Instrumento
revista de estudo e pesquisa em educação

ISSN 1984-5499

Mapeamento da performance de alunos de curso profissionalizante de informática através de predição

Performance Mapping of Students in a Professional Computer Course through Prediction

Mapeo del desempeño de los estudiantes en un curso de computación profesional a través de la predicción

Ariomar da Luz Nogueira Filho[1]
Professor titular da Secretaria de Estado de Educação do Distrito Federal, Brasília, Distrito Federal, Brasil

Recebido em: *17/02/2021*
Aceito em: *24/07/2024*

Resumo
O desempenho dos estudantes em cursos de nível técnico em informática é essencial para a formação profissional. No entanto, o baixo rendimento de estudantes nas disciplinas específicas torna-se preocupante para o mercado de trabalho, sendo necessário formular estratégias para melhorar o desempenho desses alunos. Uma maneira de mapear e identificar tais alunos com provável baixo desempenho e/ou evasão é analisar seu histórico escolar e as condições e estrutura familiar do aluno. Neste trabalho, é proposto um método de identificação de alunos propensos ao baixo rendimento no curso. Os resultados indicam que é possível identificar por meio dos dados de seu histórico escolar e da estrutura familiar, com taxa de acurácia e de precisão em 100%, utilizando o algoritmo de predição *Random Forest*.
Palavras-chave: Mineração de Dados Educacional. Aprendizado de máquina. Predição de Performance de Estudantes. Evasão Escolar.

Abstract
The performance of students in technical-level computer courses is essential for professional training. However, the low performance of students in specific subjects is a concern for the job market, necessitating strategies to improve these students' outcomes. One approach to mapping and identifying students likely to underperform and/or drop out is to analyze their academic history alongside their family conditions and structure. In this study, a method is proposed for identifying students prone to low performance in these courses. The results indicate that it is possible to identify such students using data from their academic history and family structure, achieving 100% accuracy and precision with the Random Forest prediction algorithm.
Keywords: Educational Data Mining. Machine Learning. Student Performance Prediction. School Dropout.

Resumen
El desempeño de los estudiantes en los cursos de nivel técnico en informática es fundamental para la formación profesional. Sin embargo, el bajo desempeño de los estudiantes en disciplinas específicas se convierte en una

[1] ariomar@gmail.com .

preocupación para el mercado laboral, por lo que es necesario formular estrategias para mejorar el desempeño de estos estudiantes. Una forma de mapear e identificar a los estudiantes con probable bajo rendimiento y / o abandono escolar es analizar su historial escolar y las condiciones y estructura de la familia del estudiante. En este trabajo se propone un método de identificación de alumnos con tendencia a un bajo rendimiento en el curso. Los resultados indican que es posible identificar a través de datos de su historial escolar y estructura familiar, con una tasa de precisión del 100%, utilizando el algoritmo de predicción Random Forest.
Palabras clave: Minería de datos educativos. Aprendizaje automático. Predicción del rendimiento de los estudiantes. Abandono escolar.

Introdução

Nos cursos profissionalizantes a nível médio na área de informática, é habitual que estudantes dos cursos apresentem obstáculos e complicações no aprendizado de disciplinas específicas, como linguagem de programação, originando aumento das taxas de reprovação, abandono do curso ou aprovação com baixo rendimento, segundo Queiroga; Cechinel; Araújo (2017, p. 1554). De acordo com Cardoso, Antonello (2015, p. 1255), esse fato se agrava no momento que consideramos que essas disciplinas são capitais para uma adequada formação profissional do indivíduo. Sendo assim, é conveniente e necessário fornecer meios de identificar esses indivíduos, para mapear os fatores responsáveis pelas dificuldades dos alunos, com o intuito de engendrar estratégias para superá-las.

Existem diversos estudos na literatura que corroboram este estudo, como por exemplo: Miguéis et al. (2018, p. 38), que relata que uma abordagem eficaz para segmentar os alunos com base tanto no desempenho previsto no final do curso acadêmico quanto no desempenho observado no final do primeiro ano envolve inicialmente definir métricas claras de desempenho acadêmico, como médias de notas e taxas de retenção. Os dados históricos de desempenho dos alunos são então coletados e analisados utilizando técnicas estatísticas para identificar padrões de desempenho no primeiro ano. Com base nessa análise, modelos preditivos são desenvolvidos para estimar o desempenho futuro dos alunos. Os alunos são então segmentados em grupos distintos, como "Alto Desempenho Antecipado", "Desempenho Melhorável" e "Necessidade de Intervenção", cada um recebendo estratégias educacionais personalizadas para promover seu progresso acadêmico.

Souza, Batista, Barbosa (2016, p. 45), afirma que as dificuldades na aprendizagem e na aplicação dos conceitos de programação estão ligadas a problemas motivacionais, assim como que esses problemas motivacionais estão relacionados com as dificuldades em aprender e aplicar

conceitos de programação. Castro, Siqueira (2019, p. 230) consideram que um método se refere ao percurso necessário para atingir um objetivo específico. No dia a dia, as pessoas têm metas que não se concretizam automaticamente; é preciso seguir uma série de passos para atingi-las. No estudo de Qian, Lehman (2016, p. 79) mostram que os resultados indicam que a proficiência em inglês dos estudantes teve a relação mais forte com o desempenho na aprendizagem de programação e foi o fator mais significativo para explicar as diferenças na programação.

Os principais fatores que contribuem para um baixo entendimento e assimilação do aluno nesses cursos de informática são:

Habilidade de resolução de problemas, como afirma Hinterholz, Cruz (2015, p. 144) que visa estimular alunos do Ensino Médio a pensar como um computador promove o desenvolvimento do pensamento computacional, pois receber informações, analisá-las, visualizar possíveis soluções e aplicar uma resolução é a abordagem que prepara os alunos para resolver problemas de forma eficaz.

O Domínio de interpretação e compreensão de texto Brandão, Simão, Souza (2014, p. 92), pois compreender uma ação como interpretação, onde o sujeito "atualiza" um dado ou informação, este reorganiza-se e o interpreta com base em suas experiências prévias, o que resulta na criação de conhecimento.

Os Fundamentos matemáticos, pois segundo Souza, Araújo Andrade, De Paulo Martins (2020 p.163), o conceito de letramento matemático não se limita apenas às práticas sociais de uso da matemática, mas também envolve técnicas. Por isso, o conceito de letramento não deve ser dissociado da alfabetização. No ensino da Matemática, a compreensão de ideias como contar, medir, estimar, observar padrões e regularidades, entre outras, permite que os estudantes sejam intrometidos a uma interpretação matemática do mundo. Além disso, a matemática contribui para os processos de alfabetização, sendo essas práticas chamadas de letramento matemático.

Por fim, o pré-requisitos em conhecimento de lógica de programação, computadores e softwares (Games, Internet, etc), Ribas, Dal Bianco, Lahm (2016, p. 8) afirma-se que o conhecimento em programação auxilia na compreensão de conceitos matemáticos, mas o desenvolvimento de disciplinas como esta demonstra que o contrário também é verdadeiro. Quanto mais conhecimento lógico-matemático o estudante possuir, mais facilidade terá para compreender os conceitos de computação.

De acordo com Fredenhagem (2014, p.69) e Diniz, Santos (2020, p. 44833), os fatores que

estão diretamente arrolados com os índices de abandono, reprovação de disciplinas e aprovação com resultado mínimo. As causas mais comuns são ligadas ao trabalho, como a incompatibilidade de horários, dificuldades na conciliação e a necessidade de sustentar a família.

Assim sendo, é conveniente utilizar técnicas de Mineração de Dados Educacionais *(Educational Data Mining, EDM)* como afirma Rigo, Cazella, Cambruzzi (2014, p.173), com a finalidade de aperfeiçoar a metodologia de ensino-aprendizagem nas disciplinas específicas de cursos profissionalizantes de informática. EDM consiste na aplicação de técnicas de Mineração de Dados *(Data Mining)* no campo educacional, que segundo Portal, Schlemmer (2015, p. 6), é composta por um aglomerado de procedimentos com o objetivo de extrair informações não explícitas e desconhecidas previamente de um conjunto de dados.

A compreensão e entendimento desses dados possui grande valia em diversas circunstâncias, como por exemplo: a tomada de decisões em escolas, universidades, empresas e inclusive no governo. A aplicação de técnicas de EDM vem ganhando popularidade entre instituições de ensino, que buscam melhorar a qualidade da sua formação e desempenho de seus estudantes. Aplicando técnicas de Data Mining, podem-se analisar dados educacionais e buscar padrões e relações desses dados com o intuito de detectar atributos peculiares destas técnicas, originando assim conhecimento, e consequentemente daí alcançar previsões de performance de um estudante como afirma Maschio et al. (2018, p. 1938).

Em virtude da relevância das disciplinas de programação na formação profissional de estudantes de nível médio dos cursos técnicos de Informática e da popularização da área de EDM, propõe-se neste estudo uma metodologia para conjecturar o desempenho dos estudantes no primeiro semestre dos cursos profissionalizantes, usando, como base, os dados referentes ao seu histórico escolar do ensino fundamental de séries finais e também a composição da estrutura familiar e socioeconômica do estudante.

A viabilidade do método é estudada a partir de dados de estudantes do primeiro semestre do curso técnico de informática, do CEMI-GAMA (Centro de Ensino Médio Integrado a Educação Profissional do Gama). Essa escola está localizada na Região Administrativa II, ou seja, na Cidade satélite Gama em Brasília - Distrito Federal.

Este artigo está organizado com os seguintes procedimentos: Inicialmente é apresentado a literatura correlacionada, em seguida descreve-se a metodologia empregada no desenvolvimento do estudo, logo após descreve-se os resultados propostos e discutidos e

finaliza-se com a conclusão e sugestão para trabalhos futuros.

Trabalhos relacionados

O EDM é uma área de pesquisa bastante promissora, pois apresenta um grande crescimento no Brasil e no mundo nos últimos anos segundo os estudos de Maschio et al. (2018, p.1936) e Tsiakmaki et al. (2020, p.2). De acordo com Charitopoulos, Rangoussi, Koulouriotis (2020, p. 372), o EDM pode ser separado em categorias, mas para este estudo, o foco é a predição de desempenho de alunos, pois é considerada como a mais popular aplicação de Mineração de Dados Educacionais.

Conforme o estudo de Pinto, Freitas Júnior, Costa (2020, p. 1176), a predição de desempenho de alunos pode ser aplicada por uma adequada multiplicidade de técnicas de Data Mining, como Redes Neurais Artificiais, Redes Baysianas, Análises de Regressão e também de Classificação, possuindo assim, uma vasta aplicabilidade.

Os autores Ribeiro, Maciel (2020, p. 21) e Pereira et al. (2020, p. 1681) tentam entender os fatores responsáveis pelas dificuldades, obstáculos e complicações enfrentadas pelos alunos nas disciplinas de informática de um curso de técnico ou universitário. Ponderam os obstáculos enfrentados pelos estudantes no decorrer da aprendizagem dessas disciplinas, tangenciando com distintas metodologias de ensino, com o intuito de desenvolver um melhor ambiente de ensino e aprendizagem entre professores e alunos.

O projeto, criado por Garcia, Correia e Shimabukuro (2008, p.247), foi intitulado “Ensino de Lógica de Programação e Estrutura de Dados para Alunos do Ensino Médio”. Esse projeto recrutava alunos do Ensino Médio de escolas brasileiras que tinham interesse em programação. O objetivo da iniciativa era introduzir precocemente os estudantes aos conceitos de raciocínio lógico e à tecnologia. Dessa forma, capacitar os alunos a solucionar problemas lógicos com o auxílio de algoritmos e estruturas de dados. Os resultados do projeto foram positivos: os participantes que ingressaram em cursos de informática nas universidades apresentaram bom desempenho e baixa taxa de evasão nas disciplinas relacionadas. Além disso, os alunos desenvolviam habilidades importantes para o mercado de trabalho, como pensamento crítico, resolução de problemas e trabalho em equipe. A experiência também despertou o interesse dos estudantes por outras áreas da tecnologia, como inteligência artificial, ciência de dados e desenvolvimento de software.

Outro benefício observado foi o aumento da confiança dos alunos em suas habilidades

acadêmicas e a motivação para perseguir carreiras na área de tecnologia. O projeto também serviu como uma plataforma para identificar talentos promissores que, de outra forma, poderiam não ter tido acesso a esse tipo de educação especializada. Ademais, a interação com a programação desde cedo ajudou a preparar os estudantes para as demandas do século XXI, onde a tecnologia desempenha um papel central em quase todas as profissões.

No estudo de Kovacic (2010, p. 650), foram utilizadas técnicas de *Data Mining*, como Árvores de Classificação e Regressão, onde foram analisados os dados sociais, os dados demográficos e os dados do ambiente de estudo dos alunos do Sistemas de Informações da Politécnico Aberto da Nova Zelândia, entre os anos 2006 e 2009. O intuito deste estudo foi detectar quais são os fatores fundamentais, para diferenciar um aluno de sucesso (aprovado) ou de insucesso (reprovado). Chegando à conclusão de que a etnia, o curso e o semestre que foram oferecidos são os fundamentais fatores para a prognóstico de alunos que alcançarão sucesso na disciplina.

Diversos trabalhos de EDM são aplicados em cursos de universitários, como maneira de prognóstico de desempenho de alunos. Barros et al. (2019, p.1494) avaliam a probabilidade de predição da atuação de alunos do primeiro período nas disciplinas de cálculo por meio da sua atuação no vestibular. Esses autores argumentam que os dados têm um poderoso valor acadêmico, pois o fracasso nas primeiras disciplinas do curso é proporcional ao aumento nas taxas de evasão e/ou abandono nos cursos.

O estudo de Saa (2016, p. 212) recomenda o uso do classificador *Naive Bayes* para ponderar as notas dos alunos universitários com base em dados sociais, econômicos e de desempenho no ensino médio. O *Naive Bayes*, baseado no teorema de *Bayes*, é eficiente e capaz de lidar com grandes conjuntos de dados, oferecendo resultados precisos. Ele trata tanto dados categóricos quanto contínuos, identificando padrões e tendências. Assim, essa abordagem inovadora melhora a avaliação acadêmica e fornece insights para aprimorar a educação universitária.

Metodologia

Obtenção dos dados e pré-processamento

A base de dados utilizada neste trabalho foi obtida por meio da secretaria do CEMI-Gama, contendo registros anônimos dos alunos do curso profissionalizante de Técnico em Informática, os

dados originais foram gerados entre os anos 2018 e 2020. Os registros são compostos de informações de desempenho (nota final) e quantidade de faltas dos alunos nas séries 6º ano, 7º ano, 8º ano e 9º ano do ensino fundamental nas disciplinas matemática, português, ciências naturais e Inglês. Além dos critérios sócio econômico da família e o tempo de estudo livre e orientado fora da escola. Todos os dados foram inseridos em uma tabela em formato CSV (arquivo separado por vírgula) e em seguida a tabela foi processada pelo software *Orange Data Mining.*

Este software é uma ferramenta robusta projetada para análise e mineração de dados, acessível a usuários de todos os níveis de habilidade. Com uma interface visual intuitiva, permite a construção de fluxos de trabalho de análise sem necessidade de programação. Seu principal objetivo é facilitar a exploração de dados complexos, a criação de modelos preditivos e a visualização clara dos resultados. Oferecendo uma vasta gama de técnicas, como classificação, regressão, *clustering* e análise de associação, utilizando algoritmos avançados como árvores de decisão e redes neurais. Isso o torna adaptável a diversas necessidades de análise de dados.

Os benefícios do Orange são significativos: permite explorar grandes conjuntos de dados de maneira eficiente, descobrir padrões ocultos e tomar decisões embasadas em evidências. Sua capacidade de visualização interativa facilita a interpretação dos modelos desenvolvidos, tornando a análise de dados acessível mesmo para não especialistas. Além de ser uma plataforma de código aberto, o Orange é apoiado por uma comunidade ativa que contribui com novas funcionalidades e extensões, garantindo que continue a evoluir e atender às necessidades dos usuários em ciência de dados. [2]

A Tabela 1 contém informações de quantitativo de alunos Aprovados (AP), Reprovados (RP), Evadido e/ou Transferido (ABA) e Aprovado com Dependência (APD). Os dados utilizados para realizar a predição foram as notas das disciplinas cursadas e a quantidade de faltas em cada ano pelos alunos durante o ensino fundamental, no 6º ano, 7º ano, 8º ano e 9º ano, nas disciplinas: Matemática; Português; Ciências Naturais e Língua Estrangeira Inglês. Outro critério utilizado foi a realidade sócia econômica da família de cada aluno, coletada por meio de registro da orientação educacional da escola. Felizmente, todos os dados foram completos, não existindo ruído ou necessidade de um pré-processamento das informações obtidas, pois na escola

[2] Para realizar o download, acesse https://orangedatamining.com/ .

Tabela 1

Quantitativos dos Alunos Utilizados para o Treinamento do Algoritmo de Predição

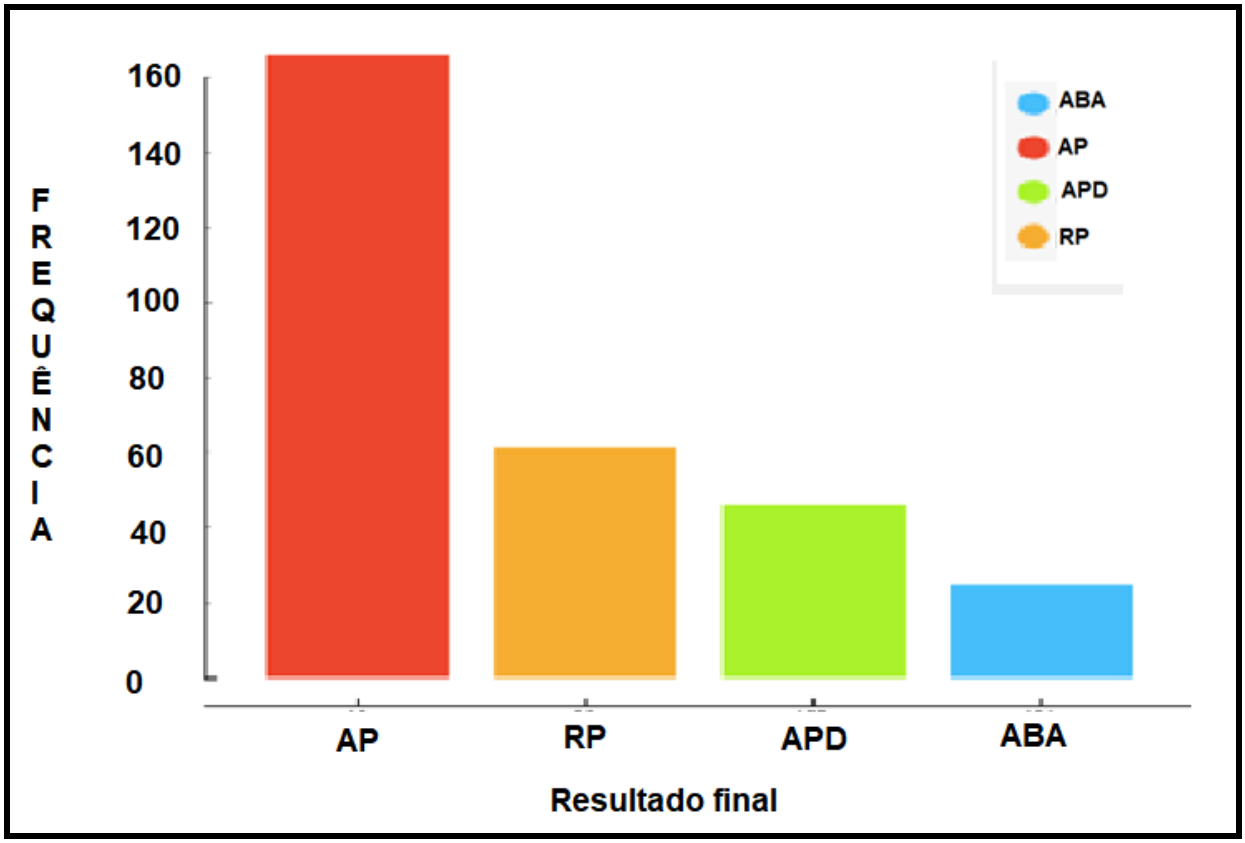
Total	Aluno (AP)	Reprovado (RP)	Evadido e/ou Transferido (ABA)	Aprovado com Dependência (APD)
298	166	61	25	46

Fonte: Elaborado pelo autor.

Na Figura 1, observa-se a distribuição do quantitativo de alunos em relação ao resultado deles, utilizado no banco de dados de treinamento do algoritmo.

Figura 1

Distribuição do quantitativo de alunos em relação ao resultado final



Fonte: Elaborado pelo autor.

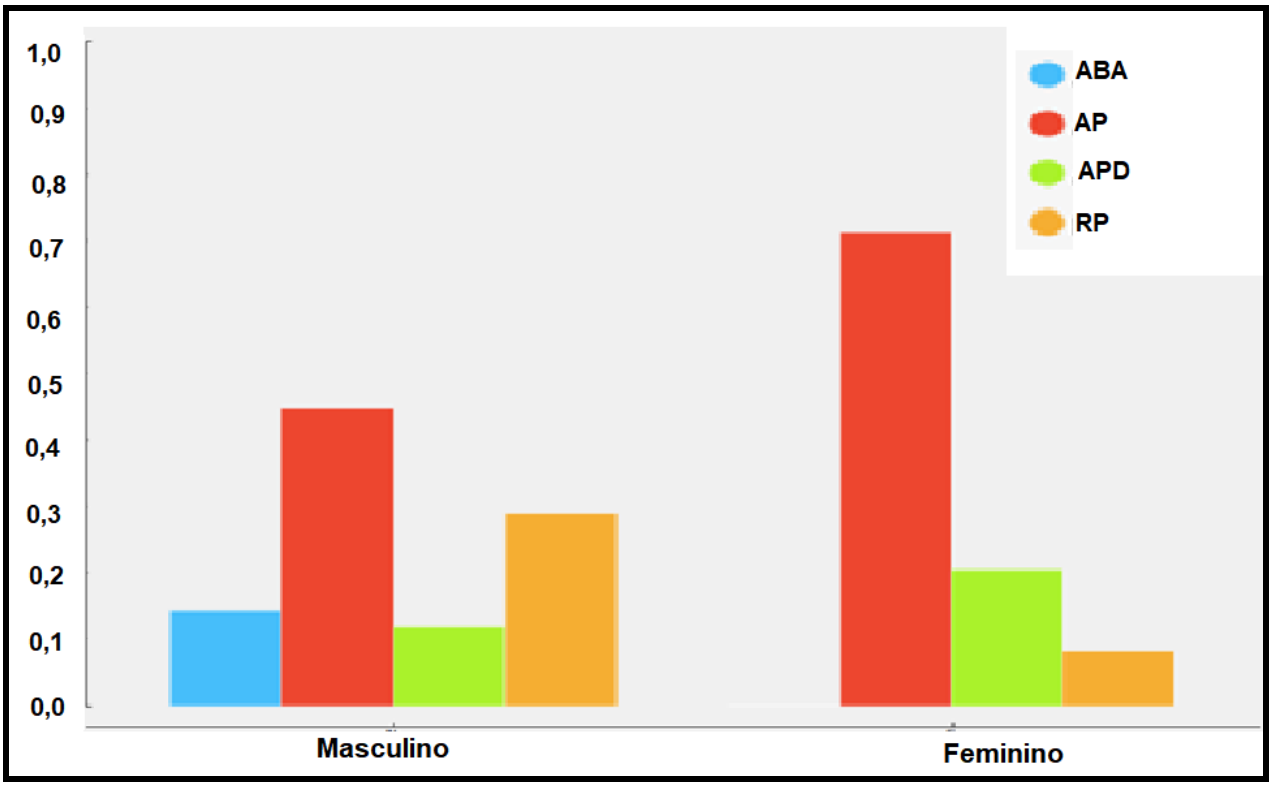
Os critérios sócio econômicos utilizados para este estudo foram os seguintes: Sexo; Idade; Zona de habitação (Rural ou Urbano); Quantitativo de irmãos; Estado civil dos responsáveis (Solteiro, Casado ou Divorciado); Grau de instrução do responsável paterno e materno (Fundamental, Médio ou Superior); Situação empregatícia dos responsáveis paterno e materno (Empregado ou Desempregado); Razão da escolha da escola; Tutor principal do aluno; Tempo de deslocamento do aluno à escola em

minutos; Tempo de estudo fora da escola por semana em horas; Tempo livre diário fora do ambiente escolar em horas; Suporte educacional extracurricular, ofertado pela família (Sim ou Não); Suporte educacional familiar (Sim ou Não); Interesse no ensino superior (Sim ou Não); Acesso à internet para estudos (Sim ou Não); Qual tipo de aparelho de informática utiliza para acessar a internet (Celular, Notebook, Desktop, Tablet);

Por meio das respostas obtidas pelo questionário, obtido pela orientação educacional da escola, foi possível identificar que o estudante masculino tem maior tendência de abandonar o curso, como pode ser verificado na Figura 2.

Figura 2

Distribuição do Resultado Final dos Alunos por Sexo

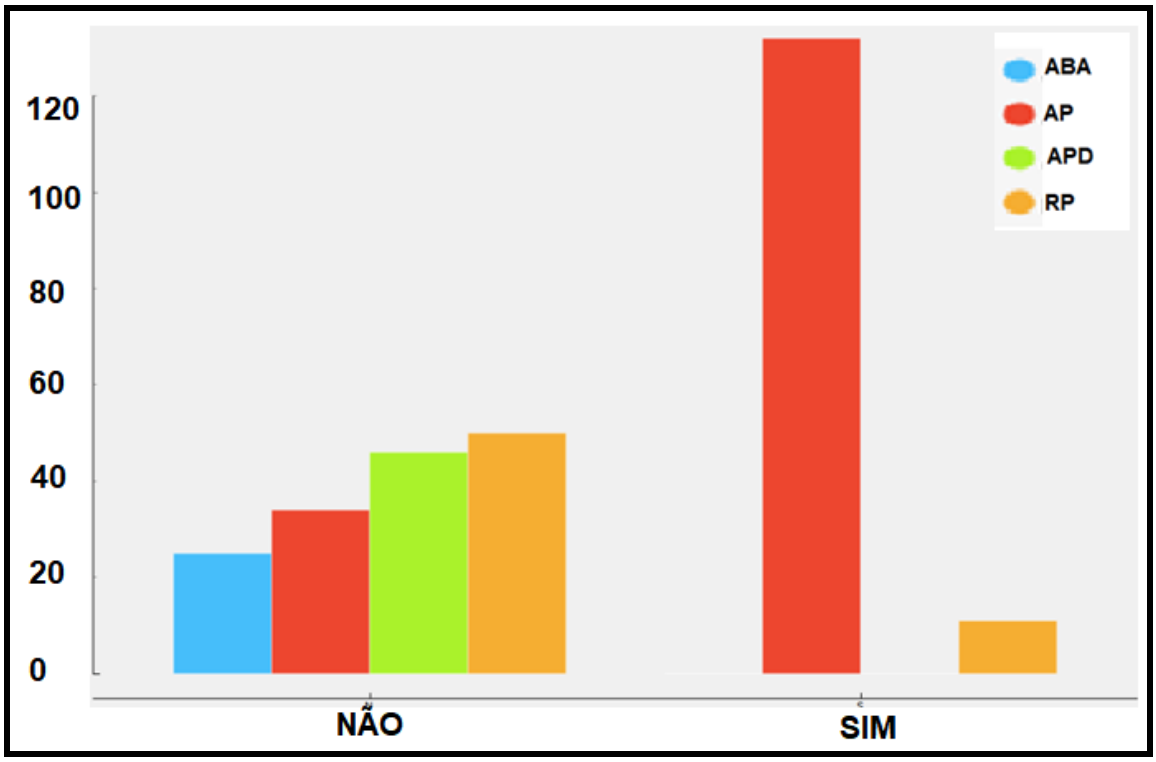


Fonte: Elaborado pelo autor.

Pode-se verificar pela Figura 3 e Figura 4 que o suporte educacional fora da escola é importante no processo de aprovação do aluno, o que remete que a presença da família no ambiente escolar, sendo extremamente importante para o mérito do aluno. Observa-se que com a existência de auxílio educacional, seja da própria família ou com outros profissionais da educação, fora da escola tem resultado positivo.

Figura 3

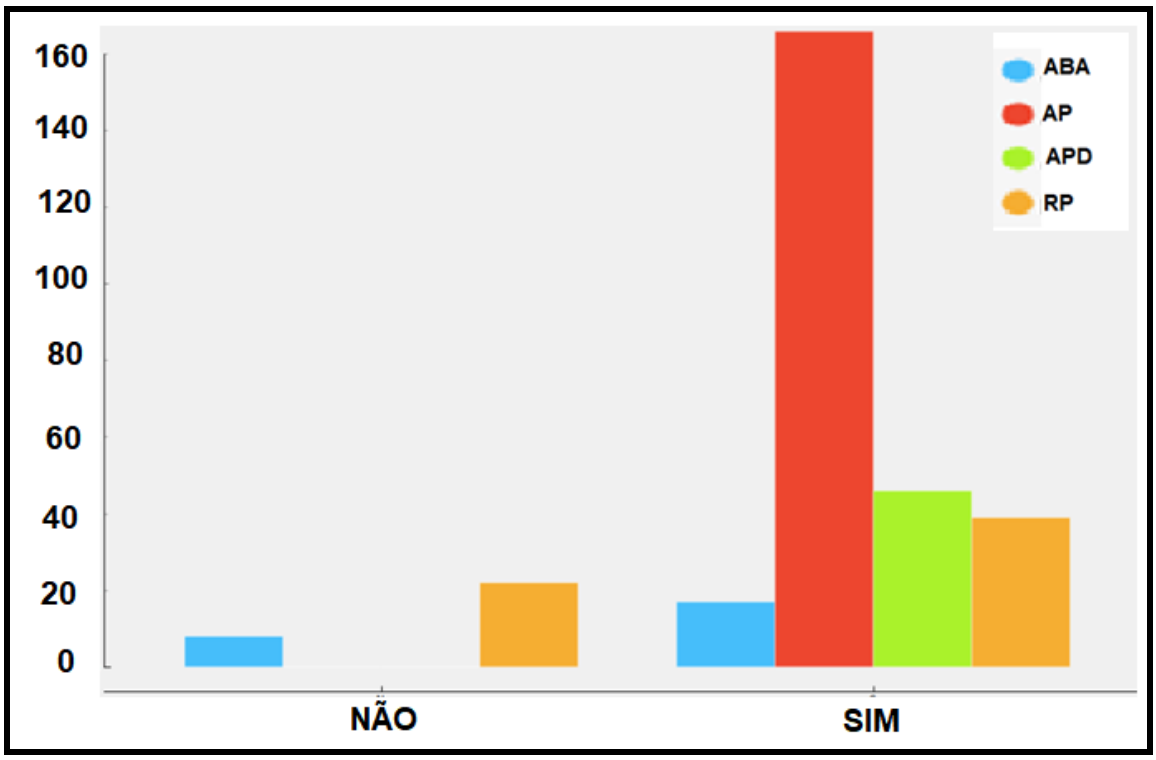
Distribuição do Suporte Educacional Extra Curricular por Resultado Final



Fonte: Elaborado pelo autor.

Figura 4

Distribuição do Suporte Educacional Familiar por Resultado Final



Fonte: Elaborado pelo autor.

Modelo proposto

Neste trabalho, é proposto um modelo para identificar, precocemente, alunos propensos ao sucesso ou ao insucesso no 1º semestre de 2020 no curso profissionalizante de Técnico em Informática. Essa técnica não necessariamente é exclusiva para curso profissionalizante, ou seja, pode ser utilizada na predição de outros tipos cursos, do ensino regular à graduação, sendo necessários alguns ajustes diante da realidade.

Como informado anteriormente, foi realizada a predição de desempenho dos estudantes para o primeiro semestre do curso profissionalizante em Técnico de Informática, com o intuito de conhecer e mapear os novos alunos, não para rotulá-lo como derrotado ou vitorioso.

A identificação precoce dos estudantes propensos ao insucesso tem a função de auxiliar os professores e coordenadores do curso, desenvolvendo ações que visam reduzir a quantidade de reprovações e, possivelmente, evasão por fracasso no primeiro semestre, ajudando a aprimorar os sistemas de ensino das escolas.

A situação do aluno nas disciplinas é demarcada a partir da sua média final. Alunos que obtiveram nota menor que 5,0 (nota mínima para aprovação em uma disciplina na instituição) são considerados como sendo da classe “Reprovados”, os alunos que alcançaram nota igual ou superior a 5,0 são considerados “Aprovados” e aquele aluno que atingiu média inferior a 5,0 em até 2 (duas) disciplinas do total cursada no semestre, pode ser Aprovado com Dependência para o próximo semestre.

Avaliação do modelo

Para investigar minuciosamente a viabilidade do método proposto na previsão de desempenho dos alunos nos cursos de Técnico em Informática, foram conduzidos diversos testes utilizando o software ORANGE Data Mining. Este software é conhecido por sua extensa coleção de algoritmos de aprendizado de máquina implementados, os quais permitem uma análise detalhada e comparativa dos resultados obtidos por meio de diferentes métodos. Tal abordagem facilita a identificação e seleção de algoritmos mais adequados para resolver o problema específico, conforme argumentado por Demšar *et al*. (2013, p. 2350).

Selecionaram-se quatro algoritmos de aprendizagem de máquina distintos: *Random Forest* (Floresta aleatória), *Artificial Neural Network* (Redes Neurais Artificiais), *AdaBoost* e *Decision Tree* (árvore de decisão).

Esses algoritmos foram processados com as seguintes características: O *Random Forest*

utilizou um total de 20 árvores de classificação. A montagem do algoritmo *Artificial Neural Network* foi com 400 neurônios em camadas ocultas, com ativação do tipo ReLu, solucionador do tipo SGD, com número máximo de iterações de 200 e com treinamento replicável. O algoritmo *AdaBoost* possui os seguintes parâmetros: estimador de base em formato de árvore, 50 estimadores, taxa de Aprendizagem igual ao multiplicador de 1, o método de impulso utilizou o algoritmo de classificação SAMME.R e função de perda de regressão em formato linear. O algoritmo *Decision Tree* possui os seguintes parâmetros: árvore binária induzida com no mínimo de duas instâncias nas folhas, a profundidade máxima da árvore foi de 100 e a classificação é finalizada apenas quando alcançar 95% de eficácia.

Para cada um deles, buscou-se, por tentativa e teste, encontrar o conjunto de parâmetros que produzisse os melhores resultados. Em ambos os casos, a divisão entre conjunto de treinamento e testes foram feitas a partir de Validação Cruzada (*Cross Validation*) com 20 *folds*. Nessa abordagem o conjunto de dados original é dividido em *k* subgrupos disjuntos (folds), onde para *k*-1 subgrupos é feito o treinamento do modelo, que é testado com o subconjunto restante. O processo é repetido usando-se subconjuntos diferentes para treinamento e teste, até completar *k* repetições. Ao término dos testes, calcula-se a média das taxas de acerto e erro de todos os testes com o intuito de se obter resultados mais consistentes nos testes.

Como forma de avaliação dos métodos, utilizou-se a acurácia (número de instâncias classificadas corretamente) e as taxas de verdadeiros positivos (VP) e verdadeiros negativos (VN), já que a primeira métrica, quando usada isoladamente, pode não nos dar informações consistentes sobre o desempenho dos classificadores. Tanto os VP quanto os VN são classificações corretas e indicam as porcentagens de classificações corretas de estudantes que reprovaram nas disciplinas e a classificações corretas de estudantes que foram aprovados nas disciplinas, respectivamente.

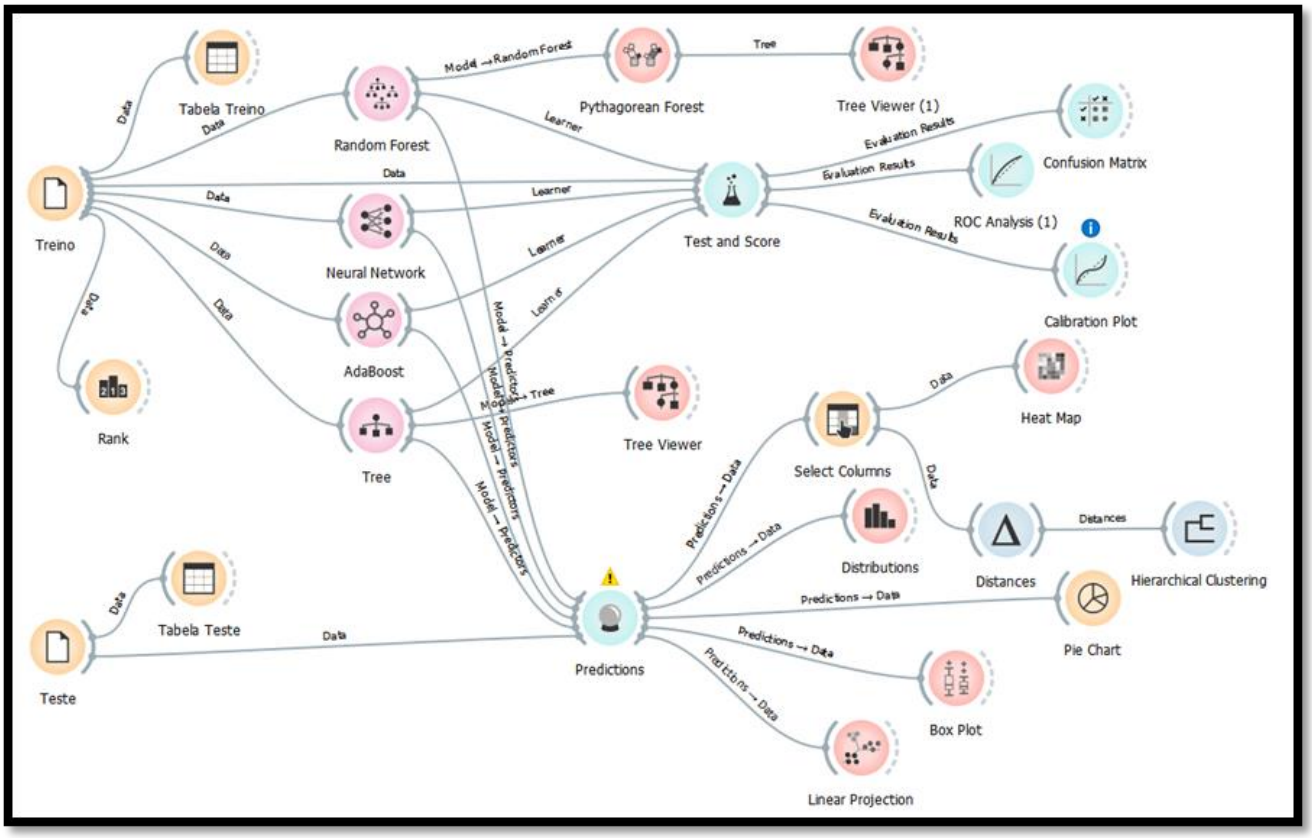
As taxas de falsos negativos (FN) e falsos positivos (FP) são complementares as medidas de VP e FP, respectivamente, e serão omitidas dos resultados por questão de simplicidade.

Apresentação do Modelo

O modelo completo foi confeccionado pelo software de mineração de dados ORANGE Data Mining ele é um programa baseado em Python por meio de linguagem visual. Podendo ser verificado na Figura 5.

Figura 5

Layout completo do Programa



Fonte: Elaborado pelo autor.

Para a execução dos testes, foi manipulado um banco de dados com 20 alunos reais do CEMI-Gama, utilizando os mesmos critérios do banco de dados contemplados no treinamento.

Resultados e Discussão

Resultado do Treinamento

Nesta seção são apresentados os resultados do método de classificação dos alunos no software ORANGE Data Mining. O primeiro teste pode ser observado na Figura 6. Refere-se ao teste de pontuação do banco de dados, treinado pelos algoritmos de precisão. Verifica-se, na Figura 6, que é possível diferenciar a evolução dos resultados pelos seguintes termos:

AUC (A área sob ROC) é a área sob a curva de operação do receptor; CA (Classificação de Acurácia) é a proporção de exemplos classificados corretamente; F1 é uma média harmônica ponderada de precisão e recordação; *Precision* (Precisão) é a proporção de verdadeiros positivos (VP) entre os casos

classificados como positivos; *Recall* (Recordação) é a proporção de verdadeiros positivos (VP) entre todas as ocorrências positivas nos dados.

Figura 6

Teste e Pontuação do Banco de Dados de Treinamento

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.997	0.990	0.990	0.990	0.990
Random Forest	1.000	1.000	1.000	1.000	1.000
Neural Network	0.996	0.997	0.997	0.997	0.997
AdaBoost	0.988	0.980	0.980	0.981	0.980

Fonte: Elaborado pelo autor.

Observa-se, na Figura 6, o *ranking* dos melhores algoritmos testados, onde do melhor para o menos equilibrado ficou na seguinte sequência: *Random Forest*; *Neural Network*; *Tree*; *AdaBoost*.

A partir da Figura 7, Figura 8, Figura 9 e Figura 10, é possível verificar os erros e acertos de cada algoritmo por meio da matriz de confusão e obter a melhor coerência nos testes do banco de dados de treinamento.

Figura 7

Matriz de confusão para o algoritmo *Randon Forest*

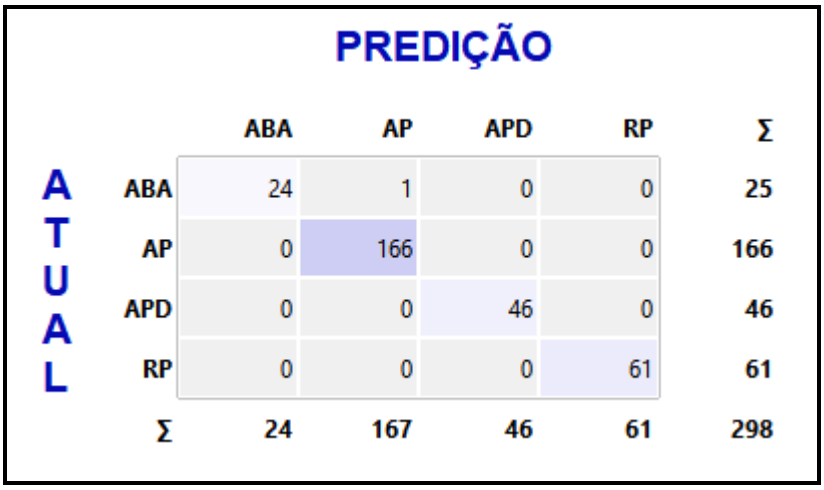
PREDIÇÃO

ATUAL	ABA	AP	APD	RP	Σ
ABA	25	0	0	0	25
AP	0	166	0	0	166
APD	0	0	46	0	46
RP	0	0	0	61	61
Σ	25	166	46	61	298

Fonte: Elaborado pelo autor.

Figura 8

Matriz de confusão para o algoritmo *Neural Network*

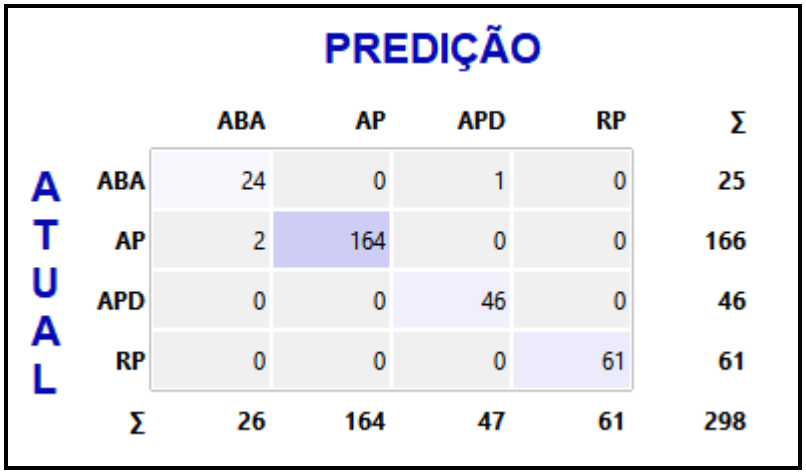


ATUAL \ PREDIÇÃO	ABA	AP	APD	RP	Σ
ABA	24	1	0	0	25
AP	0	166	0	0	166
APD	0	0	46	0	46
RP	0	0	0	61	61
Σ	24	167	46	61	298

Fonte: Elaborado pelo autor.

Figura 9

Matriz de confusão para o algoritmo *Tree*

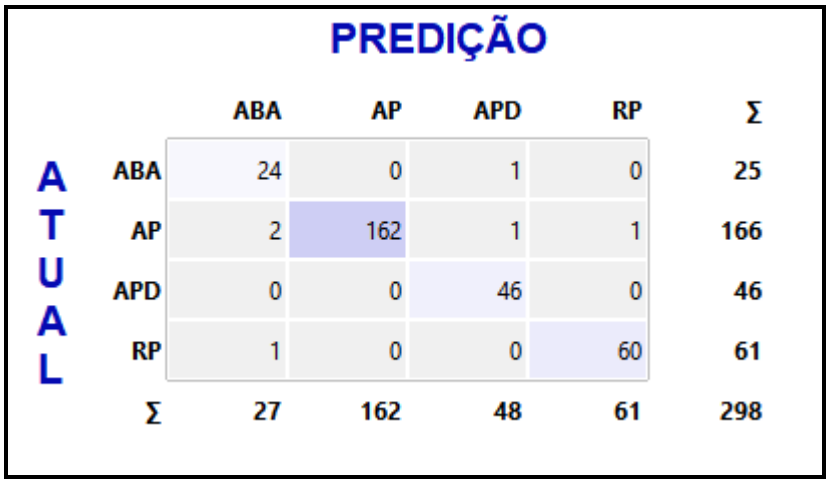


ATUAL \ PREDIÇÃO	ABA	AP	APD	RP	Σ
ABA	24	0	1	0	25
AP	2	164	0	0	166
APD	0	0	46	0	46
RP	0	0	0	61	61
Σ	26	164	47	61	298

Fonte: Elaborado pelo autor.

Figura 10

Matriz de confusão para o algoritmo *AdaBoost*



Fonte: Elaborado pelo autor.

Na Tabela 2, observam-se os valores de calibração de cada algoritmo de acordo com os: Valores Preditivos Positivos (VPP), Valores Preditivos Negativos (VPN), Taxa de Verdadeiro Positivo (TVP) e Taxa de Falso Positivo (TFP). As calibrações VPP, VPN e TVP, quanto mais próximo de 1,000 e o TFP mais próximo possível de 0,000 mais confiável será o algoritmo.

Tabela 2

Valores de Calibração de cada Algoritmo

Resultado Final	Calibração	Randon Forest	Neural Network	Tree	AdaBoost
AP Aprovado	VPP	1,000	0,994	1,000	1,000
	VPN	1,000	1,000	1,000	0,971
	TVP	1,000	1,000	1,000	0,976
	TFP	0,000	0,008	0,000	0,000
APD Aprovado com Dependência	VPP	1,000	1,000	0,979	0,958
	VPN	1,000	1,000	1,000	1,000
	TVP	1,000	1,000	1,000	1,000
	TFP	0,000	0,000	0,004	0,008
RP Reprovado	VPP	1,000	1,000	1,000	0,984
	VPN	1,000	1,000	1,000	0,996
	TVP	1,000	1,000	1,000	0,984
	TFP	0,000	0,000	0,000	0,004
ABA Evasão ou Abandono	VPP	1,000	1,000	0,923	0,889
	VPN	1,000	0,996	0,996	0,996
	TVP	1,000	0,960	0,960	0,960
	TFP	0,000	0,000	0,007	0,011

Fonte: Elaborado pelo autor.

Portanto, de acordo com os valores exposto na Tabela 2, pode-se concluir que o melhor

algoritmo para ser utilizado nesta análise é o *Random Forest*, pois os valores do VPP, VPN e TVP são todos 1,000 e o TFP são 0,000, tornando-o mais confiável para executar o banco de dados de teste. A classificação de VPP é importante, pois representa os alunos que provavelmente terão insucesso na disciplina, então a sua identificação é primordial para que ações sejam tomadas de forma a minimizar os índices de reprovação no semestre.

Resultado do Teste

O resultado final deste estudo pode ser verificado na Figura 11. Como foi esclarecido anteriormente, os estudantes enumerados na Figura 11, são reais, portanto, o resultado dos alunos foi o seguinte: Linha 1 a 10 Aprovado; Linha 11 e 12 Aprovado com Dependência; Linha 13 a 15 Evasão; Linha 16 a 20 Reprovado. Verifique que abaixo de cada coluna indicada por um algoritmo, surge uma sequência de 4 números separados por 'dois pontos' e ao final o resultado com maior probabilidade. Por exemplo, na linha 11 da Figura 11 na coluna *Random Forest* existe a seguinte sequência: 0.00 : 0.05 : 0.95 : 0.00 -> APD . Isso significa que este aluno tem 0.0% de ser evadido (ABA), 5% de ser aprovado (AP), 95% de ser aprovado com dependência (APD) e 0,0% de ser reprovado (RP).

Figura 11

Resultado Final de Predição

Predictions

Show probabilities for

ABA
AP
APD
RP

	Random Forest	Neural Network	AdaBoost	Tree
1	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
2	0.00 : 1.00 : 0.00 : 0.00 → AP	0.01 : 0.98 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
3	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
4	0.00 : 1.00 : 0.00 : 0.00 → AP	0.01 : 0.99 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
5	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
6	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 0.99 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
7	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
8	0.00 : 1.00 : 0.00 : 0.00 → AP	0.02 : 0.98 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
9	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
10	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP	0.00 : 1.00 : 0.00 : 0.00 → AP
11	0.00 : 0.05 : 0.95 : 0.00 → APD	0.02 : 0.02 : 0.89 : 0.07 → APD	0.00 : 0.00 : 1.00 : 0.00 → APD	0.00 : 0.00 : 1.00 : 0.00 → APD
12	0.02 : 0.10 : 0.79 : 0.09 → APD	0.03 : 0.04 : 0.84 : 0.08 → APD	0.00 : 0.00 : 1.00 : 0.00 → APD	0.00 : 0.00 : 1.00 : 0.00 → APD
13	0.75 : 0.25 : 0.00 : 0.00 → ABA	0.46 : 0.47 : 0.05 : 0.02 → AP	1.00 : 0.00 : 0.00 : 0.00 → ABA	1.00 : 0.00 : 0.00 : 0.00 → ABA
14	1.00 : 0.00 : 0.00 : 0.00 → ABA	0.88 : 0.02 : 0.02 : 0.08 → ABA	1.00 : 0.00 : 0.00 : 0.00 → ABA	1.00 : 0.00 : 0.00 : 0.00 → ABA
15	0.93 : 0.00 : 0.05 : 0.03 → ABA	0.95 : 0.02 : 0.00 : 0.03 → ABA	1.00 : 0.00 : 0.00 : 0.00 → ABA	0.50 : 0.50 : 0.00 : 0.00 → ABA
16	0.00 : 0.06 : 0.00 : 0.94 → RP	0.06 : 0.14 : 0.06 : 0.74 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP
17	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP
18	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.01 : 0.01 : 0.98 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP
19	0.05 : 0.00 : 0.01 : 0.94 → RP	0.03 : 0.01 : 0.13 : 0.82 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP
20	0.00 : 0.00 : 0.00 : 1.00 → RP	0.01 : 0.01 : 0.01 : 0.97 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP	0.00 : 0.00 : 0.00 : 1.00 → RP

Fonte: Elaborado pelo autor.

A Figura 12 é uma distribuição de frequência do resultado final do aluno comparado ao suporte educacional da família ao estudante, vermelho para os alunos com suporte e azul sem suporte. A Figura 13 é uma distribuição de frequência do resultado do aluno comparado ao suporte educacional extra proporcionado pela família ao estudante, vermelho para os alunos com suporte e azul sem suporte.

Os gráficos das figuras 12 e 13 indicam claramente que o suporte educacional familiar, tanto interno quanto externo, é fundamental para o bom desempenho do aluno. Esse suporte pode vir de várias formas, como o envolvimento direto dos pais nas atividades escolares, a criação de um ambiente de estudo adequado em casa, e o incentivo constante à dedicação acadêmica.

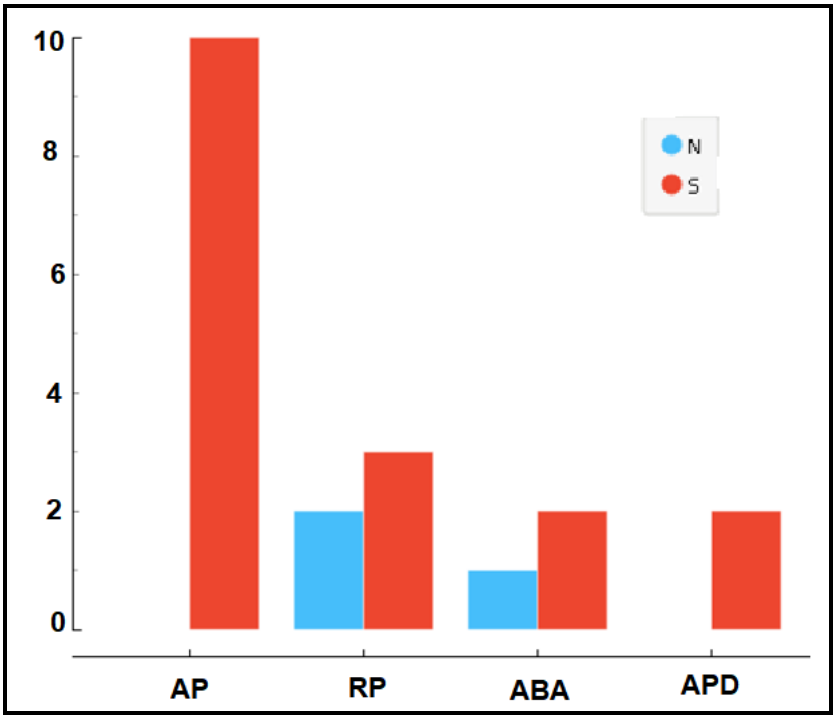
Internamente, os pais que acompanham o progresso dos filhos, ajudam nas tarefas de casa e mantêm uma comunicação regular com os professores, contribuem significativamente para o sucesso acadêmico dos alunos. Esse envolvimento demonstra para os alunos a importância da educação e os motiva a se esforçar mais.

Externamente, o acesso a recursos adicionais, como tutores, cursos extracurriculares e programas de reforço, também desempenha um papel crucial. Esses recursos complementam o aprendizado escolar e ajudam a preencher lacunas no conhecimento.

Além disso, um suporte emocional estável e positivo em casa pode aumentar a autoconfiança e a resiliência dos alunos, permitindo que eles enfrentem desafios acadêmicos com mais segurança. As famílias que incentivam a curiosidade intelectual e valorizam a educação criam um ambiente propício para o desenvolvimento cognitivo e emocional dos estudantes.

Figura 12

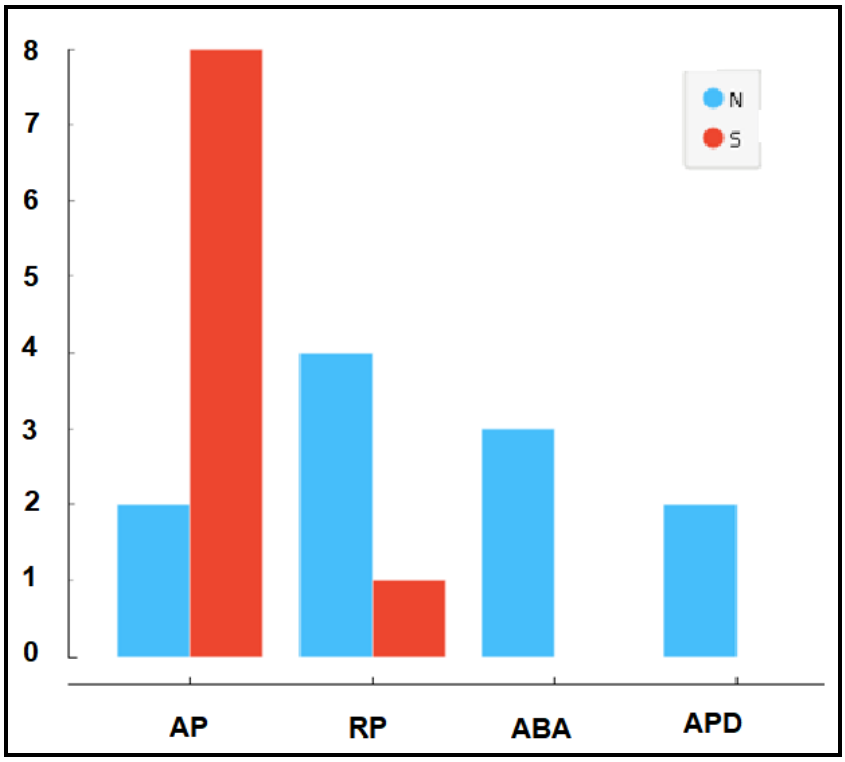
Distribuição de Frequência do Resultado Final do Aluno Comparado ao Suporte Educacional da Família



Fonte: Elaborado pelo autor.

Figura 13

Distribuição de Frequência do Resultado Final do Aluno Comparado ao Suporte Educacional Extra.



Fonte: Elaborado pelo autor

Portanto, o envolvimento ativo da família na educação não só melhora o desempenho acadêmico, mas também promove um desenvolvimento integral do aluno, preparando-o melhor para os desafios futuros. A combinação de apoio emocional, recursos educacionais e um ambiente de aprendizado positivo é essencial para que os alunos alcancem seu potencial máximo. Como a contribuição, temos que, em posse do modelo de previsão apresentado, as tomadas de decisões tanto de professores como de coordenadores serão facilitadas, pois, com a análise realizada, os docentes terão uma previsão dos alunos propensos a falhar numa disciplina, de modo que possam tomar ações educativas prévias a fim de reverter o possível insucesso do estudante nas disciplinas e, consequentemente, ajudar a diminuir os índices de reprovação e evasão em cursos profissionalizantes técnicos de informática.

Conclusão

Este trabalho apresenta um modelo moderno para identificar alunos com propensão à reprovação, aprovação, evasão e aprovação por dependência no primeiro semestre do curso técnico de informática. A validação do modelo foi realizada utilizando registros históricos de alunos que estudaram no CEMI-Gama entre 2018 e 2020, juntamente com dados socio econômico das famílias dos estudantes. A aplicação de algoritmos de predição do software ORANGE Data Mining demonstrou a eficácia das previsões realizadas.

Os resultados deste estudo são promissores e podem ser utilizados para desenvolver estratégias de intervenção personalizadas, visando reduzir as taxas de reprovação nas disciplinas de informática. Isso, por sua vez, pode contribuir significativamente para a diminuição dos índices de evasão, frequentemente exacerbados por altas taxas de reprovação. Além disso, a identificação precoce de alunos em risco permite a implementação de programas de apoio direcionados, como tutoria individualizada, recursos adicionais de aprendizagem e aconselhamento acadêmico, promovendo um ambiente de aprendizado mais inclusivo e eficaz.

Para futuros trabalhos, pretende-se expandir a análise para incluir dados adicionais que possam prever com maior precisão as áreas de conhecimento nas quais os alunos terão dificuldades. Isso envolve identificar não apenas as disciplinas, mas também os tópicos específicos que apresentam maior desafio, permitindo a criação de um modelo de classificação de estudantes ainda mais robusto. A utilização de técnicas de aprendizado de máquina avançadas e a integração de dados comportamentais e psicométricos poderão refinar ainda mais as previsões, oferecendo um mapeamento detalhado das

necessidades individuais dos alunos.

A longo prazo, espera-se que esses modelos aprimorados possam ser aplicados em diferentes contextos educacionais, adaptando-se às particularidades de cada instituição e curso. A personalização das intervenções educacionais com base em dados concretos tem o potencial de transformar a abordagem pedagógica, promovendo a retenção e o sucesso acadêmico em larga escala.

Referências

BARROS, R. et al. Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando métodos de aprendizagem de máquina. **Anais do XXX Simpósio Brasileiro de Informática na Educação** (SBIE 2019), p. 1491-1500, 2019.

BRANDÃO, Desirre Marques; SIMÃO, Flávio Pavesi; SOUZA, Marcos de. Impressões dos alunos do curso técnico em informática sobre a leitura e a produção textual utilizando Tecnologias de Informação e Comunicação. **InterScience Place**, n. 31, v. 1, p. 86-123, 2014.

CARDOSO, Rogério; ANTONELLO, Sérgio. Interdisciplinaridade, programação visual e robótica educacional: relato de experiência sobre o ensino inicial de programação. **Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação** (CBIE 2015) p. 1255-1262, 2015.

CASTRO, Ronney Moreira de; SIQUEIRA, Sean. Metodologias, Técnicas, Ambientes e Tecnologias Alternativas Utilizadas no Ensino de Algoritmos e Programação no Ensino Superior no Brasil. **Anais dos Workshops do VIII Congresso Brasileiro de Informática na Educação** (CBIE 2019), p. 228-237, 2019.

CHARITOPOULOS, Angelos; RANGOUSSI, Maria; KOULOURIOTIS, Dimitrios. On the Use of Soft Computing Methods in Educational Data Mining and Learning Analytics Research: a Review of Years 2010–2018. **International Journal of Artificial Intelligence in Education**, v. 30 p. 371-430, 2020.

DEMŠAR, J. et al. Orange: Data mining toolbox in python**.** J**ournal of Machine Learning Research,** n.14, p.2349-2353, 2013.

DINIZ, Elza Magela; SANTOS, Talitha Araújo. Retenção e evasão escolar na educação profissional de nível médio técnico: o que nos dizem as publicações da ANPED entre os anos 2012 a 2017. **Brazilian Journal of Development**, Curitiba, v. 6, n. 7, p.44829-44838, 2020.

FREDENHAGEM, Sheyla Villar. Evasão Escolar no âmbito do Instituto Federal de Brasília. **Revista EIXO**, v.3 n.2, p. 49-71, 2014.

GARCIA, Rogerio Eduardo; CORREIA, Ronaldo Celso Messias; SHIMABUKURO, Milton Hirokazu. Ensino de Lógica de Programação e Estruturas de Dados para Alunos do Ensino Médio. **Anais do XXVIII Congresso da SBC**, Belém do Pará, PA, p.246-249, 2008.

HINTERHOLZ, Lucas; CRUZ, Marcia Kniphoff da. Desenvolvimento do Pensamento Computacional: um relato de atividade junto ao Ensino Médio, através do Estágio Supervisionado em Computação III. **Anais do XXI Workshop de Informática na Escola** (WIE 2015), 2015.

KOVACIC, J. Zlatko. Early Prediction of Student Success: Mining Students Enrolment Data. **Proceedings of the 2010 InSITE Conference**, v. 10 p. 647-665, 2010.

MASCHIO, Pedro. et al. Um Panorama acerca da Mineração de Dados Educacionais no Brasil. **Anais do XXIX Simpósio Brasileiro de Informática na Educação** (SBIE 2018), p. 1936-1940, 2018.

MIGUÉIS, V. L. et al. Early segmentation of students according to their academic performance: A predictive modelling approach. **Decision Support Systems**, v 115, novembro, p. 36-51, 2018.

PEREIRA, Filipe Dwan et al. Predição de desempenho em ambientes computacionais para turmas de programação: um Mapeamento Sistemático da Literatura. **Anais do XXXI Simpósio Brasileiro de Informática na Educação** (SBIE 2020), p. 1673-1682, 2020.

PINTO, Glevson Da Silva; FREITAS JÚNIOR, Olival De Gusmão; COSTA, Evandro De Barros. Mineração de Dados Educacionais: Um Modelo de Predição do Perfil do Aluno para Melhoria do IDEB. **Anais do XXXI Simpósio Brasileiro de Informática na Educação** (SBIE 2020), p. 1172-1182, 2020.

PORTAL, Cleber; SCHLEMMER, Eliane. Estratégias Para Minimizar a Evasão Na Eduação a Distância: O Uso De Um Sistema De Mineração De Dados Educacionais E Learning Analytics. In: **21° CIAED Congresso Internacional ABED**, v. 1. p. 1-10, Bento Goncalves, 2015.

QIAN, Yizhou; LEHMAN, James D. Correlates of Success in Introductory Programming: A Study with Middle School Students. **Journal of Education and Learning**, 2016.

QUEIROGA, Emanuel; CECHINEL, Cristian; ARAÚJO, Ricardo. Predição de estudantes com risco de evasão em cursos técnicos a distância. **Anais do XXVIII Simpósio Brasileiro de Informática na Educação** (SBIE 2017), p. 1547-1556, 2017.

RIBAS, Elisângela; DAL BIANCO, Guilherme; LAHM, Regis Alexandre. Programação visual para introdução ao ensino de programação na Educação Superior: uma análise prática Palavras-chave: ensino de programação-linguagem visual-estratégias de ensino. **Revista Novas Tecnologias na Educação**, v. 14, n. 2, p. 1-10, 2016.

RIBEIRO, Karen Da Silva Figueiredo Medeiros; MACIEL, Cristiano. Um Estudo sobre o Desenvolvimento da Carreira das Estudantes do Ensino Médio Integrado em Informática. **Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação** (WCBIE 2020), p.21-30, 2020.

RIGO, Sandro José; CAZELLA, Silvio C.; CAMBRUZZI, Wagner. Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. **Revista Brasileira de Informática na Educação,** p. 168-177, 2014.

SAA, Amjad Abu. Educational Data Mining & Students' Performance Prediction. (IJACSA) **International Journal of Advanced Computer Science and Applications**, v. 7, n. 5, p. 212-220, 2016.

SOUZA, Draylson Micael; BATISTA, Marisa Helena da Silva; BARBOSA, Ellen Francine. Problemas e Dificuldades no Ensino de Programação: Um Mapeamento Sistemático. R**evista Brasileira de Informática na Educação**, v. 24, n. 1, p. 39-52, 2016.

SOUZA, Francislaine Ávila de; ARAÚJO ANDRADE, José Antônio; DE PAULO MARTINS, Francine. As práticas de letramento matemático digital e o papel mediador das tecnologias digitais: uma experiência com o software superlog na educação básica**. Revista Devir Educação**, Lavras-MG. Edição Especial –Ago p. 155-178, 2020.

TSIAKMAKI, Maria *et al.* Implementing autoML in educational data mining for prediction tasks**. Applied Sciences** (Switzerland), v. 10, n. 90, p. 1-27. 2020.

Revisão textual e de normas da ABNT realizada por: Karen Valéria Sardo Leão.