UNIVERSIDADE FEDERAL DE ALAGOAS INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ ESTEVAM VILAR BORGES

# UM ESTUDO EXPLORATÓRIO SOBRE OS ASPECTOS QUE INFLUENCIAM NA EFICIÊNCIA DA PREDIÇÃO DE SUCESSO ACADÊMICO DE ALUNOS DE PROGRAMAÇÃO INTRODUTÓRIA

MACEIÓ-ALAGOAS
2019

JOSÉ ESTEVAM VILAR BORGES

# UM ESTUDO EXPLORATÓRIO SOBRE OS ASPECTOS QUE INFLUENCIAM NA EFICIÊNCIA DA PREDIÇÃO DE SUCESSO ACADÊMICO DE ALUNOS DE PROGRAMAÇÃO INTRODUTÓRIA

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Evandro de Barros Costa

MACEIÓ-ALAGOAS
2019

**Catalogação na fonte**
**Universidade Federal de Alagoas**
**Biblioteca Central**
**Divisão de Tratamento Técnico**

Bibliotecária Responsável: Helena Cristina Pimentel do Vale – CRB4 - 661

B732e Borges, José Estevam Vilar.

Um estudo exploratório sobre os aspectos que influenciam na eficiência da predição de sucesso acadêmico de alunos de programação introdutória / José Estevam Vilar Borges. – 2019.

101f. : il.

Orientador: Evandro de Barros Costa.

Dissertação (mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2019.

Bibliografia: f. 91-93.

Apêndice: f. 94-101.

1. Mineração de dados educacionais. 2. Mineração de dados (Computação). 3. Aprendizagem por computador. 4. Seleção de atributos. 5. Algoritmos de predição. Título.

CDU: 004.5: 37.018.43

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
**Programa de Pós-Graduação em Informática – PpgI**
**Instituto de Computação**
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de José Estevam Vilar Borges intitulada: *"Um Estudo Exploratório sobre Os Aspectos que Influenciam na Eficiência da Predição de Sucesso Acadêmico de Alunos de Programação Introdutória"*, apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 07 de maio de 2019, às 16h00, na sala 15, do Instituto de Computação, da Universidade Federal de Alagoas.

**COMISSÃO JULGADORA**

**Prof. Dr. Evandro de Barros Costa**
UFAL – Instituto de Computação
Orientador

**Prof. Dr. Baldoino Fonseca dos Santos Neto** UFAL – Instituto de Computação
Examinador Interno

**Prof. Dr. Rafael Ferreira Leite de Mello**
Universidade Federal Rural de Pernambuco
Examinador Externo

*“Aprendi que deveríamos ser gratos a Deus por não nos dar tudo que lhe pedimos”*
*(Willam Shakespeare)*

## AGRADECIMENTOS

Agradeço em primeiro lugar aos meus queridos pais, Egídio José Ramos Borges (in memorian) e Maria Augusta Vilar Borges, que me proporcionaram uma excelente educação de base, mesmo diante de todas as dificuldades, sempre com muito amor, e ensinando os valores mais importantes para a vida.

Agradeço à minha amada esposa, Lisiane Maria Santos Araújo, que sempre acreditou em mim e me apoiou nos momentos mais difíceis.

Agradeço à minha linda filha, Helena Maria Araújo Borges, que só por estar presente na minha vida me deu as forças necessárias para vencer os maiores obstáculos.

Agradeço aos meus estimados professores orientadores, Evandro de Barros Costa e Baldoino F. dos Santos Neto, que me mostraram os caminhos dentro da carreira acadêmica.

Agradeço, por fim, à minha família e amigos, que contribuíram indiretamente para o meu sucesso nesta importante etapa de minha vida.

## RESUMO

Existe uma carência de ferramentas inteligentes para auxiliar os professores na análise de grande massa de dados gerada sobre os alunos de um determinado curso, permitindo mensurar o nível de aprendizado dos alunos ou ainda auxiliar na identificação daqueles que não estão conseguindo obter um bom rendimento no aprendizado. Para amenizar esse tipo de lacuna, atualmente existem diversos estudos na área de Mineração de Dados e Aprendizagem de Máquina, abordando o processo de obtenção de conhecimento para ajudar na tomada de decisões. Assim essa área, tem sido explorada para apoiar o ensino e pesquisa, permitindo a extração de informação relevante, dentre os dados disponíveis, para suportar a tomada de decisões pelos educadores, principalmente, e também gestores. Este trabalho busca explorar os fatores, dentro do processo de predição, que possuem influência na eficiência das técnicas já conhecidas de predição de sucesso dentro do âmbito educacional. Para isso, foram aplicadas em um banco de dados real quatro abordagens distintas, porém correlacionadas, para mensurar a eficiência de cada técnica. Tais abordagens consideram características particulares de cada aluno, além de explorar também a eficiência de cada técnica relacionadas à parcela amostral de treino e teste, características específicas dos algoritmos de árvore de decisão e aspectos temporais da predição.



**Palavras-chave:** Mineração de dados educacionais, aprendizagem de máquina, seleção de atributos, algoritmos de predição.

## ABSTRACT

There is a lack of intelligent tools to assist teachers in the analysis of the large amount of data generated on the students of a particular course, allowing to measure the level of learning of the students or to assist in the identification of those who are not achieving a good performance in learning . To mitigate this type of gap, there are currently several studies in the area of Data Mining and Machine Learning, addressing the process of obtaining knowledge to assist in decision making. Thus, this area has been exploited to support teaching and research, allowing the extraction of relevant information, among the available data, to support decision-making by educators, mainly as well as managers. This work seeks to explore the factors, within the prediction process, that influence the efficiency of the already known techniques of prediction of success within the educational scope. For this purpose, four different but correlated approaches were applied to a real database to measure the efficiency of each technique. These approaches take into account the particular characteristics of each student, as well as to explore the efficiency of each technique related to the training sample and test, specific characteristics of the decision tree algorithms and temporal aspects of the prediction.



**Keywords:** Educational data mining, machine learning, attribute selection, prediction algorithms.

## LISTA DE ILUSTRAÇÃO

## LISTA DE QUADROS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

É sempre um grande desafio para os professores, de todos os ramos da educação, avaliarem a eficácia de sua metodologia didática. Ultimamente, tem havido avanço no uso de ambientes online no apoio à execução dos cursos, tanto a distância, quanto presenciais. O uso de tais ambientes tem permitido a geração de grandes massas de dados sobre os alunos, possibilitando análises desses dados a fim de se obter informações valiosas no apoio ao processo de tomada de decisão MÁRQUEZ; Morales e Soto (2013). No entanto, existe uma carência significativa por ferramentas automatizadas que deem suporte na previsão do insucesso de alunos, análise de características relevantes para a aprovação, ou ainda para uma abordagem preditiva, atuando nos principais fatores que venham a promover o sucesso do aluno, ou seja, fatores que se prestem a indicar a deficiência do aluno e, assim, serem úteis na escolha ou desenvolvimento de mecanismos para contribuir na atividade de remediação. Com os avanços dos estudos da Inteligência Artificial (IA), mais especificamente a Mineração de Dados Educacionais (*Educational Data Mining* - EDM), tais ferramentas se tornam possíveis de serem desenvolvidas, o que já vem ocorrendo. Dentro da área de estudo de EDM, também existem diversos trabalhos abordando técnicas que buscam previsões de desempenhos acadêmicos de estudantes, observando-se um grande esforço por parte dos pesquisadores para se alcançar modelos eficientes em termos de tempo e capazes de gerar informações importantes a respeito dos alunos de forma antecipada (MÁRQUEZ et al., 2016).

Este trabalho busca realizar uma abordagem exploratória de possíveis fatores que possam contribuir significativamente para o processo de mineração de dados como um todo, além de dar subsídios para a criação de modelos preditivos pretensamente mais eficientes, considerando-se um comparativo entre as métricas de eficiência de cada técnica, indiretamente de cada algoritmo de classificação.

## 1.1 Problemática

Dado o contexto de predição mencionado anteriormente, a busca por eficiência dos modelos se constitui em um problema de alta relevância. Assim, o presente trabalho caracteriza e aborda esse aspecto de eficiência nos algoritmos preditivos sobre fontes de dados reais, tratando-os de forma comparativa e exploratória. Um diferencial deste trabalho é o direcionamento pormenorizado para o aspecto “Eficiência”, termo amplamente conhecido e utilizado nas mais diversas áreas de estudo, sendo definido por Chiavenato (2013) como “o uso correto dos recursos disponíveis” quando aplicado à administração. Chiavenato ainda sugere que a eficiência pode ser definida pela equação E = P / R, onde E representa a medida

quantitativa da eficiência, P é o produto resultante e R os recursos utilizados. Alguns outros autores, como também Koontz e Weihrich (1992) definiram eficiência como a realização de metas com a menor quantidade de recursos, o que remete à ideia de fazer mais com menos, ou ainda melhores resultados com a utilização de menos recursos. Poucos autores da área de EDM exploram a eficiência de forma mais específica, explícita e detalhada para a mineração de dados. Um estudo que pode ser citado envolvendo eficiência e voltado para EDM (KUMAR, 2011). Por outro lado, diversos trabalhos exploram eficiência de técnicas de predição em outras áreas como por exemplo detecção de *code smells* (HOZANO, 2017).

Quando se deseja mensurar a eficiência voltada a técnicas de predição, podemos entender como objetivo, baseando-se nos conceitos mais amplos de eficiência, explorar o quanto tais técnicas conseguem alcançar em nível de assertividade na predição, quando expostas a um número reduzido de recursos, que neste caso podemos entender como recursos: tamanho da amostra disponível para treino, ou limitações em relação ao nível profundidade da árvore de decisão, ou ainda o recurso tempo, este último bastante relevante em estudos onde o foco é conseguir melhores resultados com menos tempo. Sendo assim, a motivação para a presente pesquisa exploratória, inclusive para as questões propostas a seguir, teve como base tanto os trabalhos anteriores, que deixaram lacunas a serem exploradas, quanto os aspectos relativos à eficiência amostral, eficiência de complexidade para o caso de árvores de decisão, e ainda os aspectos temporais, idealizados a partir do conceito de eficiência. Logo, a fim de explorar adequadamente tais aspectos de eficiência nas técnicas de predição em EDM, formulou-se e procurou-se responder às seguintes questões de pesquisa:

**Questão 1.** Quais as características mais relevantes dos alunos para a predição do seu sucesso acadêmico, buscando-se eficiência voltada para a qualidade dos dados?

**Questão 2.** Quão eficientes são os algoritmos de classificação aplicados na predição do sucesso de alunos de programação introdutória no momento de início do curso, considerando-se os percentuais de instâncias utilizadas para treino e teste?

**Questão 3.** Quão eficientes são os algoritmos de classificação por árvores de decisão, no que se refere ao nível de profundidade das árvores?

**Questão 4.** Em que tempo, a partir da aplicação de atividades de aprendizagem durante o curso, é possível obter níveis de desempenho satisfatórios na predição do sucesso?

Dada a relevância significativa da Questão 4, propõe-se uma subdivisão desta em 5 questões derivadas, no intuito de enfatizar melhor sua importância para os objetivos da pesquisa, assim como permitir uma análise mais profunda deste problema especificamente. Logo, pode-se descrever as questões derivadas como:

*Questão 4.1)* Em que tempo, a partir da aplicação de atividades de aprendizagem durante o curso, é possível obter níveis de desempenho satisfatórios na predição do sucesso?
*Questão 4.2)* Quais os Algoritmos recomendados para Predição em T0?
*Questão 4.3)* Quais os Algoritmos recomendados para Predição durante o curso?
*Questão 4.4)* Quais os Dados recomendados a serem utilizados em predições em T0?
*Questão 4.5)* Quais os Dados recomendados a serem utilizados em predições durante o curso?

Vale ressaltar que a questão identificada como 4.1 representa a principal, ou seja, a questão primária relativa à análise temporal proposta, das quais derivam a 4.2, 4.3, 4.4 e 4.5.

## 1.2 Objetivos

O Objetivo principal deste trabalho é o de realizar um estudo exploratório sobre uma coleção de técnicas preditivas em EDM aplicada a uma base de dados real, analisando e buscando melhorar aspectos de eficiência em tais técnicas. Pretende-se, deste modo, conhecer os comportamentos e tendências de cada técnica e compará-las.

Para abordar este objetivo geral, estabeleceu-se os seguintes objetivos específicos de pesquisa:

a. Analisar quais atributos do conjunto de dados original possuem maior influência para o sucesso do aluno na disciplina, e, a partir desta informação obtida, deseja-se utilizar os atributos mais relevantes para alcançar um melhor desempenho dos algoritmos de Classificação, fomentando a possibilidade de criação de modelos mais assertivos de predição.
b. Explorar a eficiência amostral dos algoritmos de classificação, no que diz respeito às parcelas de instâncias utilizadas no treino e no teste, buscando assim um estudo comparativo dos níveis alcançados de acurácia na predição utilizando-se menos e mais dados para treino e/ou teste.
c. Apresentar uma análise temporal dos níveis de acurácia alcançados na predição de sucesso a partir da inclusão de dados sobre os alunos obtidos durante o curso, como notas das atividades de aprendizagem realizadas por eles.
d. Analisar a eficiência dos algoritmos de classificação baseados em árvores de decisão especificamente, comparando 5 algoritmos conhecidos, *ADTree, J48, RandomTree, REPTree, SimpleCart*, e o desempenho de cada um quando forçados a modelos com árvores menores ou maiores, em outras palavras descrever a acurácia de cada algoritmo ao mesmo tempo em que são utilizadas podas (*prunning*) nas árvores apresentadas como modelo por cada algoritmo.

e. Buscar o momento, no decorrer do curso, a partir da análise de notas das atividades realizadas, onde é possível uma predição a níveis suficientemente aceitáveis de desempenho.

### 1.3 Relevância

A principal motivação deste estudo é a atual necessidade de ferramentas eficientes de predição de insucesso no ambiente educacional. A partir desta análise comparativa e exploratória pesquisadores da área de EDM ou aprendizagem de máquinas terão a possibilidade de conhecer e aplicar a metodologia aqui apresentada para aperfeiçoamento de modelos de predição. A aplicação e análise dos resultados das técnicas no conjunto de dados reais coletados na Universidade Federal de Alagoas, e o comparativo dos resultados com os de outros trabalhos já realizados na área, trarão uma nova visão das possibilidades de ferramentas de predição para educadores, assim como contribuirão para pesquisas futuras em EDM. Estudos recentes na área mostram diversas áreas ainda não exploradas e que podem trazer avanços para as ferramentas de apoio a professores e sua aplicação nas metodologias de ensino. O Foco no aspecto eficiência das técnicas pode ser visto como um diferencial metodológico explorado neste trabalho.

### 1.4 Estrutura

Este trabalho está dividido em 6 capítulos. Após a Introdução, temos no capítulo 2 a apresentação dos trabalhos relacionados da área, com aqueles mais específicos de EDM e os voltados para aprendizagem de máquina. Em seguida, no capítulo 3, são apresentados os conceitos teóricos que estão relacionados com o presente estudo como um todo, a descrição dos algoritmos de classificação bem como os de seleção de atributos, além de descrever também as principais métricas utilizadas em EDM. O capítulo 4 é voltado completamente para a descrição detalhada da metodologia de pesquisa. No capítulo 5 temos a apresentação e discussão dos resultados obtidos nos experimentos e por fim as conclusões possíveis, apresentadas no capítulo 6.

# 2 REFERENCIAL TEÓRICO

Neste capítulo serão abordados os conceitos relativos à área de EDM em geral, assim como demais áreas correlacionadas também com o objeto de pesquisa.

## 2.1 Descoberta do conhecimento e Mineração de dados

O grande volume de dados disponíveis e gerados a cada dia proporcionado pelos avanços da tecnologia, unido à grande demanda por informações valiosas que suportem a tomada de decisões, direcionaram cada vez mais a atenção de pesquisadores e especialistas de diversos ramos para uma área conhecida como KDD - *Knowledge-Discovery in Databases*, ou Descoberta de Conhecimento em Bases de Dados, onde a Mineração de dados (DM – *Data Minning*) é apenas uma de suas etapas.

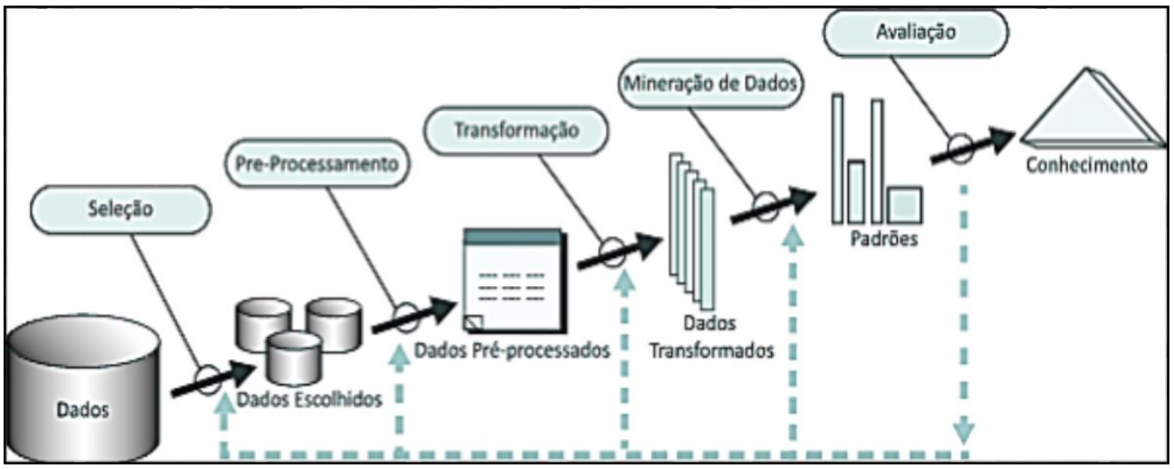
De acordo com Fayyad et al. (1996), "KDD é um processo, não trivial, de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis", em outras palavras podemos entender como um processo bem definido em etapas com o objetivo de obter informações importantes para a aquisição de conhecimento e tomada de decisão a partir de dados de diversas fontes. KDD pode ser visto como um termo utilizado por pesquisadores da área de Inteligência Artificial para o processo para extrair conhecimento em dados, sendo a mineração de dados uma parte integrante da descoberta de conhecimento em banco de dados. Portanto, a Mineração de Dados, representa uma das etapas principais do KDD. Em KDD verifica-se ainda a inclusão de mais duas grandes etapas: pré-processamento de dados (preparação de dados, abrangendo mecanismos para captação, organização e tratamento dos dados) e pós-processamento dos resultados obtidos na mineração de dados. Neste sentido, de uma definição abrangente, na qual KDD é descrito como um processo geral de descoberta de conhecimento composto pelas três grandes etapas mencionadas.

Os padrões identificados no processo de KDD devem ser novos, compreensíveis e úteis, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para uma possível tomada de decisão. Assim, esse processo consiste de uma série de passos para transformação, do pré-processamento dos dados até o pós-processamento dos resultados da mineração. As cinco etapas que constituem este processo KDD são apresentadas na Figura 1.

Há, no entanto, uma falta de consenso entre os autores sobre uma definição para o termo Mineração de Dados, dificultando a consolidação de uma definição única, o que faz muitos

pesquisadores optarem por considerar o termo mineração de dados como um sinônimo de KDD. Nesta pesquisa, após esse esclarecimento inicial, os termos KDD e Mineração de Dados serão tratados como sinônimos. No entanto, há inclusive autores que consideram Data Mining como sinônimo de KDD, referindo-se a ambas como uma disciplina que objetiva a extração automática de padrões interessantes e implícitos de grandes coleções de dados. Doravante neste texto, por simplicidade, não distinguiremos mais estes dois termos. Assim, podemos definir o processo de mineração de dados como consistindo de três passos Peña-Ayala (2014) executados em sequência: Pré-processamento de dados, análise de dados e a interpretação resultante, na qual o modelo selecionado é testado sobre novos dados para predizer o resultado esperado.

**Figura 1**- Processo KDD.




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

Inicialmente, por meio de um entendimento bem definido do domínio da aplicação, é necessário selecionar as bases de dados, bem como os dados, que serão usados no processo de descoberta de conhecimento. Em seguida, é efetuada a limpeza e o pré-processamento, uma vez que, frequentemente, os dados são encontrados com inúmeras inconsistências. Essas tarefas são fundamentais, pois o objetivo é eliminar incongruências, de modo que não influenciem o resultado dos algoritmos de mineração que serão aplicados. Posteriormente, realiza-se a transformação que consiste em reduzir ou projetar tais dados.

As três etapas iniciais podem ser agrupadas, originando uma grande fase nesse processo, conhecida como preparação dos dados. A partir da preparação apropriada dos dados inicia-se a etapa de mineração que, conforme já mencionado, consiste em escolher técnicas e algoritmos que possibilitem a extração de padrões.

Por último, efetua-se a etapa de avaliação que compreende na interpretação dos padrões minerados. Após uma análise minuciosa, usa-se o conhecimento diretamente, incorporando-o

a sistemas de apoio a decisões, ou simplesmente se documenta esse conhecimento, expondo-o às partes interessadas. Portanto, essa fase é importante, pois assegura que apenas resultados úteis e válidos sejam utilizados.

## 2.2 Mineração de dados

Segundo Ian H. Witten et al. (2011) a mineração de dados pode ser definida como o processo de descoberta de padrões nos dados. Tal processo deve ser automático ou, com é visto mais comumente, semiautomático. Os padrões descobertos a partir dos dados devem ser significativos, pois levam a alguma vantagem, geralmente econômica. Os dados estão invariavelmente presentes em quantidades substanciais (Ian H. Witten et al. (2011).

De acordo com Faceli et al., (2011), mineração de dados consiste em extrair ou "minerar" conhecimento a partir de grandes quantidades de dados, e é tratada também por outros autores como sinônimo de KDD. Percebe-se dessa forma que a mineração de dados está infimamente relacionada com o processo de obtenção de conhecimento para a tomada de decisões. O conhecimento das técnicas de mineração de dados é essencial para o estudo apresentado neste trabalho.

Foram analisadas nesta pesquisa as principais técnicas de Mineração de Dados, mais precisamente as conhecidas pela aplicação em EDM. A seguir descreve-se alguns conceitos e definições a respeito das características destas técnicas, que são descritas com mais detalhes em (Faceli et al e Ian H. Witten et al, 2011).

## 2.3 Técnicas de Seleção de Atributos

Várias aplicações reais apresentam um grande número de atributos Faceli et al., (2011), isso pode acarretar problemas de dimensionalidade, parte destes atributos podem ser irrelevantes, ou redundante, gerando ruídos. As técnicas de Seleção de atributos ajudam a lidar com estes tipos de problemas.

A seleção de Atributos consiste em uma das técnicas de realizadas na etapa de pré-processamento e que é responsável por identificar, selecionar ou ainda ordenar por nível de relevância os atributos que possuem uma maior relação com o atributo classificador, ou seja, aquele o qual pretende-se predizer. Além de ser útil para reduzir custo de processamento e simplificar os modelos gerados na classificação, a seleção de atributos representa uma das formas de melhorar o desempenho do processo de classificação (LÔBO, M. T. F, 2015). Os diversos métodos conhecidos de seleção de atributos, basicamente divididos em três categorias (FACELI et al., 2011):

i. Embutida: a seleção do subconjunto de atributos é embutida ou integrada no próprio

algoritmo de aprendizado.

ii. Baseada em filtro: realizada no pré-processamento. Filtra um subconjunto de atributos do conjunto original, sem levar em consideração o algoritmo de aprendizado que utilizará este subconjunto.

iii. Baseada em *wrapper*: utiliza o próprio algoritmo de aprendizado como uma caixa preta para a seleção. Para cada possível subconjunto, o algoritmo é consultado e o subconjunto que apresentar a melhor combinação entre redução de taxa de erro e redução do número de atributos é em geral selecionado.

Nesta pesquisa, o método de seleção de atributos que será predominantemente usado é o baseado em Filtros.

Pode-se classificar ainda os métodos de seleção de atributos em 2 dois tipos, no que se refere à forma de se avaliar e mensurar a relevância dos atributos, são eles:

a. *Attribute Subset Evaluators* (Avaliadores de Subconjunto de Atributos) –utilizam um subconjunto de atributos e retornam uma medida numérica que orienta a busca.

b. *Single-Attribute Evaluators* (Avaliadores de Atributo Único) – são usados com o método de pesquisa do tipo *Ranker* (do inglês, ranqueador ou ordenador) para gerar uma lista de classificação da qual o *Ranker* descarta um determinado número.

## 2.4 Modelos Preditivos e Algoritmos de Classificação

Segundo Faceli et al., (2011), um algoritmo de Aprendizagem de Máquina é uma função que, dado um conjunto de exemplos rotulados, constrói um estimador, também conhecido na área de Inteligência artificial como Modelo Preditivo. Caso o rótulo em que se baseia o algoritmo seja de valores nominais, temos tem-se um problema de Classificação, ou aprendizado de conceitos, e o estimador gerado é um classificador. Por sua vez se o domínio dos valores do rótulo for um conjunto infinito e ordenado de valores, tem-se um problema de Regressão, que induz um regresso.

Pode-se definir formalmente Faceli; Gama e Carvalho (2011), dado um conjunto de observações de pares $D = \{(x_i, f(x_i), i = 1,2, \dots, n\}$ , onde $f$ representa uma função desconhecida, um algoritmo preditivo aprende uma aproximação, ou estimador, $\hat{f}$, da função desconhecida *f*. Pode-se então estimar o valor de *f* a partir de $\hat{f}$ para novas observações de x. De acordo com f, temos duas possíveis situações:

1. Classificação: $y_i = f(x_i) \in \{c_1, \dots, c_m\}$, ou seja, $f(x_i)$ assume valores em um conjunto discreto, não ordenado.
2. Regressão: $y_i = f(x_i) \in \mathbb{R}$, ou seja, $f(x_i)$ assume valores em um conjunto infinito e

ordenado de valores.

Neste trabalho, como a problemática envolve a classificação de alunos em apenas 2 classes, a dos Aprovados e a dos Reprovados, serão utilizadas apenas técnicas e algoritmos de Classificação. Existem diversos métodos que permitem classificar dados, podemos descrever e diferenciar tais métodos de acordo com (Faceli et al., 2011):

a. Métodos Baseado em Distância: Estes métodos se baseiam na proximidade entre os dados da realização das predições. Consideram como hipótese base que, dados similares tendem a estar concentrados em uma mesma região em um espaço de entrada, por outro lado dados não similares estarão distantes entre si. Aqui temos um algoritmo bastante conhecido que é o k-NN, ou *k-Nearest Neighbour*, em português, o k-ésimo vizinho mais próximo, que pode utiliza a distância euclidiana para avaliar a distância entre os pontos, neste caso dados de entrada, permitindo assim classifica-los de acordo com a distância ou similaridade.
b. Métodos Probabilísticos: Utilizam como base o teorema de Bayes, bastante conhecido da Teoria da Probabilidade. Podemos citar aqui como exemplo o algoritmo Naive Bayes, que utiliza a técnica de classificação probabilística para o aprendizado de máquina baseado em modelos de probabilidade que incorporam fortes pressupostos de independência (JOHN e LANGLEY,1995). Naive Bayes tem sido usado para problemas de classificação binária (duas classes) e multi-classe. Assumindo que os valores dos atributos de um exemplo são independentes entre si, dada a classe, $P(x|y_i)$, pode ser decomposto no produto $P(x^1|y_i) \times$ ... $\times$ $P(x^d|y_i)$, em que $x^j$ é o j-ésimo atributo do exemplo x. Logo a probabilidade de um exemplo pertencer à classe $y_i$, é proporcional à expressão:

$$P(y_i|x) \propto P(y_i) \prod_{j=1}^{d} P(x^j|y_i) \quad (1)$$

Logo, a partir desta equação discriminante e de uma regra de decisão, o Naive Bayes consegue aprender e classificar dados de entrada.

c. Métodos Baseados em Procura: Aqui o objetivo é formular modelos de aprendizagem para a classificação a partir da procura em um espaço de possíveis soluções. Os algoritmos que se baseiam em procura mais conhecidos são os de Árvore de Decisão e os de Regras. Indução da Árvore de Decisão é uma técnica de classificação baseada em um processo de construção de um modelo de classes a partir de um conjunto de instâncias em uma estrutura de tabela que contém rótulos de classes. Ele tem sido usado para problemas de classificação binária (duas classes) e multi-classe. Em particular, o J48 é um algoritmo de árvore de decisão que implementa uma abordagem de modelagem preditiva, que é uma implementação Java de código aberto do conhecido algoritmo C4. (QUILAN,1993). Temos ainda estudos

consagrados em árvores de decisão que resultaram em outros algoritmos conhecidos como o CART (BREIMAN, 1984), e o ID3 (QUILAN, 1986). Estudos mais recentes e interessantes que comparam os algoritmos C4.5, CART e ID3 foi realizado por Hssina et al., (2014); Singh e Gupta, (2014). Já os algoritmos baseados em Regras, apesar de assemelharem-se com os de Árvores por também utilizarem métodos de procura, estão voltados para estruturas de decisão da forma: se A então B. Dentre eles, foram selecionados para esta pesquisa 5 bastante conhecidos: JRip, NNge, OneR, Prism e Ridor. São descritos de forma sucinta no Capítulo 4.

d. Métodos Baseados em Otimização: Os principais e amplamente conhecidos algoritmos baseados em Otimização são os SVM e Rede Neural. É uma técnica de aprendizado de máquina, como uma espécie de abordagem de aprendizado supervisionado baseada na noção de kernel, desenvolvida para resolver problemas de classificação e regressão, principalmente envolvendo problemas binários (duas classes), mas também SVMs multiclasse. Entende-se como um algoritmo embasado pela Teoria de Aprendizado Estatístico (TAE) (VAPNIK, 1995). A TAE estabelece condições matemáticas que auxiliam na escolha de um classificador particular, $\hat{h}$, a partir de um conjunto de dados de treinamento. Tais condições levam em consideração o desempenho do classificador no conjunto de treinamento e a sua complexidade, objetivando assim obter um bom desempenho para novos conjuntos de dados.

Uma definição ainda importante a respeito dos algoritmos classificadores é a distinção entre *White Box* e *Black Box*, Caixa branca e Caixa Preta respectivamente. Os algoritmos do tipo Caixa Branca são aqueles relacionados aos Testes de Caixa Branca, onde é possível se ter uma perspectiva interna do processo de realização do teste, inclusive é possível também entender, de forma relativamente clara, como foi que o algoritmo chegou a determinado resultado. Algoritmos de Regras e Árvores de Decisão são essencialmente do tipo Caixa Branca. Por outro lado, os algoritmos que realizam processamentos ou testes onde não é possível o acesso como determinado resultado foi encontrado são considerados Caixa Preta. SVM e Nayve Bayes são exemplos deste tipo de algoritmo.

## 2.5 Métricas de Desempenho em Aprendizagem de Máquina

Conforme também aplicadas em MÁRQUEZ; Morales e Soto (2013); MÁRQUEZ et al., (2016) e conceituado em Faceli; Gama e Carvalho (2011), as métricas comumente utilizadas para predições em EDM, e que também serão exploradas o decorrer desta pesquisa são:

a. Taxa de acerto ou acurácia total (*Accuracy*): calculada pela soma dos valores da diagonal

principal da matriz de confusão, dividida pela soma de todos os elementos da matriz. É a métrica principal utilizada nesta pesquisa.

$$Accurácia\ (ACC) = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population} \quad (2)$$

b. Taxa de Verdadeiros Positivos (*True Positive Rate*): Calcula a soma dos verdadeiros positivos sobre os verdadeiros.

$$TP\ rate = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Negative} \quad (3)$$

c. Taxa de Verdadeiros Negativos (*True Negative Rate*): Calcula a soma dos verdadeiros negativos sobre os negativos.

$$TN\ rate = \frac{\sum True\ Negative}{\sum True\ Negative + \sum False\ Positive} \quad (4)$$

d. Média Geométrica (*Geometric Mean*): indica o equilíbrio entre duas medidas de classificação. Representa uma medida de trade-off comumente usada com conjuntos de dados desequilibrados e calculada como:

$$GM\ (Geometric\ Mean) = \sqrt{TP\ rate\ .\ TN\ rate} \quad (5)$$

No intuito de facilitar a identificação de padrões, comparação entre algoritmos, visualização gráfica, e análise exploratória do comportamento dos atributos, neste trabalho optou-se por mensurar o desempenho das técnicas utilizando-se predominantemente a acurácia total. Pretende-se, em trabalhos futuros, explorar outras métricas importantes que venham a explicitar outras perspectivas de eficiência.

# 3 TRABALHOS RELACIONADOS

Existe um número expressivo de trabalhos acadêmicos recentes na área de EDM aplicados aos mais diversos aspectos educacionais. Este capítulo apresenta os principais trabalhos relacionados à EDM tomados como base para o planejamento e elaboração da presente pesquisa.

Este trabalho foi proposto a partir dos trabalhos realizados por Marquez  e Vera em 2013); MÁrquez; Morales e Soto (2013) e MÁRQUEZ et al., (2016). No trabalho de 2013, o referido autor aborda tanto o sucesso acadêmico quanto evasão. Foi apresentada uma maneira de selecionar os melhores atributos, trabalhando no uso de técnicas de mineração de dados focadas em regras de indução e árvores de decisão, para prever falhas acadêmicas dos alunos em escolas de ensino médio ou médio. No entanto, para alcançar esses resultados, este trabalho teve que considerar muitas variáveis diferentes de várias fontes de dados, incluindo dados não acadêmicos, tais como pessoais e familiares, de pesquisa demorada, bem como dados dos alunos de notas obtidas em vários cursos. Além disso, eles não preveem se um aluno falhará na fase inicial. Um ponto positivo é a descrição do processo de mineração de dados a partir das 3 fontes de dados, comparando os desempenhos de 10 algoritmos de classificação, sendo 5 de árvore de decisão e 5 de regras, comparando ainda o ganho obtido com a seleção de atributos. A metodologia utilizada na presente pesquisa foi fortemente inspirada no referido trabalho do Marquez Vera, considerando inclusive os mesmos algoritmos de seleção de atributos, com critério semelhante para a exclusão dos atributos irrelevantes, realizada no Experimento 1 deste trabalho, os mesmos 10 algoritmos de classificação, incluindo 4 novos do tipo caixa preta para fins de comparação de desempenho. Já o trabalho de 2016, trata de um outro ponto também explorado neste trabalho que é a questão temporal, considerada no Experimento 4, porém a pesquisa referenciada objetiva apenas a predição antecipada de evasão escolar, deixando de lado a questão aprovação, foco do estudo aqui explorado.

Um outro trabalho que teve bastante relevância para esta pesquisa foi o de Costa et al. (2017), onde foram abordadas problemáticas relacionadas à predição antecipada do sucesso acadêmico, considerando o ajuste de parâmetros no pré-processamento, incluindo seleção de atributos e o *fine tunning* (ajuste fino), com o intuito de avaliar o ganho de desempenho. Não foram utilizados neste caso dados socioeconômicos, apenas alguns poucos atributos pessoais dos alunos juntamente com dados relativos a notas das atividades realizadas durante o curso, esta abordagem também é explorada na presente pesquisa, mais precisamente na questão de pesquisa 4. Alguns diferenciais em relação ao referido trabalho foram aplicados nesta

pesquisa: a inclusão de atributos socioeconômicos para a análise dos atributos relevantes e análise da influência destes na predição, exploração mais ampla em termo de quantidade de algoritmos de classificação, além da proposta de analisar o desempenho em relação ao tempo desde o início do curso até o final do curso, diferentemente do trabalho referenciado, que buscou explorar apenas os 2 primeiros meses de curso.

O trabalho de Khobragade (2015) segue uma abordagem semelhante à do trabalho previamente discutido, para predizer o fracasso dos alunos, este trabalho investiu em alguns algoritmos de classificadores *White-Box* para induzir regras e árvore de decisão, envolvendo o uso de dois algoritmos para regras e dois para árvore de decisão. Além disso, também utilizou o algoritmo Naive Bayes. Ele selecionou os 11 melhores atributos e a maioria dos recursos selecionados foi baseada em dados reais das informações dos alunos, por exemplo, marcas dos alunos, antecedentes familiares, aspectos sociais e acadêmicos, e também seu desempenho passado, como o desempenho passado de um aluno é indicativo do seu presente e futuro desempenho. Na maioria dos casos, esses dados podem ser coletados usando relatórios e pesquisas da faculdade. O algoritmo Naive Bayes forneceu a melhor precisão com 87.12. Portanto, este trabalho obteve resultado para previsão alta, mas não investiu em previsão antecipada e utilizou muitas outras fontes de dados, além de dados acadêmicos, o que tem um potencial alto custo e é uma solução demorada.

Mais uma pesquisa bastante recente na área de EDM e que aborda a seleção de atributos como tema focal e também voltada para a predição do sucesso acadêmico, é a apresentada por Roy et al. (2018). Um diferencial desta pesquisa é a preocupação com a o direcionamento dos estudantes para a melhoria de suas capacidades e habilidades, para que possam alcançar a aprovação. Roy utiliza também atributos socioeconômicos e acadêmicos, além de técnicas de classificação conhecidas, assim como as selecionadas para este trabalho, a exemplo de SVM e Rede Neural.

Um outro trabalho que busca uma análise temporal, porém focado em um ambiente de aprendizagem do ensino à distância, o *Moodle*, foi realizado por Burgos et al., (2018). Neste, o objetivo é encontrar padrões com mineração de dados dos alunos de um curso à distância, que permitam a predição antecipada dos que irão abandonar o curso, dando assim um suporte ao professor do ensino à distância. A atenção aqui está direcionada para os resultados alcançados de desempenho das técnicas comparadas: observou-se que todas as técnicas alcançaram acurácia superior a 70% a partir da 8ª semana.

O estudo realizado por Yassein et al. (2017) trata também da seleção de atributos com técnicas de mineração de dados educacionais, com um grande diferencial que é a utilização do

software *SPSS - Statistical Package for Social Sciences*. Apesar da utilização de poucos atributos, é possível perceber que a ferramenta SPSS também é capaz de identificar aqueles mais relevantes. Também utilizando SPSS, ressalta-se um relevante trabalho realizado por Bezerra et al., (2016), porém desta vez o objetivo é a extração de conhecimento a partir dos dados do censo escolar disponibilizado pelo INEP visando identificar o perfil do aluno evasor e estimar a propensão à evasão através de Árvore de Decisão, Indução de Regras e Regressão Logística. Os resultados mostraram que fatores como idade, turno das aulas e região geográfica das escolas influenciam fortemente a evasão. É interessante comparar tais fatores com os que mostram influência para a aprovação, explorados no Experimento 1 deste trabalho.

Em relação às técnicas de Seleção de Atributos, observa-se uma descrição bastante focada e rica no trabalho de Maryam Zaffar (2017). É feita uma análise da performance da seleção de atributos em dados de estudantes, mais precisamente, o ganho de desempenho alcançado a partir da remoção dos atributos menos relevantes. Esta metodologia também é abordada no atual trabalho buscando-se o mesmo objetivo de comparação de ganho em desempenho.

Na pesquisa realizada por Reda et al. (2018) a atenção é voltada para a performance acadêmica, assim como o estudo realizado neste trabalho. Foram comparados os desempenhos das técnicas Naive Bayesian, Árvore de Decisão e *Multi-Layer Perceptron* (MLP). O comparativo entre uma técnica de árvore com uma técnica probabilística e simultaneamente com uma rede neural remete a uma metodologia semelhante à proposta neste trabalho. A partir dos resultados encontrados por Reda, observa-se um destaque para a técnica de árvore de decisão, que alcançou um patamar destacável de 97,69 em acurácia. Um ponto negativo do referido trabalho é a ausência de uma descrição detalhada a respeito dos dados utilizados.

O trabalho proposto por Ahmad, Ismail e Aziz (2018), apresenta uma abordagem para prever o desempenho acadêmico dos alunos de primeiro ano bacharel em curso de Ciência da Computação. Ele usa um framework, que suporta as técnicas de classificação Decision Tree, Naive Bayes e Rule Based, a serem aplicadas aos dados dos alunos para produzir o modelo de previsão de desempenho acadêmico dos melhores alunos. Os dados utilizados foram coletados durante oito semestres, contendo dados demográficos dos alunos, registros acadêmicos anteriores e informações sobre antecedentes familiares. Os resultados mostram que o classificador baseado em regras é o melhor modelo entre as outras técnicas, recebendo o maior valor de precisão de 71,3. Isso não é uma boa precisão, principalmente se estamos considerando que essa abordagem envolve dados de várias fontes de dados, ou seja, uma solução possível de alto custo e demorada.

O trabalho de Bydžovská et al., (2016) abordou o problema de prever as notas finais dos alunos no início do semestre, com ênfase na identificação de alunos malsucedidos. Para isso, utilizou duas abordagens diferentes, sendo a primeira baseada em algoritmos de classificação e regressão. Essa abordagem foi considerada interessante quando utilizada para a predição de séries de cursos com um pequeno número de alunos. Os algoritmos empregados foram: Support Vector Machine, Random Forest, classificador baseado em regras, Decision Tree, Part, IB1 e Naive Bayes. Neste estudo, o SVM alcançou o melhor desempenho. Os resultados foram melhorados usando também dados sobre o comportamento social dos alunos nas previsões. A segunda abordagem usada foi em uma linha diferente, considerando técnicas de filtragem colaborativa e notas previstas, com base na similaridade das realizações dos alunos. Este caminho é muito diferente daquele discutido em nossa abordagem. Incluiu dados sobre o comportamento social dos estudantes, sendo assim diferente da nossa abordagem, mas o melhor desempenho com SVM foi semelhante ao obtido em nossa abordagem.

Um outro trabalho relacionado recente é o realizado por Marbouti et al., (2016) onde são utilizados algoritmos de Regressão Logística, SVM, *Decision Tree, Multilayer Perceptron*, Naive Bayes e k-NN, com uma abordagem diferenciada onde são utilizados dados do ano de 2013 para treino e em seguida os modelos aprendidos são aplicados em 2014. A Acurácia também é a métrica mais utilizada para representar o desempenho dos algoritmos. As predições são obtidas experimentalmente com e sem Seleção de Atributos. Os algoritmos que mais se destacaram em termos de acurácia nesta abordagem apresentada de treino e teste foram k-NN e Regressão Logística. Um outro ponto de observação que pode ser comparado com o presente trabalho é que na referida pesquisa foram utilizados essencialmente como atributos para as predições dados de desempenho em tarefas de casa elaboradas e aplicadas pelos professores aos alunos, assim como as notas das atividades do Huxley utilizadas neste trabalho. A relevância das atividades do Huxley está descrita com maior detalhe no capítulo 5.

## 3.1 Limitações dos Trabalhos Relacionados

Um fator relevante para o direcionamento deste estudo é o fato de que não há trabalhos anteriores abordando de forma mais ampla os fatores relacionados com a eficiência das técnicas de EDM voltadas à predição de aprovação e/ou reprovação, o que deixa uma lacuna para um estudo exploratório que dê um suporte para a criação de modelos capazes de obter resultados satisfatórios com poucos recursos.

Existem inúmeros trabalhos recentes na área de EDM que exploram bem o processo de predição, apesar disso poucos exploram a abordagem amostral, ou seja, o comportamento da

eficiência de cada técnica ou algoritmo quando utilizam uma parcela menor de dados para treino, dentre os dados disponíveis, e parcelas maiores para o teste do modelo, ao mesmo tempo também com parcelas maiores de treino contra parcelas menores para teste. Também é visto como pouco explorada a abordagem dos algoritmos de árvore de decisão e o desempenho de cada um a partir de arvores menores, podadas, e maiores, com um maior grau de granularidade e refinamento na decisão. A análise temporal também é pouco explorada, onde busca-se conhecer em que momento do curso é possível uma acurácia satisfatória a partir do ganho de informação obtido com a inserção de dados das notas dos alunos obtidos em atividades de aprendizagem.

# 4 METODOLOGIA PROPOSTA

A metodologia proposta neste estudo para abordar problemas relacionados à tarefa de predição, consiste nas duas grandes fases descritas a seguir. Para este estudo especificamente, serão tratados os fatores que possam vir a afetar a eficiência das técnicas de EDM na predição do sucesso dos alunos. Assim, o processo principal da metodologia proposta neste estudo foi dividido em 2 fases:

1) Fase inicial – Fase primária e comum a todos os experimentos. Divide-se em:
    a. Coleta e Preparação dos Dados.
    b. Pré-processamento
2) Fase de experimentação – Esta fase contempla 4 diferentes experimentos, que estão diretamente relacionadas com os objetivos da pesquisa. São elas:
    a. Experimento 1 – Processo de Análise de Seleção de Atributos
    b. Experimento 2 – Processo de Análise da Eficiência dos Algoritmos de predição com diferentes parcelas de treino/teste
    c. Experimento 3 – Processo de Análise da Eficiência dos Algoritmos de classificação de árvores de decisão
    d. Experimento 4 – Processo de análise da Eficiência dos algoritmos de classificação no decorrer do curso.

Os experimentos foram realizados com a utilização do pacote de software Weka - *Waikato Environment for Knowledge Analysis*, uma ferramenta gratuita desenvolvida em JAVA bastante conhecida nos estudos de EDM para análise de dados, a escolha se deu ao fato de que nas principais pesquisas de referência foram utilizados também o Weka para pré-processamento e processamento dos dados e obtenção dos resultados desejados de eficiência na predição. As métricas utilizadas neste trabalho para análise dos resultados também seguiram as mesmas tendências dos trabalhos anteriores.

É de interesse desta pesquisa, como análise de eficiência, considerar em princípio as informações existentes sobre cada aluno no momento em que ele inicia o curso. A motivação para esta análise preliminar se dá justamente pelo fato de que quanto mais prévia é a detecção, por parte do professor, dos alunos predispostos à reprovação, melhor para as tomadas de decisão, direcionamento e atuação preventiva durante o curso. Apenas o Experimento 4, contido nesta pesquisa, considera informações dos alunos após o início do curso.

### 4.1 Fase Inicial

Nesta etapa, é feito todo o planejamento inicial a respeito do escopo, relevância de dados para a pesquisa, disponibilidade, procedimentos de coleta, e ainda preparação e formatação dos dados para que se tornem aptos para a aplicação das técnicas de Seleção de Atributos e Classificação. Em virtude da possibilidade de acesso a algumas fontes de dados dentro do Campus da Universidade Federal de Alagoas, a pesquisa contou com 3 fontes de dados de onde foram extraídas as amostras utilizadas neste trabalho, são elas:

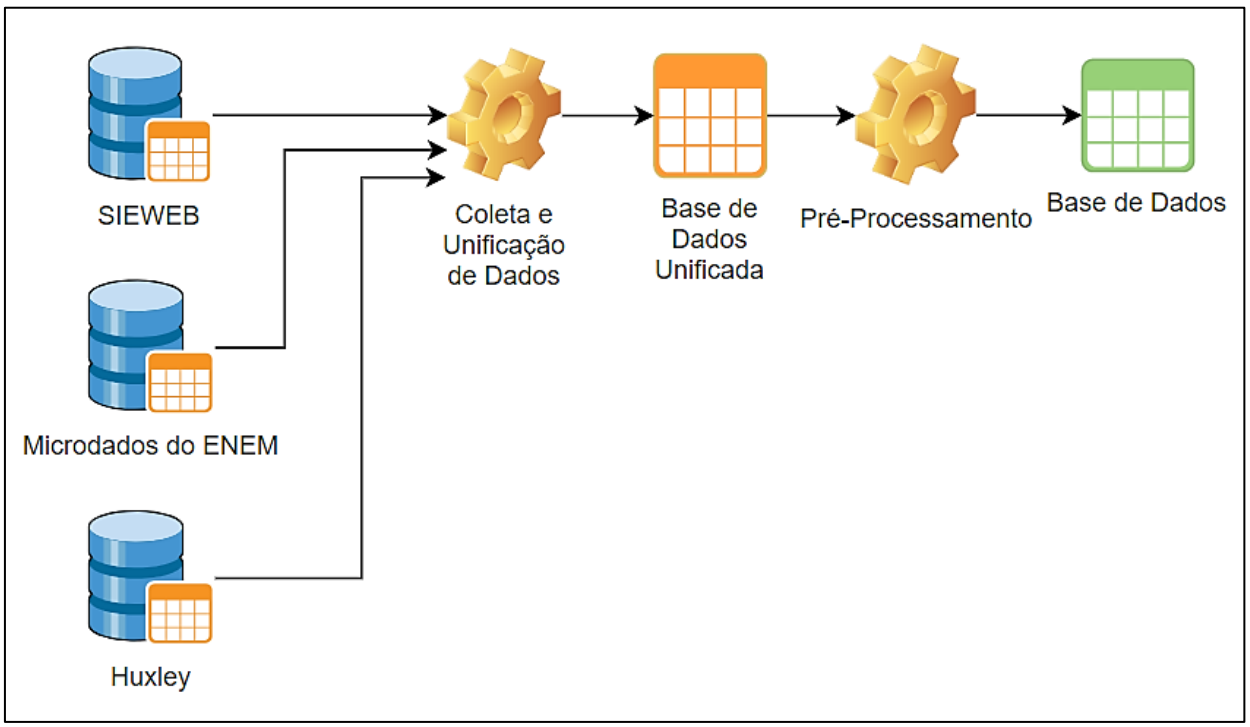
1) SIE WEB: Consiste no Sistema Acadêmico Online da UFAL, disponível em https://sistemas.ufal.br/academico. A Base de Dados do SIE WEB contém todas as informações acadêmicas dos alunos de graduação da universidade. É importante ressaltar que para acesso a uma amostra destes dados, foi necessária a solicitação formal via processo administrativo, contendo um termo de confidencialidade dos dados além do compromisso de utilização exclusiva para a pesquisa. Esta é a fonte mais importante para a pesquisa, pois é onde se encontra, dentre outros, o atributo "Nota Final", e é a partir deste atributo que podemos ter a informação de se o aluno foi aprovado ou reprovado na disciplina. Este atributo, na etapa de Pré-processamento, é transformado em categórico, e são criadas 2 classes apenas para descrevê-lo: R – Reprovado e A- Aprovado. Este atributo, por sua vez, é definido como Classificador, para o processo de Classificação e Seleção de Atributos. Podemos nomear o conjunto de dados extraídos desta fonte de dados como "Dados Acadêmicos".
2) Huxley: Uma plataforma online para aprendizagem e testes em Programação, disponível em https://www.thehuxley.com. Desenvolvida dentro da própria instituição, por docentes e discentes, o Huxley dispõe de milhares de problemas propostos na área de programação, com a possibilidade de correção automática, para que alunos possam exercitar seus conhecimentos. A plataforma também pode ser utilizada para aplicação de testes de conhecimento. Essencialmente, para a presente pesquisa, viu-se como relevante a coleta dos dados referentes às atividades realizadas pelos alunos da disciplina Programação I, que é a disciplina que está sendo utilizada nesta pesquisa. O acesso a tais dados foi possível após a autorização do professor administrador do sistema. Adotaremos o termo "Dados do Huxley" para definir o subconjunto de dados coletados desta fonte.
3) ENEM 2012: Dados essencialmente socioeconômicos dos candidatos que prestaram o ENEM - Exame Nacional do Ensino Médio, no ano de 2012, onde consequentemente estão contidas as informações dos alunos que ingressaram na Instituição em 2013. Dados disponíveis em http://portal.inep.gov.br/web/guest/microdados. Como a disciplina de

Programação I, alvo desta pesquisa, é uma disciplina inicial do curso, todos os alunos ingressantes obrigatoriamente devem se matricular, por isso foi possível a localização dos alunos para a unificação das 3 sub amostras das 3 fontes de dados. No Anexo I é apresentado um Dicionário de Dados referente a esta fonte de dados, permitindo assim conhecer melhor cada um dos atributos coletados e seus respectivos valores ou classes categóricas. É importante ressaltar que dentro dos dados do ENEM, existem 3 grupos básicos de atributos, que podem ser classificados ou subdivididos como: A) Dados Pessoais; B) Dados Socioeconômicos e C) Notas na prova do ENEM. No decorrer deste estudo é possível que se façam referências aos dados do ENEM de forma completa, ou seja, todos os atributos disponíveis nesta base, ou ainda podem existir referências a apenas uma parte desta base de dados, como por exemplo “Dados das Notas dos alunos na Prova do ENEM” ou “Dados socioeconômicos do ENEM”. É importante tornar explícita e bem definida esta distinção para que se tornem claras as referências aos dados do ENEM no decorrer do trabalho, evitando assim a má interpretação.

A partir do Planejamento, foram definidas as fontes de dados, quais os dados de interesse de cada fonte, procedimentos para extração e unificação, além do modelo final da base de dados a ser utilizada nos experimentos da pesquisa. A Figura 2 ilustra o processo de que compõe a fase inicial.

Um fator relevante que representou uma possível ameaça ao presente estudo foi a limitação referente aos quantitativo de instâncias disponíveis de alunos para serem utilizados na pesquisa. Tal limitação ocorreu por 2 motivos: 1) Devido à ausência de base de dados do Huxley para os anos anteriores a 2013.2, pois foi o ano em que este entrou em operação; 2) Também devido ao fato da inexistência da informação da “matrícula do aluno” nas bases de dados do ENEM disponibilizadas no portal do INEP, nos anos de 2014 em diante, que por sua vez inviabilizou a identificação dos alunos para a integração com os dados do SIE WEB, tal informação foi disponibilizada apenas nos anos 2012 e 2013. Diante destas limitações, o único período onde foi possível integrar dados das três bases foi o segundo semestre de 2013, que por sua vez foi o período utilizado nesta pesquisa.

**Figura 3** – Processos da Fase Inicial.




**Fonte:** José E. V. Borges (2019).

### 4.1.1 Processo de Coleta de Dados e Unificação

Para a extração dos dados de cada fonte foram necessários procedimentos diferentes. Os Dados Acadêmicos, extraídos do SIE WEB, foram disponibilizados pela equipe da UFAL que administra o sistema em um arquivo formato .csv contendo todos os registros acadêmicos do período selecionado, 2013.2. A partir disso, foram filtradas, com a utilização do Microsoft Excel, apenas as instâncias dos alunos dos cursos de interesse, Ciência da Computação e Engenharia da Computação, e em seguida da disciplina de interesse, Programação I. Além disso também foram necessários alguns tratamentos na base de dados para que atendesse ao objetivo da pesquisa, foram descartados atributos não importantes, como endereço, e-mail, e outros, e foram mantidos apenas os dados acadêmicos de notas, além dos dados pessoais Idade, Sexo, Etnia e Deficiência.

Para a extração dos dados do Huxley, foram necessários apenas consultas SQL diretamente ao banco do sistema, realizadas pela equipe responsável pena manutenção do sistema, onde como resultado tivemos um arquivo .csv com os dados de cada aluno e suas notas nas atividades realizadas durante o curso de Programação I, período 2013.2. A periodicidade das atividades realizadas no Huxley durante o curso depende do professor que está ministrando a disciplina, para este caso, verificamos que foram aplicadas atividades de aprendizagem a cada 15 dias de curso.

Quanto à extração dos dados do ENEM 2012, devido ao fato da fonte de dados disponível no site do INEP encontrar-se unificada a nível Brasil, ou seja, em um só arquivo de dados estão armazenados os dados de todos os candidatos que se inscreveram no ENEM no ano de 2012 de todo o Brasil, foi necessária a utilização do software RStudio, software de desenvolvimento integrado para a linguagem de programação bastante conhecida atualmente para gráficos e cálculos estatísticos R, apenas para a conversão do arquivo em um banco de dados relacional, permitindo assim uma consulta direta SQL apenas pelos alunos que estavam matriculados na disciplina Programação I, período 2013.2.

A partir da coleta das amostras específicas de dados das 3 fontes, a tarefa agora é a unificação em uma só base de dados. Essa tarefa foi possível através dos campos utilizados como chave: número de inscrição e nome dos alunos. Para esta tarefa foram utilizados Microsoft Excel e RStudio.

### 4.1.2 Pré-processamento

A partir dos dados unificados, para que seja possível aplicar os algoritmos de Classificação e Regressão, ou de Seleção de Atributos, são necessárias algumas alterações na base de dados, que consistem na etapa de pré-processamento de dados. Limpeza de dados inconsistentes, integração, categorização, balanceamento de dados e transformação de variáveis são algumas das alterações necessárias. Tais alterações foram realizadas em parte utilizando o Microsoft Excel e outra parte utilizando o software Weka.

Primeiramente, foram removidos os atributos que não continham informações relevantes, dados sem integridade ou ainda atributos com instâncias sem resposta. Tais informações poderiam comprometer o resultado dos experimentos por não agregar informação.

O segundo passo, considerado relevante para os algoritmos de classificação e seleção de atributos, foi a normatização dos atributos que continham valores numéricos. Como os dados numéricos representam atributos de grandezas distintas, consequentemente de escalas numéricas diferentes, para que não haja um tratamento diferenciado pelos algoritmos, onde possivelmente os atributos com maior valor numérico seriam privilegiados, foi utilizado a normatização de todos os atributos numéricos, através da métrica *Standard Score*, onde *x* é o valor numérico do dado ou amostra, μ é a média amostral e σ o desvio padrão amostral.

$$Standard\ Score = \ Z\ score = \frac{x - \mu}{\sigma} \tag{6}$$

Outra transformação necessária para os experimentos é a definição do atributo

Classificador, atributo este que vai ser utilizado pelo algoritmo de classificação para a aprendizagem e criação do modelo além de também ser necessário para o teste do modelo nos dados e obtenção das métricas de desempenho do algoritmo. Os algoritmos de Seleção de Atributos também necessitam do atributo classificador para s Como a proposta é a predição de sucesso acadêmico, ou seja, se o aluno será ou não aprovado, o atributo que será classificador neste caso será o "Conceito", que consiste no conceito final do aluno e deve ser categórico com apenas duas categorias: A = Aprovado e R = Reprovado. Temos dessa forma todas as instâncias da base de dados dividida em 2 classe, a dos alunos aprovados na disciplina e alunos reprovados.

Em seguida, um procedimento importante e utilizado nos estudos de predição, é o balanceamento dos dados. Consiste na equalização das classes contidas na base de dados para que fiquem iguais em número de instâncias. O balanceamento pode ser feito através de 2 procedimentos conhecidos: *oversampling* ou *undersampling*. *Oversampling* é utilizado quando, para o balanceamento, são criadas sinteticamente novas instâncias para a classe que contém menor número de instâncias fique com número igual à classe de maior número. Já o *undersampling* consiste na redução do número de instâncias, através de uma escolha aleatória, da classe de maior número de forma que fique com mesmo quantitativo de instâncias da classe de menor número. Como a base obtida encontra-se levemente desbalanceada, com 34 Reprovados e 33 Aprovados, optou-se pelo *oversampling* para que o tamanho da amostra não seja reduzido. Uma técnica bastante utilizada e conhecida de *oversampling* é o SMOTE - *Synthetic Minority Over-sampling Technique* (Chawla et al., 2002). O Weka já possui dentro do seu pacote de algoritmos o SMOTE para essa finalidade, na parte de pré-processamento de dados.

Vale ressaltar que, apenas para a execução de um dos algoritmos de classificação, o Prism, devido ao fato de que ele aceita apenas dados categóricos como entrada, viu-se necessário a transformação de todos os atributos numéricos em categóricos. Isso foi feito através do filtro *NumericToNominal*, existente no Weka.

Após a unificação, foi obtido como resultado uma base de dados final com 59 atributos no total, e 66 instâncias de alunos. O Quadro 1 descreve os atributos coletados e mantidos de cada fonte de dados coletada. Vale ressaltar que os dados do Huxley foram considerados apenas no Experimento 4, ou seja, foram removidos da base nos Experimentos 1, 2 e 3. No Anexo I está apresentado o dicionário de dados referente aos atributos do ENEM.

**Quadro 1-**Conjunto de Dados Coletados.

| Fonte de Dados | Atributos |
|---|---|
| **Acadêmicos – SIEWEB (10)** | AV1; AV2; Reavaliação; Prova Final; Média Final; Conceito; Idade; Sexo; Etnia; Deficiência. |
| **Huxley (6)** | Notas da Atividade 1, Notas da Atividade 2, Notas da Atividade 3, Notas da Atividade 4, Notas da Atividade 5, Notas da Atividade 6. |
| **ENEM (43)** | Idade; Sexo; Etnia; Deficiência; Município de Residência; UF de Residência; Nota na prova de Linguagens e Códigos (ENEM); Nota na prova de Ciências Humanas (ENEM); Nota na prova de Ciências da Natureza (ENEM); Nota na prova de Matemática (ENEM); Nota na prova de Redação (ENEM); Nota geral na prova do ENEM; Com quantas pessoas mora; Nível de escolaridade do Pai; Nível de escolaridade da Mãe; Renda familiar; Situação da residência onde mora; Zona onde está localizada a sua residência; Fez ENEM para testar conhecimentos; Fez ENEM para obter uma bolsa de estudos; Quantos anos levou para a conclusão do ensino fundamental; Deixou de estudar durante o ensino fundamental; Tipo de escola em que cursou o ensino fundamental; Quantos anos levou para a conclusão do ensino médio; Deixou de estudar durante o ensino médio; Tipo de escola em que cursou o ensino médio; Cursou o programa de Educação de Jovens e Adultos; Cursou o ensino regular; Pretende aderir ao FIES; Pretende aderir ao PROUNI; Pretende aderir bolsa de estudos da instituição; Pretende aderir bolsa de estudos da empresa onde trabalha; Possui TV; Possui DVD/vídeo cassete em casa; Possui rádio em casa; Possui computador; Possui automóvel; Possui máquina de lavar; Possui geladeira; Possui freezer; Possui telefone fixo; Possui telefone celular; Possui acesso à internet em casa; Possui tv por assinatura; Possui aspirador de pó; Possui empregada mensalista; Quantos Banheiros possui em casa. |

**Fonte:** José E. V. Borges (2019).

## 4.2 Experimento 1: Processo de Análise de Seleção de Atributos

O primeiro experimento proposto nesta metodologia baseia-se em um dos objetivos do trabalho: o conhecimento acerca dos fatores, ou atributos, que mais influenciam no sucesso dos alunos. Para este objetivo, foi utilizada uma metodologia semelhante à utilizada por Marquez Vera MÁRQUEZ et al., (2016) porém com pequenas alterações em seu procedimento. Através da utilização de 10 algoritmos de seleção de atributos, os mesmos utilizados por Marquez Vera, obtém-se como resultado de cada algoritmo um subconjunto de atributos considerados relevantes. Aqueles atributos que foram considerados com relevantes por 6 ou mais algoritmos dentre os 10, ou seja, mais de 50%, são considerados para fins desta pesquisa como influentes no sucesso dos alunos. Tal metodologia permite que os atributos analisados sejam avaliados por mais de um algoritmo de seleção, apontando dessa forma aqueles que foram selecionados com maior frequência, evitando assim o risco de direcionar o resultado da seleção a apenas um algoritmo. Apesar deste processo de seleção ser semelhante ao adotado pelo Marquez Vera, inclusive os algoritmos de seleção são exatamente os mesmos, existe uma diferença marcante

no que diz respeito à regra de aceitação ou exclusão do atributo: na pesquisa do Marquez Vera, são mantidos e considerados relevantes os atributos que foram selecionados por pelo menos 2 algoritmos, diferentemente da regra adotada nesta pesquisa, que considera atributos selecionados por 6 ou mais algoritmos. Esta adaptação na regra de aceitação do atributo para o subconjunto foi realizada no intuito de refinar um pouco mais a seleção, ou seja, aplicando-se um critério mais restritivo, o subconjunto selecionado de atributos no processo se torna menor, reduzindo mais significativamente a dimensionalidade da base de dados. Acredita-se que dessa forma haja uma melhoria no processo de seleção, através da redução a um conjunto menor e mais fortemente relevante de atributos.

Conforme visto no capítulo 2, existem diversas maneiras de se selecionar atributos, inclusive existem diversos algoritmos já implementados no Weka já prontos para executar tal função. Os algoritmos selecionados para serem utilizados nesta pesquisa foram os mesmos utilizados por Marquez Vera em MÁRQUEZ; Morales e Soto (2013) e estão descritos de forma sucinta conforme WITTEN; Ian e Hall, (2011); Bouckaert *et al.*, (2010); Frank, Hall e Witten (2010, p. 128) no Quadro 2. Busca-se basicamente com este experimento encontrar a resposta para a questão de pesquisa 1. A Figura 3 descreve o processo metodológico do Experimento 1.

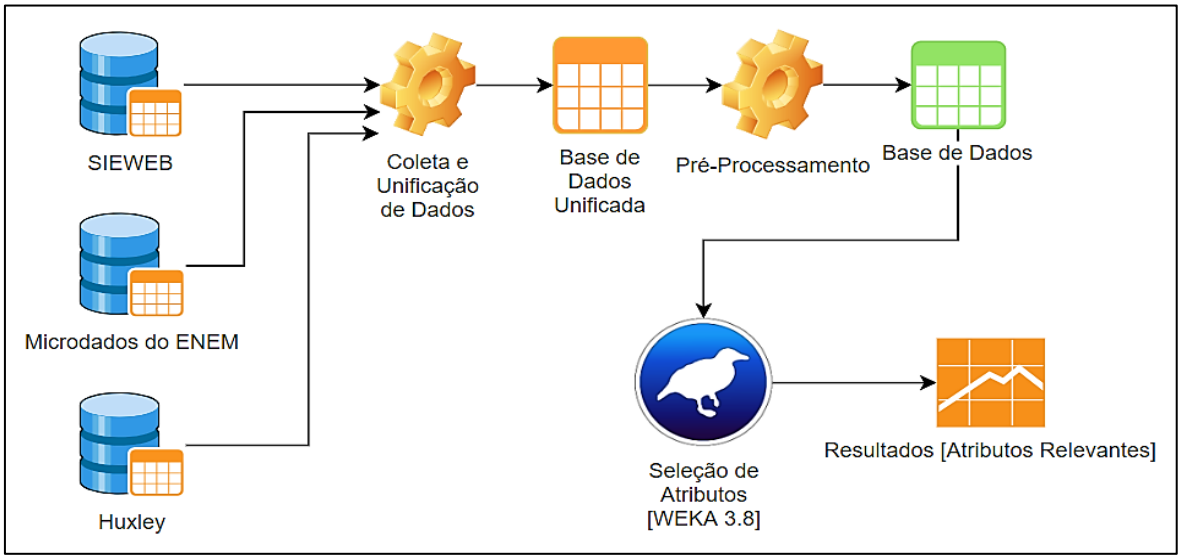
**Quadro 2**-Descrição dos Algoritmos de Seleção de Atributos adotados na pesquisa e implementados no Weka.

| **Algoritmo** | **Função** |
|---|---|
| CfsSubsetEval | Considera o valor preditivo de cada atributo individualmente, juntamente com o grau de redundância entre eles |
| ChiSquared-AttributeEval. | Calcular qui-quadrado estatístico de cada atributo em relação à classe |
| Consistency-SubsetEval | Projeta um grupo de trino dentro do conjunto de atributos e mede a consistência nos valores de classe |
| Filtered-AttributeEval | Aplica um avaliador de atributo aos dados filtrados. Versão de atributo único do filtro baseadas em subconjunto |
| FilteredSubsetEval | Aplica um filtro aos dados de treinamento antes que a seleção de atributos seja executada |

| | |
|---|---|
| GainRatio-AttributeEval | Avalia atributos medindo sua razão de ganho em relação à classe. |
| InfoGain-AttributeEval | Avalia atributos medindo seu ganho de informação com relação à classe. |
| OneRAttributeEval | Usa a medida de precisão (acurácia) simples adotada pelo classificador OneR |
| ReliefFAttributeEval | Baseado em instância: faz amostragem de instâncias aleatoriamente e verifica instâncias vizinhas das mesmas e diferentes classes |
| SymmetricalUncertAttributeEval | Avalia atributo com base na incerteza simétrica |

**Fonte:** José E. V. Borges (2019).

**Figura 4 -** Processo de Análise de Seleção de Atributos.



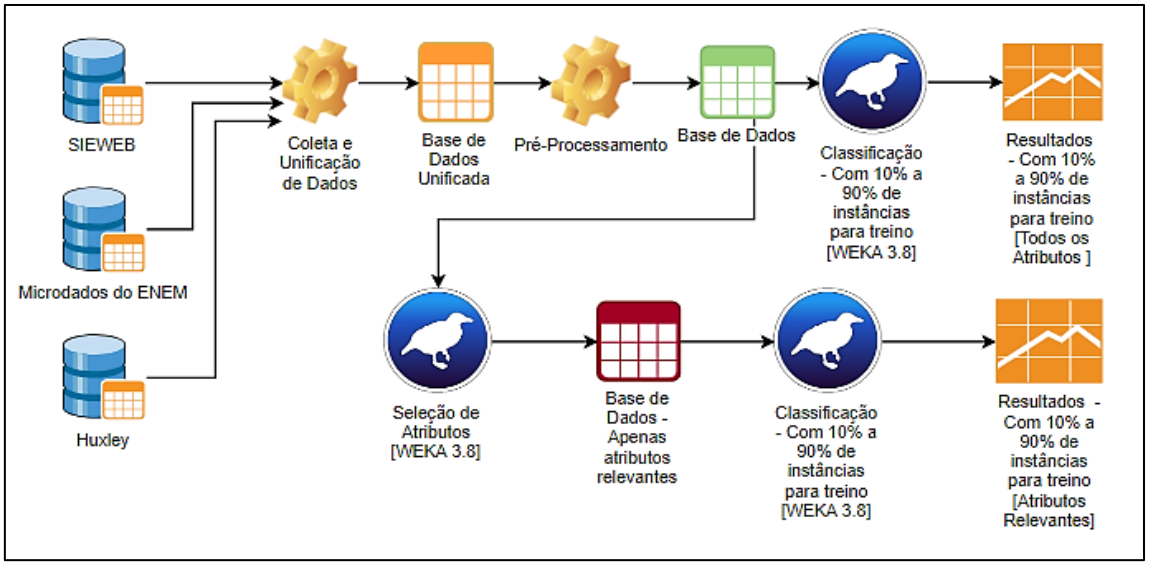

**Fonte:** José E. V. Borges (2019).

## 4.3 Experimento 2: Processo de Análise da Eficiência dos Algoritmos de predição com diferentes parcelas de treino/teste

O objetivo do Experimento 2 é, conforme proposta de pesquisa, analisar a eficiência de cada algoritmo de classificação quando utilizados na predição de sucesso de alunos. Em outras palavras, busca-se responder à questão de pesquisa 2. Como o termo Eficiência, conceituado anteriormente, significa em outras palavras “fazer mais com menos”, a metodologia de medição

da eficiência neste experimento pretende mostrar uma análise de quais algoritmos possuem melhor desempenho, ou acurácia, com menos dados de treino. Para isso, foram utilizados 14 algoritmos de classificação, dentre eles 10 do tipo Caixa Branca e 4 Caixa Preta, para realizar a classificação dos alunos, utilizando, de forma progressiva, uma parcela de 10% a 90% do total de instâncias da base para treino, aprendizagem e criação do modelo, e respectivamente de 90% a 10% da base para teste e classificação. Os Experimentos do 2 ao 4, foram realizados simultaneamente utilizando os dados da base sem alterações após o pré-processamento e utilizando a base de dados após processo de seleção de atributos, onde apenas os atributos mais relevantes foram mantidos, com o objetivo de comparar os desempenhos. Vale lembrar também que nos Experimentos 2 e 3, foram utilizados apenas os atributos dos alunos condizentes com o seu momento de entrada no curso, não considerando, por exemplo, atributos como "Nota da Atividade '" ou "Nota da Prova Final", simulando desta forma uma situação de predição real, onde o professor dispõe apenas de dados passados dos alunos, como os socioeconômicos e notas do ENEM, por exemplo. No Experimento 4, por outro lado, as notas das atividades realizadas pelos alunos durante o curso são consideradas, porém inseridas quinzenalmente permitindo uma análise temporal. A Figura 4 ilustra o processo, que é semelhante aos Experimentos 3 e 4, com a alteração apenas na abordagem da execução do algoritmo de classificação.

**Figura 5** - Processo de Análise da Eficiência dos Algoritmos de predição com diferentes parcelas de treino/teste.




**Fonte:** José E. V. Borges (2009).

No processo de Classificação realizado nos Experimentos de 2 a 4 foram utilizados os mesmos algoritmos de regras e árvore de decisão apresentados também por Marquez Vera em MÁRQUEZ; Morales e Soto (2013), e foram adicionados 4 algoritmos do tipo "caixa preta"

conhecidos: *SVM, Naive Bayes, Multilayer Perceptron e k-NN*. A relação dos algoritmos utilizados e suas respectivas características descritas com maior riqueza de detalhes por WITTEN; Ian e Hall, (2011); Bouckaert et al., (2010); Ghatak, (2010) encontram-se no Quadro 3. As quatro métricas selecionadas para a apresentação dos resultados no artigo do Marquez Vera também foram mantidos nesta pesquisa, são elas: *True Positive Rate* (TPR), *True Negative Rate* (TNR), *Accuracy* (Acc) e *Geometric Mean* (GM), muito embora a Acurácia seja aqui tratada de forma predominante.

**Quadro 3**-Algoritmos de Classificação do Weka abordados.

| **Nome do algoritmo implementado no Weka** | **Função** | **Categoria** | **Método de Busca** | **Tipo** |
|---|---|---|---|---|
| JRip | implementa *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER), incluindo a otimização global heurística do conjunto de regras. | Regras | Baseado em Procura | Caixa Branca |
| NNge | *nearest-neighbor method*, é um método de vizinho mais próximo para gerar regras usando exemplares generalizados não aninhados | | | |
| OneR | é o classificador 1R com um parâmetro – o tamanho mínimo do *bucket* para discretização. | | | |
| Prism | implementa o algoritmo de cobertura elementar para regras | | | |
| Ridor | *ripple-down rule learner* (Ridor) aprende regras com exceções gerando a regra padrão, usando a redução incremental de erro reduzido para localizar exceções com a menor taxa de erro, encontrando as melhores exceções para cada exceção e iterando. | | | |
| ADTree | *alternating decision trees*, constrói uma árvore de decisão alternativa para problemas de duas classes usando *boosting* | Árvore de Decisão | | |
| J48 | utiliza C4.5 árvore de decisão para aprendizagem. Implementa o C4.5 revisão 8. | | | |
| RandomTree | constrói uma árvore que considera um dado número de características aleatórias a cada nó. | | | |
| REPTree | *reduced-error pruning*. constrói uma árvore de decisão ou regressão usando a redução de ganho / variação de | | | |

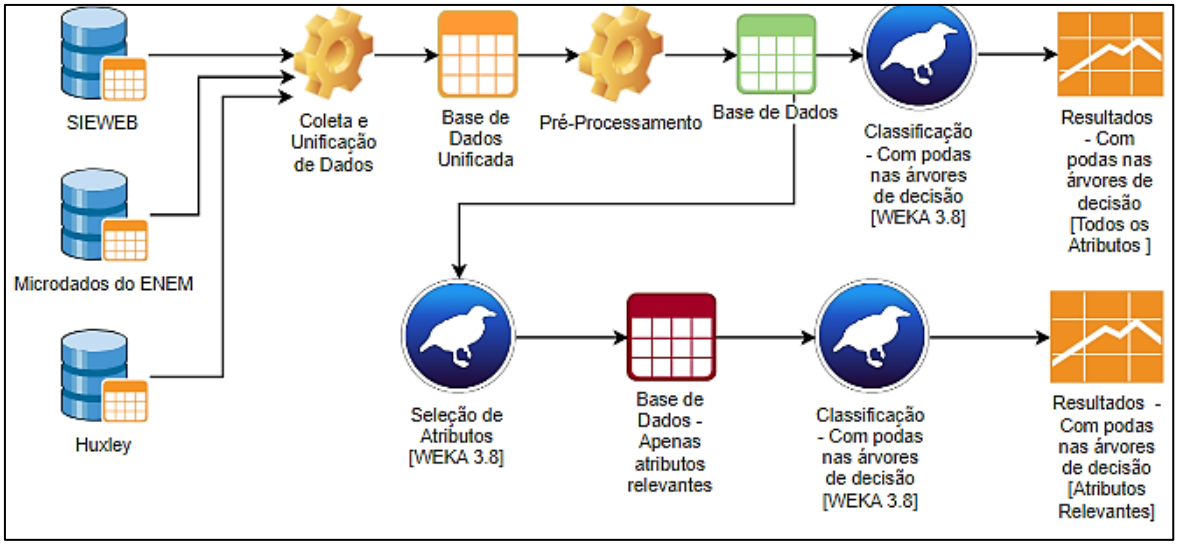
| | | | | |
|---|---|---|---|---|
| | informação e remove-a usando a podagem de erro reduzido. | | | |
| SimpleCart | árvore de decisão para classificação que emprega a estratégia de remoção de complexidade de custo mínimo CART (*classification and regression tree*) | | | |
| LibSVM | biblioteca que inclui classificadores do tipo *wrapper* que permitem implementações de máquinas de vetores de suporte (SVM) e regressão logística de terceiros para o Weka. | SVM | Baseado em Otimização | Caixa Preta |
| Naive Bayes | implementa o classificador probabilístico Naïve Bayes padrão | Naive Bayes | Probabilístico | |
| MultilayerPerceptron | são uma rede neural que treina usando "retropropagação" (*backpropagation*) | Rede Neural | Baseado em Otimização | |
| iBk | *k-nearest-neighbor* (k-NN), é um classificador de k-vizinhos mais próximos. | k-NN | Baseado em Distância | |

**Fonte:** José E. V. Borges (2019).

## 4.4 Experimento 3: Processo de Análise da Eficiência dos Algoritmos de Classificação por Árvore de Decisão

Neste experimento, o foco está voltado à eficiência dos algoritmos de Árvore de Decisão. Conforme proposto inicialmente, o objetivo é analisar os 5 algoritmos de árvore de decisão, *ADTree, J48, RandomTree, REPTree, SimpleCart*, e como eles se comportam quando possuem alterações de profundidade, podas, nas árvores criadas nos modelos. Como a maioria das árvores apresentadas como resultado do processo de aprendizagem realizados neste experimento possuíam profundidade máxima igual a 5, quando não era aplicada nenhuma poda, ou seja, não havia limitação aplicada à profundidade das árvores, optou-se por medir a eficiência de cada algoritmo com árvores reduzidas às profundidades: 1, 2, 3, 4, 5, com podas, e também sem podas. Nas configurações do teste, no Weka, foi utilizada Validação Cruzada, com 10 *folds*, parâmetro bastante utilizado nos estudos de EDM para configurar a forma de aprendizagem do algoritmo. Este experimento busca responder à questão de pesquisa 3. A Figura 5 ilustra o processo para análise de eficiência dos algoritmos de árvores.

**Figura 6** - Processo de Análise da Eficiência dos Algoritmos de Classificação por Árvore de Decisão.

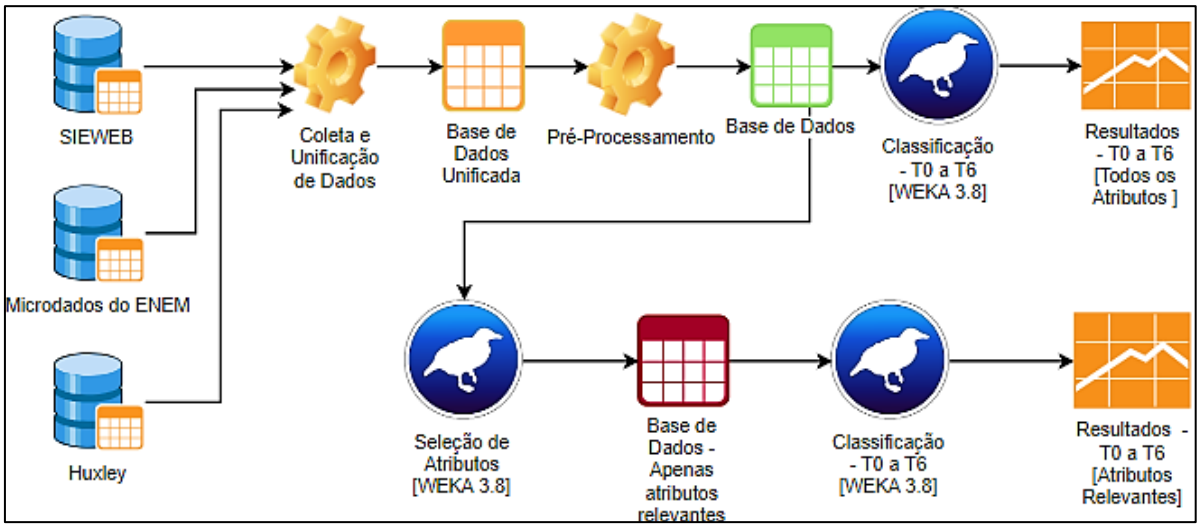



**Fonte:** José E. V. Borges (2019).

## 4.5 Experimento 4: Processo de análise da Eficiência dos algoritmos de classificação no decorrer do curso.

No Experimento 4 temos uma análise de eficiência temporal relacionada a um dos objetivos propostos: mensurar a eficiência de cada um dos algoritmos de classificação quando, no decorrer do curso, são inseridas novas informações e respeito dos alunos, neste caso especificamente as notas dos alunos nas atividades de aprendizagem de programação realizadas na plataforma Huxley. A Figura 6 mostra o processo do experimento.

**Figura 7** - Processo de análise da Eficiência dos algoritmos de classificação no decorrer do curso.

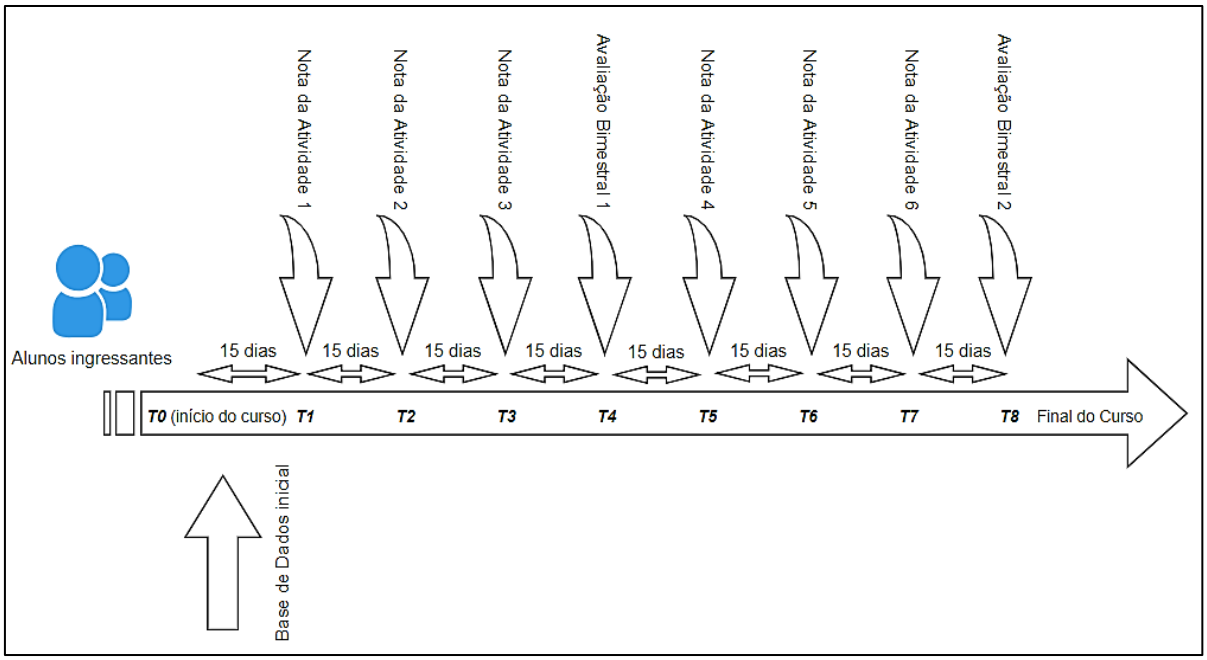



**Fonte:** José E. V. Borges (2019).

Busca-se, em outras palavras, comparar a acurácia de cada algoritmo quando são incluídos como dados de entrada apenas informações do aluno obtidas no momento T0, ou seja,

no momento em que o aluno inicia a disciplina. Neste momento sabe-se pouco sobre o aluno, praticamente apenas dados socioeconômicos e pessoais, porém no decorrer do curso mais informações são obtidas sobre o aluno, que são as notas do desempenho de cada aluno nas atividades aplicadas. Acredita-se obter um aumento da acurácia na predição dos algoritmos com a inserção dos dados das atividades. Para a nossa amostra de dados, extraída da plataforma Huxley, foram aplicadas atividades quinzenais. Na Figura 7 está ilustrado em um diagrama temporal, em quais momentos são inseridas as novas parcelas de dados contendo as notas dos alunos nas atividades de programação. Vale ressaltar que para a execução dos algoritmos de classificação no Weka, nas opções de teste, também foi utilizada Validação Cruzada, com 10 *folds*, assim como no Experimento 3, parâmetro também utilizado na maioria dos trabalhos de referência.

**Figura 8** - Diagrama da análise temporal com dados das Notas dos alunos no Huxley.




**Fonte:** José E. V. Borges (2019).

Pode-se dizer que o Experimento 4 busca encontrar as respostas para as questões propostas, derivadas da questão de pesquisa 4, apresentadas no capítulo 1.1 e que estão relacionadas à análise exploratória sob o aspecto temporal nas predições. São elas:

*Questão 4.1)* Em que tempo, a partir da aplicação de atividades de aprendizagem durante o curso, é possível obter níveis de desempenho satisfatórios na predição do sucesso?

*Questão 4.2)* Quais os Algoritmos recomendados para Predição em T0?

*Questão 4.3)* Quais os Algoritmos recomendados para Predição durante o curso?

*Questão 4.4)* Quais os Dados recomendados a serem utilizados em predições em T0?

*Questão 4.5)* Quais os Dados recomendados a serem utilizados em predições durante o curso?

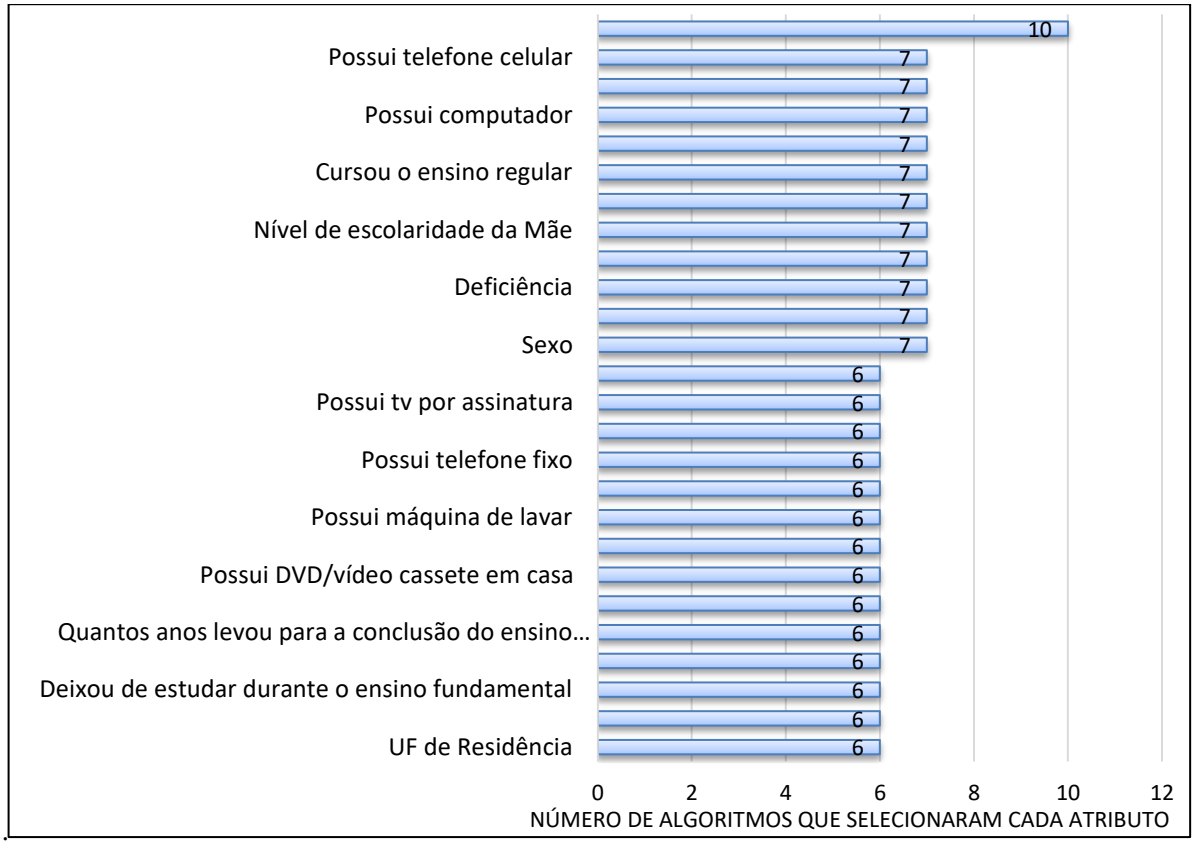
# 5 RESULTADOS E DISCUSSÕES

Neste capítulo serão apresentados os resultados dos experimentos após a execução de todos os processos conforme metodologia proposta e descrita no capítulo anterior.

## 6.1 Questão 1. Quais as características mais relevantes dos alunos para a predição do seu sucesso acadêmico, buscando-se eficiência voltada para a qualidade dos dados?

Uma vez executados os 10 algoritmos de seleção de atributos, conforme metodologia proposta, foram selecionados 10 subconjuntos contendo os atributos apontados como relevantes por cada algoritmo, e em seguida foi feita a análise de frequência de cada atributo dentro dos subconjuntos. Aqueles atributos que foram apontados como relevantes por 6 ou mais atributos, ou seja, mais de 50% dos algoritmos o selecionaram, foram considerados significativamente relevantes para o estudo. Considera-se consequentemente que estes atributos, por serem mais relevantes para a predição do sucesso do aluno, também influenciam mais fortemente da eficiência da predição dos algoritmos classificadores. A Figura 8 mostra graficamente os atributos mais relevantes por ranking.

**Figura 9** - Atributos considerados como mais relevantes pelos algoritmos de Seleção de Atributos




**Fonte:** José E. V. Borges (2019).

Algumas considerações importantes podem ser feitas a partir dos resultados deste experimento, uma delas é sobre o atributo mais relevante dentre todos, Município de Residência. Nota-se que este atributo foi selecionado como relevante por todos os 10 algoritmos de seleção, o que nos leva a concluir que há uma relação muito forte entre a aprovação dos alunos e o município onde aquele aluno reside. Um ponto importante a frisar é que os algoritmos de seleção de atributos sempre buscam, através de técnicas de aprendizagem, as relações existentes nos dados entre todos os atributos e o atributo classificador, que neste caso é Conceito, onde existe a informação de se o aluno foi realmente aprovado ou não na disciplina. Temos como concluir, com base no atributo mais relevante, que de fato o município onde o aluno reside interfere bastante na sua chance de aprovação, podemos relacionar isso ao fato da necessidade de se deslocar muitas horas por dia de casa até a instituição, para alguns alunos.

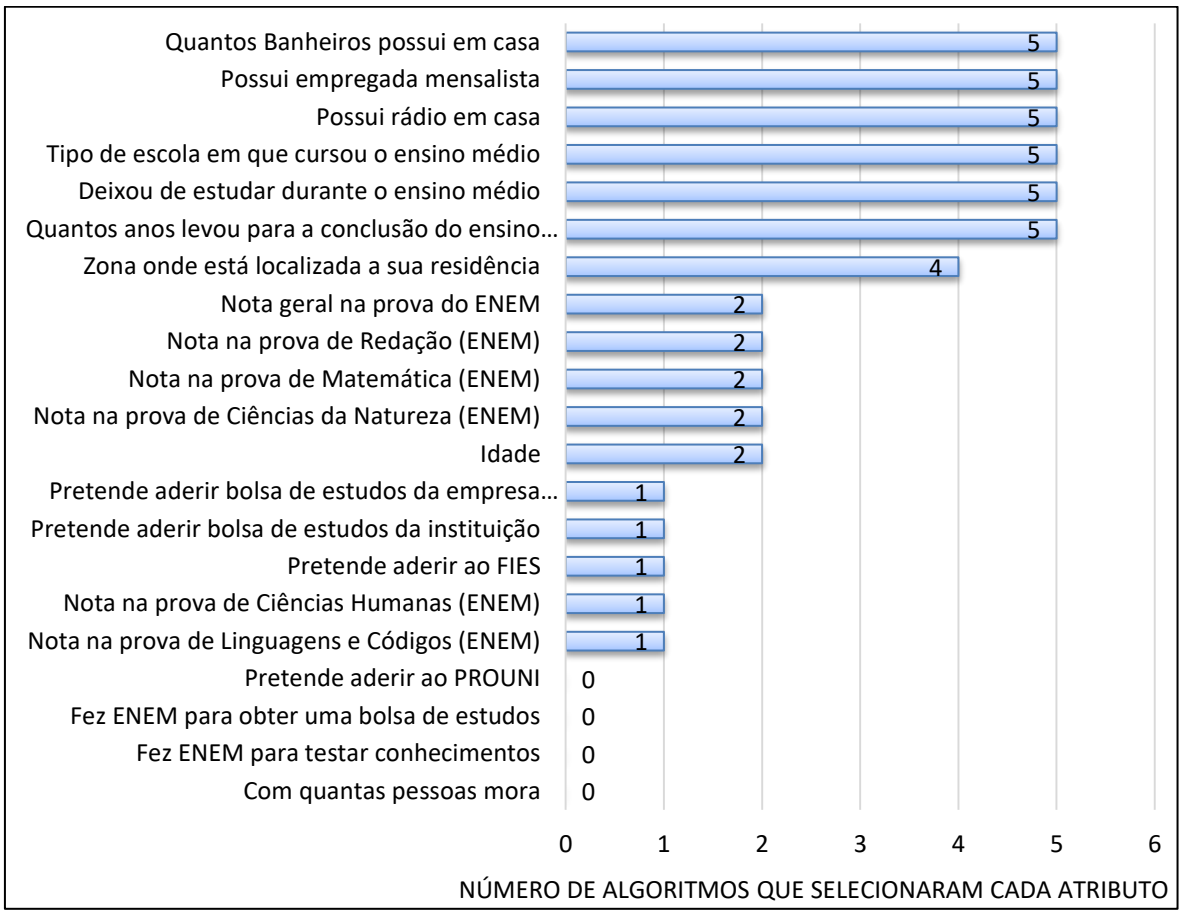
Os resultados mostram também que alguns dos principais atributos socioeconômicos, como por exemplo “Possui Celular”, “Possui Computador” ou “Renda Familiar”, também aparecem dentre os mais relevantes para o sucesso acadêmico. É notório que a classe social do aluno o privilegia em relação ao desempenho acadêmico, neste caso mais precisamente no desempenho em programação. Tal fato apesar de ser esperado pôde ser confirmado através da técnica de seleção.

Dentre os atributos considerados relevantes em nível intermediário, podemos considerar como sendo aqueles que foram apontados por exatamente 6 algoritmos, encontramos dois que estão conceitualmente relacionados por representarem o acesso à informação do aluno, são eles: “Possui acesso à internet” e “Possui TV por assinatura”. Nota-se que o acesso à informação é apontado como um fator relevante para o sucesso acadêmico do aluno na disciplina. Vale ressaltar que no Experimento 1 apenas foram utilizados como dados de entrada os atributos dos alunos relacionados ao momento em que ele inicia o curso, ou seja, dados passados dos alunos.

A Figura 9, por sua vez, relaciona o ranking dos atributos menos significativos para a aprovação dos alunos em Programação I. Vemos que as pontuações dos alunos na Prova do ENEM não são consideradas relevantes, pois todos os atributos de nota no ENEM se encontram com frequência menor ou igual a 5 nas seleções. Este é um resultado curioso e que de certa forma um fator surpresa apresentado por este experimento, uma vez que é esperado que alunos que obtiveram boas notas na prova do ENEM sejam mais preparados e consequentemente sejam mais propensos à aprovação. Além disso, os 4 atributos considerados como menos relevantes em igual peso, ou seja, não foram apontados como relevantes por nenhum dos 10 algoritmos, estão relacionados com a intenção do aluno para o ENEM, intenção para bolsa de estudos e com quantas pessoas o aluno reside.

Em resposta à questão de pesquisa 1, conclui-se que atributos relativos à localização da residência do aluno, juntamente com os atributos relativos à classe social, direta ou indiretamente, além da questão acesso à informação, foram as características apresentadas como mais relevantes para a eficiência na predição do sucesso acadêmico de alunos de programação.

**Figura 10** - Atributos Considerados menos relevantes para o sucesso acadêmico em programação introdutória.




**Fonte:** José E. V. Borges (2019).

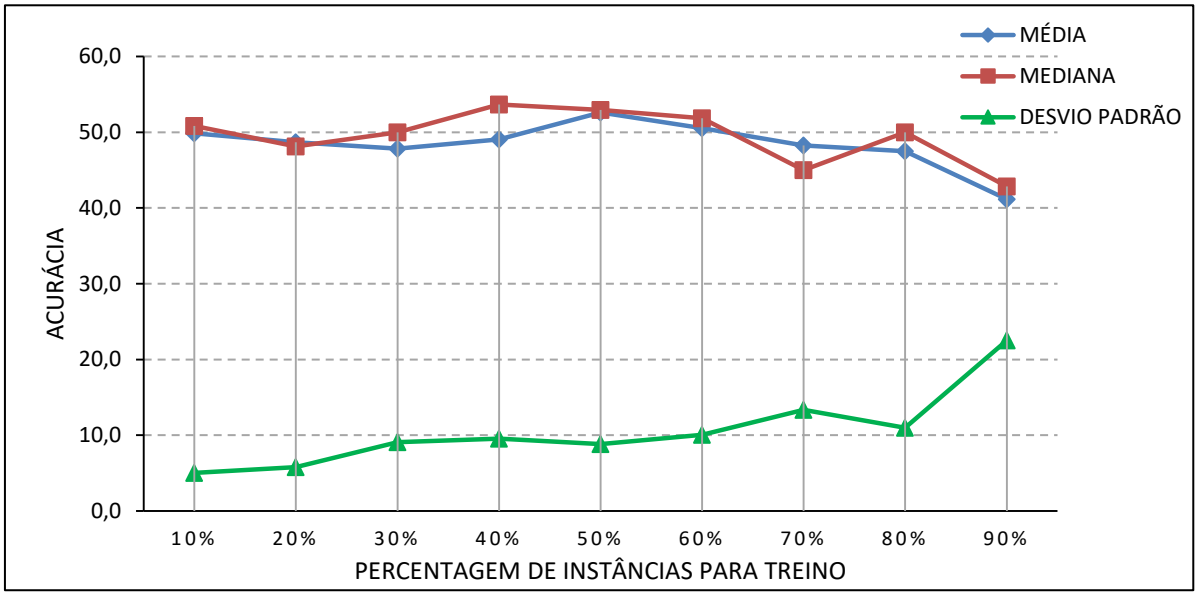
## 6.2 Questão 2. Quão eficientes são os algoritmos de classificação aplicados na predição do sucesso de alunos de programação introdutória no momento de início do curso, considerando-se os percentuais de instâncias utilizadas para treino e teste?

Neste Experimento 2 o objetivo é responder à questão proposta sobre a eficiência dos algoritmos de classificação quando utilizamos menos dados para treino e mais dados para teste, e vice-versa, no processo de classificação para predição. A métrica principal utilizada para a comparação da eficiência dos algoritmos de classificação foi a Acurácia, encontrada nos trabalhos de referência da área. A Acurácia caracteriza-se por mensurar, para este caso

especificamente, quantos alunos foram classificados corretamente, sejam eles aprovados ou reprovados, sobre o número total de alunos utilizados no teste do modelo criado pelo algoritmo. Como o principal interesse neste estudo é identificar aqueles alunos que correm risco de ser reprovados, a métrica Recall, também conhecida como Sensibilidade ou revocação, representa uma medida importante para o estudo pois direciona à medição do nível de acerto relativo a uma das classes, que neste caso seria a classe dos alunos Reprovados, em outras palavras, trata da razão entre os elementos TP e os (TP + FN), onde TP + FN representam todos os alunos que realmente estão em direção ao fracasso.

Primeiramente, foi realizada uma análise geral com todos os 14 algoritmos de classificação, mensurando média, mediana e desvio padrão para cada uma das 9 execuções, ou seja, a classificação foi executada com cada algoritmo segmentando o total de instâncias 10% para treino e 90% para teste, 20% para treino e 80% para teste, e assim sucessivamente até chegar a 90% de treino para 10% apenas para teste. Os resultados são apresentados na Figura 10. Vale lembrar que neste experimento e também no Experimento 3, são considerados apenas os atributos do aluno disponíveis no instante T0, ou seja, no momento de início do curso: dados socioeconômicos e pessoais, extraídos da fonte de dados ENEM.

**Figura 11** - Média, Mediana e Desvio Padrão das Acurácias de todos os Algoritmos de Classificação relacionados com o percentual de instâncias de Treino utilizado.




**Fonte:** José E. V. Borges (2019).

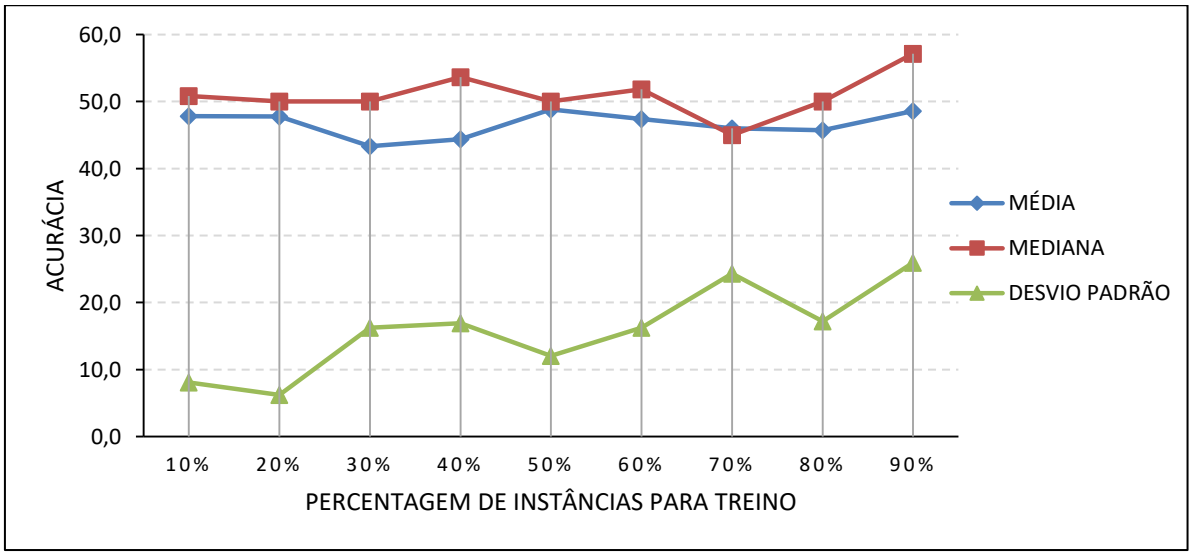
Percebemos pela Figura 10 que a média e mediana dentre as acurácias dos 14 algoritmos classificadores, apresentaram uma leve queda geral ao chegar ao nível de 90% para treino e 10% para teste. Ao mesmo tempo, o desvio padrão aponta uma maior variabilidade nos

resultados dos classificadores quando utilizando parcelas menores de teste, isso se dá devido ao peso percentual de cada erro ou acerto na classificação quando se possui poucas instâncias a testar. Justifica-se utilizar as medidas de média, mediana e desvio padrão nas exibições dos resultados, tanto em gráfico como em tabelas, para que se possa, primariamente, perceber as tendências gerais de desempenhos dos grupos de algoritmos, além da presença ou não de valores discrepantes, apresentados pela mediana, e ainda o nível de dispersão ou variabilidade dos resultados, perceptível a partir do desvio-padrão.

Uma outra análise que este experimento possui é quanto ao comportamento dos algoritmos por categorias: Caixa Branca e Caixa Preta. Cabe também testar o desempenho comparativamente entre os algoritmos de Regras contra os de Árvore de Decisão. Podemos explorar tais análises através das Figuras 11, 12 e 13.

Quando a análise se restringe apenas aos algoritmos de Caixa Branca, de Regras, não é possível observar uma tendência geral de acurácias com mais ou menos instâncias para treino. Apesar das amostras dos resultados apresentarem uma oscilação, não mostram crescimento ou redução gradativos. Apenas o desvio padrão segue um aumento gradativo pelos mesmos motivos discutidos anteriormente para o gráfico da Figura 10. Isso pode ser observado através do gráfico da Figura 11.

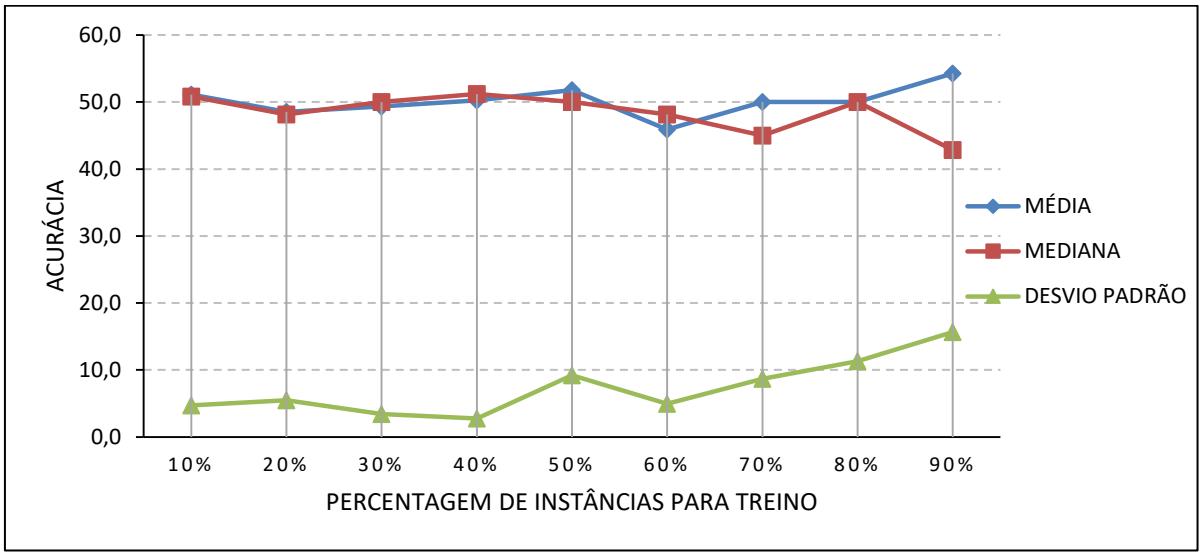
**Figura 12** - Média, Mediana e Desvio Padrão dos Algoritmos de Classificação do Tipo Caixa Branca - Regras relacionados com o percentual de instâncias de Treino utilizado.




**Fonte:** José E. V. Borges (2019).

Os algoritmos de Caixa Branca, Árvores de Decisão, também não apresentaram tendência perceptível quando analisados separadamente em gráfico, conforme podemos visualizar na Figura 12.

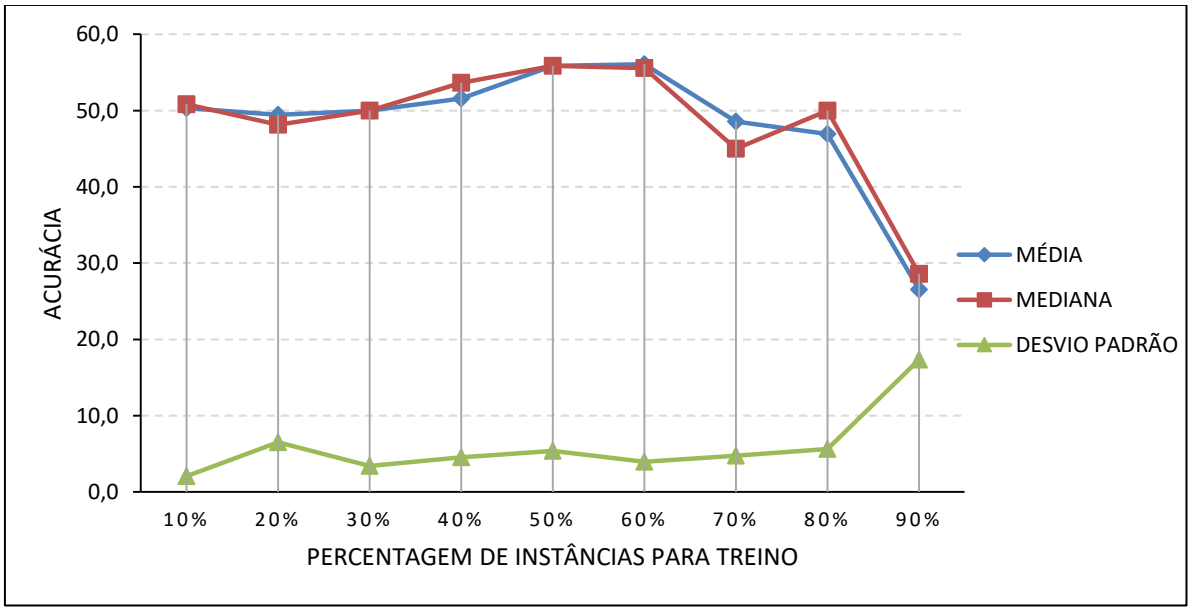
**Figura 13** - Média, Mediana e Desvio Padrão dos Algoritmos de Classificação do Tipo Caixa Branca - Árvores de Decisão relacionados com o percentual de instâncias de Treino utilizado.




**Fonte:** José E. V. Borges (2019).

Quando agrupamos os resultados das acurácias dos algoritmos de Caixa Preta, percebemos que houve uma queda significativa geral quando utilizando parcelas de poucas instâncias para teste e parcelas maiores para treino. Podemos visualizar através da Figura 13.

**Figura 14** - Média, Mediana e Desvio Padrão dos Algoritmos de Classificação do Tipo Caixa Preta relacionados com o percentual de instâncias de Treino utilizado.




**Fonte:** José E. V. Borges (2019).

Nas Tabelas 1 e 2, estão listados todos os resultados das acurácias individuais de cada algoritmo para o Experimento 2 com e sem seleção de atributos. A visualização em números

facilita a interpretação das tendências de cada algoritmo individualmente. Ao mesmo tempo, nas Figuras de 14 a 17 estão apresentados os gráficos resultantes da execução dos algoritmos de classificação com a Seleção de Atributos.

Na Tabela 1 encontram-se os resultados das acurácias quando utilizados todos os atributos da base de dados principal, já na Tabela 2 temos os resultados das acurácias quando apenas os atributos selecionados através do Experimento 1, considerados como mais relevantes, são utilizados. Estão destacados nas 2 tabelas os valores máximos e mínimos, em linha e em coluna. Observa-se uma concentração dos piores resultados de acurácias, em vermelho, nas execuções com 80% e 90% para os algoritmos de Árvores de Decisão e também para os do tipo Caixa Preta. Pode-se constatar também que o algoritmo que apresentou pior eficiência geral foi o Prisma, chegando a uma média de 22% dentre todos os resultados obtidos em suas execuções com as 9 parcelas de instâncias para treino. É importante ressaltar que o Prisma possui algumas características peculiares observadas através da matriz de confusão gerada nos experimentos: 1) Tal algoritmo sofre muito com alta dimensionalidade, ou seja, não lida muito bem com muitos atributos. 2) Possui a limitação de só trabalhar com dados categóricos, o que torna necessária a discretização dos dados numéricos de entrada antes de sua execução. 3) Sua acurácia é comprometida pelo fato de que ele deixa algumas de fora durante a sua execução, ou seja, instâncias muito próximas ao limiar de classificação de seu modelo não são classificadas nem como aprovado nem reprovado, o que consequentemente conta negativamente para o cálculo de acurácia, uma vez que são considerados com instâncias para o total de instâncias de entrada, porém não são consideradas nem como TP nem TN. A Discretização dos dados para permitir a execução deste algoritmo foi realizada diretamente no Weka a partir do filtro já implementado *NumericToDecimal*.

**Tabela 1**-Acurácia dos Algoritmos de Classificação com parcelas de treino de 10% a 90%.

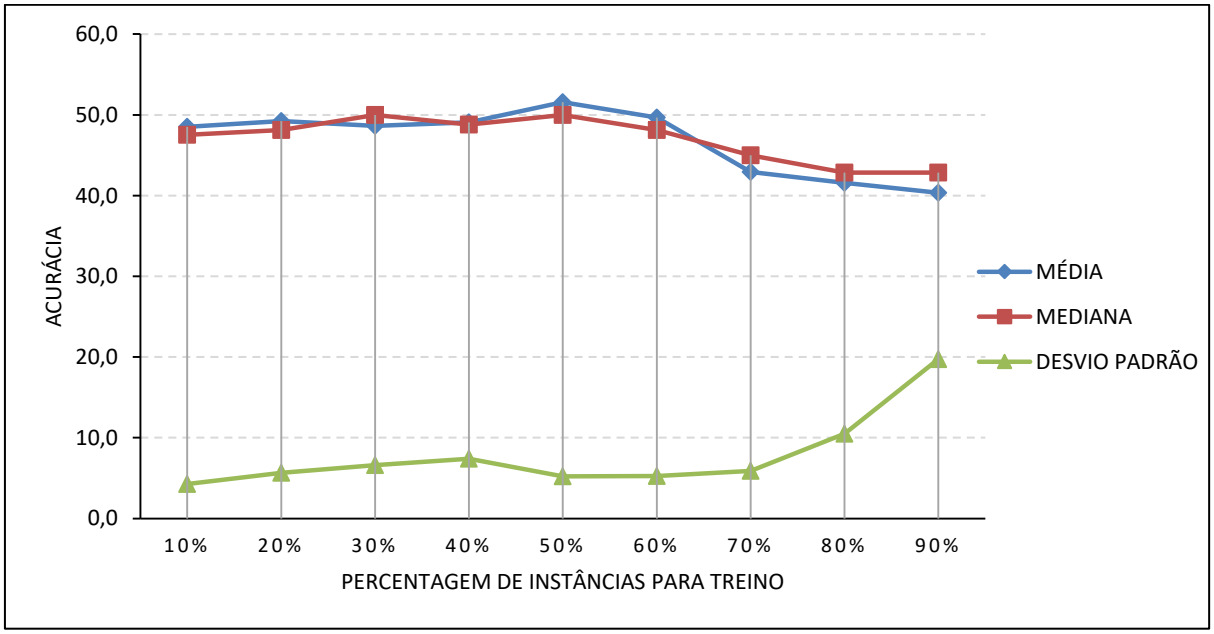
| **Algoritmo** | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | 50,8 | 48,1 | **47,9** | 53,7 | 58,8 | 51,9 | 60,0 | 64,3 | **71,4** | **56,3** | **53,7** |
| NNge | 55,7 | 51,9 | 50,0 | 53,7 | 58,8 | **66,7** | 40,0 | 35,7 | **28,6** | **49,0** | **51,9** |
| OneR | 50,8 | 51,9 | 54,2 | 53,7 | 50,0 | **44,4** | 45,0 | 57,1 | **71,4** | **53,2** | **51,9** |
| Prism | 34,4 | **37,0** | 14,6 | 14,6 | 29,4 | 22,2 | **10,0** | 21,4 | 14,3 | **22,0** | **21,4** |
| Ridor | 47,5 | 50,0 | 50,0 | **46,3** | 47,1 | 51,9 | **75,0** | 50,0 | 57,1 | **52,8** | **50,0** |
| ADTree | 50,8 | 46,3 | 48,8 | 51,2 | 67,6 | 48,1 | 45,0 | **35,7** | **71,4** | **51,7** | **48,8** |
| J48 | 50,8 | 48,1 | 43,8 | 48,8 | 44,1 | 48,1 | **50,0** | 50,0 | **42,9** | **47,4** | **48,1** |
| RandomTree | 59,0 | 42,6 | 50,0 | 51,2 | 50,0 | **37,0** | 65,0 | 57,1 | **71,4** | **53,7** | **51,2** |
| REPTree | 47,5 | 48,1 | **52,1** | 46,3 | 47,1 | 48,1 | 45,0 | **42,9** | **42,9** | **46,7** | **47,1** |
| SimpleCart | 47,5 | 57,4 | 52,1 | 53,7 | 50,0 | 48,1 | 45,0 | **64,3** | **42,9** | **51,2** | **50,0** |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SVM - kernel:linear | 50,8 | 57,4 | 50,0 | 53,7 | 55,9 | **59,3** | 45,0 | 50,0 | **14,3** | **48,5** | **50,8** |
| SVM - kernel:polynomial | 52,5 | 57,4 | 52,1 | 53,7 | 55,9 | **59,3** | 55,0 | 42,9 | **14,3** | **49,2** | **53,7** |
| SVM - kernel:RBF | 47,5 | 48,1 | 47,9 | 53,7 | 58,8 | **59,3** | 50,0 | 50,0 | **28,6** | **49,3** | **50,0** |
| SVM - kernel:sigmoid | 47,5 | 48,1 | 47,9 | 46,3 | 47,1 | 48,1 | 45,0 | **50,0** | **42,9** | **47,0** | **47,5** |
| Naive Bayes | 52,5 | 50,0 | 50,0 | **56,1** | 52,9 | 55,6 | 45,0 | 50,0 | **42,9** | **50,5** | **50,0** |
| Rede Neural | 50,8 | 46,3 | **56,3** | 53,7 | 55,9 | 55,6 | 45,0 | 35,7 | **0,0** | **44,3** | **50,8** |
| KNN | 50,8 | **38,8** | 45,8 | 43,9 | **64,7** | 55,6 | 55,0 | 50,0 | 42,9 | **49,7** | **50,0** |
| **MÉDIA** | **49,8** | **48,7** | **47,8** | **49,1** | **52,6** | **50,5** | **48,2** | **47,5** | **41,2** | | |
| **MEDIANA** | **50,8** | **48,1** | **50,0** | **53,7** | **52,9** | **51,9** | **45,0** | **50,0** | **42,9** | | |
| **DESVIO PADRÃO** | **5,03** | **5,78** | **9,07** | **9,56** | **8,82** | **10,05** | **13,34** | **10,99** | **22,52** | | |

**Fonte:** José E. V. Borges (2019).

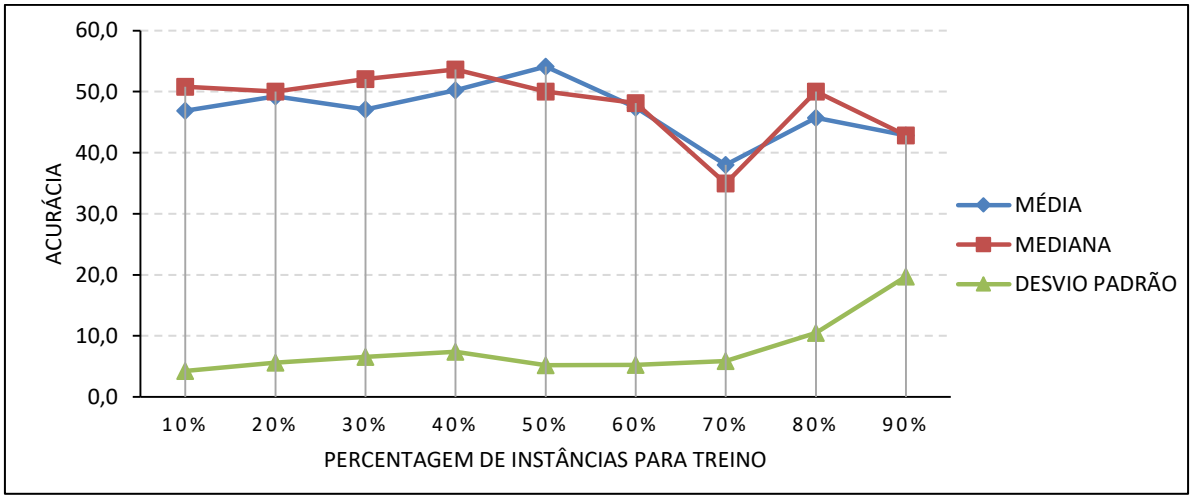
Analisando os resultados da Tabela 2, ondo foram considerados apenas os atributos mais relevantes, observamos que a tendência de redução de acurácia com o aumento de percentual de dados para treino persiste, confirmando o resultado anterior. Neste caso o decréscimo foi ainda mais acentuado, chegando a um valor médio geral próximo aos 40%.

**Figura 15** - Média, Mediana e Desvio Padrão das Acurácias de todos os Algoritmos de Classificação relacionados com o percentual de instâncias de Treino utilizado - com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

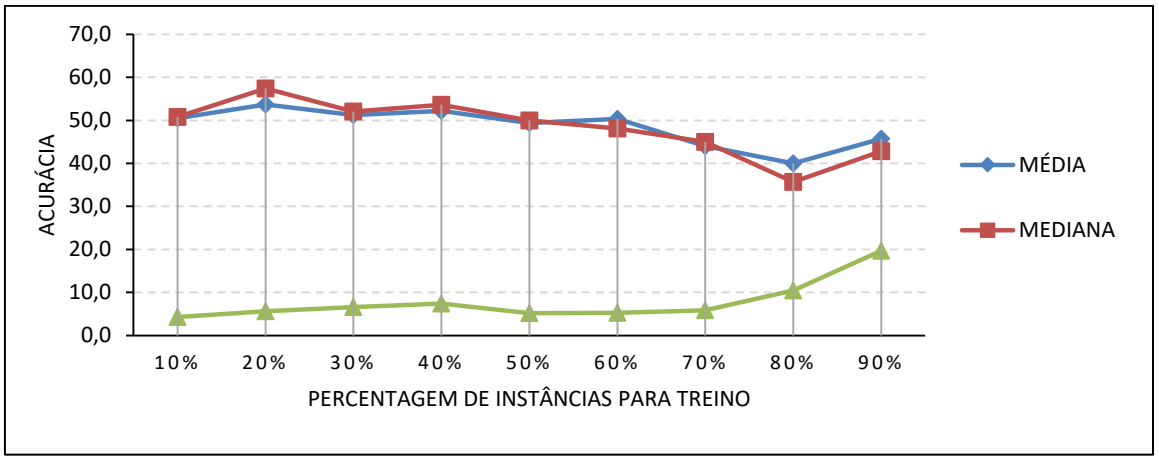
**Figura 16** - Média, Mediana e Desvio Padrão dos Algoritmos de Classificação do Tipo Caixa Branca - Regras relacionados com o percentual de instâncias de Treino utilizado - com Seleção de Atributos.




Fonte: José E. V. Borges (2019).

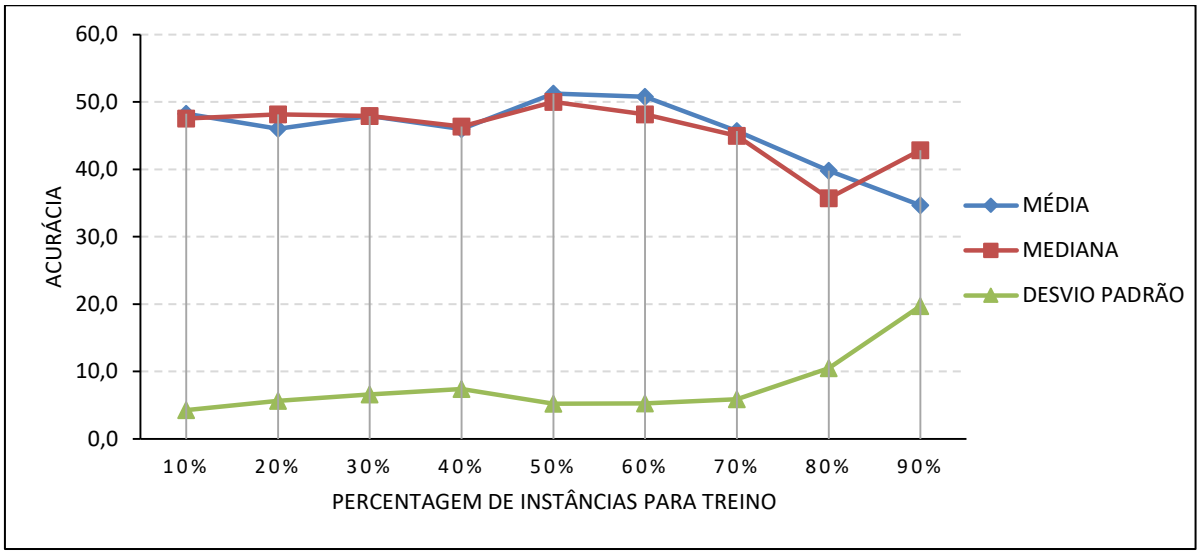
A partir das Figuras 15, 16 e 17, é possível perceber uma tendência que não pôde ser percebida nos gráficos dos resultados sem seleção de atributos: A tendência de decréscimo de acurácia dos algoritmos com o aumento do percentual de treino dessa vez também foi apresentada pelos grupos de algoritmos de Regras e Árvore de Decisão, o que leva a concluir mais seguramente que se trata de uma tendência geral de comportamento. Pode-se então responder à questão de pesquisa 2 a partir desta análise geral que, independente do algoritmo, seu desempenho tende a ser menor quando sujeitos a parcelas grandes de treino, como 80% ou 90%, e parcelas reduzidas de teste.

**Figura 17** - Média, Mediana e Desvio Padrão dos Algoritmos de Classificação do Tipo Caixa Branca – Árvores de Decisão relacionados com o percentual de instâncias de Treino utilizado - com Seleção de Atributos - com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

**Figura 18** - Média, Mediana e Desvio Padrão dos Algoritmos de Classificação do Tipo Caixa Preta relacionados com o percentual de instâncias de Treino utilizado - com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

Outras conclusões podem ser obtidas através de uma análise mais precisa de eficiência na Tabela 2:

1) Os algoritmos que apresentaram maior eficiência foram NNge e SVM – kernel: linear, chegando a 61,8% de acurácia a 50% de dados de treino;
2) As maiores acurácias gerais foram alcançadas pelo Ridor e RandomTree, 71,4% ambos. Em outras palavras, estes seriam os algoritmos mais indicados para a utilização como preditores, mesmo em casos onde há diferença significativa entre o percentual de instâncias para treino e teste. O Ridos ainda se mostra superior para a aplicação pois além de apresentar maior acurácia no geral, também alcança a maior média quando submetido às variações de instâncias de treino e teste;
3) A Seleção de atributos elevou a menor acurácia média, apresentada pelo Prism, de 22% para 38,6%, comparando-se as Tabelas 1 e 2. Este comportamento também foi apresentado nos resultados da pesquisa de Marquez Vera 2013 MÁRQUEZ; Morales e Soto (2013) onde as piores acurácias apresentadas tiveram um acréscimo significativo após a seleção de atributos;
4) As Acurácias máximas e superiores à média não foram afetadas significativamente pela Seleção de Atributos neste caso, resultado semelhante ao apresentado por Marquez Vera. Alguns inclusive apresentaram acurácia inferior após a seleção;

5) A Proximidade entre os valores Média e Mediana, tanto da tabela quanto dos gráficos, em geral, dá a informação de que existem poucos valores discrepantes ou concentrações de amostras de resultados para mais ou para menos, o que torna segura a utilização da média para as conclusões das tendências apresentadas;
6) Resposta à Questão 1: Conforme comportamento geral apresentado, quanto menor o percentual de teste disponibilizado ao algoritmo de Classificação, menor o desempenho da predição.

**Tabela 2**-Acurácia dos Algoritmos de Classificação com parcelas de treino de 10% a 90% - com Seleção de Atributos.

| **Algoritmo** | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | 50,8 | 50,0 | 52,1 | 48,8 | **58,8** | 44,4 | 30,0 | 50,0 | **28,6** | **45,9** | **50,0** |
| NNge | 50,8 | 48,1 | 52,1 | 56,1 | **61,8** | 55,6 | 35,0 | 35,7 | **28,6** | **47,1** | **50,8** |
| OneR | 50,8 | 51,9 | **54,2** | 53,7 | 50,0 | 48,1 | 45,0 | 50,0 | **42,9** | **49,6** | **50,0** |
| Prism | 34,4 | 40,7 | **29,2** | 31,7 | **50,0** | 40,7 | 35,0 | 42,9 | 42,9 | **38,6** | **40,7** |
| Ridor | 47,5 | 55,6 | 47,9 | 61,0 | 50,0 | 48,1 | 45,0 | 50,0 | **71,4** | **52,9** | **50,0** |
| ADTree | 50,8 | 57,4 | 56,3 | **58,5** | 50,0 | 40,7 | 35,0 | 35,7 | **28,6** | **45,9** | **50,0** |
| J48 | 50,8 | 48,1 | **41,7** | 48,8 | 44,1 | **55,6** | 50,0 | 50,0 | 42,9 | **48,0** | **48,8** |
| RandomTree | 55,7 | 57,4 | 54,2 | 53,7 | 55,9 | 59,3 | 40,0 | **21,4** | **71,4** | **52,1** | **55,7** |
| REPTree | 47,5 | 48,1 | **52,1** | 46,3 | 47,1 | 48,1 | 45,0 | **35,7** | 42,9 | **45,9** | **47,1** |
| SimpleCart | 47,5 | **57,4** | 52,1 | 53,7 | 50,0 | 48,1 | 50,0 | 57,1 | **42,9** | **51,0** | **50,0** |
| SVM - kernel:linear | 49,2 | 44,4 | 50,0 | 48,8 | **61,8** | 55,6 | 45,0 | 50,0 | **0,0** | **45,0** | **49,2** |
| SVM - kernel:polynomial | 47,5 | **48,1** | 47,9 | 46,3 | 47,1 | 48,1 | 45,0 | **35,7** | 42,9 | **45,4** | **47,1** |
| SVM - kernel:RBF | 47,5 | **48,1** | 47,9 | 46,3 | 47,1 | 48,1 | 45,0 | **28,6** | 42,9 | **44,6** | **47,1** |
| SVM - kernel:sigmoid | 47,5 | **48,1** | 47,9 | 46,3 | 47,1 | 48,1 | 45,0 | **28,6** | 42,9 | **44,6** | **47,1** |
| Naive Bayes | 50,8 | 50,0 | 50,0 | 53,7 | 52,9 | 51,9 | 50,0 | **35,7** | **57,1** | **50,2** | **50,8** |
| Rede Neural | 47,5 | 46,3 | 52,1 | 43,9 | **52,9** | 48,1 | 45,0 | 42,9 | **0,0** | **42,1** | **46,3** |
| KNN | 47,5 | 37,0 | 39,6 | **36,6** | 50,0 | 55,6 | 45,0 | **57,1** | **57,1** | **47,3** | **47,5** |
| **MÉDIA** | **48,5** | **49,2** | **48,6** | **49,1** | ***51,6*** | **49,7** | **42,9** | **41,6** | **40,3** | | |
| **MEDIANA** | **47,5** | **48,1** | ***50,0*** | **48,8** | ***50,0*** | **48,1** | **45,0** | **42,9** | **42,9** | | |
| **DESVIO PADRÃO** | **4,26** | **5,63** | **6,59** | **7,41** | **5,21** | **5,25** | **5,88** | **10,49** | **19,71** | | |

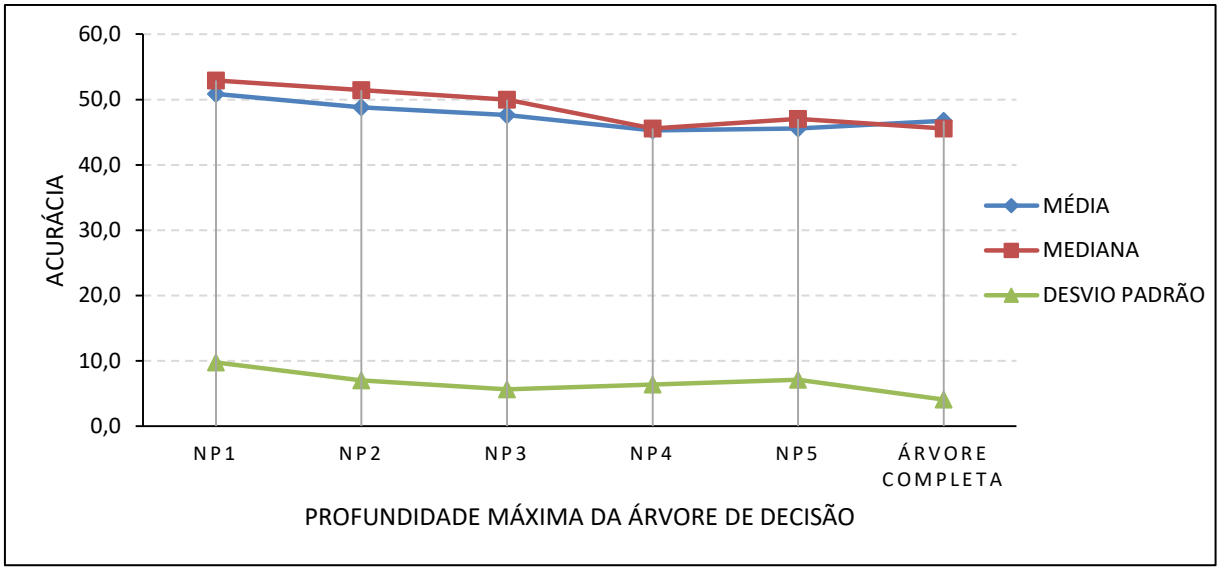
**Fonte:** José E. V. Borges (2019).

### 6.3 Questão 3. Quão eficientes são os algoritmos de classificação por árvores de decisão, no que se refere ao nível de profundidade das árvores, no momento de início do curso?

Na execução dos algoritmos de classificação para o Experimento 3, foi utilizada a técnica de Validação Cruzada, com 10 *folds*, conforme metodologia proposta baseada nos

trabalhos de referência. Aqui o objetivo é responder à questão de pesquisa 3, que sugere uma análise das árvores com ou sem podas, assim como com podas gradativas. Foram utilizados aqui apenas os 5 algoritmos de classificação de árvore de decisão. Vale salientar que para cada algoritmo de árvore, existem parâmetros distintos e diversos para o processo de poda da árvore, assim como para cada algoritmo a árvore se comporta de forma diferente quando podada. Porém, para todos os resultados apresentados, as árvores mantiveram nível de profundidade máximo de 1, 2, 3, 4, 5 e ilimitado. Os resultados das acurácias para cada condição de altura máxima da árvore estão apresentados na Figura 18. Entende-se por NP como sendo "Nível de Profundidade" da árvore.

**Figura 19** - Média, Mediana e Desvio padrão do nível de acurácia geral dos modelos de árvore de decisão relativos à profundidade da árvore.




**Fonte:** José E. V. Borges (2019).

Observa-se uma leve tendência geral de queda no desempenho geral, gradativamente com o aumento do tamanho das árvores. O Desvio Padrão também de mostrou menor para árvores maiores, utilizando-se todos os atributos disponíveis em T0. O fato de que a média e mediana não apresentarem distâncias muito grandes leva a crer que não houve valores discrepantes de acurácia durante a análise gradativa com podas.

Na Tabela 3 verificamos o nível de acurácia alcançado por cada algoritmo em cada situação de poda. Os valores em azul e vermelho indicam respectivamente os maiores e menores valores alcançados, em linha e em coluna. O algoritmo REPTree se mostrou uniforme em acurácia, para todos os limites de tamanho de árvore, o que mostra que ele não sofre influência significativa com as podas, além disso ele também manteve o maior nível médio de acurácia

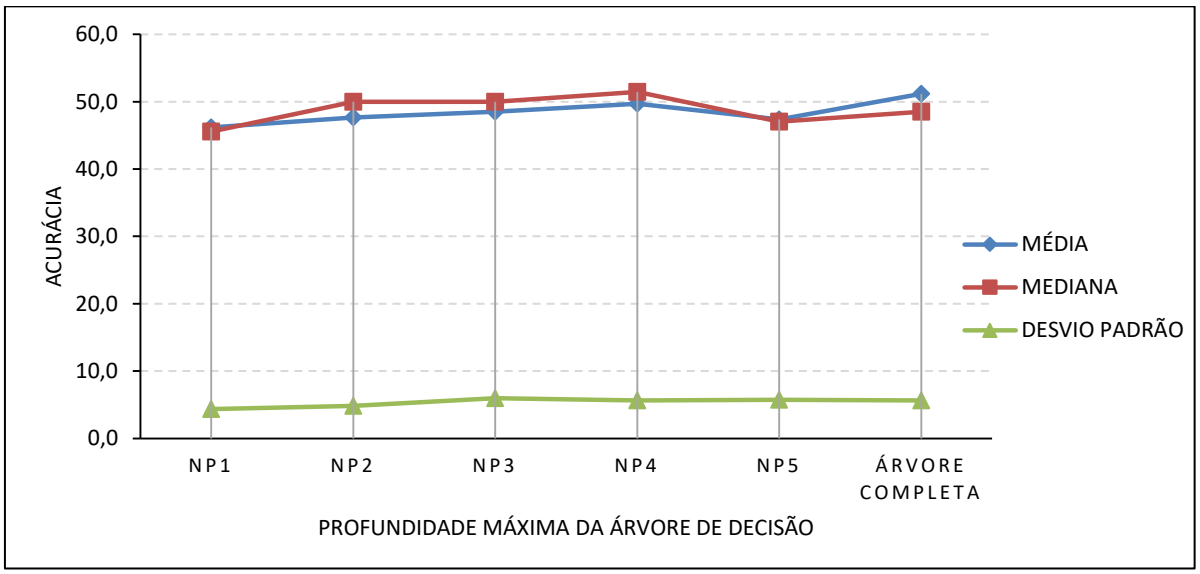
quando comparado às outras árvores. Além disso, em termos de eficiência, vale reconhecer que o ADTree e o RandomTree se mostraram melhores, entregando as maiores acurácias com árvores de menor profundidade, quando utilizando dados sem Seleção de Atributos, respondendo primeiramente à questão de pesquisa 3.

**Tabela 3**-Acurácia dos Modelos de Predição em Árvore de Decisão relativos à profundidade da árvore.

| | NP 1 | NP 2 | NP 3 | NP 4 | NP 5 | Árvore Completa | MÉDIA | MEDIANA |
|---|---|---|---|---|---|---|---|---|
| ADTree | **58,8** | 51,5 | 50,0 | 47,1 | 48,5 | **45,6** | **50,2** | **49,3** |
| J48 | **35,3** | 38,2 | 38,2 | 35,3 | 33,8 | **42,6** | **37,3** | **36,8** |
| RandomTree | **58,8** | 55,9 | 50,0 | 45,6 | 47,1 | **44,1** | **50,2** | **48,5** |
| REPTree | 52,9 | 52,9 | 52,9 | 52,9 | 52,9 | 52,9 | **52,9** | **52,9** |
| SimpleCart | **48,5** | **45,6** | 47,1 | **45,6** | **45,6** | **48,5** | **46,8** | **46,3** |
| **MÉDIA** | **50,9** | **48,8** | **47,6** | **45,3** | **45,6** | **46,8** | | |
| **MEDIANA** | **52,9** | **51,5** | **50,0** | **45,6** | **47,1** | **45,6** | | |
| **DESVIO PADRÃO** | **9,73** | **7,01** | **5,66** | **6,36** | **7,13** | **4,08** | | |

**Fonte:** José E. V. Borges (2019).

**Figura 20** - Média, Mediana e Desvio padrão do nível de acurácia geral dos modelos de árvore de decisão relativos à profundidade da árvore com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

Os Resultados apresentados após o processo de Seleção de Atributos mostraram curiosamente um comportamento geral inverso ao apresentado quando não é feita a seleção. Na Figura 19 pode-se perceber desta vez uma tendência, embora leve, de crescimento de desempenho proporcionalmente à profundidade da árvore.

**Tabela 4**-Acurácia dos Modelos de Predição em Árvore de Decisão relativos à profundidade da árvore com Seleção de Atributos.

| | NP 1 | NP 2 | NP 3 | NP 4 | NP 5 | Árvore Completa | MÉDIA | MEDIANA |
|---|---|---|---|---|---|---|---|---|
| ADTree | 47,1 | 50,0 | **45,6** | 51,5 | 52,9 | **60,3** | **51,2** | **50,7** |
| J48 | 41,2 | 41,2 | **39,7** | **39,7** | **39,7** | **48,5** | **41,7** | **40,4** |
| RandomTree | **44,1** | **44,1** | **54,4** | 51,5 | **44,1** | 47,1 | **47,5** | **45,6** |
| REPTree | 52,9 | 52,9 | 52,9 | 52,9 | 52,9 | 52,9 | **52,9** | **52,9** |
| SimpleCart | **45,6** | 50,0 | 50,0 | **52,9** | 47,1 | 47,1 | **48,8** | **48,5** |
| **MÉDIA** | **46,2** | **47,6** | **48,5** | **49,7** | **47,3** | **51,2** | | |
| **MEDIANA** | **45,6** | **50,0** | **50,0** | **51,5** | **47,1** | **48,5** | | |
| **DESVIO PADRÃO** | **4,36** | **4,84** | **5,98** | **5,64** | **5,74** | **5,64** | | |

**Fonte:** José E. V. Borges (2019).

Na Tabela 4 confirma-se tal tendência a partir da observação da concentração dos resultados de maior acurácia nas árvores mais profundas, Tamanho 3 ou maior. Por sua vez as piores acurácias se encontram concentradas nas árvores menores. Pode-se concluir como fator relevante para este comportamento o fato de que dados mais "ruidosos", em outras palavras, que possuem atributos que não são relevantes para a aprovação do aluno, tendem a contribuir negativamente para os processos de aprendizagem do algoritmo, direcionando à criação de modelos de árvore "poluídos", que utilizam estes atributos pouco relevantes como nós decisores da árvore. Quando mais próximo da raiz estiver um atributo irrelevante, maior será o impacto negativo na acurácia obtida para a predição. Como resposta adicional à Questão 3, pode-se considerar este comportamento apresentado, acrescentando que a eficiência das árvores será afetada pelo processo de poda de acordo com a qualidade dos atributos de entrada.

Mais uma vez observa-se que a seleção de atributos elevou as acurácias inferiores: J48 melhorou de 37,3% para 41,7%. Observa-se também que, utilizando-se apenas atributos mais relevantes, o algoritmo REPTree se mostra desta vez mais eficiente: além de entregar a maior acurácia com árvore de profundidade 1, não sofre alteração de acurácia com as podas.

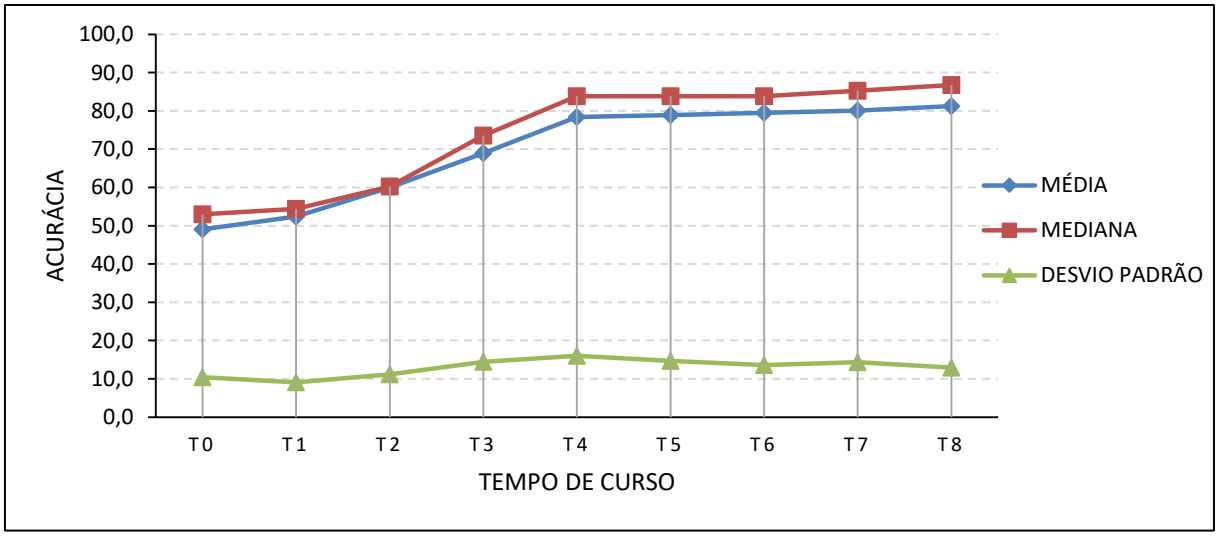
Uma conclusão essencial que pode ser obtida com os Experimentos 2 e 3, que possuem a característica em comum de utilizar apenas os dados dos alunos disponíveis em T0, ou momento inicial do curso, é que os níveis de acurácia dos algoritmos de classificação, quando submetidos a atributos dos alunos disponíveis para o professor no momento chamado T0, não são confiáveis o suficiente, apresentando uma maior concentração de amostras dos resultados de acurácias entre 45% e 55%. Isso leva a crer que uma predição realizada no momento T0, apenas com os dados socioeconômicos, não é segura o suficiente para dar suporte ao professor,

sistema ou especialista em EDM. Acredita-se que com uma amostra maior de dados para treino e teste permita uma predição mais acurada em T0.

## 6.4 Questão 4. Em que tempo, a partir da aplicação de atividades de aprendizagem durante o curso, é possível obter níveis de desempenho satisfatórios na predição do sucesso?

O Experimento 4 destina-se a responder à questão de pesquisa 4, assim como as suas questões derivas 4.1, 4.2, 4.3, 4.4 e 4.5, através de uma análise temporal da eficiência dos 14 algoritmos no decorrer do curso, buscando encontrar o momento mais breve possível onde uma predição satisfatória pode ser realizada. Para este experimento exclusivamente, foram mantidos os dados das notas das atividades dos alunos no Huxley. A Figura 20 ilustra o comportamento geral dos algoritmos, apresentando as médias, medianas e desvio padrão das acurácias de todos os algoritmos, quando submetidos a uma progressão temporal com inserções dos dados das atividades realizadas no decorrer do curso.

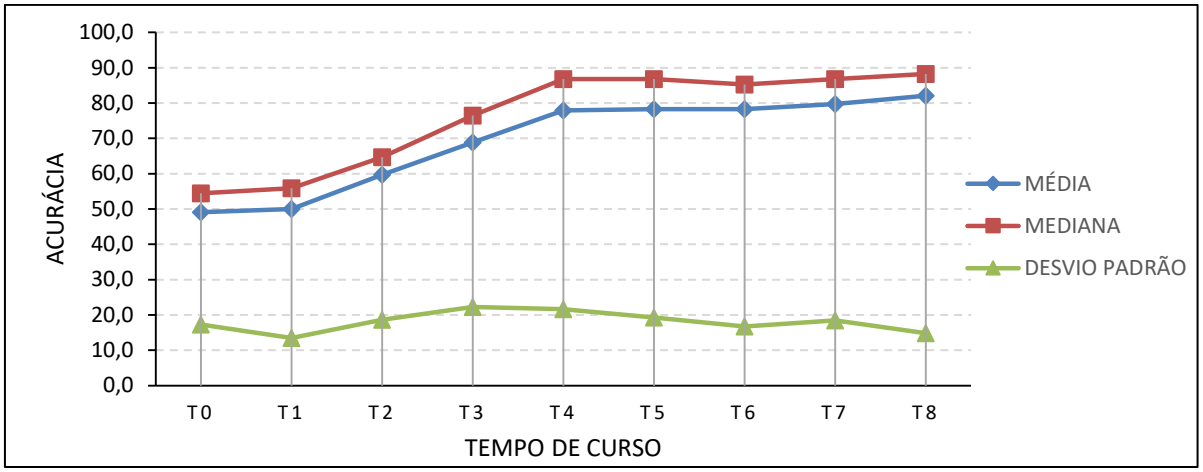
**Figura 21** - Média, Mediana e Desvio padrão das acurácias dos atributos em cada parcela de tempo no decorrer do curso.




**Fonte:** José E. V. Borges (2019).

O gráfico da Figura 20 juntamente com a Tabela 5 confirmam de certa forma o que era esperado: com a adição das informações das notas dos alunos nas atividades, independente do algoritmo de classificação utilizado, aos dados iniciais dos alunos disponíveis em T0, sempre existe incremento na acurácia, proporcional ao tempo de curso e consequentemente proporcional ao número de atividades realizadas. Percebe-se também uma constância no desvio padrão ao longo de todo o período, indicando variabilidade constante dos resultados. É possível perceber ainda que após a tempo T4, o nível geral de acurácia na predição de todos os algoritmos se mante praticamente constante ou com pequena variação até o final do curso.

**Figura 22** - Média, Mediana e Desvio padrão das acurácias dos algoritmos de Regras em cada parcela de tempo no decorrer do curso

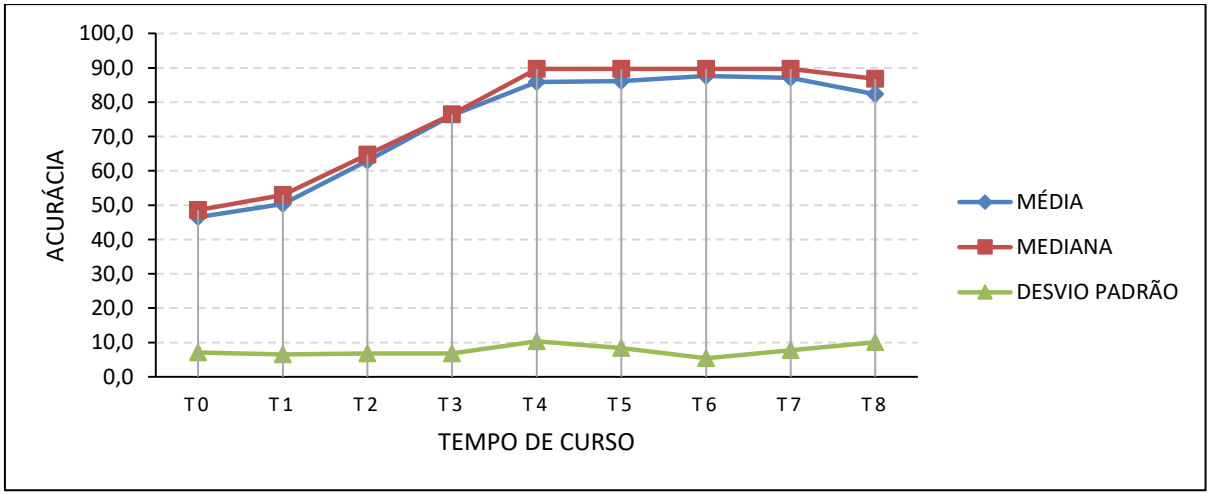



**Fonte:** José E. V. Borges (2019).

Os gráficos expostos nas Figuras 21, 22 e 23 praticamente a mesma tendência do gráfico geral, contido na Figura 20, ou seja, analisando eficiência dos algoritmos de classificação Caixa Branca-Regra, Caixa Branca-Árvore e Caixa preta em grupos separados, todos eles apresentam crescimento significativo na acurácia até T4 e pouca variação a partir de então.

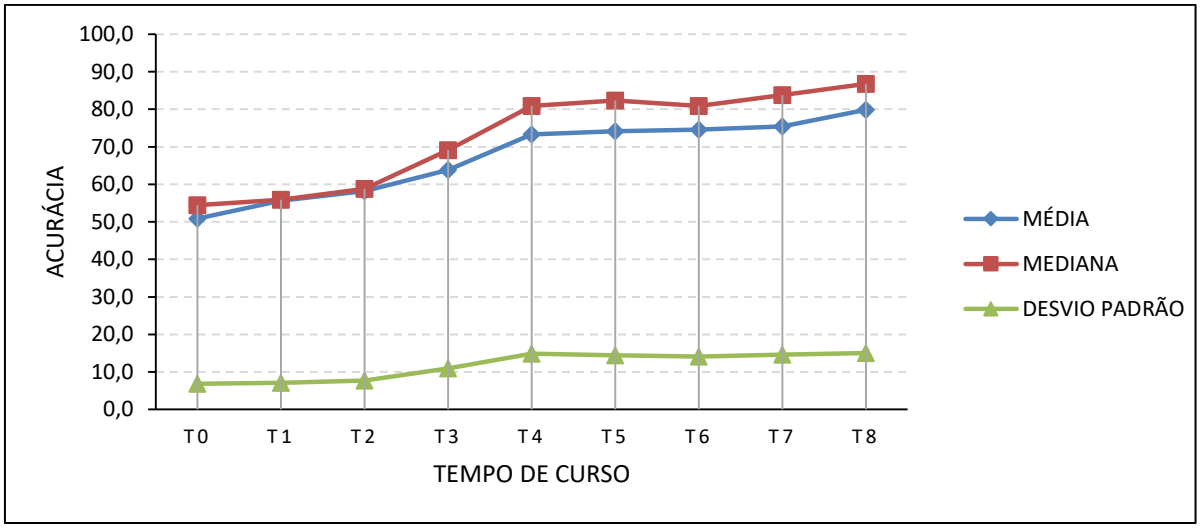
Destaca-se nos gráficos das Figuras 21 e 23, a diferença significativa entre média e mediana, apresentadas nas análises dos algoritmos de Regras e Caixa Preta. Justifica-se esta diferença essencialmente pelos resultados expressivamente baixos dos algoritmos Prism e SVM – kernel:sigmoid, que apresentaram média de 37,6% e 50,3% respectivamente, influenciando para a redução da média geral. Apesar disso, a análise das tendências de cada grupo de algoritmo não foi comprometida.

**Figura 23** - Média, Mediana e Desvio padrão das acurácias dos algoritmos de Árvore de Decisão em cada parcela de tempo no decorrer do curso.




**Fonte:** José E. V. Borges (2019).

**Figura 24** - Média, Mediana e Desvio padrão das acurácias dos algoritmos Caixa Preta em cada parcela de tempo no decorrer do curso.




**Fonte:** José E. V. Borges (2019).

Analisando os valores individuais de acurácia de cada atributo, apresentados na Tabela 5, percebe-se uma tendência geral de crescimento, reforçando o que foi evidenciado nos gráficos das Figuras 20, 21, 22 e 23. A máxima acurácia apresentada foi do algoritmo *SimpleCart*, a partir do instante T4, com 92,6% de acurácia.

**Tabela 5**-Acurácias da predição dos algoritmos de classificação relativos ao tempo de curso e informações adquiridas nas atividades de aprendizagem.

| **Algoritmo** | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | **55,9** | 57,4 | 64,7 | 75,0 | **91,2** | 89,7 | 85,3 | **91,2** | 85,3 | **77,3** | **85,3** |
| NNge | **54,4** | 55,9 | 60,3 | 80,9 | 82,4 | 82,4 | 82,4 | 83,8 | **88,2** | **74,5** | **82,4** |
| OneR | 63,2 | **51,5** | 75,0 | 82,4 | 86,8 | 86,8 | 86,8 | 86,8 | **89,7** | **78,8** | **86,8** |
| Prism | **19,1** | 26,5 | 27,9 | 29,4 | 39,7 | 44,1 | 48,5 | 47,1 | **55,9** | **37,6** | **39,7** |
| Ridor | **52,9** | 58,8 | 70,6 | 76,5 | 89,7 | 88,2 | 88,2 | 89,7 | **91,2** | **78,4** | **88,2** |
| ADTree | **51,5** | 57,4 | 67,6 | 82,4 | **91,2** | 85,3 | 85,3 | 88,2 | 86,8 | **77,3** | **85,3** |
| J48 | **35,3** | 42,6 | 64,7 | 76,5 | 89,7 | 91,2 | **91,2** | **91,2** | 83,8 | **74,0** | **83,8** |
| RandomTree | **44,1** | 54,4 | 55,9 | 66,2 | 67,6 | 72,1 | **79,4** | 73,5 | 64,7 | **64,2** | **66,2** |
| REPTree | 52,9 | **44,1** | 55,9 | 73,5 | 88,2 | **89,7** | **89,7** | **89,7** | 86,8 | **74,5** | **86,8** |
| SimpleCart | **48,5** | 52,9 | 70,6 | 82,4 | **92,6** | **92,6** | **92,6** | **92,6** | 89,7 | **79,4** | **89,7** |
| SVM - kernel:linear | **55,9** | 60,3 | 64,7 | 70,6 | 85,3 | 82,4 | 82,4 | 85,3 | **88,2** | **75,0** | **82,4** |
| SVM - kernel:polynomial | **54,4** | 60,3 | 60,3 | 69,1 | 83,8 | 82,4 | 80,9 | 83,8 | **86,8** | **73,5** | **80,9** |
| SVM - kernel:RBF | **58,8** | 66,2 | 69,1 | 73,5 | 80,9 | 83,8 | 83,8 | 85,3 | **89,7** | **76,8** | **80,9** |
| SVM - kernel:sigmoid | 45,6 | 45,6 | 45,6 | **44,1** | 45,6 | 47,1 | 47,1 | 47,1 | **48,5** | **50,3** | **45,6** |
| Naive Bayes | **47,1** | 51,5 | 58,8 | 73,5 | 83,8 | 85,3 | 86,8 | 85,3 | **88,2** | **73,4** | **83,8** |

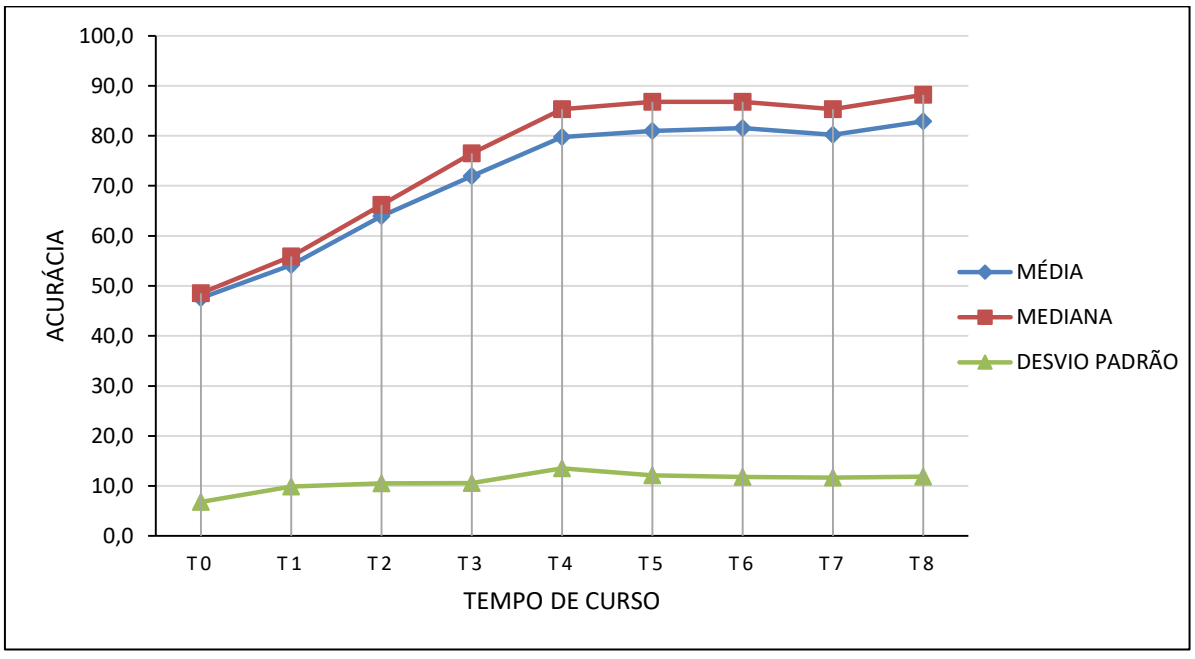
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rede Neural | **39,7** | 50,0 | 54,4 | 57,4 | 72,1 | 76,5 | 76,5 | 76,5 | **85,3** | **65,4** | **72,1** |
| KNN | **54,4** | 55,9 | 54,4 | 58,8 | 61,8 | 61,8 | 64,7 | 64,7 | **72,1** | **60,9** | **61,8** |
| **MÉDIA** | **49,0** | **52,4** | **60,0** | **68,9** | **78,4** | **78,9** | **81,7** | **80,1** | **81,2** | | |
| **MEDIANA** | **52,9** | **54,4** | **60,3** | **73,5** | **83,8** | **83,8** | **83,8** | **85,3** | **86,8** | | |
| **DESVIO PADRÃO** | **10,39** | **9,09** | **11,15** | **14,41** | **15,98** | **14,65** | **10,68** | **14,33** | **12,90** | | |

**Fonte:** José E. V. Borges (2019).

Praticamente todos os algoritmos apresentaras como seus piores resultados de acurácia o desempenho em T0, ou seja, no início do curso, quando não se tem nenhuma avaliação de conhecimento do aluno. Já as melhores acurácias individuais de cada algoritmo se encontram a partir de T4.

Os resultados exibidos nos gráficos das Figuras 24, 25, 26 e 27 são as medidas de desempenho calculadas incluindo no pré-processamento a Seleção de Atributos, ou seja, foram considerados nos dados de entradas dos classificadores apenas os atributos mais relevantes, dentre os atributos disponíveis do aluno no início do curso, conforme Experimento 1, adicionados quinzenalmente os dados das notas dos alunos nas atividades do Huxley.

**Figura 25** - Média, Mediana e Desvio padrão das acurácias dos atributos em cada parcela de tempo no decorrer do curso com Seleção de Atributos.



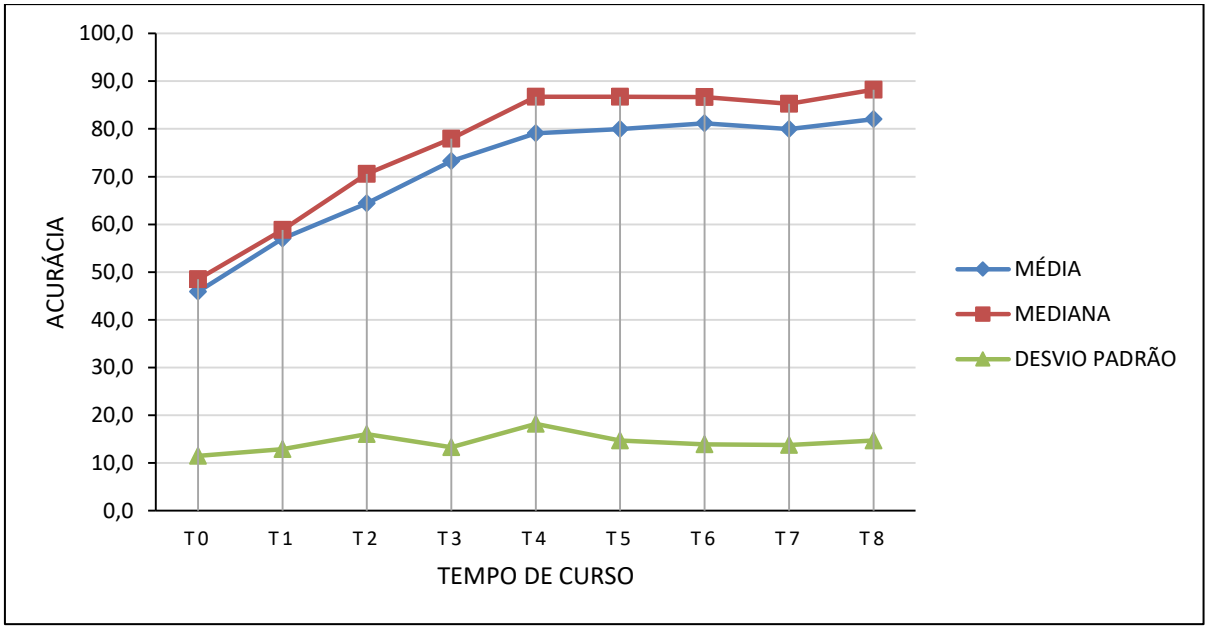

**Fonte:** José E. V. Borges (2019).

Observa-se como tendência geral que, mesmo quando aplicada a seleção de atributos, também em T4 a média entre as acurácias dos algoritmos de classificação chega próximo de 80% e continua até T8 com pouco crescimento, basicamente flutuando entre 80% e 90% de

média. Podemos perceber também comparando as Tabelas 5 e 6 que a seleção de atributos elevou a acurácia mínima de 19,1% para 26,5% e a média das acurácias obtidas em T8 de 81,2% para 82,9%.

Na análise dos resultados por tipo de algoritmo, percebe-se que os algoritmos caixa branca de árvore de decisão se destacaram em eficiência temporal, alcançando uma média de acurácia bem próximo a 90% em T4. O Algoritmo SimpleCart se destacou dentre todos, chegando a exatos 92,6% de T4 a T7.

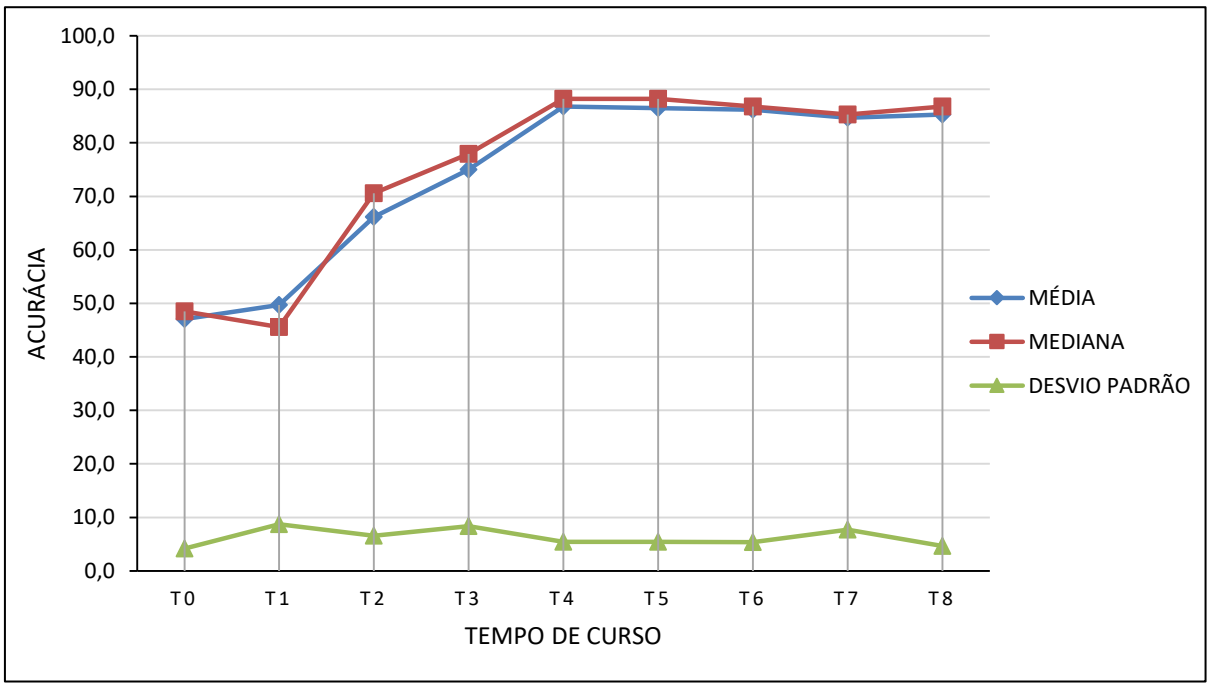
**Figura 26** - Média, Mediana e Desvio padrão das acurácias dos algoritmos de Regras em cada parcela de tempo no decorrer do curso com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

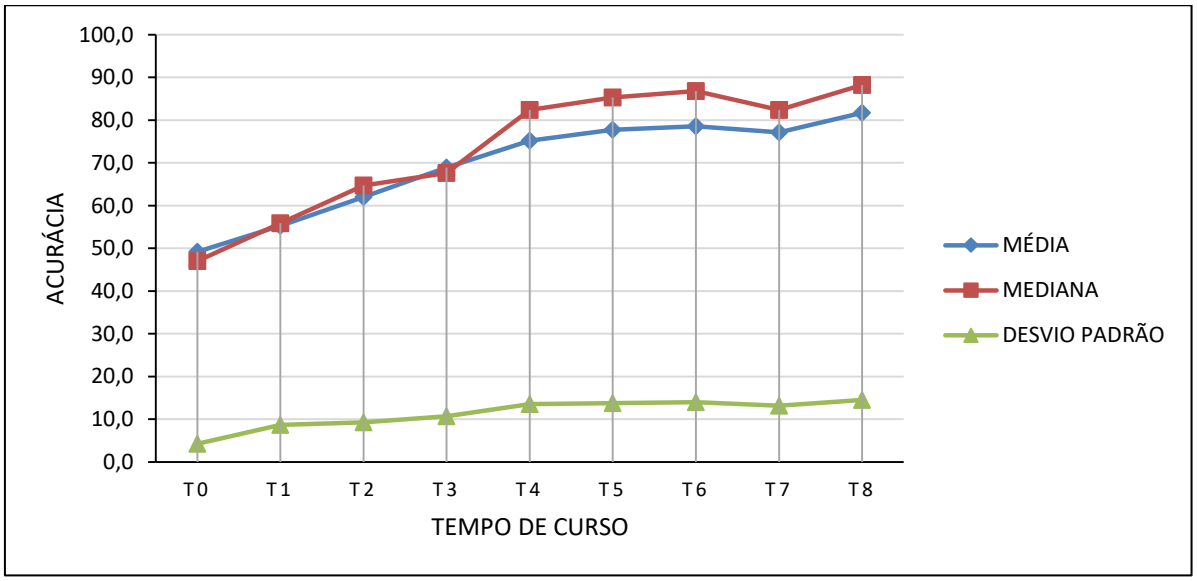
Comparando-se os resultados dos grupos de algoritmos, Regras, Árvores de Decisão e Caixa Preta, apresentados nos gráficos das Figuras 25, 26 e 27, observa-se tendência geral semelhante, com pouco ou nenhum crescimento após T4.

**Figura 27** - Média, Mediana e Desvio padrão das acurácias dos algoritmos de Árvore de Decisão em cada parcela de tempo no decorrer do curso com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

**Figura 28** - Média, Mediana e Desvio padrão das acurácias dos algoritmos Caixa Preta em cada parcela de tempo no decorrer do curso com Seleção de Atributos.




**Fonte:** José E. V. Borges (2019).

Na Tabela 6 é possível confirmar os algoritmos mais eficientes, que neste caso foram os de árvore de decisão, que obtiveram suas melhores acurácias já em T4. JRip também se destacou nesta situação, apresentado 91,2 em T4, superando os demais algoritmos de regras. A Seleção de Atributos elevou a Média geral de acurácia em T3 de 68,9% para 72,0%, em T4 de 78.4% para 79.8% e em T8 de 81,2% para 82,9%.

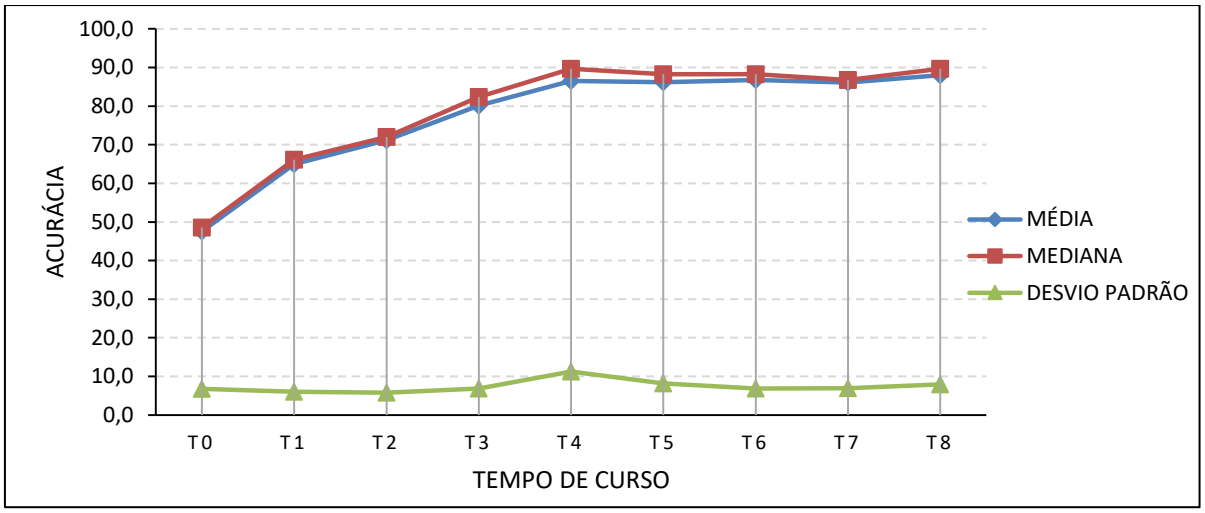
**Tabela 6**-Acurácias da predição dos algoritmos de classificação relativos ao tempo de curso e informações adquiridas nas atividades de aprendizagem com Seleção de Atributos.

| **Algoritmo** | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | **45,6** | 69,1 | 70,6 | 80,9 | **91,2** | 88,2 | **91,2** | 85,3 | 86,8 | **78,8** | **85,3** |
| NNge | **54,4** | 58,8 | 64,7 | 75,0 | 82,4 | 80,9 | 80,9 | 82,4 | **88,2** | **74,2** | **80,9** |
| OneR | **48,5** | 63,2 | 75,0 | 82,4 | 86,8 | 86,8 | 86,7 | 86,8 | **89,7** | **78,4** | **86,7** |
| Prism | **26,5** | 35,3 | 36,8 | 50,0 | 47,1 | 54,4 | **57,4** | 55,9 | 55,9 | **46,6** | **50,0** |
| Ridor | **54,4** | 58,8 | 75,0 | 77,9 | 88,2 | **89,7** | **89,7** | **89,7** | **89,7** | **79,2** | **88,2** |
| ADTree | **48,5** | 61,8 | 70,6 | 80,9 | 86,8 | **88,2** | 86,8 | 85,3 | **88,2** | **77,4** | **85,3** |
| J48 | 42,6 | **41,2** | 70,6 | 77,9 | **88,2** | 85,3 | 85,3 | 85,3 | 83,8 | **73,4** | **83,8** |
| RandomTree | 51,5 | **45,6** | 63,2 | 61,8 | **77,9** | **77,9** | **77,9** | 72,1 | **77,9** | **67,3** | **72,1** |
| REPTree | 50,0 | **44,1** | 55,9 | 72,1 | **88,2** | **88,2** | **88,2** | **88,2** | 86,8 | **73,5** | **86,8** |
| SimpleCart | **42,6** | 55,9 | 70,6 | 82,4 | **92,6** | **92,6** | **92,6** | **92,6** | 89,7 | **79,1** | **89,7** |
| SVM - kernel:linear | **45,6** | 55,9 | 58,8 | 67,6 | 82,4 | 86,8 | 86,8 | 86,8 | **91,2** | **73,5** | **82,4** |
| SVM - kernel:polynomial | 45,6 | **44,1** | 45,6 | **50,0** | 48,5 | **50,0** | **50,0** | **50,0** | **50,0** | **48,2** | **50,0** |
| SVM - kernel:RBF | **47,1** | 64,7 | 70,6 | 79,4 | 85,3 | 86,8 | 86,8 | 85,3 | **88,2** | **77,1** | **85,3** |
| SVM - kernel:sigmoid | 47,1 | **44,1** | 72,1 | 79,4 | 82,4 | 85,3 | **88,2** | 82,4 | **88,2** | **74,3** | **82,4** |
| Naive Bayes | **51,5** | 63,2 | 66,2 | 76,5 | 86,8 | 88,2 | 88,2 | 86,8 | **89,7** | **77,4** | **86,8** |
| Rede Neural | **50,0** | 52,9 | 55,9 | 61,8 | 72,1 | 75,0 | 76,5 | 75,0 | **85,3** | **67,2** | **72,1** |
| KNN | **57,4** | 61,8 | 64,7 | 67,6 | 69,1 | 72,1 | 73,5 | 73,5 | **79,4** | **68,8** | **69,1** |
| **MÉDIA** | **47,6** | **54,1** | **63,9** | **72,0** | **79,8** | **81,0** | **81,6** | **80,2** | **82,9** | | |
| **MEDIANA** | **48,5** | **55,9** | **66,2** | **76,5** | **85,3** | **86,8** | **86,8** | **85,3** | **88,2** | | |
| **DESVIO PADRÃO** | **6,80** | **9,90** | **10,52** | **10,60** | **13,53** | **12,12** | **11,79** | **11,71** | **11,87** | | |

**Fonte:** José E. V. Borges (2019).

Ainda no Experimento 4, e referente à questão de pesquisa 4, indo mais além no que diz respeito à qualidade dos dados de entrada, buscou-se uma análise temporal mais ousada e ao mesmo tempo experimental, utilizando-se como dados de entrada em cada marco de tempo de T0 a T8 apenas os dados das atividades realizadas no Huxley, ou seja, foram deixados de lado todos os dados socioeconômicos a partir de T0. Uma abordagem semelhante foi feita por MÁRQUEZ et al., (2016) porém com o objetivo de predizer evasão de alunos. Em T0, para os resultados que serão mostrados a seguir, foram utilizados os dados socioeconômicos após seleção de atributos, a fim de permitir um comparativo entre T0 com atributos socioeconômicos mais relevantes e T1 em diante sem socioeconômicos e apenas com informações obtidas através das atividades realizadas.

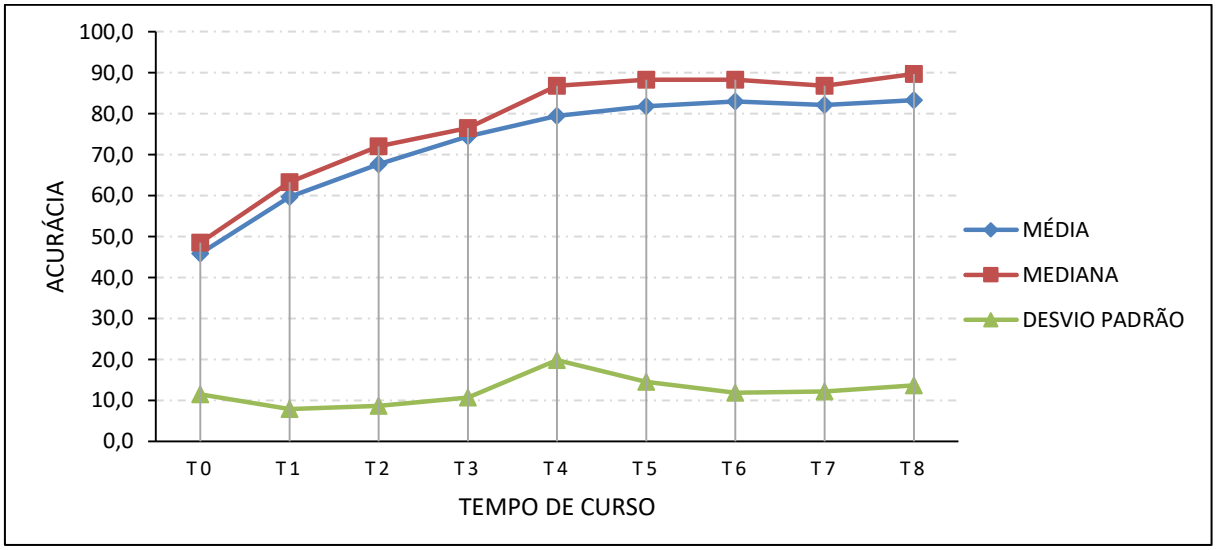
**Figura 29** - Média, Mediana e Desvio padrão das acurácias dos atributos em cada parcela de tempo no decorrer do curso, utilizando apenas atributos das atividades do Huxley.




**Fonte:** José E. V. Borges (2019).

No gráfico apresentado na Figura 28 já é possível observar que a média geral, considerando todos os algoritmos, alcançou patamares maiores em T4, chegando na casa dos 85%. A tendência após de T3 se manteve a mesma dos casos anteriores, onde os atributos socioeconômicos estavam presentes, ou seja, houve pouca variação, flutuando desta vez sobre os 85%. Tal melhoria de desempenho pode ser direcionada a 2 fatores correlacionados: 1) O ruído gerado pelos atributos socioeconômicos nas predições temporais e 2) O fato de que as atividades de aprendizagem são um elemento valioso para a tarefa de predição, podendo por si só permitir a extração de informações relevantes para a tomada de decisão dos educadores.

**Figura 30** - Média, Mediana e Desvio padrão das acurácias dos algoritmos de Regras em cada parcela de tempo no decorrer do curso utilizando apenas atributos das atividades do Huxley.

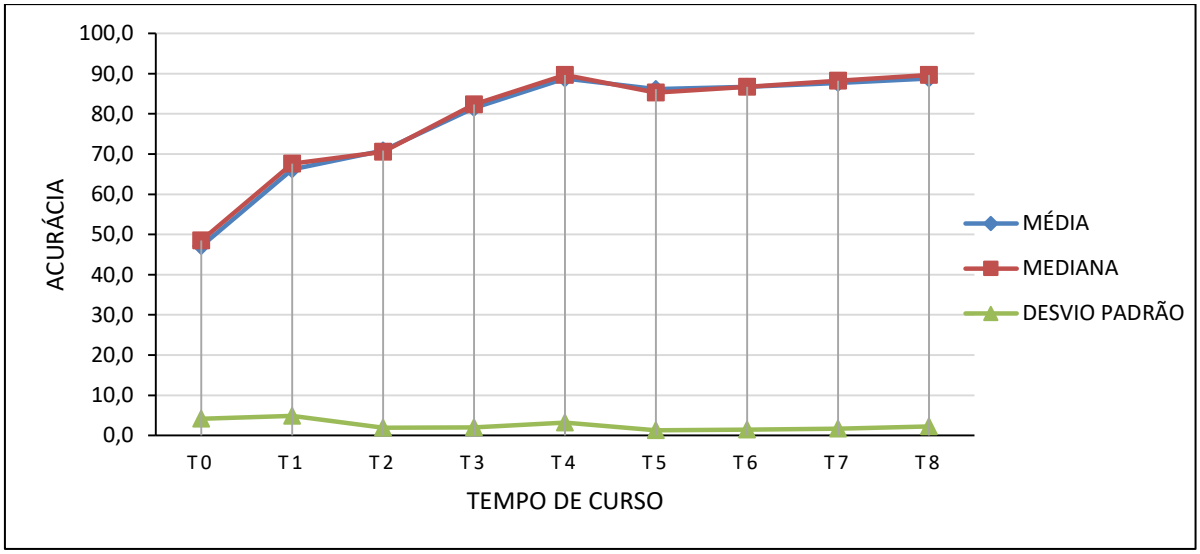



**Fonte:** José E. V. Borges (2019).

Na Figura 29, onde estão sendo considerados apenas os algoritmos de Regras, observa-se que a média atingida por estes algoritmos superou os 80% em T4, e se mantendo acima deste patamar durante todo o restante do curso. Pode-se observar os resultados de desempenho dos algoritmos de árvore na Figura 30, e dos algoritmos Caixa Preta na Figura 31.

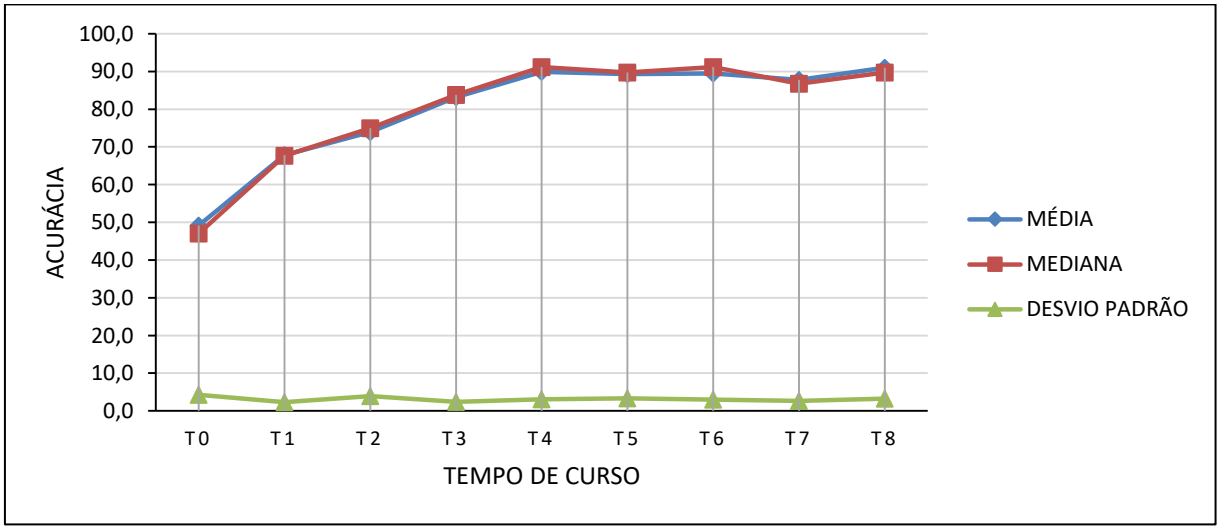
Os algoritmos de árvore de decisão e os de caixa preta também apresentaram um desvio padrão inferior em termos de médias dos resultados das acurácias quando comparados aos algoritmos de Regras, sinalizando pouca variação entre as acurácias.

**Figura 31** - Média, Mediana e Desvio padrão das acurácias dos algoritmos de Árvore de Decisão em cada parcela de tempo no decorrer do curso utilizando apenas atributos das atividades do Huxley.




**Fonte:** José E. V. Borges (2019).

**Figura 32** - Média, Mediana e Desvio padrão das acurácias dos algoritmos Caixa Preta em cada parcela de tempo no decorrer do curso utilizando apenas atributos das atividades do Huxley.




**Fonte:** José E. V. Borges (2019).

As melhores acurácias mais uma vez foram alcançadas pelos algoritmos de caixa preta. Embora alguns resultados máximos foram alcançados apenas em T4, observa-se desempenhos bastante altos em acurácia já no momento T3: a exemplo do algoritmo SVM com kernel = *sigmoid*, que chegou a 86,8% de acurácia. Pode-se traduzir em outras palavras que a cada 10 alunos, 8 o professor pode ter uma predição assertiva em 45 dias de curso, ainda antes da primeira avaliação bimestral, e considerando apenas a aplicação de 3 atividades. Atividades com intervalos de tempo menores entre elas, ou seja, atividades aplicadas semanalmente, pode-se acreditar a partir desta análise que é possível alcançar predições ainda mais eficientes com menos tempo.

**Tabela 7**-Acurácias da predição dos algoritmos de classificação relativos ao tempo de curso e informações adquiridas nas atividades de aprendizagem com Seleção de Atributos utilizando apenas atributos das atividades do Huxley.

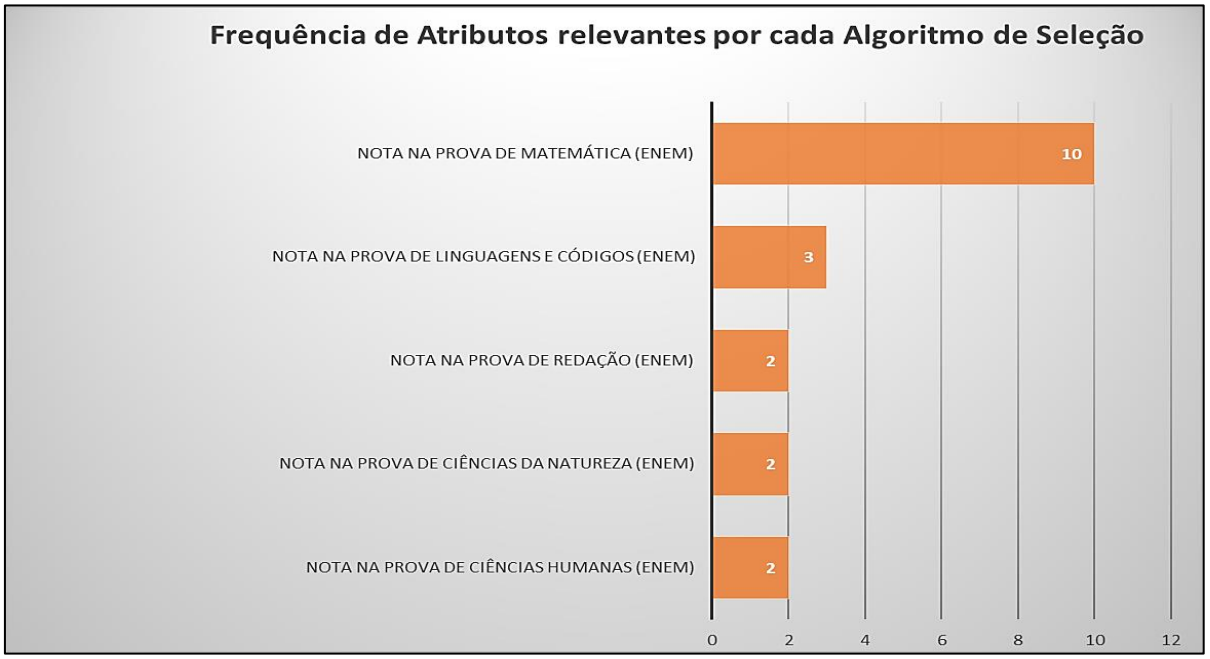
| **Algoritmo** | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | **45,6** | 66,2 | 73,5 | 85,3 | **89,7** | 88,2 | 88,2 | 86,8 | **89,7** | **79,2** | **86,8** |
| NNge | **54,4** | **54,4** | 63,2 | 69,1 | 85,3 | **89,7** | **89,7** | 88,2 | 88,2 | **75,8** | **85,3** |
| OneR | **48,5** | 63,2 | 75,0 | 82,4 | 86,8 | 86,8 | 86,8 | 86,8 | **89,7** | **78,4** | **86,8** |
| Prism | **26,5** | 48,5 | 54,4 | 58,8 | 44,1 | 55,9 | **61,8** | 60,3 | 58,8 | **52,1** | **55,9** |
| Ridor | **54,4** | 66,2 | 72,1 | 76,5 | **91,2** | 88,2 | 88,2 | 88,2 | 89,7 | **79,4** | **88,2** |
| ADTree | **48,5** | 67,6 | 69,1 | 82,4 | 89,7 | 85,3 | 88,2 | 88,2 | **91,2** | **78,9** | **85,3** |
| J48 | **42,6** | 61,8 | 72,1 | 82,4 | **89,7** | 85,3 | 85,3 | 85,3 | 85,3 | **76,6** | **85,3** |
| RandomTree | **51,5** | 70,6 | 69,1 | 82,4 | 83,8 | 86,8 | 86,8 | **89,7** | **89,7** | **78,9** | **83,8** |
| REPTree | **50,0** | 60,3 | 70,6 | 77,9 | **88,2** | **88,2** | **88,2** | **88,2** | **88,2** | **77,8** | **88,2** |
| SimpleCart | **42,6** | 70,6 | 73,5 | 82,4 | **92,6** | 85,3 | 85,3 | 86,8 | 89,7 | **78,8** | **85,3** |
| SVM - kernel:linear | **45,6** | 67,6 | 77,9 | 83,8 | 92,6 | 91,2 | 91,2 | 91,2 | **95,6** | **81,9** | **91,2** |
| SVM - kernel:polynomial | **45,6** | 69,1 | 67,6 | 83,8 | **86,8** | 82,4 | 83,8 | 83,8 | 86,8 | **76,6** | **83,8** |
| SVM - kernel:RBF | **47,1** | 66,2 | 75,0 | 80,9 | **92,6** | 91,2 | 91,2 | 86,8 | 89,7 | **80,1** | **86,8** |
| SVM - kernel:sigmoid | **47,1** | 70,6 | 77,9 | 86,8 | 92,6 | 89,7 | 91,2 | 86,8 | **94,1** | **81,9** | **86,8** |
| Naive Bayes | **51,5** | 66,2 | 72,1 | 80,9 | 88,2 | 89,7 | 88,2 | 86,8 | **92,6** | **79,6** | **86,8** |
| Rede Neural | **50,0** | 64,7 | 76,5 | 85,3 | 91,2 | **92,6** | **92,6** | 91,2 | 89,7 | **81,5** | **89,7** |
| KNN | **57,4** | 70,6 | 70,6 | 80,9 | 85,3 | **88,2** | **88,2** | **88,2** | **88,2** | **79,7** | **85,3** |
| **MÉDIA** | **47,6** | **65,0** | **71,2** | **80,1** | **86,5** | **86,2** | **86,8** | **86,1** | **88,1** | | |
| **MEDIANA** | **48,5** | **66,2** | **72,1** | **82,4** | **89,7** | **88,2** | **88,2** | **86,8** | **89,7** | | |
| **DESVIO PADRÃO** | **6,80** | **6,02** | **5,74** | **6,82** | **11,29** | **8,22** | **6,86** | **6,90** | **7,93** | | |

**Fonte:** José E. V. Borges (2019).

Após dos resultados da Tabela 7, buscando-se ainda uma eficiência maior nas predições a partir da seleção dos dados utilizados, foram aplicadas mais duas Seleção de Atributos: 1)

Apenas sobre as Notas do ENEM, ou seja, apenas sobre os dados de notas dos alunos que são disponíveis em T0, e 2) Sobre as Notas do Huxley com Notas do ENEM, desta vez buscando analisar quais são os atributos mais relevantes dentre este conjunto de notas, que por sua vez são disponíveis apenas em T8, ou seja, próximo ao final do curso. Os resultados destas duas análises encontram-se na Figura 32 e 33.

**Figura 33** - Seleção de Atributos considerando apenas os dados das Notas do ENEM, em T0.

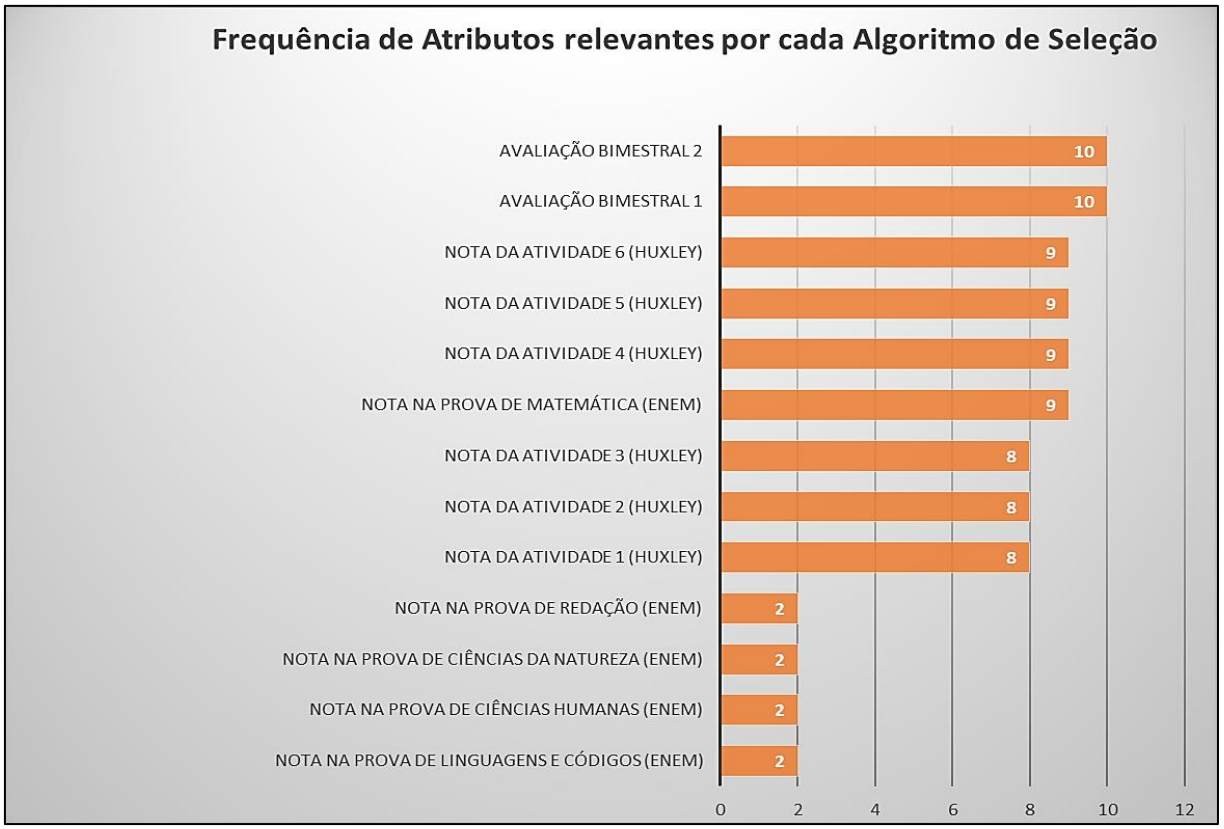



**Fonte:** José E. V. Borges (2019).

Observa-se através da Figura 32 que o atributo “Nota na Prova de Matemática (ENEM)” desta vez se destacou dentre as demais notas. Vale ressaltar que esse destaque não pôde ser observado quando a Seleção de Atributos foi realizada sobre todos os atributos disponíveis em T0, no Experimento 1. Dentre os 10 algoritmos de Seleção de Atributos executados, todos os 10 consideraram o atributo “Nota na Prova de Matemática (ENEM)” como relevante, enquanto apenas 3 consideraram “Nota na Prova de Linguagens e Códigos (ENEM)” e apenas 2 selecionaram as demais notas.

Por sua vez, a Figura 33 mostra um grande destaque, na Seleção de atributos, para as notas do Huxley e o atributo “Nota na Prova de Matemática (ENEM)”, sendo estes selecionados como relevantes por pelo menos 8 dos 10 algoritmos de Seleção. Tais resultados já dão indícios de que possivelmente com a utilização de apenas estes atributos relevantes, é possível uma maior acurácia na predição quando comparadas com os resultados anteriores, onde foram utilizados os 26 atributos selecionados dentro o conjunto total de atributos.

**Figura 34** - Seleção de Atributos considerando apenas os dados das Notas do ENEM juntamente com Huxley, em T8.




**Fonte:** José E. V. Borges (2019).

A Tabela 7, 8 e 9 resumem a última parte da análise do Experimento 4, onde a busca continua pelos os melhores algoritmos para as predições em cada momento T, e ao mesmo tempo também pretende-se encontrar quais os dados ou atributos recomendados para se obter uma predição mais eficiente. A partir dos resultados apresentados nas Figuras 32 e 33, viu-se a possibilidade de se explorar a eficiência nas predições quando utilizando apenas notas do ENEM e Notas do ENEM com notas do Huxley.

Na Tabela 7, apresentada anteriormente, estão listadas as acurácias de cada algoritmo em cada momento T, quando utilizados apenas dados das notas do Huxley. Vale salientar neste caso que em T0, ou seja, no início do curso, quando não existiam notas do Huxley disponíveis, foram utilizados os dados socioeconômicos do ENEM após processo de seleção de atributos conforme Experimento 1, para fins comparativos.

As Tabelas 8 e 9 apresentam respectivamente os resultados das acurácias quando utilizando os dados Notas do ENEM juntamente com Notas do Huxley, e utilizando apenas "Nota na Prova de Matemática (ENEM)" juntamente com Notas do Huxley.

**Tabela 8**-Acurácias da predição dos algoritmos de classificação relativos ao tempo de curso utilizando apenas atributos das Notas do ENEM e Notas das atividades do Huxley.

| Algoritmo | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | MÉDIA | MEDIANA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | 58,8 | 51,5 | 75,0 | 76,5 | 85,3 | 89,7 | 89,7 | 86,8 | 85,3 | 77,6 | 85,3 |
| NNge | 58,8 | 57,4 | 66,2 | 77,9 | 91,2 | 91,2 | 91,2 | 89,7 | 91,2 | 79,4 | 89,7 |
| OneR | 54,4 | 51,5 | 75,0 | 82,4 | 86,8 | 86,8 | 86,8 | 86,8 | 89,7 | 77,8 | 86,8 |
| Prism | 10,3 | 23,5 | 27,9 | 30,9 | 42,6 | 50,0 | 55,9 | 54,4 | 58,8 | 39,4 | 42,6 |
| Ridor | 60,3 | 61,8 | 72,1 | 79,4 | 91,2 | 88,2 | 88,2 | 88,2 | 89,7 | 79,9 | 88,2 |
| ADTree | 58,8 | 55,9 | 67,6 | 76,5 | 89,7 | 91,2 | 91,2 | 91,2 | 89,7 | 79,1 | 89,7 |
| J48 | 54,4 | 58,8 | 67,6 | 72,1 | 88,2 | 85,3 | 85,3 | 85,3 | 85,3 | 75,8 | 85,3 |
| RandomTree | 57,4 | 50,0 | 52,9 | 67,6 | 83,8 | 86,8 | 83,8 | 82,4 | 89,7 | 72,7 | 82,4 |
| REPTree | 51,5 | 64,7 | 58,8 | 76,5 | 88,2 | 88,2 | 88,2 | 88,2 | 86,8 | 76,8 | 86,8 |
| SimpleCart | 55,9 | 51,5 | 70,6 | 80,9 | 92,6 | 92,6 | 92,6 | 92,6 | 89,7 | 79,9 | 89,7 |
| SVM - kernel:linear | 58,8 | 66,2 | 70,6 | 80,9 | 91,2 | 95,6 | 95,6 | 91,2 | 94,1 | 82,7 | 91,2 |
| SVM - kernel:polynomial | 52,9 | 64,7 | 69,1 | 75,0 | 77,9 | 85,3 | 85,3 | 82,4 | 88,2 | 75,7 | 77,9 |
| SVM - kernel:RBF | 57,4 | 55,9 | 64,7 | 80,9 | 89,7 | 89,7 | 89,7 | 88,2 | 92,6 | 78,8 | 88,2 |
| SVM - kernel:sigmoid | 63,2 | 66,2 | 72,1 | 83,8 | 91,2 | 88,2 | 91,2 | 88,2 | 95,6 | 82,2 | 88,2 |
| Naive Bayes | 60,3 | 64,7 | 72,1 | 80,9 | 88,2 | 89,7 | 88,2 | 88,2 | 92,6 | 80,6 | 88,2 |
| Rede Neural | 58,8 | 50,0 | 63,2 | 70,6 | 89,7 | 91,2 | 89,7 | 86,8 | 89,7 | 76,6 | 86,8 |
| KNN | 50,0 | 57,4 | 60,3 | 73,5 | 79,4 | 83,8 | 88,2 | 85,3 | 89,7 | 74,2 | 79,4 |
| **MÉDIA** | **54,2** | **56,0** | **65,0** | **74,5** | **85,1** | **86,7** | **87,1** | **85,6** | **88,1** | | |
| **MEDIANA** | **57,4** | **57,4** | **67,6** | **76,5** | **88,2** | **88,2** | **88,2** | **88,2** | **89,7** | | |
| **DESVIO PADRÃO** | **11,83** | **10,21** | **11,25** | **12,07** | **11,69** | **9,90** | **8,55** | **8,52** | **8,04** | | |

**Fonte:** José E. V. Borges (2019).

**Tabela 9**-Acurácias da predição dos algoritmos de classificação relativos ao tempo de curso utilizando apenas atributos Nota na Prova de Matemática (ENEM) e Notas das atividades do Huxley.

| Algoritmo | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | MÉDIA | MEDIANA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | 64,7 | 54,4 | 72,1 | 77,9 | 89,7 | 89,7 | 88,2 | 88,2 | 86,8 | 79,1 | 86,8 |
| NNge | 60,3 | 69,1 | 69,1 | 77,9 | 91,2 | 91,2 | 91,2 | 89,7 | 91,2 | 81,2 | 89,7 |
| OneR | 57,4 | 57,4 | 75,0 | 82,4 | 86,8 | 86,8 | 86,8 | 86,8 | 89,7 | 78,8 | 86,8 |
| Prism | 8,8 | 14,7 | 20,6 | 23,5 | 44,1 | 52,9 | 58,8 | 57,4 | 58,8 | 37,7 | 44,1 |
| Ridor | 57,4 | 66,2 | 72,1 | 80,9 | 91,2 | 88,2 | 88,2 | 88,2 | 89,7 | 80,2 | 88,2 |
| ADTree | 60,3 | 63,2 | 64,7 | 80,9 | 85,3 | 86,8 | 86,8 | 88,2 | 88,2 | 78,3 | 85,3 |
| J48 | 58,8 | 60,3 | 67,6 | 79,4 | 89,7 | 86,8 | 86,8 | 86,8 | 85,3 | 77,9 | 85,3 |
| RandomTree | 61,8 | 66,2 | 63,2 | 72,1 | 83,8 | 83,8 | 80,9 | 86,8 | 88,2 | 76,3 | 80,9 |
| REPTree | 66,2 | 72,1 | 69,1 | 77,9 | 88,2 | 88,2 | 88,2 | 88,2 | 88,2 | 80,7 | 88,2 |
| SimpleCart | 63,2 | 61,8 | 72,1 | 82,4 | 92,6 | 92,6 | 92,6 | 92,6 | 89,7 | 82,2 | 89,7 |
| SVM - kernel:linear | 58,8 | 69,1 | 75,0 | 80,9 | 92,6 | 91,2 | 91,2 | 89,7 | 92,6 | 82,3 | 89,7 |
| SVM - kernel:polynomial | 52,9 | 70,6 | 70,6 | 80,9 | 86,8 | 83,8 | 86,8 | 85,3 | 88,2 | 78,4 | 83,8 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SVM - kernel:RBF | **66,2** | 70,6 | 73,5 | 80,9 | **92,6** | **92,6** | 91,2 | 89,7 | 89,7 | **83,0** | **89,7** |
| SVM - kernel:sigmoid | **55,9** | 60,3 | 73,5 | 85,3 | 92,6 | 89,7 | 91,2 | 88,2 | **95,6** | **81,4** | **88,2** |
| Naive Bayes | **60,3** | 69,1 | 73,5 | 85,3 | 88,3 | 89,7 | 88,2 | 86,8 | **92,6** | **81,5** | **86,8** |
| Rede Neural | **61,8** | 70,6 | 72,1 | 82,4 | 88,2 | 89,7 | **91,2** | 89,7 | 89,7 | **81,7** | **88,2** |
| KNN | **61,8** | 69,1 | 64,7 | 70,6 | 79,4 | **89,7** | 88,2 | 85,3 | 88,2 | **77,4** | **79,4** |
| **MÉDIA** | **57,4** | **62,6** | **67,6** | **76,6** | **86,1** | **86,7** | **86,8** | **86,3** | **87,8** | | |
| **MEDIANA** | **60,3** | **66,2** | **72,1** | **80,9** | **88,3** | **89,7** | **88,2** | **88,2** | **89,7** | | |
| **DESVIO PADRÃO** | **13,01** | **13,41** | **12,63** | **14,21** | **11,39** | **9,07** | **7,74** | **7,69** | **7,85** | | |

**Fonte:** José E. V. Borges (2019).

A partir das análises dos resultados obtidos no Experimento 4, pode-se concluir que a Seleção de Atributos se confirmou como uma técnica que melhora a eficiência, a partir da melhoria geral de desempenho, relativa à qualidade dos dados. Houve melhoria significativa e gradativa à medida foram utilizados apenas os atributos mais relevantes, mesmo que estes representem conjuntos de dados bem menores quando comparados aos conjuntos de dados iniciais, que por sua vez são mais completos.

Como resposta à questão de pesquisa 4, mais precisamente a questão derivada 4.1, T4 seria o tempo mais marcante em termos de Eficiência, pois a partir dele a melhoria é pequena e irrelevante dado o tempo de curso. Ao mesmo tempo em T3, ou seja, a partir de aproximadamente 3 atividades de aprendizagem, já é possível obter uma acurácia significativa, nos melhores casos acima de 85%, e considerando a média geral, acima de 80% na predição dos alunos que se encontram em risco de serem reprovados na disciplina. Isso leva a crer de a parcela de informação agregada com a realização de atividades rotineiras durante o curso se mostra bastante relevante e auxilia diretamente na predição por parte dos algoritmos de classificação, permitindo inclusive a criação de modelos capazes de apoiar o professor na identificação dos alunos e atuação preventiva.

No que diz respeito às questões 4.2, 4.3, 4.4 e 4.5, pode-se concluir de forma sucinta que o modelo recomendado, que concentra os algoritmos mais eficientes e ao mesmo tempo os dados mais relevantes, tanto em T0 quanto durante o curso, pode ser mapeado da seguinte forma:

1. Para o Momento T0 – Utilizar apenas o atributo “Nota na Prova de Matemática (ENEM)”,e os seguintes algoritmos: JRip (Regra); REPTree (Árvore); ou SVM (kernel: Radial Basis Function).
2. Para o Momento >T0 (durante o curso) = Utilizar Apenas Notas das atividades no Huxley e os seguintes algoritmos: JRip (Regra); ADTree ou J48 ou RandomTree ou CART

(Árvore); ou SVM (kernel: Sigmoid).

A Tabela 10 a seguir resume o modelo apresentado como conclusão do Experimento 4. Pode-se observar adicionalmente que resultados médios de acurácia relacionados com o modelo proposto se encontram destacados na tabela.

**Tabela 10**-Resumo comparativo das Médias das acurácias de T0 a T8 utilizando conjuntos de dados diferentes, de acordo com análise do Experimento 4 e modelo sugerido.

| | Tempo de Curso | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dados Utilizados no processo | **T0** | **T1** | **T2** | **T3** | **T4** | **T5** | **T6** | **T7** | **T8** | **MÉDIA** |
| **Todos os Dados Socioeconômicos de T0 a T8 + Notas das Atividades Huxley de T1 a T8** | 49,0 | 52,4 | 60,0 | 68,9 | 78,4 | 78,9 | 79,5 | 80,1 | 81,2 | 69,8 |
| **Dados Socioeconômicos mais relevantes (Seleção de Atributos) de T0 a T8 + Notas das Atividades Huxley de T1 a T8** | 47,6 | 54,1 | 63,9 | 72,0 | 79,8 | 81,0 | 81,6 | 80,2 | 82,9 | 71,4 |
| **Dados Socioeconômicos mais relevantes (Seleção de Atributos) apenas em T0 + Notas das Atividades Huxley de T1 a T8** | 47,6 | ***65,0*** | ***71,2*** | ***80,1*** | ***86,5*** | ***86,2*** | ***86,8*** | ***86,1*** | ***88,1*** | ***77,5*** |
| **Dados das Notas do ENEM de T0 a T8 + Notas das Atividades Huxley de T1 a T8** | 54,2 | 56,0 | 65,0 | 74,5 | 85,1 | 86,7 | 87,1 | 85,6 | 88,1 | 75,8 |
| **Dados das Notas de Matemática do ENEM de T0 a T8 + Notas das Atividades Huxley de T1 a T8** | ***57,4*** | 62,6 | 67,6 | 76,6 | 86,1 | 86,7 | 86,8 | 86,3 | 87,8 | 77,5 |

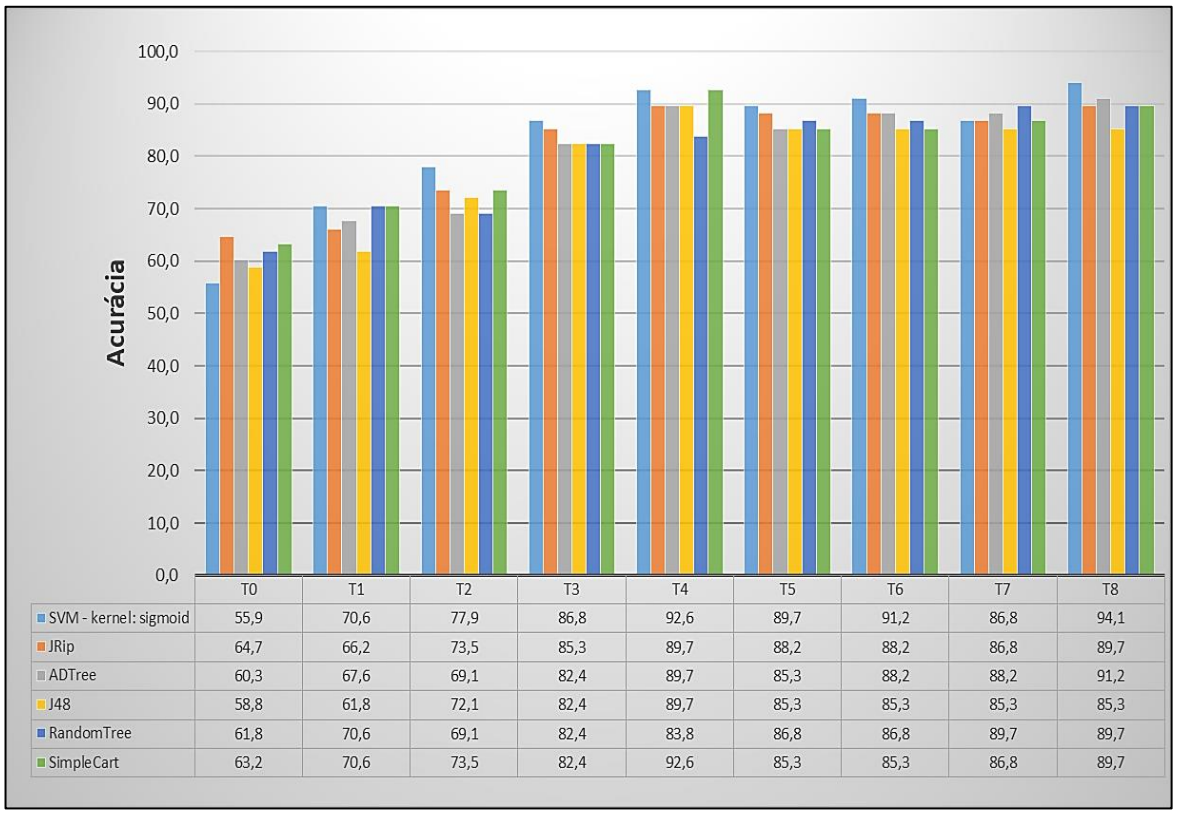
**Fonte:** José E. V. Borges (2019).

Uma vez definido um modelo de recomendação que atende aos objetivos das questões de pesquisa relacionadas ao Experimento 4, no intuito de explorar melhor as métricas de desempenho em aprendizagem de máquina, foi realizado um experimento específico em T3 utilizando outras métricas diferentes da Acurácia. O momento T3 foi escolhido por ser um momento marcante no que diz respeito às predições, devido ao fato de ser antes da primeira avaliação de aprendizagem e também pelo fato de apresentar acurácias satisfatórias, atingindo números superiores a 80% na média total dos resultados de acurácia alcançados pelos algoritmos quando utilizados apenas os dados das notas nas atividades do Huxley.

A Figura 34 são apresentados os resultados de acurácia individual dos algoritmos que se apresentaram mais eficientes em T3, utilizando apenas o atributo “Nota na Prova de Matemática (ENEM)” em T0 e apenas as notas das atividades do Huxley de T1 a T8. Foram selecionados 6 algoritmos mais eficientes citados no modelo de recomendação do Experimento 4, dentre eles o algoritmo mais eficiente do tipo Regra, o mais eficiente do tipo Árvore de

Decisão e também o mais eficiente Caixa Preta. Como ocorreu empate para os algoritmos de Árvore de Decisão, ou seja, 4 algoritmos empataram como mais eficientes, foram plotados todos os 4 no gráfico, totalizando 6 melhores resultados. Observa-se que em T0 os algoritmos Caixa Branca obtiveram destaque.

**Figura 35**- Acurácias individuais e comparativo dos 6 algoritmos que apresentaram maior eficiência em T3




| | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 |
|---|---|---|---|---|---|---|---|---|---|
| SVM - kernel: sigmoid | 55,9 | 70,6 | 77,9 | 86,8 | 92,6 | 89,7 | 91,2 | 86,8 | 94,1 |
| JRip | 64,7 | 66,2 | 73,5 | 85,3 | 89,7 | 88,2 | 88,2 | 86,8 | 89,7 |
| ADTree | 60,3 | 67,6 | 69,1 | 82,4 | 89,7 | 85,3 | 88,2 | 88,2 | 91,2 |
| J48 | 58,8 | 61,8 | 72,1 | 82,4 | 89,7 | 85,3 | 85,3 | 85,3 | 85,3 |
| RandomTree | 61,8 | 70,6 | 69,1 | 82,4 | 83,8 | 86,8 | 86,8 | 89,7 | 89,7 |
| SimpleCart | 63,2 | 70,6 | 73,5 | 82,4 | 92,6 | 85,3 | 85,3 | 86,8 | 89,7 |

**Fonte:** José E. V. Borges (2019).

A Tabela 11 apresenta as 4 métricas mais comuns em trabalhos na área de EDM, inclusive são as métricas utilizadas por Marquez; Veras e Soto (2013). São elas *True Positive Rate* (TP Rate), *True Negative Rate* (TP Rate), *Accuracy* (Acc) *e Geometric Mean* (GM), todas descritas no Capítulo 2. Os melhores resultados dentre cada métrica estão destacados em negrito. De forma geral pode-se perceber que as TP Rate superam as TN Rate, em outras palavras os algoritmos costumam acertar mais os casos onde o aluno foi classificado como aprovado quando ele realmente foi aprovado na disciplina. Já a TN Rate deixa explicito a taxa de acertos dentre aqueles alunos que foram classificados como reprovados e realmente foram reprovados na disciplina, o que é preocupante quando nosso objetivo principal em EDM é encontrar os alunos com dificuldade de aprendizado ou deficiência acadêmica. Baixas taxas de TN Rate mostram que o algoritmo está pouco eficiente na detecção dos alunos propensos à reprovação. É possível concluir, diante desta análise, que para EDM existem uma atenção maior

para as TN Rate. Nota-se também os destaques tanto em TP Rate quanto em TN Rate para o algoritmo SVM.

**Tabela 11-** TP Rate, TN Rate, Acurácia e *Geometric Mean* de cada algoritmo em T3.

| Algoritmo | TP Rate | TN Rate | Acc | GM |
|---|---|---|---|---|
| JRip | **94,1** | 76,5 | 85,3 | 84,8 |
| NNge | 58,8 | 79,4 | 69,1 | 68,3 |
| OneR | 82,4 | 82,4 | 82,4 | 82,4 |
| Prism | 90,6 | 34,4 | 58,8 | 55,8 |
| Ridor | 76,5 | 73,5 | 76,5 | 75,0 |
| ADTree | 88,2 | 76,5 | 82,4 | 82,1 |
| J48 | 88,2 | 76,5 | 82,4 | 82,1 |
| RandomTree | 88,2 | 76,5 | 82,4 | 82,1 |
| REPTree | 79,4 | 76,5 | 77,9 | 77,9 |
| SimpleCart | 88,2 | 76,5 | 82,4 | 82,1 |
| SVM - kernel:linear | 91,2 | 76,5 | 83,8 | 83,5 |
| SVM - kernel:polynomial | 79,4 | **88,2** | 83,8 | 83,7 |
| SVM - kernel:RBF | 85,3 | 76,5 | 80,9 | 80,8 |
| SVM - kernel:sigmoid | **94,1** | 79,4 | **86,8** | **86,4** |
| Naive Bayes | 82,4 | 79,4 | 80,9 | 80,9 |
| Rede Neural | 91,2 | 79,4 | 85,3 | 85,1 |
| KNN | 88,2 | 73,5 | 80,9 | 80,5 |

**Fonte:** José E. V. Borges (2019).

Dando continuidade à análise sobre a perspectiva dos alunos propensos ao fracasso, é interessante tentar extrair informações deste estudo exploratório que possam ser utilizadas pelos professores diretamente. Em outras palavras, o que o professor pode tirar de proveito deste estudo exploratório? Apesar de ser um estudo direcionado não só ao professor, mas também gestores, pesquisadores, desenvolvedores, especialistas, ou quaisquer grupos que possuam interesse na área de EDM, faz-se útil um breve levantamento sobre a contribuição deste trabalho exclusivamente para o professor.

É fato que a maioria das técnicas utilizadas e apresentadas nesta pesquisa podem não ser facilmente implementadas pelo professor sem um suporte computacional que faça o papel de classificador. A partir da aprendizagem de máquina é possível para um professor que disponha de uma ferramenta que implemente algum ou alguns dos algoritmos estudados para, a partir de dados de entrada fornecidos pelo próprio professor, gerem um relatório dos alunos

classificados como tendenciosos ao fracasso. Tal ferramenta precisa ser desenvolvida, o que denota que seria papel de um especialista da área de desenvolvimento de software. Uma das contribuições deste trabalho seria fornecer as informações necessárias para um possível desenvolvedor de ferramentas de software que deem este tipo de suporte a professores. O modelo recomendado no Experimento 4 poderia servir como base para esta ferramenta, fornecendo assim ao professor resultados mais precisos nos relatórios sobre os alunos. Uma segunda contribuição, desta vez de forma mais direta, seria a análise dos atributos relevantes apresentada no Experimento 1, pois, a partir das características dos alunos, um professor que possua acesso aos dados dos alunos poderia de certa forma identificar aqueles predispostos à reprovação através dos atributos mais relevantes, a exemplo da Nota de Matemática obtida pelo aluno na prova de seleção de ingresso à instituição, que se mostrou altamente relevante durante todo o estudo.

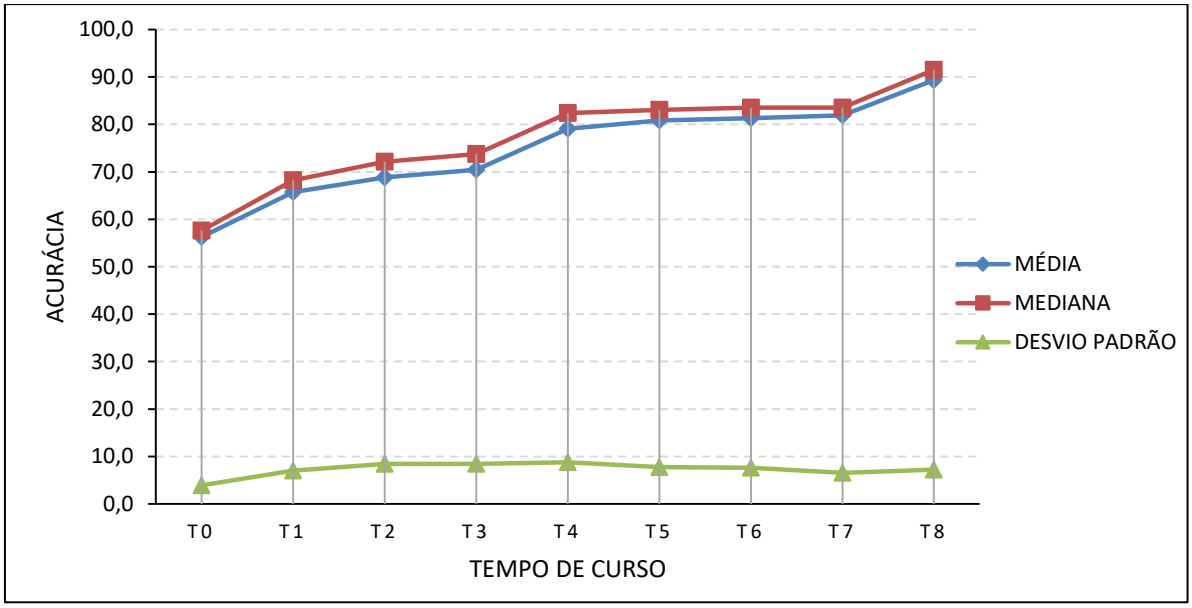
## 6.5 Estudo Comparativo

Dando continuidade ao estudo realizado no Experimento 4, buscando uma maior exploração à resposta da questão de pesquisa 4, pode-se concluir, a partir de todos os resultados obtidos no Experimento 4, que com a utilização de atividades de fixação e avaliativas quinzenais, um período de tempo que ficou bem marcado como tendo uma predição a níveis satisfatórios é o de 45 dias, ou seja, entre T0 e T3. Fundamenta-se tal conclusão a partir da exploração média considerada de todos os algoritmos, onde comparando-se os de Regras, Árvores de Decisão e Caixa preta, todos apresentaram uma mesma tendência, com leves variações. Uma outra conclusão obtida de forma paralela foi a relevância dos atributos das atividades aplicadas com os alunos, que se mostraram significativamente mais valiosos do que os socioeconômicos. Mesmo quando utilizados de forma conjunta com as notas das atividades do Huxley, os atributos socioeconômicos não agregaram informação relevante ao treino e teste dos algoritmos, de forma geral, desfavorecendo a eficiência alcançada pelos algoritmos quando utilizados tendo apenas dados das atividades como entrada.

Considerando que os atributos socioeconômicos não representam dados que agreguem informação adicional ao processo de predição quando utilizando atividades de aprendizagem, um estudo mais representativo, relacionado também à questão de pesquisa 4, foi realizado utilizando um conjunto maior de instâncias, buscando dessa forma aproximar-se do quantitativo de instâncias utilizados por Marquez Vera em [1], [2]. Apenas dados do Huxley e 5 socioeconômicos contidos no Sistema Acadêmico foram utilizados como atributos de entrada, porém referente ao período de 2013.2 a 2017.1, das mesmas disciplinas, Engenharia da

Computação e Ciência da Computação. No total desta experimentação, foram utilizados apenas 14 atributos, sendo 6 de atividades, 5 socioeconômicos extraídos do sistema acadêmico da Ufal SIEWEB, e 3 atributos acadêmicos: Notas das Avaliações Bimestrais 1 e 2 e Conceito Final. No total foram 544 instâncias de entrada, considerando os alunos matriculados nas 2 disciplinas no período mencionado, de 4 anos. Os resultados são exibidos na Figura 35 e Tabela 12. A partir de T1, os 5 atributos socioeconômicos foram desconsiderados, passando a ser considerado apenas as notas das atividades do Huxley de T1 em diante.

**Figura 36-** Média, Mediana e Desvio padrão das acurácias dos algoritmos em cada parcela de tempo no decorrer do curso utilizando apenas atributos das atividades do Huxley e referente aos períodos 2013.2 a 2017.1.




**Fonte:** José E. V. Borges (2019).

**Tabela 12-** Acurácias da predição dos algoritmos de classificação relativos ao tempo de curso e informações adquiridas nas atividades de aprendizagem utilizando apenas atributos das atividades do Huxley e referente aos períodos de 2013.2 a 2017.1

| **Algoritmo** | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| JRip | **59,1** | 71,0 | 73,9 | 75,1 | 83,4 | 84,4 | 84,5 | 83,6 | **92,5** | **78,6** | **83,4** |
| NNge | **55,9** | 61,9 | 68,2 | 72,5 | 79,2 | 82,7 | 85,0 | 85,3 | **92,3** | **75,9** | **79,2** |
| OneR | **56,0** | 68,4 | 75,2 | 75,2 | 83,6 | 83,6 | 83,6 | 82,6 | **91,0** | **77,7** | **82,6** |
| Prism | **44,0** | 46,9 | 46,4 | 49,2 | 47,6 | 52,0 | 53,1 | 58,0 | **62,7** | **51,1** | **49,2** |
| Ridor | **58,3** | 68,1 | 72,1 | 73,8 | 78,2 | 82,1 | 82,9 | 82,9 | **91,7** | **76,7** | **78,2** |
| ADTree | **58,8** | 70,5 | 73,6 | 75,4 | 82,7 | 82,6 | 82,7 | 85,5 | **91,5** | **78,2** | **82,6** |
| J48 | **57,7** | 70,5 | 72,6 | 76,5 | 83,9 | 81,8 | 82,2 | 82,9 | **90,9** | **77,7** | **81,8** |
| RandomTree | **55,9** | 66,1 | 64,8 | 65,5 | 77,0 | 82,7 | 81,1 | 82,6 | **92,4** | **74,2** | **77,0** |
| REPTree | **59,6** | 69,7 | 71,7 | 75,6 | 82,6 | 83,2 | 83,9 | 83,9 | **91,2** | **77,9** | **82,6** |
| SimpleCart | **58,8** | 70,0 | 71,8 | 75,2 | 84,2 | 83,7 | 84,0 | 83,1 | **92,0** | **78,1** | **83,1** |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SVM - kernel:linear | **58,0** | 69,2 | 73,5 | 73,1 | 83,6 | 84,2 | 84,7 | 85,3 | **92,0** | **78,2** | **83,6** |
| SVM - kernel:polynomial | **55,5** | 68,2 | 73,3 | 74,8 | 81,6 | 83,1 | 83,1 | 83,6 | **91,2** | **77,1** | **81,6** |
| SVM - kernel:RBF | **58,3** | 69,1 | 73,9 | 74,3 | 83,4 | 84,5 | 84,9 | 85,7 | **92,8** | **78,5** | **83,4** |
| SVM - kernel:sigmoid | **49,5** | 49,5 | 49,5 | 49,5 | 71,7 | 74,9 | 75,2 | 76,1 | **83,1** | **64,3** | **71,7** |
| Naive Bayes | **57,3** | 66,8 | 71,2 | 71,5 | 79,6 | 81,4 | 81,9 | 82,7 | **89,3** | **75,7** | **79,6** |
| Rede Neural | **58,5** | 68,1 | 73,1 | 72,1 | 82,4 | 84,4 | 84,4 | 84,7 | **92,0** | **77,7** | **82,4** |
| KNN | **55,9** | 63,0 | 65,0 | 68,7 | 80,1 | 83,4 | 85,0 | 84,5 | **90,7** | **75,1** | **80,1** |
| **MÉDIA** | **56,3** | **65,7** | **68,8** | **70,5** | **79,1** | **80,9** | **81,3** | **81,9** | **89,4** | | |
| **MEDIANA** | **57,7** | **68,2** | **72,1** | **73,8** | **82,4** | **83,1** | **83,6** | **83,6** | **91,5** | | |
| **DESVIO PADRÃO** | **3,94** | **7,06** | **8,41** | **8,41** | **8,75** | **7,77** | **7,63** | **6,55** | **7,22** | | |

**Fonte:** José E. V. Borges (2019).

Tanto o gráfico da Figura 35 quanto a Tabela 12, enfatizam a tendência geral apresentada anteriormente e confirmam o resultado obtido para 2013.2. As acurácias alcançadas ficaram bem próximas numericamente, devido à relação de número de instâncias alto para poucos atributos, o que tende a resultados mais estáveis e com menos variações no decorrer do curso. Curiosamente, um fato que chamou atenção é o nível da acurácia alcançado por Marquez Vera (2013) com o algoritmo Prism: no referido trabalho ele alcança a marca de 94,7% de acurácia, enquanto a maior acurácia alcançada por este algoritmo neste estudo realizado é de 65,7%.

**Tabela 13**- Comparativo entre os resultados obtidos com os dados da UFAL no período de 2013.2 a 2017.1, utilizando apenas dados do Huxley, e os resultados obtidos por Marquez Vera (2013).

| | **Dados UFAL - 2013.2 a 2017.1** | | | | **Marquez Vera** | | | |
|---|---|---|---|---|---|---|---|---|
| **Algoritmo** | **TP Rate** | **TN Rate** | **Acc** | **GM** | **TP Rate** | **TN Rate** | **Acc** | **GM** |
| JRip | 92,2 | 93,2 | 92,7 | 92,7 | 97,3 | 65,0 | 94,8 | 78,8 |
| NNge | 91,2 | 93,2 | 92,2 | 92,2 | 98,7 | 78,3 | 96,9 | 87,1 |
| OneR | 89,6 | 91,9 | 90,7 | 90,7 | 88,8 | **88,3** | 88,8 | 88,3 |
| Prism | 93,0 | 82,6 | 63,0 | 87,6 | **99,8** | 37,1 | 94,7 | 59,0 |
| Ridor | 90,6 | **94,5** | 92,5 | 92,5 | 97,9 | 70,0 | 95,4 | 81,4 |
| ADTree | 91,5 | 90,2 | 90,9 | 90,8 | 98,2 | 86,7 | **97,2** | **92,1** |
| J48 | **93,5** | 92,2 | **92,8** | **92,8** | 96,7 | 75,0 | 94,8 | 84,8 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| RandomTree | 92,5 | 91,5 | 92,0 | 92,0 | 96,1 | 68,3 | 93,6 | 79,6 |
| REPTree | 90,6 | 92,8 | 91,7 | 91,7 | 96,5 | 75,0 | 94,6 | 84,6 |
| SimpleCart | 91,9 | 92,2 | 92,0 | 92,0 | 96,4 | 76,7 | 94,6 | 85,5 |

**Fonte:** José E. V. Borges (2019).

O Fato que mais aproxima e ao mesmo tempo reforça os resultados encontrados mor Marquez Vera (2013) é a conclusão permitida a partir da questão de pesquisa 4 deste estudo, onde o T4 é o momento mais favorável para a realização de uma predição através de EDM. Marquez Vera identificou e recomendou em seu trabalho anterior que predições confiáveis podem ser alcançadas dentro da faixa de 4 a 6 semanas de curso, com a aplicação de atividades semanais para o ganho de informação sobre os alunos. Considerando que se as atividades propostas pelo professor ao invés de serem aplicadas quinzenalmente fossem aplicadas semanalmente, considerando ainda que praticamente toda a informação relevante para a predição está concentrada nas atividades periódicas realizadas durante o curso, pode-se afirmar seguramente que os resultados satisfatórios encontrados para T4 também seriam alcançados em 4 semanas, ao invés de 4 quinzenas. Este momento encontrado converge para o período encontrado pelo Marquez Vera. Na Tabela 13 é apresentado um resumo comparativo entre os resultados obtidos neste estudo e os resultados alcançados pelo Marquez Vera.

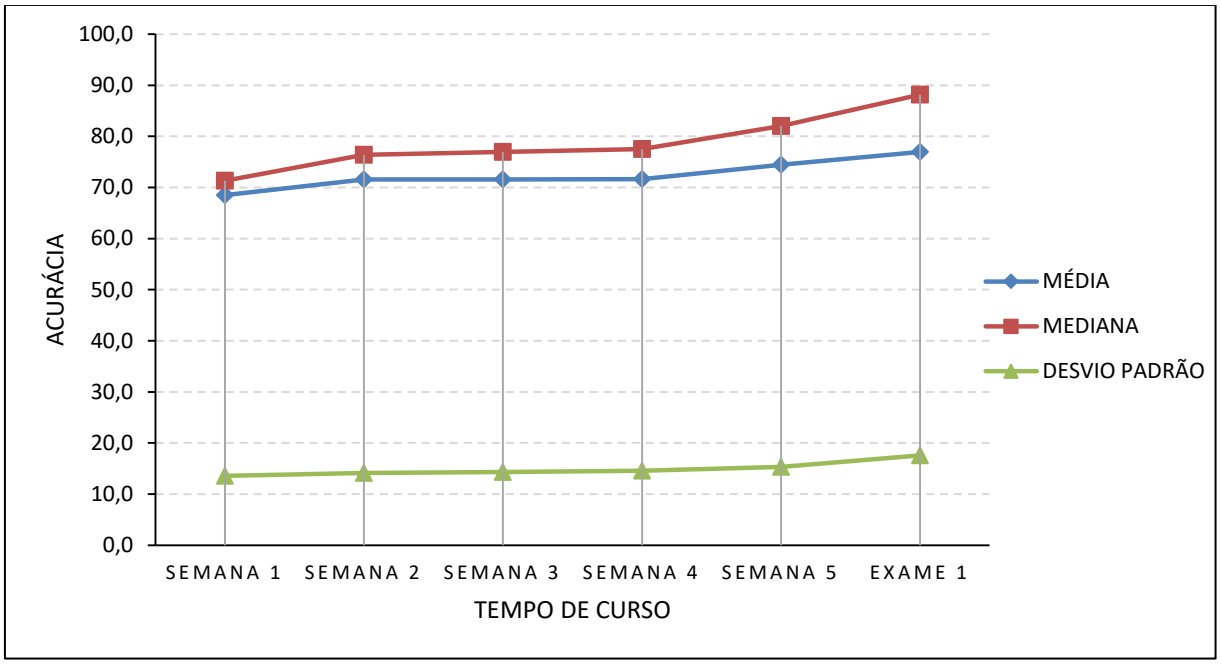
Um outro estudo comparativo foi realizado ainda para reforçar a resposta à questão 4: Baseado no trabalho de Costa et al., (2017) foram utilizados na presente pesquisa os mesmos dados explorados na referida pesquisa de 2017:

a. EAD: contém informações sobre 262 estudantes de graduação que fizeram o curso introdutório de programação na modalidade de educação a distância em nossa universidade em 2013, durante 10 semanas. Neste curso os alunos foram avaliados semanalmente de acordo com suas atividades e mais dois exames aplicados na quinta e última semana do curso Costa et al., (2017). Essas atividades e exames foram aplicados por meio da Plataforma *Moodle* ao invés da Plataforma Huxley.
b. Presencial: A segunda fonte de dados contém informações sobre 161 alunos que fizeram o curso de programação introdutória realizado no campus em nossa universidade em 2014, durante 16 semanas Costa et al., (2017). Neste curso, os alunos foram avaliados semanalmente por atividades na Plataforma *Huxley*.

Pode-se observar que também há tendência de crescimento de acurácia, embora seja um crescimento mais sutil em relação ao tempo, vistos de forma mais explícita nas Tabelas 14 e 15 e Figuras 36 e 37. Um ponto importante a salientar são os níveis de acurácia superiores alcançados já na primeira semana de curso, equivalente a T1, tanto presencial quanto EAD,

chegando em média próximo a 70%. Dá-se isso ao fato de que nestes dados estão presentes informações de grande relevância dos alunos, como por exemplo quantidade de postagens feitas pelo aluno no ambiente online. Outro fator que contribuiu para a eficiência temporal com estes dados é a quantidade de instâncias.

**Figura 37**- Média, Mediana e Desvio Padrão das acurácias dos 14 Algoritmos de classificação aplicados aos dados da turma EAD 2013




**Fonte:** José E. V. Borges (2019).

**Tabela 14**- Acurácias dos 14 Algoritmos de Classificação aplicados aos dados EAD2013.

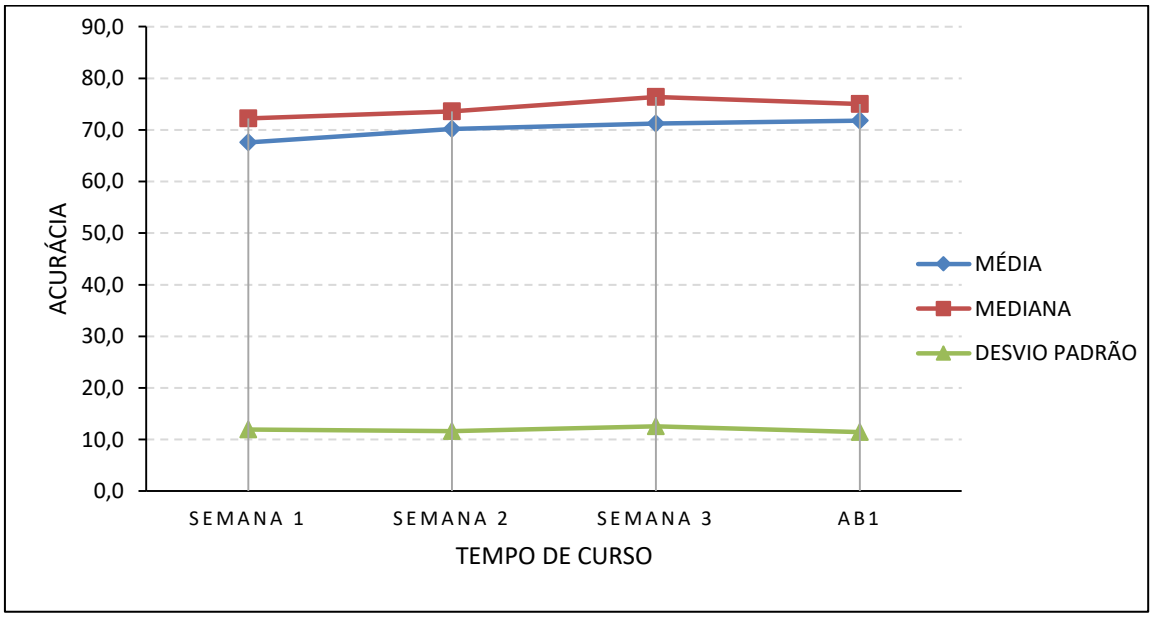
| Algoritmo | Semana 1 | Semana 2 | Semana 3 | Semana 4 | Semana 5 | Exame 1 | MÉDIA | MEDIANA |
|---|---|---|---|---|---|---|---|---|
| JRip | **71,3** | 79,8 | 73,6 | 81,5 | 79,8 | **90,4** | **68,1** | **79,8** |
| NNge | **71,3** | 74,2 | 77,3 | 78,1 | 83,1 | **88,8** | **67,5** | **77,3** |
| OneR | 49,4 | 49,4 | 49,4 | 49,4 | 49,4 | 49,4 | **42,4** | **49,4** |
| Prism | **52,8** | 57,9 | 57,9 | 60,7 | 64,0 | **65,7** | **51,3** | **57,9** |
| Ridor | **75,3** | **75,8** | 78,7 | 78,7 | 82,0 | **91,0** | **68,8** | **78,7** |
| ADTree | **75,3** | 80,9 | 77,5 | 84,3 | 86,5 | **89,3** | **70,5** | **80,9** |
| J48 | **65,7** | 76,4 | 75,8 | 77,5 | 84,8 | **89,3** | **67,1** | **76,4** |
| RandomTree | 63,5 | **78,1** | 73,6 | **61,8** | 74,2 | 65,2 | **59,5** | **65,2** |
| REPTree | 52,8 | 52,8 | 52,8 | 52,8 | 52,8 | 52,8 | **45,3** | **52,8** |
| SimpleCart | 60,7 | 60,7 | 60,7 | 60,7 | 60,7 | 60,7 | **52,0** | **60,7** |
| SVM - kernel:linear | 85,4 | 87,1 | 86,0 | **84,8** | 87,6 | **89,9** | **74,4** | **86,0** |
| SVM - kernel:polynomial | 76,4 | 74,7 | 77,0 | **74,2** | 82,6 | **89,9** | **67,8** | **76,4** |
| SVM - kernel:RBF | 78,7 | **77,0** | 77,5 | 77,5 | 79,2 | 80,9 | **67,3** | **77,5** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SVM - kernel:sigmoid | **39,3** | **39,3** | **39,3** | **39,3** | **39,3** | **38,8** | **33,6** | **39,3** |
| Naive Bayes | **80,9** | 82,6 | **86,0** | 83,1 | 84,8 | 85,4 | **71,8** | **83,1** |
| Rede Neural | **85,4** | **85,4** | 86,0 | 86,5 | 89,3 | **92,7** | **75,0** | **86,0** |
| KNN | **80,9** | 84,8 | **88,2** | 86,5 | 86,0 | **88,2** | **73,5** | **86,0** |
| **MÉDIA** | **68,5** | **71,6** | **71,6** | **71,6** | **74,5** | **77,0** | | |
| **MEDIANA** | **71,3** | **76,4** | **77,0** | **77,5** | **82,0** | **88,2** | | |
| **DESVIO PADRÃO** | **13,59** | **14,14** | **14,35** | **14,58** | **15,34** | **17,59** | | |

**Fonte:** José E. V. Borges (2019).

Uma lacuna encontrada na análise temporal da pesquisa citada é o fato de que é explorado apenas o período até a primeira avaliação bimestral, ficando desconhecidos os resultados de desempenho até o final do curso. Uma análise mais abrangente pôde ser apresentada no presente trabalho, estendendo a análise comparativa e confirmando o que acontece após a primeira avaliação bimestral: para todos os grupos de algoritmos estudados, o desempenho permanece praticamente constante após T4.

**Figura 38**- Média, Mediana e Desvio Padrão das acurácias dos 14 Algoritmos de classificação aplicados aos dados da turma Presencial 2014.




**Fonte:** José E. V. Borges (2019).

**Tabela 15-** Acurácias dos 14 Algoritmos de Classificação aplicados aos dados Presencial 2014.

| **Algoritmo** | Semana 1 | Semana 2 | Semana 3 | AB1 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|
| JRip | **72,2** | 73,6 | **83,3** | 77,8 | **76,7** | **75,7** |
| NNge | **72,2** | 79,2 | **80,6** | 76,4 | **77,1** | **77,8** |
| OneR | **79,2** | **79,2** | **79,2** | **73,6** | **77,8** | **79,2** |

| | | | | | | |
|---|---|---|---|---|---|---|
| Prism | 37,5 | 44,4 | 44,4 | **47,2** | **43,4** | **44,4** |
| Ridor | **72,2** | **72,2** | 76,4 | 75,0 | **74,0** | **73,6** |
| ADTree | 77,8 | 77,8 | **83,3** | **76,4** | **78,8** | **77,8** |
| J48 | **73,6** | **81,9** | 77,8 | **73,6** | **76,7** | **75,7** |
| RandomTree | 79,2 | **75,0** | 79,2 | **83,3** | **79,2** | **79,2** |
| REPTree | **75,0** | 79,2 | 79,2 | **81,9** | **78,8** | **79,2** |
| SimpleCart | **73,6** | **79,2** | **79,2** | 75,0 | **76,7** | **77,1** |
| SVM - kernel:linear | **59,7** | 68,1 | 65,3 | **80,6** | **68,4** | **66,7** |
| SVM - kernel:polynomial | **69,4** | 72,2 | 73,6 | **76,4** | **72,9** | **72,9** |
| SVM - kernel:RBF | 56,9 | 56,9 | 56,9 | 56,9 | **56,9** | **56,9** |
| SVM - kernel:sigmoid | 44,4 | 44,4 | 44,4 | 44,4 | **44,4** | **44,4** |
| Naive Bayes | **61,1** | 62,5 | **61,1** | **68,1** | **63,2** | **61,8** |
| Rede Neural | **72,2** | 75,0 | 76,4 | **79,2** | **75,7** | **75,7** |
| KNN | 72,2 | 72,2 | **70,8** | **75,0** | **72,6** | **72,2** |
| **MÉDIA** | **67,6** | **70,2** | **71,2** | **71,8** | | |
| **MEDIANA** | **72,2** | **73,6** | **76,4** | **75,0** | | |
| **DESVIO PADRÃO** | **11,91** | **11,60** | **12,53** | **11,43** | | |

**Fonte:** José E. V. Borges (2019).

Uma terceira análise comparativa pôde ser realizada ainda como contribuição ao presente estudo exploratório: a metodologia adotada nesta pesquisa foi também aplicada a uma outra base de dados de exemplo real, permitindo assim uma comparação de resultados. Trata-se de dados de 21 alunos da disciplina de Programação I, ministrada no curso de Técnico em Informática do Instituto Federal de Alagoas - IFAL, no ano de 2017. Como finalidade, pretende-se avaliar se as conclusões obtidas a partir da pesquisa podem ser transferidas para outras bases.

Por se tratar de um curso Técnico de Nível Médio, algumas diferenças marcantes devem ser levadas em consideração:

1) Trata-se de uma realidade de alunos mais jovens, entre 17 e 20 anos;
2) As Turmas são anuais, com avaliações de aprendizagem bimestrais;
3) Há 7 atributos Pessoais dos alunos disponíveis em T0: Idade, Sexo, Cotista, Posição do aluno na classificação da seleção de ingresso, Nota de Português na Seleção de ingresso, Nota de Matemática na Seleção de ingresso, Nota Geral na seleção de ingresso.
4) Além destes, também existem atividades de fixação aplicadas com periodicidade aproximadamente quinzenal através da mesma plataforma Huxley. Em média, por bimestre, são aplicadas 5 atividades de fixação.

De forma semelhante aos experimentos realizados nesta pesquisa, pretende-se analisar particularmente 2 momentos: T0 ou início do Curso, e T1 em diante, que representa o período

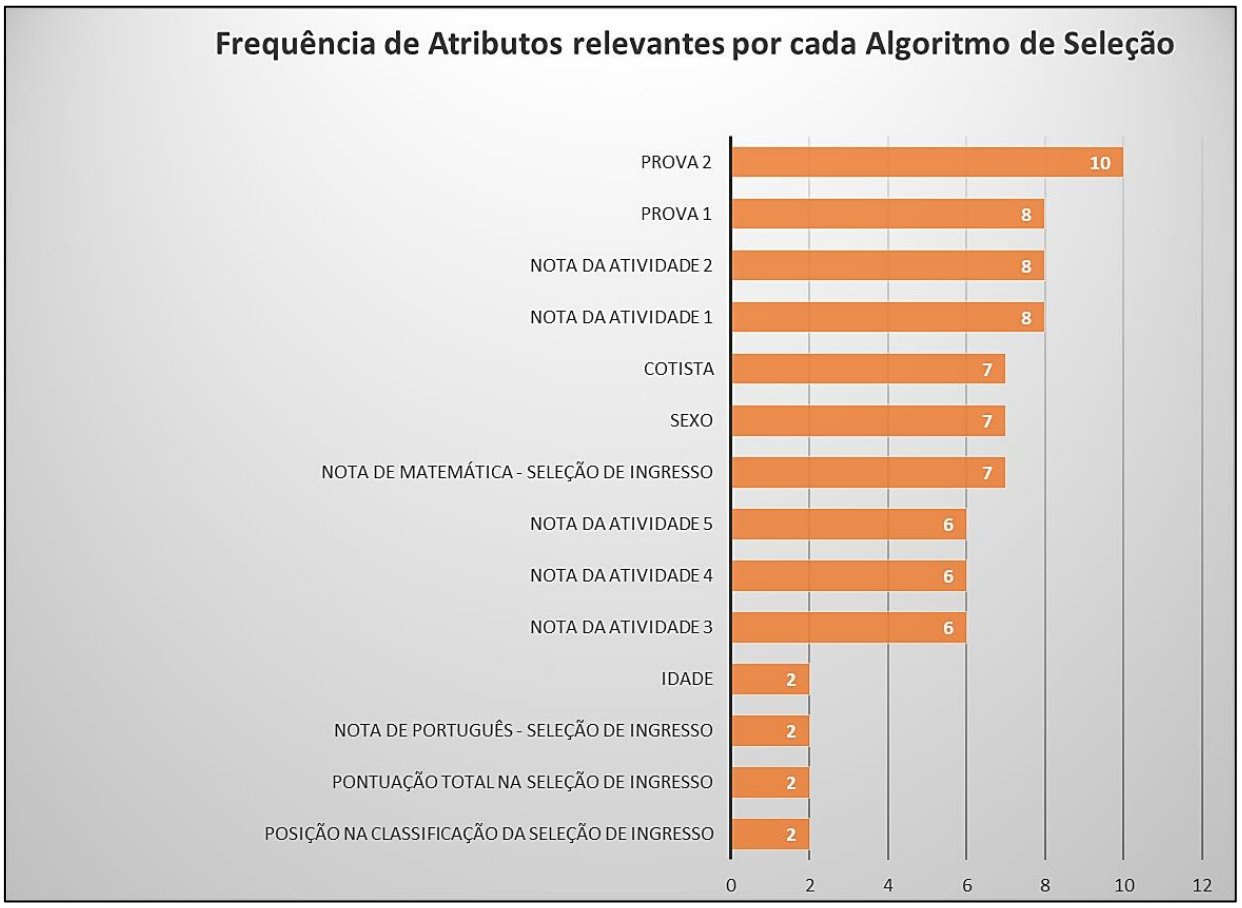
durante o curso. Inicialmente, mede-se a relevância de cada um dos atributos através da técnica de Seleção de Atributos.

Em seguida aplica-se o modelo encontrado na conclusão do Experimento 4, mencionado anteriormente, para comprovar se realmente trata-se de um modelo capaz de promover a Eficiência na predição.

A Figura 38 apresenta graficamente um ranking dos atributos presentes na base de dados do IFAL, classificado do mais relevante para o menos relevante. O Processo de Seleção de Atributos foi idêntico ao descrito no Experimento 1 e realizado nesta pesquisa até então.

As tabelas a seguir apresentam os resultados das acurácias individuais de todos os algoritmos, de T0 a T7, referente à análise com os dados do IFAL utilizando todos os atributos como entrada, na Tabela 16, e apenas a Nota da Prova de Matemática para a seleção juntamente com dados das atividades do Huxley, na Tabela 17. Estas duas tabelas permitem uma análise comparativa sobre o modelo sugerido no subcapítulo 5.4.

**Figura 39**- Resultado do processo de Seleção de Atributos utilizando os dados do IFAL.




**Fonte:** José E. V. Borges (2019).

**Tabela 16**-Acurácia dos algoritmos utilizando dados do IFAL, com dados pessoais, dados da seleção de ingresso e socioeconômicos de T0 a T8, e Notas do Huxley de T1 a T8.

| **Algoritmo** | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|
| JRip | **45,5** | 86,4 | 86,4 | 81,8 | **90,9** | **90,9** | **90,9** | 86,4 | **73,2** | **86,4** |
| NNge | **54,5** | 77,3 | 86,4 | 86,4 | 90,9 | 90,9 | 90,9 | **95,5** | **74,7** | **86,4** |
| OneR | **45,5** | 72,7 | 81,8 | 81,8 | **90,9** | **90,9** | **90,9** | **90,9** | **71,7** | **81,8** |
| Prism | **22,7** | 59,1 | **72,7** | **72,7** | **72,7** | **72,7** | **72,7** | **72,7** | **57,6** | **72,7** |
| Ridor | **40,9** | **86,4** | **86,4** | **86,4** | **86,4** | **86,4** | **86,4** | **86,4** | **71,7** | **86,4** |
| ADTree | **54,5** | 81,8 | **90,9** | 86,4 | **90,9** | **90,9** | **90,9** | **90,9** | **75,2** | **90,9** |
| J48 | **77,3** | 81,8 | 77,3 | 77,3 | **90,9** | **90,9** | **90,9** | **90,9** | **75,2** | **81,8** |
| RandomTree | **50,0** | 77,3 | 72,7 | 72,7 | 95,5 | 81,8 | 77,3 | **100,0** | **69,7** | **77,3** |
| REPTree | **50,0** | 72,7 | **90,9** | **90,9** | **90,9** | **90,9** | **90,9** | **90,9** | **74,2** | **90,9** |
| SimpleCart | **45,5** | 86,4 | 81,8 | 81,8 | **90,9** | **90,9** | **90,9** | **90,9** | **73,2** | **86,4** |
| SVM - kernel:linear | **59,1** | 77,3 | 90,9 | **95,5** | **95,5** | **95,5** | 90,9 | 90,9 | **77,3** | **90,9** |
| SVM - kernel:polynomial | **50,0** | 86,4 | 86,4 | **90,9** | **90,9** | **90,9** | 86,4 | **90,9** | **74,7** | **86,4** |
| SVM - kernel:RBF | **36,4** | 63,6 | 68,2 | 68,2 | 68,2 | 68,2 | 63,6 | **72,7** | **56,6** | **68,2** |
| SVM - kernel:sigmoid | 45,5 | 45,5 | 45,5 | 45,5 | 45,5 | 45,5 | 45,5 | 45,5 | **40,4** | **45,5** |
| Naive Bayes | **50,0** | 72,7 | 81,8 | 81,8 | **90,9** | **90,9** | **90,9** | **90,9** | **72,2** | **81,8** |
| Rede Neural | **68,2** | 72,7 | 81,8 | 81,8 | 81,8 | 81,8 | 86,4 | **90,9** | **71,7** | **81,8** |
| KNN | **72,7** | 81,8 | 77,3 | 81,8 | 77,3 | 77,3 | **81,8** | **81,8** | **70,2** | **77,3** |
| **MÉDIA** | **51,1** | **75,4** | **79,9** | **80,2** | **84,8** | **84,0** | **83,4** | **85,8** | | |
| **MEDIANA** | **50,0** | **77,3** | **81,8** | **81,8** | **90,9** | **90,9** | **90,9** | **90,9** | | |
| **DESVIO PADRÃO** | **13,21** | **11,02** | **11,14** | **11,36** | **12,75** | **12,46** | **12,54** | **12,54** | | |

**Fonte:** José E. V. Borges (2019).

**Tabela 17**-Acurácia dos algoritmos utilizando dados do IFAL, com apenas Nota de Matemática na seleção de ingresso em T0, e Notas do Huxley de T1 a T8.

| **Algoritmo** | T0 | T1 | T2 | T3 | T4 | T5 | T6 | T7 | **MÉDIA** | **MEDIANA** |
|---|---|---|---|---|---|---|---|---|---|---|
| JRip | **45,5** | 86,4 | 77,3 | 86,4 | 90,9 | 86,4 | **100,0** | 86,4 | **73,2** | **86,4** |
| NNge | **50,0** | 72,7 | 90,9 | 90,9 | 95,5 | 95,5 | 95,5 | **95,5** | **76,3** | **90,9** |
| OneR | **54,5** | 72,7 | 81,8 | 81,8 | **90,9** | **90,9** | **90,9** | **90,9** | **72,7** | **81,8** |
| Prism | **27,3** | 50,0 | **86,4** | **86,4** | 81,8 | 81,8 | 81,8 | 77,3 | **63,6** | **81,8** |
| Ridor | **50,0** | 81,8 | **86,4** | **86,4** | **86,4** | **86,4** | **86,4** | **86,4** | **72,2** | **86,4** |
| ADTree | **50,0** | 86,4 | **90,9** | 86,4 | **90,9** | **90,9** | **90,9** | **90,9** | **75,2** | **90,9** |
| J48 | **59,1** | 86,4 | 81,8 | 81,8 | **90,9** | **90,9** | **90,9** | **90,9** | **74,7** | **86,4** |
| RandomTree | **50,0** | 86,4 | 90,9 | 86,4 | 90,9 | 81,8 | **100,0** | 95,5 | **75,8** | **86,4** |
| REPTree | **50,0** | 72,7 | **90,9** | **90,9** | **90,9** | **90,9** | **90,9** | **90,9** | **74,2** | **90,9** |
| SimpleCart | **63,6** | 86,4 | 81,8 | 81,8 | **90,9** | **90,9** | **90,9** | **90,9** | **75,2** | **86,4** |
| SVM - kernel:linear | **63,6** | 77,3 | 90,9 | 95,5 | 95,5 | 95,5 | **100,0** | **100,0** | **79,8** | **95,5** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SVM - kernel:polynomial | **68,2** | 90,9 | 90,9 | **100,0** | **100,0** | **100,0** | **100,0** | **100,0** | **83,3** | **100,0** |
| SVM - kernel:RBF | **45,5** | 90,9 | 95,5 | **100,0** | **100,0** | 90,9 | 68,2 | 72,7 | **73,7** | **90,9** |
| SVM - kernel:sigmoid | **54,5** | **77,3** | **77,3** | 63,6 | 59,1 | 59,1 | 59,1 | 59,1 | **56,6** | **59,1** |
| Naive Bayes | **50,0** | 77,3 | 90,9 | 95,5 | **100,0** | 95,5 | 95,5 | 95,5 | **77,8** | **95,5** |
| Rede Neural | **63,6** | 86,4 | 90,9 | 90,9 | 95,5 | 90,9 | **95,5** | **95,5** | **78,8** | **90,9** |
| KNN | **50,0** | 86,4 | **90,9** | **90,9** | **90,9** | 86,4 | **90,9** | **90,9** | **75,2** | **90,9** |
| **MÉDIA** | **52,7** | **80,5** | **87,4** | **88,0** | **90,6** | **88,5** | **89,8** | **88,8** | | |
| **MEDIANA** | **50,0** | **86,4** | **90,9** | **86,4** | **90,9** | **90,9** | **90,9** | **90,9** | | |
| **DESVIO PADRÃO** | **9,51** | **10,00** | **5,46** | **8,50** | **9,44** | **8,96** | **11,20** | **10,43** | | |

**Fonte:** José E. V. Borges (2019).

A partir dos resultados apresentados utilizando-se a base de dados do IFAL, é possível concluir que o modelo criado e sugerido na pesquisa realizada com os dados da UFAL, como resposta à Questão 4, foi adequado também para se obter maior Eficiência na predição de sucesso dos alunos do IFAL. A Nota de Matemática mais uma vez, quando analisada de forma mais criteriosa, representa um fator relevante para a predição em T0. Na análise de T1 em diante, observa-se que quando utilizando apenas notas das atividades de fixação do Huxley os resultados de acurácia também são melhores, assim como visto no Experimento 4 utilizando os dados da UFAL, chegando a uma média de 90,6% de acurácia em T4, quando considerando a média de acurácia de todos os algoritmos.

JRip atingiu mais uma vez acurácias acima de 80.0% e 85.0% em T3 e T4 respectivamente. SVM mais uma vez se destacou dentre todos, atingindo em alguns casos assertividade máxima, 100%. A maioria dos Algoritmos de árvore também apresentou resultado de destaque em T3 e T4.

## 6.6 Ameaças de Pesquisa

Um fator que representa uma ameaça ao estudo realizado neste trabalho é quantitativo de instâncias utilizados para as análises dos atributos oriundos da fonte ENEM. Devido às limitações encontradas no decorrer do processo da pesquisa e apresentadas no capítulo 4, para a coleta e unificação dos dados, de forma que os dados das 3 fontes compreendam aos mesmos períodos, não foi possível a utilização de uma amostra maior em instâncias para o estudo em T0, ou seja, com os dados socioeconômicos. Ainda assim foi possível a observação de tendências entre os diversos algoritmos estudados, algumas confirmando resultados obtidos por outros trabalhos relacionados. Trabalhos futuros podem estender a metodologia aqui aplicada vislumbrando a possibilidade de exploração das mesmas análises com amostras maiores.

Pode-se citar também como ameaça ao trabalho, em menor grau de criticidade, a

ausência da aplicação de um teste estatístico que venha a reforçar as hipóteses aqui levantadas. Apesar da proposta da pesquisa ser de um estudo descritivo e exploratório, que dê evidências com dados numéricos para mensurar os desempenhos das predições realizadas em cada situação, explorando os aspectos que possam vir a influenciar da eficiência, a utilização de ferramentas da Estatística Inferencial poderiam apresentar conclusões mais direcionadas à tomada de decisão pelos professores, gestores e especialistas.

A utilização da média aritmética como métrica principal para a análise dos comportamentos dos grupos de algoritmos e comparativos, representa também uma limitação aos resultados apresentados. A média aritmética possui a característica de sofrer influências dos valores discrepantes, o que pode acarretar na distorção de alguns resultados. A utilização conjunta da mediana e a apresentação dos resultados também individuais de cada atributo, foram as medidas aplicadas para suprir tal ameaça.

# 6 CONCLUSÃO

Após as análises realizadas nesta pesquisa, de forma geral, foi possível compreender um pouco mais sobre os processos e técnicas de Predição aplicado em EDM, dando uma visão nova e diferenciada dos comportamentos e tendências, no que se refere à eficiência, dos algoritmos classificadores. A Análise de Seleção de Atributos para a identificação dos fatores mais relevantes também mostrou na prática, em uma situação de dados reais de alunos, algumas tendências que até então só se conseguia perceber de forma intuitiva, ou através de outros trabalhos específicos semelhantes, levando a concluir que para alguns fatores ou atributos já era esperado que sempre influenciarão no desempenho acadêmico dos alunos, ao mesmo tempo que outros, que acreditava-se serem relevantes, não foram considerados como relevantes nesta realidade dos cursos de Programação I, na instituição UFAL.

Uma outra conclusão bastante evidente apresentada através dos resultados dos experimentos foi a baixa performance na predição quando se possui apenas atributos do aluno no momento de início do curso. A menos que exista alguma outra ferramenta, ou dispositivo qualquer, que possibilite a obtenção de mais informações passadas ou presentes, sobre os alunos, fica praticamente inviável se prever aqueles alunos que estão predispostos à reprovação, utilizando-se apenas de informações socioeconômicas e pessoais, como as obtidas através do ENEM, como mostrado na pesquisa. Conclui-se que não é viável a predição de performance acadêmica de alunos, mesmo com seleção e filtragem de atributos relevantes, quando existem apenas dados socioeconômicos disponíveis. Como um potencial trabalho futuro pode-se buscar enfatizar ferramentas que possibilitem a obtenção de mais informações sobre os alunos, possibilitando assim uma aprendizagem e teste mais expressivos por parte das técnicas de predição existentes.

A análise realizada no Experimento 3, com as dimensões das árvores, mostrou uma tendência que, apesar de pouco expressiva numericamente, levantou informações pretendidas para a questão de pesquisa 3: quando as árvores de decisão são expostas ao *prunning*, a eficiência apresentada pelos algoritmos de árvores decresce gradativamente e inversamente proporcional à profundidade das arvores de decisão apresentadas, quando são considerados todos os atributos. O comportamento se inverte quando é realizada a seleção de atributos, as árvores apresentam um sutil crescimento de acurácia diretamente proporcional à profundidade da mesma, ou seja, melhores desempenhos com menos podas. Tal comportamento se dá ao fato de que na existência de ruído, em outras palavras quando tem-se atributos que não tem correlação com a aprovação inseridos nos dados de aprendizagem, os algoritmos de árvore de

decisão criam modelos mais propenso a erros quando livres para a criação de árvores mais profundas. Por sua vez, quando os atributos utilizados na entrada dos algoritmos são os mais relevantes, as árvores mais profundas se tornam mais precisas, com níveis de assertividade maiores.

Como informação valiosa obtida e confirmada com esta pesquisa, ressalta-se a importância do ganho em desempenho apresentado pela inserção de novos dados, mais especificamente os dados das avaliações regulares dos alunos, onde ficou muito claro o crescimento praticamente unânime de todos os algoritmos. T4 é o momento buscado pela terceira questão de pesquisa, considerando atividades quinzenais, ou seja, após a avaliação bimestral 1, porém em T3 já é possível realizar predições com acurácia acima de 85% quando considerando os algoritmos mais eficientes. Considerando ainda que apenas com as notas das atividades de aprendizagem, desconsiderando dados socioeconômicos e seleção de atributos, é possível um nível de eficiência ainda superior na predição. Períodos interatividades menores também tendem à melhoria da predição antecipada.

Um outro aspecto a ser explorado em possíveis trabalhos futuros é a interpretabilidade dos modelos, mais precisamente os de caixa branca, quando analisados sobre a perspectiva do educador ou gestor na educação. Existe uma forte relação entre a abordagem do Experimento 3, que trata da profundidade das árvores de decisão, com a questão da interpretabilidade.

# REFERÊNCIAS


MÁRQUEZ-VERA, C; C. Romero Morales, and S. Ventura Soto, "**Predicting school failure and dropout by using data mining techniques,**" *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 8, no. 1, pp. 7–14, 2013.

MÁRQUEZ-VERA, C; A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "**Early dropout prediction using data mining: A case study with high school students**," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, 2016.

**Chiavenato,** I., **Introducción a la Teoría General de la Administración.** 2013.

WEIHRICH, H. and H. KOONTZ, "**Management: A Global Perspective**," in *Management: A Global Perspective*, 1992.

KUMAR, S. A. and Vijayalakshmi, M.N, "**Efficiency of Decision Trees in Predicting Student** ' S Academic Performance," © *Cs It-Cscp 2011*, 2011.

HOZANO, M.; A. Garcia, N. Antunes, B. Fonseca, and E. Costa, "**Smells Are Sensitive to Developers! on the Efficiency of (Un)Guided Customized Detection**," *IEEE Int. Conf. Progr. Compr.*, pp. 110–120, 2017.

FAYYAD, U; G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge **Discovery in Databases,**" *Al Mag.*, 1996.

PEÑA-AYALA, A; "**Educational data mining: A survey and a data mining-based analysis of recent works**," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1432–1462, 2014.

WITTEN, E. F. Ian H. and M. A. Hall, **Data Mining Practical Machine Learning Tools and Techniques Third Edition,** vol. 1, no. 5. 2011.

FACELI, K; A. C. Lorena, J. Gama, and A. C. P. L. F. de Carvalho, **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. 2011.

LÔBO, M. T. F.; "**Contribuições ao Problema de Seleção de Atributos Mineração de Dados**," pp. 1–3, 2015.

JOHN, G. H. and P. Langley, **Estimating Continuous Distributions in Bayesian Classifiers George**. 1995.

QUINLAN, J. R., ***C4.5: Programs for Machine Learning***. 1993.

BREIMAN, L., "**CART: Classification and Regression Trees," Top Ten Algorithms Data Min.**, 1984.

QUINLAN, J. R., **"Induction of Decision Trees,**" Mach. Learn., 1986.

HSSINA, B; A. MERBOUHA, H. EZZIKOURI, and M. ERRITALI, "**A comparative study of decision tree ID3 and C4.5," Int. J. Adv. Comput. Sci. Appl.**, 2014.

SINGH, S. And p. Gupta, "**comparative study id3 , cart and c4 . 5 decision tree algorithm : a survey,**" int. J. Adv. Inf. Sci. Technol., 2014.

CORTES, C. and V. Vapnik, "**Support-Vector Networks**," Mach. Lear*n.*, 1995.

COSTA, E. B. B; Fonseca M. A. Santana, F. F. de Araújo; and J. Rego **"Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses.**," Computers in Human Behavior, vol. 73. pp. 247–256, 2017.

KHOBRAGADE, L. P., "**Predicting Students' Academic Failure Using Data Mining Techniques**," Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 3, no. 5, pp. 2321–7782, 2015.

ROY, S. and A. Garg, "**Predicting academic performance of student using classification techniques**," in 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017, 2018, vol. 2018–Janua, pp. 568–572.

BURGOS, C; M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "**Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout**," Comput. Electr. Eng., vol. 66, pp. 541–556, 2018.

YASSEIN, N. A; R. G. M Helali, and S. B. Mohomad, "**Predicting Student Academic Performance in KSA using Data Mining Techniques,**" J. Inf. Technol. Softw. Eng., vol. 07, no. 05, 2017.

BEZERRA, C; R. Scholz, P. Adeodato, T. Lucas, and I. Ataide, "**Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes**," *An. do XXVII Simpósio Bras. Informática na Educ. (SBIE 2016)*, vol. 1, no. Cbie, p. 1096, 2016.

ZAFFAR, M; M. A. Hashmani, and K. S. Savita, "**Performance analysis of feature selection algorithm for educational data mining,**" in *2017 IEEE* Conference *on* Big Data and *Analytics,* ICBDA *2017*, 2018.

R. M., N. F., and A. A., "**Predicting and Analysis of Students' Academic Performance using Data Mining Techniques**," *Int. J. Comput. Appl.*, vol. 182, no. 32, pp. 1–6, 2018.

AHMAD, F; N. H. Ismail, and A. A. Aziz, "**The prediction of students' academic performance using classification data mining techniques,**" Appl. Math. Sci., vol. 9, no. 129, pp. 6415–6426, 2015.

BYDŽOVSKÁ, H. and others, "**A Comparative Analysis of Techniques for Predicting Student Performance**," Proc. 9th Int. Conf. Educ. Data Min. *2016*, vol. 2, pp. 306–311, 2016.

MARBOUTI, F; H. A. Diefes-Dux, and K. Madhavan, "**Models for early prediction of at-**

**risk students in a course using standards-based grading**," *Comput. Educ.*, 2016.

CHAWLA, N. V; K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "**SMOTE: Synthetic minority over-sampling technique**," *J. Artif. Intell. Res.*, 2002.

BOUCKAERT *ET AL*, R. R.., "**WEKA Manual for Version 3-6-5,**" Interface, 2010.

FRANK, E., M. A. Hall, and I. H. Witten, "**The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques**," *Morgan Kaufmann, Fourth Ed.*, p. 128, 2016.

GHATAK, A., **Machine Learning with R**. 2017.

# ANEXOS

## ANEXO A – Dicionário de Dados

| Atributo | Tipo de Dado | Descrição | Fonte de Dados |
|---|---|---|---|
| Idade | Numérico | Idade do aluno em anos. | Sistema Acadêmico UFAL (SIEWEB) |
| Sexo | Categórico | Sexo do Aluno.<br>M=Masculino<br>F=Feminino | |
| Etnia | Categórico | Etnia do Aluno. Valores Possíveis: (Branca, Amarela, Parda, Preta, Não Declarado) | |
| Deficiência | Categórico | Tipo de Deficiência do aluno, caso possua. Valores Possíveis: (Física, Mental, Múltipla, Não Declarado) | |
| Tipo de Escola | Categórico | Tipo de Escola que o aluno estudou antes de ingressar na Universidade. Valores Possíveis: (Pública, Privada, Não Informado) | |
| Avaliacao Bimestral 1 | Numérico | Nota do aluno na Avaliação Bimestral 1. De 0 a 10. | |
| Avaliacao Bimestral 2 | Numérico | Nota do aluno na Avaliação Bimestral 2. De 0 a 10. | |
| Reavaliacao | Numérico | Nota do aluno na Reavaliação. De 0 a 10. | |
| Prova Final | Numérico | Nota do aluno na Prova Final. De 0 a 10. | |
| Media Final | Numérico | Nota do aluno na Média Final. De 0 a 10. | |
| Conceito | Categórico | Conceito Final do aluno na disciplina. Representa o PREDITOR.<br>A=Aprovado<br>R=Reprovado | |
| Nota da Atividade 1 | Numérico | Nota do aluno na Atividade 1 do Huxley. De 0 a 10. | Huxley* |
| Nota da Atividade 2 | Numérico | Nota do aluno na Atividade 2 do Huxley. De 0 a 10. | |
| Nota da Atividade 3 | Numérico | Nota do aluno na Atividade 3 do Huxley. De 0 a 10. | |

| | | | |
|---|---|---|---|
| Nota da Atividade n* | Numérico | Nota do aluno na Atividade n do Huxley. De 0 a 10. | |
| Município de Residência | Categórico | Município onde o aluno mora atualmente. | ENEM |
| UF de Residência | Categórico | Sigla da UF da cidade onde o aluno mora atualmente | |
| Nota na prova de Linguagens e Códigos (ENEM) | Numérico | Nota do aluno em Linguagens | |
| Nota na prova de Ciências Humanas (ENEM) | Numérico | Nota do aluno em Ciências Humanas | |
| Nota na prova de Ciências da Natureza (ENEM) | Numérico | Nota do aluno em Ciências da Natureza | |
| Nota na prova de Matemática (ENEM) | Numérico | Nota do aluno em Matemática | |
| Nota na prova de Redação (ENEM) | Numérico | Nota do aluno em Redação | |
| Nota geral na prova do ENEM | Numérico | Nota Geral do Aluno (Média das Notas por área) | |
| Com quantas pessoas mora | Numérico | Quantas pessoas moram em sua casa (incluindo o aluno) | |
| Nível de escolaridade do Pai | Categórico | Nível de Escolaridade do Pai do aluno<br>(A) Da 1ª à 4ª série do Ensino Fundamental (antigo primário)<br>(B) Da 5ª à 8ª série do Ensino Fundamental (antigo ginásio)<br>(C) Ensino Médio (antigo 2º grau)<br>(D) Ensino Superior<br>(E) Especialização<br>(F) Mestrado<br>(G) Doutorado<br>(H) Não estudou.<br>(I) Não sei | |
| Nível de escolaridade da Mãe | Categórico | Nível de Escolaridade da Mãe do aluno<br>(A) Da 1ª à 4ª série do Ensino Fundamental (antigo primário)<br>(B) Da 5ª à 8ª série do Ensino Fundamental (antigo ginásio)<br>(C) Ensino Médio (antigo 2º grau) | |

| | | | |
|---|---|---|---|
| | | (D) Ensino Superior<br>(E) Especialização<br>(F) Mestrado<br>(G) Doutorado<br>(H) Não estudou.<br>(I) Não sei | |
| Renda familiar | Categórico | Renda mensal da família (Some a renda do aluno com a dos seus familiares.)<br>a) Nenhuma renda.<br>b) Até um salário mínimo (até R$ 622,00).<br>c) Mais de um até 1,5 (até R$ 933,00).<br>d) Mais de 1,5 e até 2 (de R$ 933,01 até R$ 1.244,00).<br>e) Mais de 2 e até 2,5 (de R$ 1.244,01 até R$ 1.555,00).<br>f) Mais de 2,5 e até 3 (de R$ 1.555,01 até R$ 1.866,00).<br>g) Mais de 3 e até 4 (de 1.866,01 até R$ 2.488,00).<br>h) Mais de 4 e até 5 (de 2.488,01 até R$ 3.110,00).<br>i) Mais de 5 e até 6 (de 3.110,01 até R$ 3.732,00).<br>j) Mais de 6 e até 7 (de 3732,01 até R$ 4.354,00).<br>k) Mais de 7 e até 8 (de 4354,01 até R$ 4.976,00).<br>l) Mais de 8 e até 9 (de 4.976,01 até R$ 5.598,00).<br>m) Mais de 9 e até 10 (de 5.598,01 até R$ 6.220,00).<br>n) Mais de 10 e até 12 (de 6.220,01 até R$ 7.464,00).<br>o) Mais de 12 e até 15 (de 7.464,01 até R$ 9.330,00).<br>p) Mais de 15 e até 20 (de 9.330,01 até R$ 12.440,00). | |

| | | | |
|---|---|---|---|
| | | q) Acima 20 salários mínimos (mais de 12.440,01). | |
| Situação da residência onde mora | Categórico | A casa onde você mora é:<br>a) Própria e quitada;<br>b) Própria e em pagamento (financiada);<br>c) Alugada;<br>d) Cedida. | |
| Zona onde está localizada a sua residência | Categórico | Localização da casa do aluno<br>(A)Zona rural.<br>(B)Zona urbana<br>(C)Comunidade indígena.<br>(D)Comunidade quilombola. | |
| Fez ENEM para testar conhecimentos | Numérico | O que o levou a participar do ENEM: Testar meus conhecimentos (0 , 1 , 2 , 3 , 4 , 5) | |
| Fez ENEM para obter uma bolsa de estudos | Numérico | O que o levou a participar do ENEM: Conseguir uma bolsa de estudos (ProUni, outras) (0 , 1 , 2 , 3 , 4 , 5) | |
| Quantos anos levou para a conclusão do ensino fundamental | Categórico | Quantos anos o aluno levou para concluir o Ensino Fundamental.<br>(A) Menos de 8 anos.<br>(B) 8 anos.<br>(C) 9 anos.<br>(D) 10 anos.<br>(E) 11 anos.<br>(F) Mais de 11 anos.<br>(G) Não conclui. | |

| | | | |
|---|---|---|---|
| Deixou de estudar durante o ensino fundamental | Categórico | Se o aluno deixou de estudar durante o Ensino Fundamental.<br>(A) Não.<br>(B) Sim, por um ano.<br>(C) Sim, por dois anos.<br>(D) Sim, por três anos.<br>(E) Sim, por quatro anos ou mais. | |
| Tipo de escola em que cursou o ensino fundamental | Categórico | Em que tipo de escola cursou o Ensino Fundamental.<br>a) Somente em escola pública;<br>b) Maior parte em escola pública;<br>c) Somente em escola particular;<br>d) Maior parte em escola particular;<br>e) Somente em escola indígena;<br>f) Maior parte em escola indígena;<br>g) Somente em escola situada em comunidade quilombola;<br>h) Maior parte em escola situada em comunidade quilombola;<br>i) Não frequentei a escola. | |
| Quantos anos levou para a conclusão do ensino médio | Categórico | Quantos anos levou para concluir o Ensino Médio.<br>(A) Menos de 3 anos<br>(B) 3 anos<br>(C) 4 anos<br>(D) 5 anos<br>(E) 6 anos ou mais<br>(F) Não conclui | |
| Deixou de estudar durante o ensino médio | Categórico | Se o aluno deixou de estudar durante o Ensino Médio.<br>(A) Não.<br>(B) Sim, por um ano.<br>(C) Sim, por dois anos.<br>(D) Sim, por três anos.<br>(E) Sim, por quatro anos ou mais. | |

| | | | |
|---|---|---|---|
| Tipo de escola em que cursou o ensino médio | Categórico | Em que tipo de escola você cursou o Ensino Médio.<br>a) Somente em escola pública;<br>b) Maior parte em escola pública;<br>c) Somente em escola particular;<br>d) Maior parte em escola particular;<br>e) Somente em escola indígena;<br>f) Maior parte em escola indígena;<br>g) Somente em escola situada em comunidade quilombola;<br>h) Maior parte em escola situada em comunidade quilombola;<br>i) Não frequentei a escola. | |
| Cursou o programa de Educação de Jovens e Adultos | Categórico | Se cursa ou já cursou a Educação de Jovens e Adultos (EJA)<br>(A) Sim<br>(B) Não<br>Observação: 0 - Não; 1 - Sim | |
| Cursou o ensino regular | Categórico | Se já frequentou o ensino regular<br>(A) Sim.<br>(B) Não.<br>Observação: 0 - Não; 1 - Sim | |
| Pretende aderir ao FIES | Booleano | Caso o alno ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades: FIES (Programa de Financiamento Estudantil) Sim(A) Não(B)<br>Observação: 0 - Não; 1 - Sim | |
| Pretende aderir ao PROUNI | Booleano | Caso o alno ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades: Pró-Uni (Programa Universidade para Todos) Sim(A) Não(B)<br>Observação: 0 - Não; 1 - Sim | |
| Pretende aderir bolsa de estudos da instituição | Booleano | Caso o alno ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades: Bolsa de estudos da | |

| | | | |
|---|---|---|---|
| | | própria Instituição de Ensino Superior Sim(A) Não(B)<br>Observação: 0 - Não; 1 - Sim | |
| Pretende aderir bolsa de estudos da empresa onde trabalha | Booleano | Caso o alno ingresse no Ensino Superior privado pretende recorrer aos auxílios abaixo para custeio das mensalidades:Bolsa de estudos da empresa onde trabalho Sim(A) Não(B)<br>Observação: 0 - Não; 1 - Sim | |
| Possui TV | Categórico | Se tem em casa:<br>TV em cores<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui DVD/vídeo cassete em casa | Categórico | Se tem em casa: Videocassete e/ou DVD<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui rádio em casa | Categórico | Se tem em casa: Rádio<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui computador | Categórico | Se tem em casa: Microcomputador<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui automóvel | Categórico | Se tem em casa: Automóvel<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui máquina de lavar | Categórico | Se tem em casa: Máquina de lavar roupa<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui geladeira | Categórico | Se tem em casa: Geladeira<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui freezer | Categórico | Se tem em casa: Freezer (aparelho independente ou parte da geladeira duplex)<br>(A)- 1 (B)- 2 (C)- 3 (D)- Não tenho | |

| | | | |
|---|---|---|---|
| Possui telefone fixo | Categórico | Se tem em casa: Telefone fixo<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui telefone celular | Categórico | Se tem em casa: Telefone celular<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui acesso à internet em casa | Categórico | Se tem em casa: Acesso à Internet<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui tv por assinatura | Categórico | Se tem em casa: TV por assinatura<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui aspirador de pó | Categórico | Se tem em casa: Aspirador de pó<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Possui empregada mensalista | Categórico | Se tem em casa: Empregada mensalista<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |
| Quantos Banheiros possui em casa | Categórico | Se tem em casa: Banheiro<br>(A)- 1 (B)- 2 (C)- 3 ou mais (D)- Não tenho | |