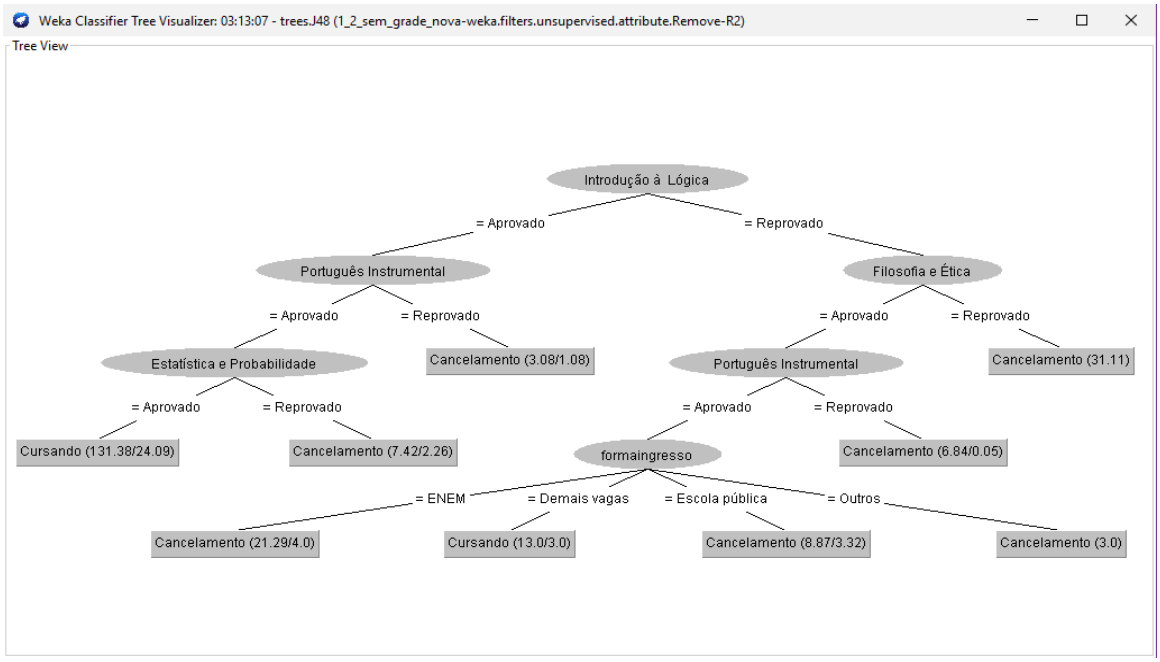
A seguir nas Figuras 38 e 39 serão apresentadas as árvores de decisão geradas pelo algoritmo na grade antiga e nova respectivamente.

Figura 38 - Árvore de decisão gerada no experimento D (grade antiga)



Fonte: adaptado Weka (2022)

Figura 39 - Árvore de decisão gerada no experimento D (grade nova)



Fonte: adaptado Weka (2022)

Já na árvore de decisão da grade antiga, pode-se observar que a matéria Introdução ao Cálculo continua como principal fator de cancelamento ou continuidade do aluno na graduação nos dados do primeiro e segundo semestre seguido de matérias como Cálculo Diferencial e Integral I, Introdução à Tecnologia da Informação, Programação de Computadores I e Introdução à lógica. Enquanto na grade nova, a matéria Introdução a Lógica aparece como a

matéria mais relevante para essa análise derivando-se para Filosofia e Ética e Português Instrumental ou Português Instrumental e Estatística e Probabilidade.

Com base nessa analise as seguintes regras mais relevantes foram extraídas das arvores de decisão:

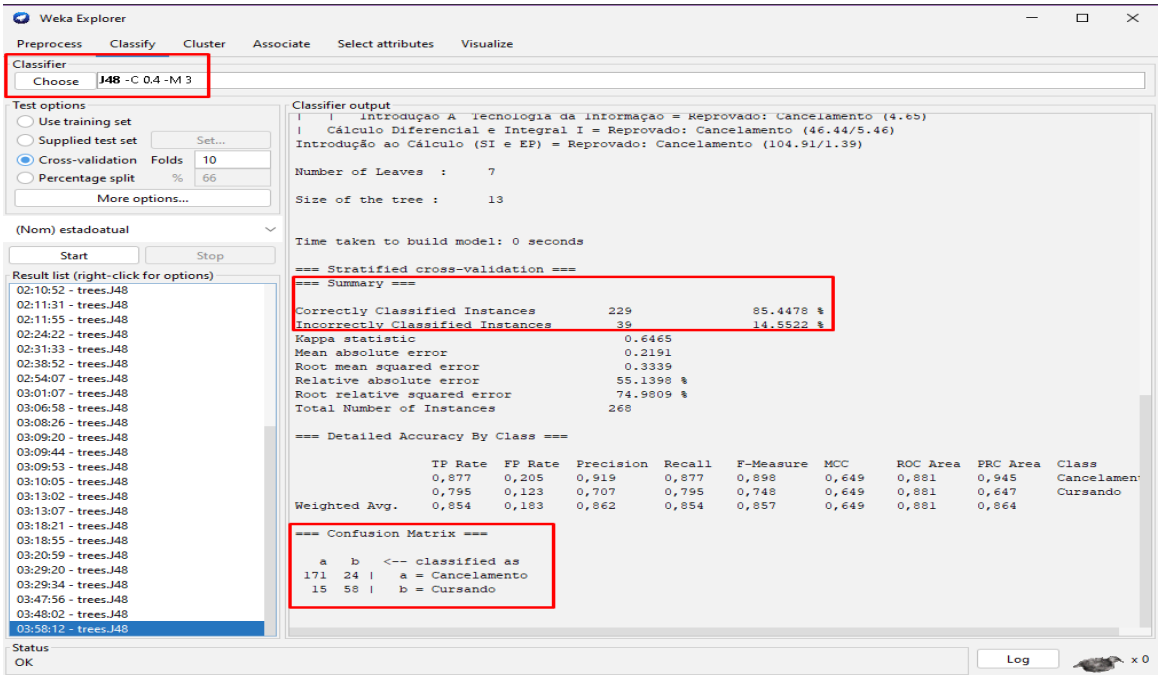
- Grade Antiga
  - SE Introdução a Cálculo = reprovado ENTÃO estadoatual = cancelamento
  - SE Introdução a Cálculo = aprovado E Cálculo 1 = reprovado ENTÃO estadoatual = cancelamento
  - SE Introdução a Cálculo = aprovado E Cálculo 1 = aprovado E Introdução a TI = aprovado E Prog 1 = aprovado E Introdução à Lógica = aprovado ENTÃO estadoatual = cursando
  - SE Introdução a Cálculo = aprovado E Cálculo 1 = aprovado E Introdução a TI = aprovado E Prog 1 = aprovado E Introdução à Lógica = reprovado ENTÃO estadoatual = cancelamento
- Grade Nova
  - SE Introdução à Lógica = reprovado E Filosofia e Ética = reprovado ENTÃO estadoatual = cancelamento
  - SE Introdução à Lógica = aprovado E Português Instrumental = aprovado E Estatística e Probabilidade = aprovado ENTÃO estadoatual = cursando
  - SE Introdução à Lógica = aprovado E Português Instrumental = aprovado E Estatística e Probabilidade = reprovado ENTÃO estadoatual = cancelamento

De acordo com as regras extraídas no experimento D, pode-se checar as principais matérias que influenciam na permanência ou não do estudante no acumulado do primeiro e segundo semestre.

5.5 Resultados do experimento E

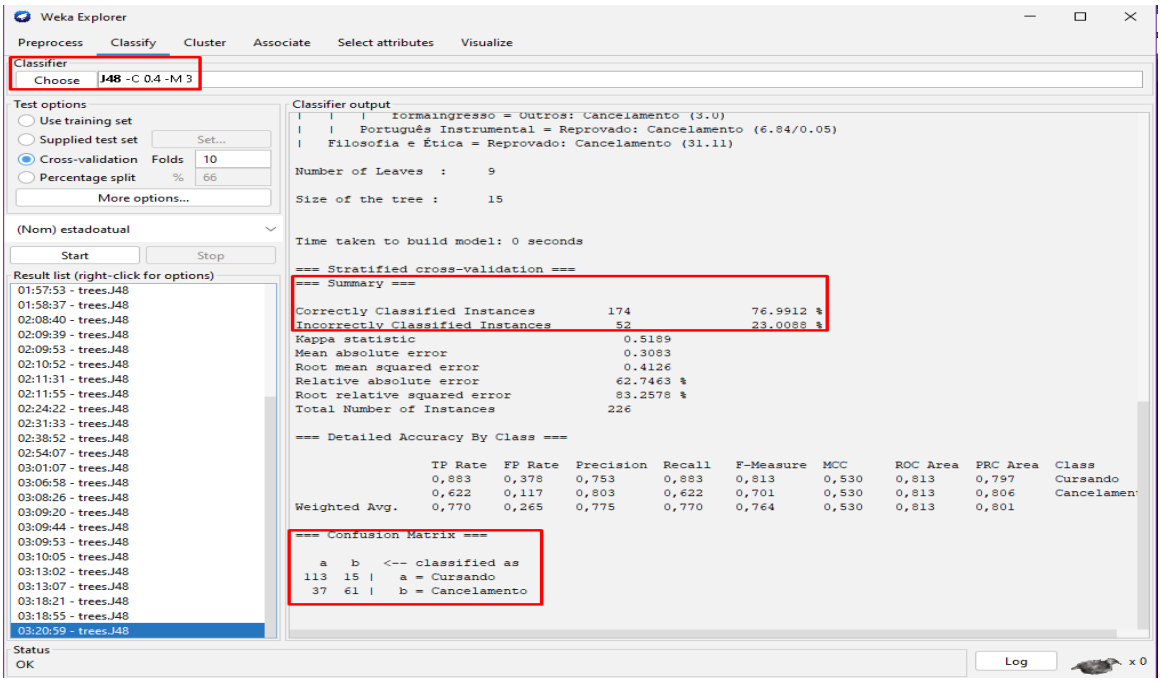
Classificação alunos do primeiro, segundo e terceiro semestre, grade antiga na Figura 40 e grade nova na Figura 41.

Figura 40 - Saídas do Algoritmo J48 no experimento E no Weka (grade antiga)



Fonte: adaptado Weka (2022)

Figura 41 - Saídas do Algoritmo J48 no experimento E no Weka (grade nova)

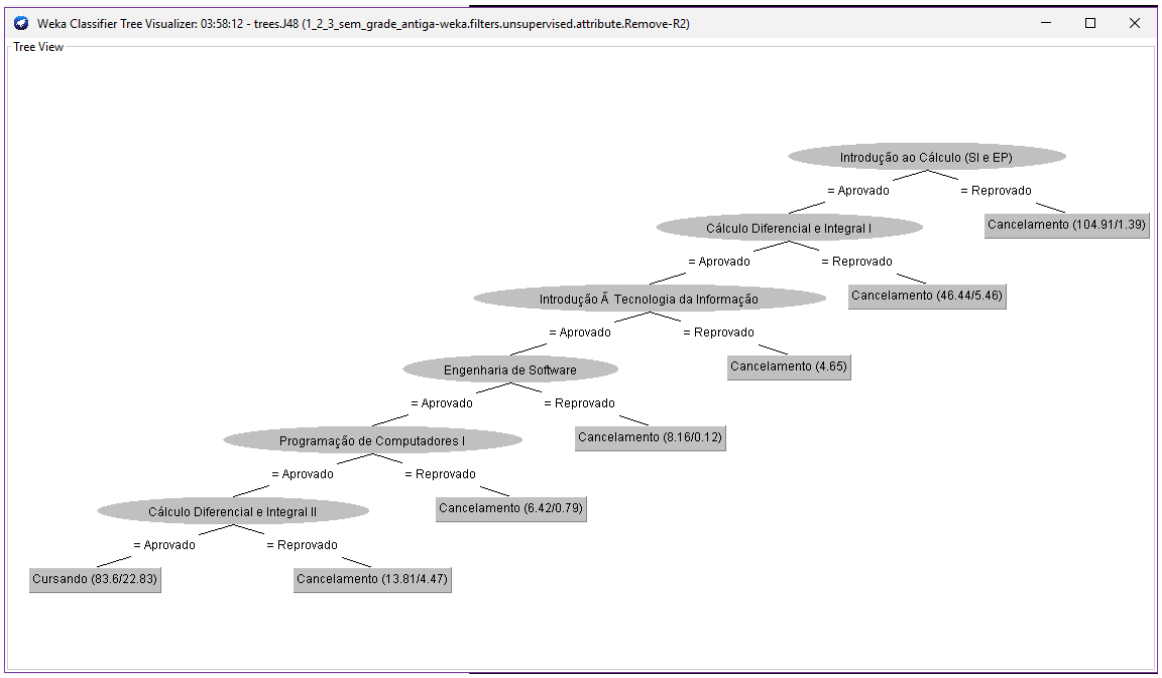


Fonte: adaptado Weka (2022)

Ao analisar os resultados do algoritmo de classificação J48 nos dados cumulativos do primeiro, segundo e terceiro semestre, pode-se observar que as classificações da grade antiga atingiram 85% de precisão, enquanto o cumulativo para a grade nova não apresentou diferença para o cumulativo apresentado no primeiro e segundo semestre. A precisão se manteve em 76%, como já visto anteriormente e, consequentemente, a árvore de decisão se manteve a mesma conforme mostrado adiante na Figura 43.

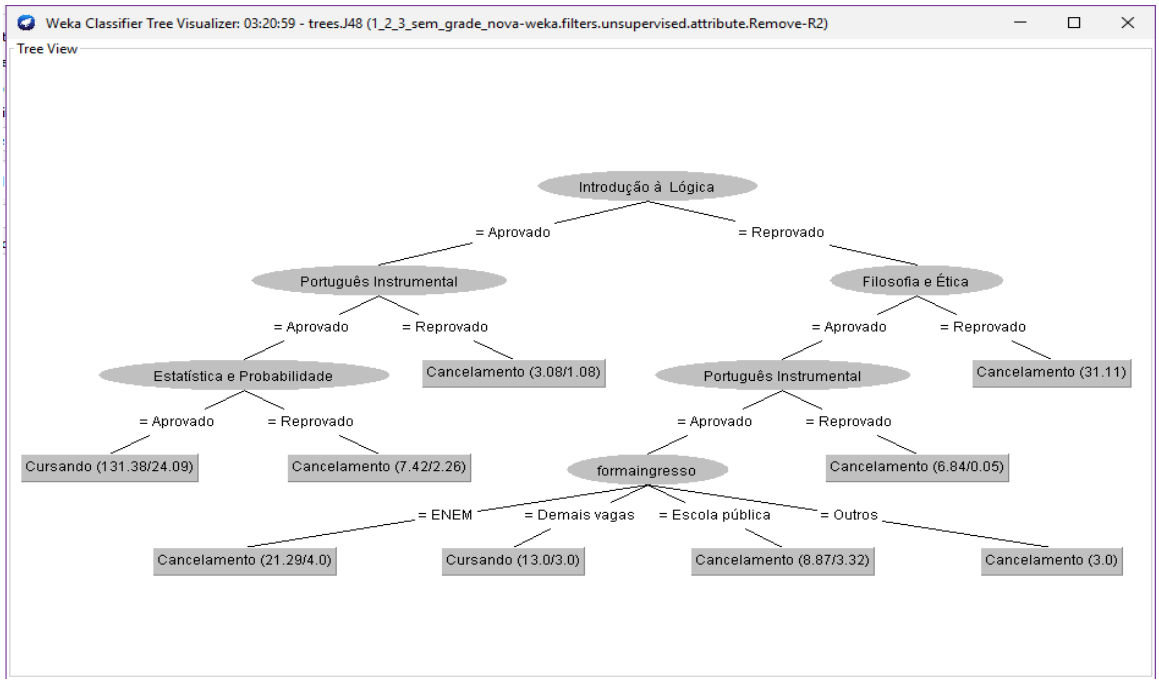
A seguir nas Figuras 42 e 43 serão apresentadas as árvores de decisão geradas pelo algoritmo na grade antiga e nova respectivamente.

Figura 42 - Árvore de decisão gerada no experimento E (grade antiga)



Fonte: adaptado Weka (2022)

Figura 43 - Árvore de decisão gerada no experimento E (grade nova)



Fonte: adaptado Weka (2022)

Já nas árvores de decisão, pode-se observar que a matéria Introdução ao Cálculo permanece como principal fator de cancelamento ou continuidade do aluno na graduação no cumulativo do primeiro, segundo e terceiro semestre da grade antiga seguido de matérias como Cálculo Diferencial e Integral I, Introdução a Tecnologia da Informação, Engenharia de

Software, Programação de Computadores I e Cálculo Diferencial e Integral II. Enquanto na grade nova a árvore de decisão se repete aos dados apresentados no experimento D.

Com base nessa analise as seguintes regras mais relevantes foram extraídas das árvores de decisão:

- Grade Antiga
  - SE Introdução a Cálculo = reprovado ENTÃO estadoatual = cancelamento
  - SE Introdução a Cálculo = aprovado E Cálculo 1 = reprovado ENTÃO estadoatual = cancelamento
  - SE Introdução a Cálculo = aprovado E Cálculo 1 = aprovado E Introdução a TI = aprovado E Engenharia de *Software* = aprovado E Prog 1 = aprovado E Cálculo 2 = aprovado ENTÃO estadoatual = cursando
  - SE Introdução a Cálculo = aprovado E Cálculo 1 = aprovado E Introdução a TI = aprovado E Engenharia de *Software* = aprovado E Prog 1 = aprovado E Cálculo 2 = reprovado ENTÃO estadoatual = cancelamento
- Grade Nova (mesmas regras do experimento D)
  - SE Introdução à Lógica = reprovado E Filosofia e Ética = reprovado ENTÃO estadoatual = cancelamento
  - SE Introdução à Lógica = aprovado E Português Instrumental = aprovado E Estatística e Probabilidade = aprovado ENTÃO estadoatual = cursando
  - SE Introdução à Lógica = aprovado E Português Instrumental = aprovado E Estatística e Probabilidade = reprovado ENTÃO estadoatual = cancelamento

No último experimento, o autor pode observar que algumas matérias relevantes no experimento anterior se mantiveram relevantes no acumulativo com o terceiro semestre da grade antiga. Já o cumulativo da grade nova não apresentou nenhum dado novo relevante e a árvore de decisão gerada foi idêntica a árvore do experimento D. Desta forma não houve novas regras sendo descoberta.

6 CONSIDERAÇÕES FINAIS

A utilização de técnica de classificação em *Machine Learning* mostrou-se útil para a descoberta de conhecimentos ocultos na base de dados da instituição analisada. Apesar da evasão ainda ter outros fatores importantes a serem mapeados, o objetivo geral desta pesquisa foi cumprido no que tange a análise da evasão através de um algoritmo de classificação.

O *software* utilizado na pesquisa se mostrou de fácil uso e fundamental para que a pesquisa fosse realizada dentro do prazo. As ferramentas de visualização gráficas foram de grande valia para uma melhor compreensão dos resultados, assim contribuindo para o cumprimento dos objetivos específicos deste trabalho.

Assim foi possível visualizar quais variáveis tiveram um maior peso na continuação ou interrompimento do aluno com o curso de Sistemas de Informação. Também pode-se avaliar que outros fatores, além das matérias cursadas, podem ser influenciadores nas estatísticas de evasão.

Com a aplicação de técnicas de *Machine Learning* e a obtenção dos dados de forma a produzir conhecimento, ficaram evidentes os benefícios oriundos do uso de técnicas como essa para análise e desenvolvimento de modelos, que auxiliariam no processo de decisão da gestão institucional.

Desta forma, estudos como este podem ser de grande valia à gestão da IES para agir previamente não só no curso de Sistemas de Informação, mas aplicar as mesmas técnicas para os outros cursos, a fim de identificar previamente o perfil de evasão dos alunos e trabalhar em ações que possam reduzir esses números.

Dentre as dificuldades enfrentadas pelo autor no desenvolvimento deste trabalho, destacam-se a divisão entre a grade curricular nova e a grade curricular antiga e ausência de valores nas variáveis de disciplina.

Pelas grades curriculares conterem variáveis diferentes para o mesmo período analisado, foi necessário realizar a separação da base de dados para que não houvesse inconsistências nos dados analisados.

Já para os valores ausentes nas matérias, não foi possível com a base de dados extraída chegar à conclusão se o aluno apenas não cursou aquela matéria em nenhum momento do curso. Desta forma, por mais que o algoritmo J48 desconsidere os valores nulos, a base fica desbalanceada, afetando assim a precisão do algoritmo em classificar os dados.

Estudos futuros poderão contemplar outros algoritmos de ML, que possam ser comparados e talvez indicar qual seria mais efetivo para o estudo de evasão. Também

agregariam outros conhecimentos ser possível identificar os alunos evadidos e aplicar um questionário para mapear os motivos que levaram os alunos a evadir, assim cobrindo variáveis que talvez não estejam contempladas na base de dados da faculdade.

REFERÊNCIAS

BAKER, R.; ISOTANI, S.; CARVALHO, A. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, v. 19, n. 02, 2011. Sociedade Brasileira de Computação - SB.

BALDASSO, Rafael Oliveira. **Aplicação de algoritmo de *machine learning* na identificação de alunos em risco de evasão**. 2019. TCC (Graduação) - Curso de Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

BARBOSA, Denise Chaves Carvalho; MACHADO, Maria Augusta. **Mineração de Dados usando o *software* WizRule em Base de Dados de Compras de TI**. Revista Eletrônica de Sistemas de Informação, v. 6, n. 1, jun. 2007.

CASTANHEIRA, Luciana Gomes. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. 2008. Disponível em: <https://www.ppgee.ufmg.br/defesas/349M.PDF>. Acesso em: 17 abr. 2019.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. *From **Data Mining to Knowledge Discovery in Databases***. AI Magazine, v. 17, n. 3, 2021. Disponível em: <https://ojs.aaai.org//index.php/aimagazine/article/view/1230>. Acesso em: 15 ago. 2022.

FERNANDES, Warley Leite. **Aplicação do algoritmo de classificação associativa (CBA) em bases educacionais para predição de desempenho**. Ufvjm.edu.br, 2017. Disponível em: <http://acervo.ufvjm.edu.br/jspui/handle/1/1726>. Acesso em: 20 nov. 2022.

FILHO, R. L. L. S. A. **Evasão No Ensino Superior Brasileiro – Novos Dados**. 2017. Disponível em: <http://www.institutolobo.org.br/imagens/pdf/artigos/art_088.pdf>. Acesso em: 17/mar./19.

GOLDSCHMIDT, R. **Uma Introdução à Inteligência Computacional: Fundamentos, Ferramentas e Aplicações**, Rio de Janeiro, 2010. Disponível em: http://www.boente.eti.br/fuzzy/ebook/ebook-fuzzy-goldschmidt.pdf.

GOLDSCHMIDT, R.; PASSOS, E. ***Data Mining*: Um guia prático**, Rio de Janeiro, 2005.

GRANDO, Nei. **A Essência do Aprendizado de Máquina**. Blog do Nei. Disponível em: <https://neigrando.com/2022/05/04/a-essencia-do-aprendizado-de-maquina/>. Acesso em: 22 nov. 2022.

HABOWSKI, Adilson Cristiano; BRANCO, Lílian Soares Alves; CONTE, Elaine. **Evasão na EAD: perspectivas de prevenção.** Perspectiva, v. 38, n. 3, 2020. Disponível em: <https://periodicos.ufsc.br/index.php/perspectiva/article/view/62978>. Acesso em: 30 nov. 2022.

HAN, Jiawei W.; KAMBER, Micheline; PEI, Jian. ***Data Mining: Concepts and techniques Third Edition***. 3 ed. 2011

HOED, Raphael Magalhães. **Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação**. Dissertação (Mestrado) - Curso de Computação Aplicada, Ciência da Computação, Universidade de Brasília, Brasília, 2016.

HOFFMANN, Ivan Londero. **Metodologia para identificação de fatores estratégicos para acompanhamento sistemático da evasão em cursos de graduação**. Ufsm.br, 2016. Disponível em: <https://repositorio.ufsm.br/handle/1/8385>. Acesso em: 07 abr. 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, vol 1. N°103, 2013.

LOBO, M. B. C. M. **Panorama da evasão no ensino superior brasileiro: Aspectos gerais das causas e soluções**. 2012. Disponível em: <http://www.institutolobo.org.br/imagens/pdf/artigos/art_087.pdf>. Acesso em: 09/abr./19.

MANHÃES, L. M. B. et al. **Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados**. *Brazilian Symposium on Computers in Education* (Simpósio Brasileiro de Informática na Educação - SBIE), v. 1, n. 1, 2011.

MINISTÉRIO DA EDUCAÇÃO. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. 1997. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=27010>. Acesso em: 10/abr./19

OLIVEIRA, Carlos Henrique Mendes de; SANTOS, Francisco Raul Teixeira; LEITINHO, Janaina Lopes; et al. **Busca dos fatores associados à evasão**. Revista Internacional de Educação Superior, v. 5, p. e019006, 2019. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/riesup/article/view/8652897>. Acesso em: 2 out. 2022.

Prefeitura Municipal de Macaé. **FeMASS forma 44 alunos para o mercado de trabalho**. Disponível em: <http://www.macae.rj.gov.br/semed/leitura/noticia/femass-forma-44-alunos-para-o-mercado-de-trabalho>. Acesso em: 09/abr./19.

Prefeitura Municipal de Macaé. **Faculdade Professor Miguel Ângelo da Silva Santos: Apresentação**. Disponível em: <https://macae.rj.gov.br/femass/conteudo/titulo/apresentacao>. Acesso em: 19/nov./22.

PRIMÃO, Aline Pacheco. **Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior: um estudo no instituto federal de Santa Catarina**. Dissertação (Mestrado) - Curso de Administração Universitária, Universidade Federal de Santa Catarina, Florianópolis, 2022.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico** – 2. ed. – Novo Hamburgo: Feevale, 2013.

QUINLAN, J. R. ***Discovering rules by induction from large collection of examples***. *Expert Systems in the Micro Electronic Age*. Edinburgh, UK: Edinburgh University Press, 1979.

QUINLAN, J. R. **C4.5: *Programs for Machine Learning***. São Francisco: Morgan Kaufmann, 1993.

QUINLAN, J. R. ***Improved use of continuous attributes in* c4.5.** *Journal of Artificial Intelligence Research*, 4:77-90, 1996.

RIGO, Sandro; BARBOSA, Jorge; CAMBRUZZI, Wagner. **Educação em Engenharia e Mineração de Dados Educacionais: oportunidades para o tratamento da evasão**. Revista: EaD & Tecnologias Digitais na Educação, Dourados, MS, n. 3, v. 1, 2014.

SILVA, Glauco Carlos**. Mineração de regras de associação aplicada a dados da Secretaria Municipal de Saúde de Londrina PR**. Dissertação (Mestrado) - Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

SMOLA, Alex; S. VISHWANATHAN. ***Introduction to Machine Learning***. undefined. Disponível em: <https://alex.smola.org/drafts/thebook.pdf>. Acesso em: 28 nov. 2022.

SOUZA, J. F. de; NÓBREGA, A. C. S.; AMORIM, B. M. O. de. **Evasão escolar e psicologia educacional: Reflexões sobre a realidade brasileira**. 2017. Disponível em: <https://www.editorarealize.com.br/revistas/conedu/trabalhos/TRABALHO_EV073_MD1_SA4_ID2111_13102017131956.pdf>. Acesso em: 17 abr. 2019.

VARGAS, Hustana; HERINGER, Rosana. **Políticas de Permanência no Ensino Superior Público em Perspectiva Comparada: Argentina, Brasil e Chile**. *Education Policy Analysis Archives/*Archivos Analíticos de Políticas Educativas, v. 25, 2017. Disponível em: <https://www.redalyc.org/articulo.oa?id=275050047114>. Acesso em: 10 jun. 2019.

WITTEN, Ian H. *et al*. ***Data Mining: Practical Machine Learning Tools and Techniques***. 4 ed. 2017.