Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 12/Jul/2023; 1st round notif.: 17/Oct/2023; New version: 17/Oct/2023; 2nd round notif.: 06/Dec/2023
Camera ready: 20/Dec/2023; Edition review: 09/Jan/2024; Available online: 09/Jan/2024; Published: 09/Jan/2024

Modelo de Predição de Evasão Escolar com Base em Dados de Autoavaliação de Cursos de Graduação

Title: School Dropout Prediction Model Based on Undergraduate Course Self-Assessment Data

Título: Modelo de Predicción de Abandono Escolar basado en Datos de Autoevaluación de Cursos de Grado

Ronei dos Santos Oliveira
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, campus João Pessoa
ORCID: 0009-0008-3261-437X
ronei.santos@academico.ifpb.edu.br

Francisco Petronio Alencar de Medeiros
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, campus João Pessoa
ORCID: 0000-0003-2955-6785
petronio@ifpb.edu.br

Resumo
A evasão escolar é um desafio diário para instituições de ensino, no caso específico do ensino superior as altas taxas acarretam perdas financeiras e escassez de profissionais no mercado. Esta pesquisa teve como objetivo desenvolver e avaliar um modelo preditivo para identificar alunos propensos à evasão, utilizando dados de um modelo semestral de autoavaliação dos cursos de graduação da Universidade Federal da Paraíba (UFPB). Utilizando a mineração de dados educacionais e a metodologia CRISP-EDM, o estudo analisou a relação entre evasão escolar e autoavaliação institucional, seguido de análise exploratória e preparação dos dados para classificação. Diversas técnicas de modelagem, como Árvore de Decisão, Floresta Aleatória e Máquinas de Vetores de Suporte, foram aplicadas, sendo os modelos avaliados por métricas de desempenho, revelando uma acurácia de 87,97%, precisão de 91,72%, recall de 91,67% e medida F de 91,57% na identificação de alunos com alta probabilidade de evasão. Cerca de 59% dos alunos ativos da UFPB admitidos a partir de 2017 demonstraram probabilidade de abandonar seus cursos nos testes do modelo preditivo proposto. Essas informações podem embasar decisões institucionais e a implementação de políticas e ações eficazes contra a evasão, visando melhorar os resultados acadêmicos. O estudo contribui para avanços na predição de evasão escolar, fornecendo insights valiosos para decisões e estratégias preventivas na UFPB e outras instituições de ensino superior.
Palavras-Chave: *Evasão escolar; Mineração de dados educacionais; Modelo preditivo; Autoavaliação.*

Abstract
School dropout is a daily challenge for educational institutions. In the specific case of higher education, high dropout rates lead to financial losses and a shortage of professionals in the market. This research aimed to develop and evaluate a predictive model to identify students prone to dropout, using data from a semester-based self-assessment model of undergraduate courses at Federal University of Paraíba (UFPB). The study analyzed the relationship between school dropout and institutional self-assessment by utilizing educational data mining and the CRISP-DM methodology, followed by exploratory analysis and data preparation for classification. Various modeling techniques, such as Decision Trees, Random Forest, and Support Vector Machines, were applied, with the models evaluated using performance metrics, revealing an accuracy of 87.97%, precision of 91.72%, recall of 91.67%, and F-measure of 91.57% in identifying students with a high probability of dropout. Approximately 59% of active students at UFPB, admitted from 2017 onwards, showed a high probability of abandoning their courses in the proposed predictive model tests. This information can support institutional decisions and the implementation of effective policies and actions against dropouts, aiming to improve academic outcomes. The study contributes to advancements in predicting school dropout, providing valuable insights for decision-making and preventive strategies in UFPB and other higher education institutions.
Keywords: *School dropout; Educational data mining; Predictive model; Self-evaluation.*

Cite as: Oliveira, R. S. & Medeiros, F. P. A. (2024). Modelo de Predição de Evasão Escolar com Base em Dados de Autoavaliação de Cursos de Graduação. Revista Brasileira de Informática na Educação, 32, 01-21. https://doi.org/10.5753/rbie.2024.3542

Resumen
La deserción escolar es un desafío diario para las instituciones educativas, y en el caso específico de la educación superior, las altas tasas conllevan pérdidas financieras y escasez de profesionales en el mercado. El objetivo de esta investigación fue desarrollar y evaluar un modelo predictivo para identificar a los estudiantes propensos a la deserción, utilizando datos de un modelo de autoevaluación semestral de los cursos de grado en Universidad Federal de Paraíba (UFPB). Mediante la minería de datos educativos y la metodología CRISP-EDM, el estudio analizó la relación entre la deserción escolar y la autoevaluación institucional, seguida de un análisis exploratorio y preparación de los datos para la clasificación. Se aplicaron diversas técnicas de modelado, como árboles de decisión, bosques aleatorios y máquinas de vectores de soporte, y los modelos se evaluaron mediante métricas de rendimiento, revelando una precisión del 87,97%, una precisión del 91,72%, una sensibilidad del 91,67% y una medida F del 91,57% en la identificación de estudiantes con alta probabilidad de deserción. Aproximadamente el 59% de los estudiantes activos en UFPB, admitidos a partir de 2017, mostraron una alta probabilidad de abandonar sus cursos en las pruebas del modelo predictivo propuesto. Esta información puede respaldar las decisiones institucionales y la implementación de políticas y acciones eficaces contra la deserción, con el objetivo de mejorar los resultados académicos. El estudio contribuye al avance en la predicción de la deserción escolar, proporcionando conocimientos valiosos para la toma de decisiones y estrategias preventivas en UFPB y otras instituciones de educación superior.
Palabras clave: *Deserción escolar; Minería de datos educativos; Modelo predictivo; autoevaluación.*

1 Introdução

A evasão escolar é um problema antigo e complexo que afeta todos os níveis de ensino. Além dos impactos individuais, a evasão escolar também causa graves prejuízos econômicos e sociais para os estudantes, as instituições e a sociedade em geral (Prestes & Fialho, 2018). No Ensino Superior, a evasão escolar tem consequências na escassez de profissionais em várias áreas, comprometendo todo um ecossistema necessário (Saccaro et al., 2019).

Para uma compreensão mais clara da evasão escolar no contexto da educação superior, uma comissão formada pelo MEC (ANDIFES et al., 1996) definiu diferentes categorias de evasão, como apresentado na Tabela 1. No entanto, este trabalho se concentrará apenas na evasão do curso, pois as categorias de evasão da instituição e do sistema possuem características distintas que não serão objeto de estudo nesta pesquisa.

Tabela 1: Definições de evasão no âmbito da educação superior.

	Definição
Evasão do curso	Quando o estudante se desliga do curso superior em situações diversas tais como: • Abandono (deixa de matricular-se); • Desistência (oficial); • Transferência ou reopção (mudança de curso); • Exclusão por norma institucional.
Evasão da instituição	Quando o estudante se desliga da instituição na qual está matriculado.
Evasão do sistema	Quando o estudante abandona de forma definitiva ou temporária o ensino superior.

Ao analisar os problemas decorrentes da evasão escolar nas instituições públicas brasileiras, é importante considerar que essas instituições desempenham um papel fundamental no setor produtivo, sendo responsáveis por uma grande parte das produções científicas e registros de patentes no país (Gamba & Righetti, 2022; Lousrhania, 2021). Nesse contexto, a evasão escolar prejudica o progresso acadêmico do aluno, interrompendo sua permanência no curso e comprometendo todo o ecossistema educacional. Além disso, a evasão escolar resulta em perdas econômicas tanto para o Estado quanto para a gestão universitária.

Apesar das várias observações e concepções sobre a evasão escolar, compreender as suas razões continua sendo um desafio para a maioria das instituições. Para obter um maior entendimento, a Universidade Federal da Paraíba (UFPB) realiza, a cada semestre, uma autoavaliação compulsória e anônima dos cursos de graduação por meio do Sistema Integrado de

Gestão de Atividades Acadêmicas (SIGAA). Essa avaliação fornece uma ampla gama de dados que podem ser explorados para compreender as especificidades da instituição.

A autoavaliação das Instituições de Ensino Superior (IES) é um processo de reflexão contínua sobre todas as ações institucionais, que vai além da prestação de contas ao Ministério da Educação. A relação entre avaliação e gestão varia de acordo com a missão e características de cada IES. No caso da UFPB, a utilização dos dados da autoavaliação dos cursos pode subsidiar a implementação de ações específicas para melhorar a qualidade do ensino e combater a evasão escolar (Baggi & Lopes, 2011).

De acordo com Baker et al. (2011), a Mineração de Dados Educacionais (MDE) consiste na aplicação de métodos de mineração de dados e Aprendizado de Máquina (AM) na área da educação. Seu objetivo é descobrir conhecimento em bases de dados relacionadas a contextos educacionais (Saraiva et al., 2019). Por meio da análise e fusão dos diferentes aspectos encontrados nos dados, conhecidos como variáveis preditoras, é possível prever informações a partir de aspectos específicos dos dados, chamados de variáveis preditivas. Apesar de existirem diversos estudos sobre a predição de evasão escolar utilizando MDE, existe uma predominância no uso de dados como fatores pessoais, acadêmicos, econômicos, sociais e institucionais (Alban & Mauricio, 2019).

A motivação para esse estudo reside na possibilidade de utilizar a MDE sobre os dados da autoavaliação dos cursos de graduação como uma ferramenta importante para subsidiar a alta gestão na implementação de ações específicas que possam ajudar os estudantes a concluírem o curso no tempo estabelecido pelo projeto pedagógico. Isso, por sua vez, contribui para que os alunos tenham uma formação de qualidade, sejam absorvidos pelo mercado de trabalho ou prossigam seus estudos acadêmicos.

Este trabalho também busca oferecer contribuições nos campos educacional, científico-tecnológico e social. No campo educacional, visa fornecer informações para a gestão institucional, permitindo ações que melhorem a qualidade e eficácia do ensino oferecido. Do ponto de vista científico-tecnológico, a pesquisa e o uso de recursos computacionais no desenvolvimento do modelo preditivo deixam um legado para esse campo. Já em termos sociais, a redução dos índices de evasão escolar contribui para que os alunos cumpram sua jornada acadêmica, atendendo às expectativas da sociedade e da comunidade acadêmica.

Além disso, a autoavaliação dos cursos de graduação pode contribuir para os processos acadêmicos e administrativos, fornecendo um instrumento para a correção de metas e objetivos. No entanto, é importante ressaltar que a simples coleta e divulgação de dados não são suficientes para melhorar a qualidade educacional. É necessário processar esses dados utilizando inteligência cognitiva e computacional, a fim de gerar informações valiosas para a instituição. Dentro desse contexto, os objetivos deste trabalho é desenvolver um modelo preditivo que utilize os dados da autoavaliação dos cursos de graduação da UFPB para identificar precocemente a evasão escolar, possibilitando a adoção de medidas preventivas por parte dos stakeholders, como a alta gestão, coordenadores de cursos e professores. Para atingir esses objetivos, a estrutura do trabalho está organizada da seguinte forma: Na Seção 2, são apresentados os trabalhos relacionados, com o intuito de identificar o estado da arte e trabalhos relevantes alinhados com o tema de pesquisa. Na Seção 3, é apresentado o método utilizado pela pesquisa para propor o modelo de previsão. Na Seção 4, são apresentados os resultados obtidos. Finalmente, na Seção 5, são realizadas as considerações finais e são fornecidas indicações para trabalhos futuros.

2 Trabalhos Relacionados

Lottering et al. (2020) conduziram uma pesquisa que aplicou técnicas de Mineração de Dados Educacionais (MDE) e Aprendizado de Máquina (AM) para identificar alunos em risco de evasão em uma universidade de tecnologia na África do Sul. Eles utilizaram as notas finais dos alunos, enriquecendo o conjunto de dados com 19 atributos derivados do banco de dados da instituição. Para abordar o desbalanceamento dos dados, realizaram uma subamostragem, resultando em um conjunto de dados com 1.156 registros. Em seguida, aplicaram diversos algoritmos de classificação supervisionada, com a Máquina de Vetores de Suporte (*Support Vector Machine* – SVM) demonstrando o melhor desempenho, alcançando uma pontuação F-Measure de 99,32%. Essa metodologia contribui para a pesquisa atual ao fornecer insights sobre modelos com dados balanceados e destacar o desempenho superior do SVM.

A pesquisa realizada por Santos et al. (2021.1) utilizou dados coletados de uma instituição de ensino superior brasileira, sendo selecionados sete atributos para análise. A peculiaridade deste estudo reside na construção de dez modelos de dados, cada um correspondendo a diferentes semestres cursados, variando desde o primeiro ao décimo semestre. Utilizando o algoritmo de Árvore de Decisão (DT), foram obtidos resultados de acurácia satisfatórios, variando de 79,31% (3º semestre) a 98,25% (9º semestre). Nota-se que alguns modelos apresentaram 100% de precisão para a classe de Evasão, indicando que todas as previsões de evasão estavam corretas. No entanto, no caso do modelo do 3º semestre, somente 80,35% das previsões de evasão estavam corretas. Em relação à classe de Formatura, a precisão variou de 75,68% a 97,22%. É relevante mencionar que alguns modelos atingiram 100% de recall para a classe de Formatura, ou seja, todas as formaturas foram previstas corretamente. No entanto, no modelo do 3º semestre, apenas 68,57% das formaturas foram previstas com precisão. No que concerne à classe de Evasão, o recall variou de 82,22% a 96,43%. É importante salientar que o poder preditivo dos modelos aumentou à medida que os semestres avançaram, devido à inclusão progressiva de atributos (disciplinas) nos modelos dos semestres mais avançados. Esses resultados contribuem significativamente para o contexto da pesquisa sobre o desenvolvimento de modelos de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação, fornecendo informações sobre a relação entre o desempenho acadêmico dos alunos e a probabilidade de evasão ou formatura ao longo dos diferentes semestres.

O estudo de Manrique et al. (2019) desempenha um papel de relevância ao explorar três abordagens distintas na predição de evasão em uma instituição de ensino superior no Brasil. A primeira abordagem se baseia em um modelo de previsão que utiliza variáveis genéricas, aplicáveis a qualquer curso, enquanto a segunda utiliza dados específicos do curso para a predição exclusiva deste curso. Ambas as abordagens empregam algoritmos de Aprendizado de Máquina, incluindo Gradient Boosting Tree, SVM, Floresta Aleatória (Random Forest – RF) e Naive Bayes. A terceira abordagem analisa a evolução dos dados dos alunos ao longo do curso, utilizando o algoritmo K-Nearest Neighbors. Os resultados indicam que o modelo com recursos globais, em conjunto com o algoritmo RF, obteve os melhores resultados nas métricas de acurácia, recall, precisão e F-Measure. Essas descobertas representam uma contribuição significativa, destacando o desempenho superior da abordagem de recursos globais com variáveis genéricas independentes do curso.

Este trabalho avança frente aos trabalhos relacionados, principalmente com relação ao contexto dos dados utilizados. A pesquisa de predição de evasão escolar com base em dados de um instrumento único de autoavaliação de cursos de graduação pode trazer avanços científicos significativos ao proporcionar uma compreensão mais profunda dos fatores e padrões que levam à evasão. Ao utilizar técnicas de mineração de dados e modelos preditivos, essa pesquisa pode identificar correlações ocultas e fatores de risco, fornecendo insights valiosos para o desenvolvimento de estratégias de intervenção mais eficazes. Esses avanços científicos

contribuem para o campo da educação, podendo melhorar a retenção dos alunos e ajudando a desenvolver abordagens mais personalizadas e baseadas em evidências para enfrentar o desafio da evasão escolar.

3 Método

Para atingir os objetivos estabelecidos, foi realizado um processo de Mineração de Dados Educacionais (MDE), que envolve a aplicação de métodos de mineração de dados e aprendizado de máquina no contexto educacional, visando descobrir conhecimentos em bases de dados educacionais (SARAIVA et al., 2019). Para auxiliar nesse processo, esta pesquisa utilizou a metodologia CRISP-EDM (acrônimo de CRoss-Industry Standard Process for Educational Data Mining), uma adaptação da metodologia CRISP-DM (acrônimo de CRoss-Industry Standard Process for Data Mining) para o contexto educacional (RAMOS et al., 2020).

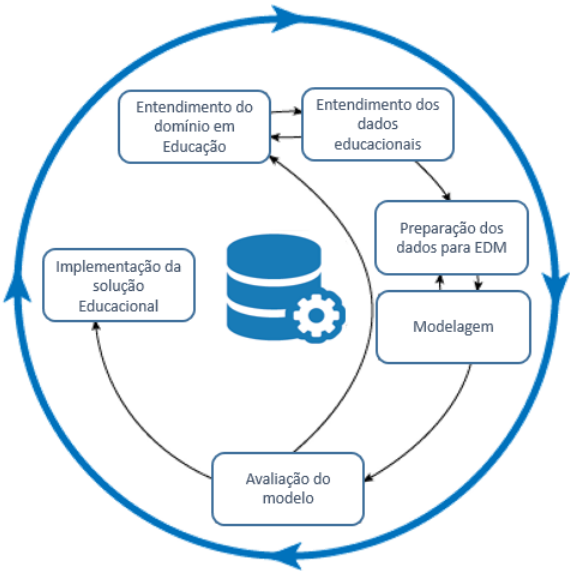

Figura 1: Metodologia CRISP-EDM.

Cada etapa do CRISP-EDM foi desenvolvida com técnicas e abordagens adequadas ao domínio educacional em análise. Dessa forma, a pesquisa foi conduzida em seis etapas, conforme ilustrado na Figura 1.

3.1 Instrumento de autoavaliação

A primeira etapa do processo foi entender o domínio da aplicação, ou seja, qual o contexto dos dados educacionais que seriam minerados e a sua correlação com a evasão escolar. Os dados brutos obtidos são frutos da autoavaliação dos cursos de graduação que é realizada pelos alunos da UFPB a cada início de semestre. A autoavaliação é realizada com base no semestre anteriormente cursado, de forma compulsória e como pré-requisito para a obtenção de matrícula para o semestre subsequente.

O instrumento de avaliação da educação superior pelo discente possui quatro dimensões distintas (Costa & Dias, 2020). Na dimensão discente, o aluno realiza uma autoavaliação do seu desempenho em cada disciplina, atribuindo uma nota de 0 a 10 para seu comprometimento e motivação. Na dimensão disciplina, o aluno avalia a importância e dificuldade percebidas em cada disciplina cursada, também utilizando uma escala de 0 a 10. Na dimensão docente, é realizada uma avaliação sobre ajustes necessários por parte dos professores, considerando aspectos como cumprimento do plano de curso, relacionamento com a turma, comparecimento às aulas, entre

outros. Além disso, o aluno avalia sua satisfação geral com o desempenho de cada professor, atribuindo uma nota de 0 a 10. Por fim, na dimensão curso, o aluno pode indicar a probabilidade de recomendar o curso para um amigo ou parente próximo, utilizando uma escala de 0 a 10, e expressar seu interesse em evadir do curso atual, utilizando a mesma escala.

Outro fato relevante para o desenvolvimento do modelo é que esse instrumento só passou a ser aplicado nesse formato a partir do primeiro semestre de 2017, para todos os cursos de graduação da instituição. Nesse contexto, os dados analisados só contaram com alunos matriculados a partir do referido semestre, essa abordagem se deve pelo fato de que alunos que não realizaram avaliações durante todo o seu ciclo acadêmico, produziram dados enviesados que prejudicaram a eficiência do modelo. Esse fato foi constatado perfazendo todo o ciclo do CRISP-EDM e realizando refinamentos no modelo.

3.2 Conjunto de dados

Nesta etapa, foi primordial entender os dados brutos obtidos de forma a realizar uma preparação adequada dos dados na etapa subsequente. Na Tabela 2, podemos observar as informações sobre as variáveis do conjunto de dados brutos que foram utilizadas nesta etapa da metodologia.

Tabela 2: Conjunto de dados brutos.

Variável	Definição	Tipo
ANO	Ano de referência que está sendo avaliado.	Discreta
PERIODO	Período de referência que está sendo avaliado.	Discreta
MATRICULA	Número da matrícula do aluno.	Contínua
STATUS_DISCENTE	Situação atual do aluno no momento da extração dos dados.	Categórica
CENTRO	Nome do Centro do curso do aluno.	Categórica
DEPARTAMENTO	Nome do Departamento do curso do aluno.	Categórica
CODIGO	Código da disciplina.	Contínua
DISCIPLINA	Título da disciplina.	Categórica
CODIGO_TURMA	Código da turma cursada.	Contínua
HORARIO	Horário da turma.	Categórica
LOCAL	Local da turma.	Categórica
CURSO	Nome do curso do aluno.	Categórica
TURNO	Turno do curso.	Categórica
MEDIA_FINAL	Média final obtida na disciplina.	Discreta
SITUACAO_MATRICULA	Status obtido na disciplina.	Categórica
QUARTA_PROVA	Informa se a disciplina possui quarta prova.	Discreta
FALTAS	Quantidade de faltas do aluno durante o período.	Contínua
1.1.1	Nota (de 0 - muito ruim, a 10 - muito bom) para o desempenho pessoal na disciplina em termos de comprometimento e motivação.	Discreta
2.1.1	Nível de importância (de 0 - sem importância, a 10 - extremamente importante) das disciplinas cursadas.	Discreta
2.2.1	Nível de dificuldade dos conteúdos das disciplinas cursadas (de 0 - muito fácil, a 10 - muito difícil).	Discreta
3.1.1.A	Professor precisa ajustar o cumprimento do plano de curso (sim ou não).	Discreta
3.1.1.B	Professor precisa ajustar o relacionamento com a turma (sim ou não).	Discreta
3.1.1.C	Professor precisa ajustar o comparecimento às aulas (sim ou não).	Discreta
3.1.1.D	Professor precisa ajustar o cumprimento do horário de início e de término das aulas (sim ou não).	Discreta
3.1.1.E	Professor precisa ajustar a atualização dos conteúdos (sim ou não).	Discreta
3.1.1.F	Professor precisa ajustar a clareza na exposição dos conteúdos (sim ou não).	Discreta
3.1.1.G	Professor precisa ajustar a disponibilidade para atendimento fora da sala de aula (sim ou não).	Discreta

Continua na próxima página.

Tabela 3: Conjunto de dados brutos. *(continuação)*

Variável	Definição	Tipo
3.1.1.H	Professor precisa ajustar a qualidade da bibliografia (sim ou não).	Discreta
3.1.1.I	Professor precisa ajustar a qualidade das avaliações (sim ou não).	Discreta
3.2.1	Satisfação geral (de 0 - totalmente insatisfeito, a 10 - totalmente satisfeito) com o desempenho do professor.	Discreta
4.1.1	Probabilidade de recomendar o curso para um amigo ou parente próximo (de 0 - muito improvável, a 10 - muito provável).	Discreta
4.2.1	Interesse em sair de curso (mudar de curso na UFPB ou para outra instituição, parar de estudar etc.) atualmente (de 0 - muito baixo, a 10 - muito alto).	Discreta
OBSERVACOES	Texto livre para qualquer manifestação adicional.	Categórica
QUANTIDADE_ TRANCAMENTOS	Número de trancamentos realizado no período.	Contínua
ANO	Ano de referência que está sendo avaliado.	Discreta

Para entender melhor os dados brutos descritos acima, foi realizada uma análise exploratória utilizando a linguagem de programação Python. Foram analisados 1.156.891 registros das avaliações dos cursos de graduação na modalidade presencial. O principal ponto de atenção dessa etapa se concentrou na variável alvo. Ela está localizada na coluna STATUS_DISCENTE e tem como possíveis valores as categorias descritas na Tabela 3.

Tabela 4: Valores assumidos pela variável alvo.

	Definição
ATIVO	Esse status é associado ao aluno que possui vínculo em vigor com a instituição, que não se encontra com status "Trancado" ou "Formando" e que está matriculado no conjunto mínimo de componente curriculares do seu curso.
TRANCADO	É o status do aluno que realizou a suspensão de programa. Nesse caso, o discente tem vínculo “Ativo”, porém, solicitou a interrupção temporária.
CONCLUIDO	É o status do aluno que concluiu todas as pendências acadêmicas exigidas pela sua estrutura curricular, que já recebeu o grau acadêmico e teve seu diploma registrado.
ATIVO – FORMANDO	É o aluno que tem condições para a conclusão de seu curso no período atual. Ou seja, estudante que está cursando os últimos componentes curriculares para finalizar a carga horária mínima de seu curso.
ATIVO – CONCLUINTE	É o aluno que já concluiu a carga horária mínima de seu curso, porém ainda não recebeu o grau acadêmico.
CANCELADO	É a situação do aluno que teve seu vínculo finalizado como evadido, seja por desistência, insuficiência de rendimento acadêmico, decurso de prazo máximo etc.

Um fato relevante sobre a variável alvo, é que o valor dela corresponde ao status do discente no momento da extração dos dados no sistema, ou seja, o status não corresponde ao que era no momento da avaliação. Dessa forma, existem períodos que possuem avaliações de um aluno que não evadiu naquele momento, por exemplo, se um aluno cursou cinco períodos até evadir, então todas as avaliações de disciplinas dos cinco períodos terão como valor de variável alvo o status CANCELADO.

3.3 Preparação dos dados

Com base nas análises prévias realizadas e refinamentos após perfazer alguns ciclos do CRISP-EDM, foram realizadas algumas tomadas de decisões para gerar um subconjunto de dados capaz de performar na etapa de modelagem.

3.3.1 Filtragem

Foi realizada uma filtragem com relação ao ano de matrícula dos alunos. Como a autoavaliação só passou a ser aplicada a partir do primeiro semestre do ano de 2017, optou-se por utilizar apenas as avaliações de alunos que ingressaram nesse período em diante, pois todo o ciclo acadêmico do

estudante, até o momento da extração dos dados, está representado. O mesmo não acontece com as matrículas antes de 2017, pois os registros atuais não contêm as avaliações de várias disciplinas já cursadas antes aplicação do instrumento de autoavaliação. Dessa forma, o novo conjunto de dados brutos passou a ter 683.634 registros, sendo desconsiderado 473.257 registros referentes as matrículas antes de 2017.

Outra filtragem que foi realizada foi com relação à nossa variável alvo STATUS_DISCENTE. Como os status ATIVO e TRANCADO são de alunos que ainda possuem vínculo com a instituição, não é possível atestar se esse aluno irá evadir ou não. Diante desse quadro, consideramos apenas os valores CONCLUIDO, ATIVO – FORMANDO, ATIVO – CONCLUINTE (não evadido) e CANCELADO (evadido), para realizar os treinamentos e testes dos modelos desenvolvidos, pois esses status são finalizadores e atestam se o aluno concluiu a sua jornada acadêmica ou saiu do curso. Com a aplicação desse filtro nosso conjunto de dados brutos passou a ter 128.235 registros. Os outros 555.399 registros referentes aos status não finalizadores não são utilizados na modelagem. Apesar da diferença entre os dados utilizados na modelagem e não utilizados, ainda temos uma quantidade significativa de dados para a obtenção de um modelo bastante representativo.

3.3.2 Remoção de variáveis

Algumas variáveis foram descartadas por não fazer sentido para o contexto da pesquisa. Apesar de elas possuírem informações relevantes e que poderiam levar a alguns subconjuntos de dados que podem ser estudados sob outra perspectiva, a proposta foca em um modelo genérico, independente de curso, disciplina, turno, entre outras características, que contemple todos os cursos da UFPB e sirva de baseline para pesquisas futuras. Diante desse contexto, as variáveis ANO, PERIODO, CENTRO, DEPARTAMENTO, CODIGO, DISCIPLINA, CODIGO_TURMA, HORARIO, LOCAL, CURSO, TURNO, SITUACAO_MATRICULA, QUARTA_PROVA e OBSERVACOES não foram utilizadas para a criação do modelo preditor.

3.3.3 Remoção de valores nulos e outliers

Todas as variáveis foram verificadas quanto a incidência de valores nulos e outliers. Para a variável MATRICULA não foi encontrado nenhum registro com valores nulos ou outliers. Quanto a variável STATUS_DISCENTE não foi encontrado nenhum valor nulo e foi considerado como outlier qualquer valor diferente dos apresentados na Tabela 3. Foram removidos todos os valores nulos nas variáveis MEDIA_FINAL, FALTAS, 1.1.1, 2.1.1, 2.2.1, 3.2.1, 4.1.1 e 4.2.1, para detectar os outliers foi levado em consideração que essas podem assumir valores entre 0 e 10, dessa forma qualquer valor fora desse intervalo é considerado um outlier. Com relação as variáveis 3.1.1.A, 3.1.1.B, 3.1.1.C, 3.1.1.D, 3.1.1.E, 3.1.1.F, 3.1.1.G, 3.1.1.H e 3.1.1.I, verificou-se que elas assumem o valor "X" quando o aluno assinala que o item precisar ser ajustado pelo professor e nenhum valor é atribuído (nulo) quando o aluno não assinala o item, dessa forma foi considerado como outlier qualquer valor que não fosse "X" ou nulo para essas variáveis, nesse caso os valores nulos não foram removidos. No caso da variável QUANTIDADE_TRANCAMENTOS, notou-se que eram informados a quantidade de trancamentos apenas nos casos em que ocorreram trancamentos, já nos casos que não ocorriam trancamento nenhum valor era informado (nulo), dessa forma foi considerado como outlier valores que não fossem inteiros ou nulos, aqui também não foram removidos os valores nulos.

Dessa forma, após a remoção de todos os *outliers*, nosso subconjunto de dados passou a ter 120.341 registros de avaliações. Deve-se observar que esses dados ainda possuem inconsistências. Uma única matrícula possui vários registros nesse conjunto de dados, ou seja, para cada disciplina cursada em diferentes períodos representam um registro. No entanto, como mencionado anteriormente, todos os registros possuem como valor da sua variável alvo o status final do

discente no momento da extração dos dados. Diante desse contexto foi necessário realizar transformações nesses dados.

3.3.4 Transformações

A variável MATRICULA foi anonimizada para que se torna-se impossível saber qual os alunos utilizados para a elaboração do modelo. A variável alvo STATUS_DISCENTE foi transformada em valores numéricos, para que os algoritmos de aprendizado de máquina pudessem ter um melhor desempenho. Os status CONCLUIDO, ATIVO – FORMANDO e ATIVO – CONCLUINTE assumiram o valor 0 e CANCELADO assumiu o valor 1. Com relação as variáveis 3.1.1.A, 3.1.1.B, 3.1.1.C, 3.1.1.D, 3.1.1.E, 3.1.1.F, 3.1.1.G, 3.1.1.H e 3.1.1.I, o valor "X" assumiu o valor 1 e os valores nulos assumiram o valor 0. Já a variável QUANTIDADE_TRANCAMENTOS passou a ter os valores nulos representados por 0.

Concluindo as modificações mencionadas acima, surgiu a necessidade de abordar a presença de diversos registros para uma mesma matrícula, correspondendo a diferentes autoavaliações para cada disciplina por período. Considerando que a predição é executada para cada entrada no conjunto de dados, o modelo preditivo poderia deduzir situações tanto de evasão quanto de permanência para uma única matrícula. Para resolver esse desafio, uma abordagem foi adotada: uma média foi calculada para cada variável preditiva, abrangendo todas as autoavaliações associadas a uma matrícula específica. Com isso, conseguimos consolidar um único registro para cada aluno, contendo a média que encapsula todo o seu percurso acadêmico. Esse procedimento proporcionou uma base individualizada para a previsão de evasão. Adicionalmente, um passo crucial envolveu a anonimização de todas as matrículas. Essa ação foi implementada para salvaguardar a privacidade dos alunos que contribuíram para a construção do modelo de previsão.

Como resultado dessas transformações, obtivemos um subconjunto final contendo 6.138 registros. Esses registros representam a média aritmética por matrícula, proveniente dos 120.341 registros resultantes do processamento dos dados iniciais.

3.4 Modelagem

Nesta fase, foram definidas as técnicas de modelagem de dados, especificamente um conjunto de algoritmos de aprendizagem de máquina, bem como seus ajustes de parâmetros. Existem vários algoritmos capazes de realizar a tarefa de classificação supervisionada, porém seria extremamente custoso criar um modelo para cada algoritmo existente. Com base nesse contexto, foram selecionados os algoritmos SVM, RF e DT, que obtiveram os melhores resultados na RSL proposta por (dos Santos et al., 2021).

Uma SVM é um modelo muito poderoso e versátil de AM capaz de realizar classificações lineares ou não lineares, de regressão e até mesmo detecção de outliers. É um dos algoritmos de AM mais populares e está presente em diversos trabalhos. As SVMs são particularmente adequadas para a classificação de conjuntos de dados complexos, porém de pequeno ou médio porte (Gerón, 2019). Como as SVMs, as DTs são algoritmos versáteis de AM que podem executar tarefas de classificação e regressão (Gerón, 2019). Além disso, DT possuem fácil interpretação quanto às suas regras de predição (Louppe, 2014). Estas particularidades fazem desse modelo um algoritmo de aprendizado popular e muito difundido para a predição da evasão escolar (Pereira & Zambrano, 2017; Sukhbaatar et al., 2018). A RF é um modelo baseado em árvores de decisão, que lida bem com conjunto de dados de alta dimensão (Hastie et al., 2009). Este tipo de modelo é usualmente utilizado não apenas para classificação, mas também para regressão, estudo de importância, seleção de variáveis, e detecção de *outlier* (Verikas et al., 2011). Para implementar

as modelagens com os algoritmos selecionados, foi utilizada a biblioteca *Scikit-learn* para a linguagem de programação Python.

3.4.1 Seleção de atributos

Durante o processo de seleção de atributos, foram utilizados os métodos *Filter*, *Wrapper* e *Embedded* durante vários ciclos da metodologia CRISP-EDM. O método *Embedded* utilizando o algoritmo *Random Forest* foi o que apresentou os melhores resultados para todos os modelos desenvolvidos. Para exemplificar, a Figura 2 mostra a média de importância dos atributos selecionados utilizando o método *embedded* com o algoritmo *Random Forest*.

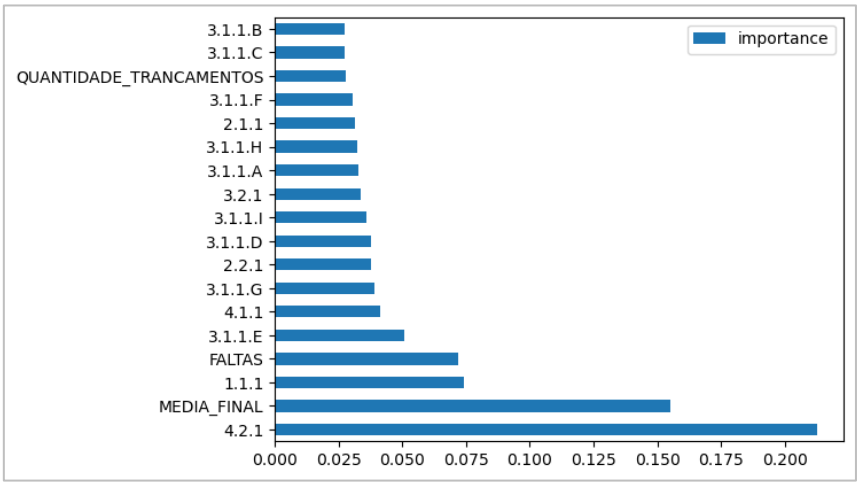

Figura 2: Média de importâncias de atributos.

Como podemos observar, o atributo de maior importância corresponde a própria manifestação do aluno com relação ao seu desejo de evadir, no entanto, apenas essa informação seria insuficiente para prever a evasão escolar, pois todos os atributos possuem algum grau de importância na tarefa de classificação. A definição dos pesos/importâncias é determinante para que o algoritmo consiga treinar o modelo de AM de forma mais eficaz. Na fase de avaliação dos resultados da metodologia é realizada a validação do modelo treinado e constatado se a seleção de atributos refletiu em um bom desempenho do modelo classificador na predição da evasão. Esse processo é cíclico e refinado constantemente como demonstra a metodologia CRISP-EDM.

3.4.2 Separação dos dados

Uma forma simples de validar a capacidade de generalização de um modelo a partir de um conjunto de dados é dividir os dados em conjuntos de treinamento e teste. Dessa forma, é possível avaliar o comportamento do modelo em dados que não foram utilizados na etapa de treinamento, estipulando a estimativa de erro do modelo no conjunto de dados desconhecidos. Esse método é denominado *Holdout*. O problema com o *Holdout* é que não é possível afirmar que o subconjunto de treinamento é representativo para o conjunto total da base de dados (Géron, 2019).

Para evitar "desperdiçar" muitos dados de treinamento em conjuntos de validação e gerar modelos representativos para todos os dados, uma técnica comum é utilizar a validação cruzada: o conjunto de treinamento é dividido em subconjuntos complementares e cada modelo é treinado com uma combinação diferente desses subconjuntos e validado em relação às partes restantes. Uma vez selecionados o tipo de modelo e os hiperparâmetros, um modelo final é treinado com a utilização desses hiperparâmetros no conjunto completo de treinamento e o erro generalizado é medido no conjunto de testes (Géron, 2019). A validação cruzada é uma técnica fundamental em aprendizado de máquina e estatística para avaliar o desempenho de um modelo e estimar sua precisão em dados não vistos. Ela ajuda a verificar como um modelo treinado pode generalizar

para novos conjuntos de dados. Existem várias abordagens de validação cruzada, sendo a validação cruzada k-fold uma das mais comuns. Nesse método, os dados são divididos em k conjuntos (ou folds) de tamanho aproximadamente igual. O modelo é treinado k vezes, cada vez utilizando k-1 folds como conjunto de treinamento e 1 fold como conjunto de validação. Esse processo é repetido até que cada fold tenha sido usado como conjunto de validação, e então as métricas de desempenho são geralmente combinadas (como a média) para fornecer uma estimativa geral do desempenho do modelo. A validação cruzada ajuda a mitigar problemas como overfitting, onde um modelo se ajusta muito aos dados de treinamento específicos e tem dificuldades em generalizar para novos dados. Além disso, ela ajuda a reduzir o viés na estimativa do desempenho do modelo, já que diferentes partições dos dados são usadas tanto para treinamento quanto para validação. Uma ótima alternativa de divisão de subconjuntos é definir 10 para o valor de k (GÉRON, 2019).

Para análise dos modelos desenvolvidos, foi utilizado a técnica *Holdout* no conjunto de dados, pois, a partir do recorte da avaliação é possível entender como o modelo se comporta de forma mais controlada. Para essa análise, os dados foram separados em um subconjunto aleatório de teste com 30% dos registros (1.842) e um de treinamento com 70% dos registros (4.296). Dos 4.296 exemplos do subconjunto de dados de treinamento, 3.026 correspondem a classe majoritária (CANCELADO) e 1.270 representam a classe minoritária (CONCLUIDO). Já os dados de teste são representados por 1.288 (CANCELADO) e 554 (CONCLUIDO). Para validar o poder de generalização do modelo desenvolvido, foi realizada uma validação cruzada sobre todo o conjunto de dados, dessa forma diferentes configurações dos dados de treinamento e teste foram garantidas, e uma média de todas as avaliações foi realizada.

3.4.3 Desbalanceamento dos dados

Um ponto de atenção durante a tarefa de classificação foi a verificação dos dados de treinamento quanto ao seu desbalanceamento. Existem mais registros de alunos evadidos (3.026) do que alunos formados (1.270). A maioria dos algoritmos de AM assume que os seus conjuntos de dados de treino estão balanceados, isto é, que o volume de amostras está distribuído de igual forma por cada categoria que está a ser analisada. No entanto, isto nem sempre acontece no mundo real, ou seja, pode acontecer que o número de instâncias correspondente a uma determinada classe é muito diferente do número de instâncias correspondente a outra classe. Neste caso, estamos perante um problema de desequilíbrio de classes, o que é algo bastante comum e constitui por vezes um obstáculo para a obtenção de bons índices de classificação por parte dos algoritmos (BATISTA *et al*., 2004).

Neste trabalho, avaliamos diferentes métodos de subamostragem e sobreamostragem para balancear a distribuição de classes nos dados de treinamento. Dois desses métodos, sobreamostragem aleatória e subamostragem aleatória, são métodos não heurísticos que foram inicialmente incluídos nesta avaliação como métodos de linha de base. Diante desse contexto, foram implementados modelos com os dados de treinamento balanceados através das técnicas de subamostragem (*undersampling*) e sobreamostragem (*oversampling*) utilizando a biblioteca *Imbalanced-learn* e modelos utilizando os dados desbalanceados. No caso da subamostragem, foi gerada uma subamostra aleatória (*Random Undersampling*) da classe de maior frequência e foi mantida a classe de menor frequência. No caso da sobre amostragem, foi gerada uma sobre amostra replicando aleatoriamente os registros da classe de menor frequência (*Random Oversampling*), também foi gerada duas outras sobreamostras da classe de menor frequência utilizando as técnicas SMOTE e ADASYN. Com esse cenário, foi possível avaliar como os algoritmos se comportavam e qual abordagem seria mais coerente para o objetivo da pesquisa, que é detectar os alunos com maior risco de evasão escolar.

4 Resultados e Discussões

Para os testes que foram efetuados neste trabalho, recorrendo aos diferentes algoritmos de AM, foram utilizadas diferentes métricas de avaliação de resultados. Quando se analisam os resultados de um algoritmo, em especial relacionados com sistemas de detecção de evasão escolar, todas as métricas estão relacionadas com o número de previsões. Essas previsões são também habitualmente representadas sob a forma de uma matriz denominada de matriz de confusão (Figura 3), onde o "Sim" seria o CANCELADO (evadiu do curso) e o "Não" CONCLUIDO (não evadiu do curso).

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 3: Matriz de confusão.

A acurácia é uma métrica calculada a partir da divisão entre o número de previsões verdadeiras (VP e VN) e o número total de previsões. A acurácia é mais aconselhada para conjuntos de dados balanceados (JOSHI, 2020).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

O valor de Precisão é calculado a partir da divisão entre as previsões positivas verdadeiras (VP) e o total de previsões positivas (VP + FP). Um resultado elevado, indica a presença de um valor baixo de previsões positivas falsas (FP) (JOSHI, 2020).

$$Precisão = \frac{VP}{VP + FP} \quad (2)$$

A métrica *Recall* é calculada a partir da divisão das previsões positivas verdadeiras (VP) e a soma das previsões positivas verdadeiras (VP) e as previsões negativas falsas (FN). Analisando as possíveis variações de resultados, é possível verificar que um alto valor de *Recall* indica a presença de um baixo número de FN (JOSHI, 2020).

$$Recall = \frac{VP}{VP + FN} \quad (3)$$

O *F-Measure* é um valor ponderado que resulta da divisão do dobro da multiplicação entre a *Recall* e a Precisão com a soma destas duas métricas (JOSHI, 2020).

$$F - Measure = \frac{2 \times (Recall \times Precisão)}{Recall + Precisão} \quad (4)$$

4.1.1 Dados desbalanceados

O primeiro modelo desenvolvido utilizou os dados de treinamento desbalanceados, dessa forma foi possível observar como os algoritmos de AM se comportavam quanto ao desbalanceamento.

Tabela 5: Métricas com dados desbalanceados utilizando um conjunto de dados aleatório.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.82193268	0.87209302	0.87344720	0.87276958
Floresta Aleatória	**0.88599348**	**0.89574155**	**0.94720496**	**0.92075471**
Máquinas de Vetores de Suporte	0.83713355	0.88413685	0.88276397	0.88344988

Como podemos observar na Tabela 4, o algoritmo Floresta Aleatória obteve os melhores resultados em todas as métricas, utilizando a técnica Holdout para um único conjunto de dados de treinamento e teste escolhido de forma aleatória. Para entender melhor as métricas apresentadas, a Figura 4 apresenta as matrizes de confusão obtidas nos modelos utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.

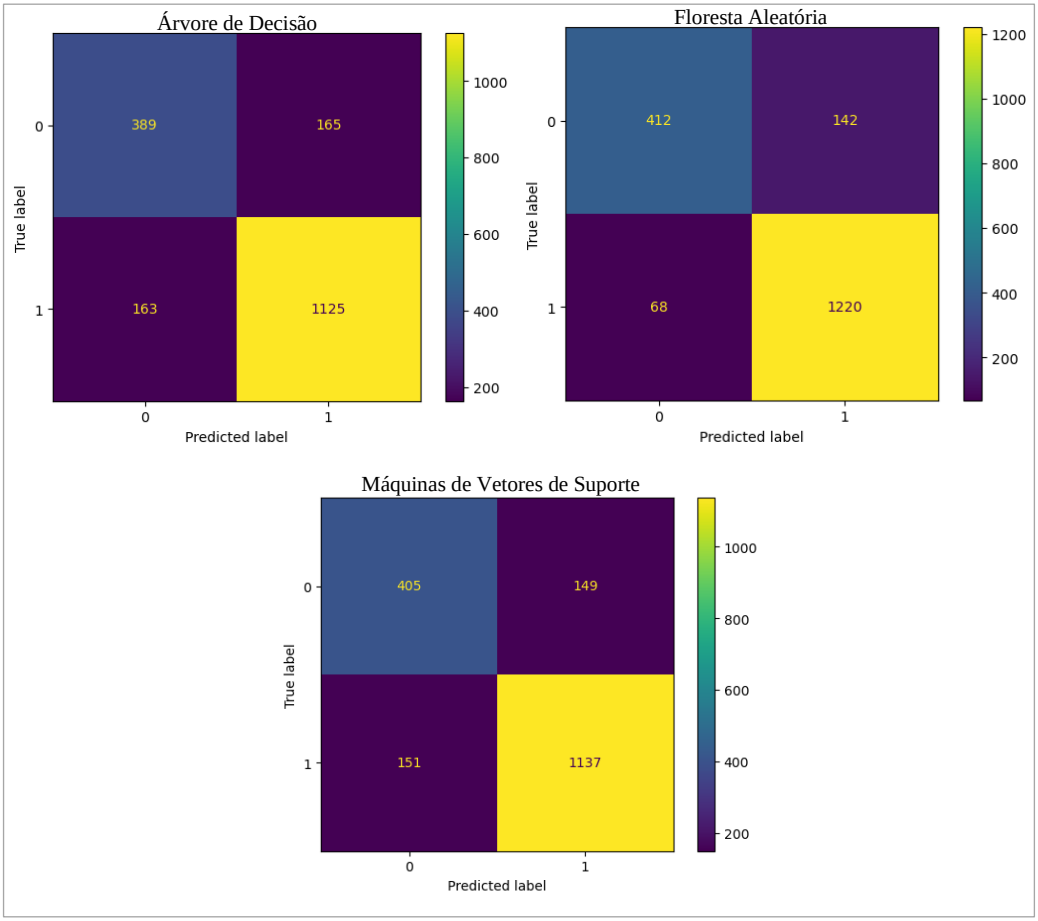

Figura 4: Matrizes de confusão dos modelos desbalanceados utilizando um conjunto de dados aleatório

Podemos observar no algoritmo Floresta Aleatória, que em um universo de teste com dados reais de 1.288 alunos evadidos (CANCELADO), o modelo obteve apenas 68 Falsos Negativos (FN), atingindo um *recall* de 94,72%. O modelo de Árvore de Decisão classificou 163 como FN e apresentou um aumento de Falsos Positivos (FP), com 165 casos. Já o modelo SVM teve melhor desempenho que o modelo de Árvore de Decisão, com 151 FN e 149 FP.

Para uma validação mais precisa, foi aplicada a técnica de validação cruzada com a utilização de 10 dobras para o algoritmo *StratifiedKFold*. Neste método os dados são divididos em k=10 subconjuntos. Em seguida, o método *holdout* é repetido k vezes, de tal forma que, a cada vez, um dos k subconjuntos é usado como set de validação e os outros subconjuntos k-1 são colocados juntos para formar um set de treinamento. A média, dos k resultados das avaliações realizadas, pode ser observada na Tabela 5, sendo o algoritmo Floresta Aleatória o que obteve os melhores resultados em todas as métricas.

Tabela 6: Média dos resultados das métricas com dados de treinamento desbalanceado utilizando a técnica de validação cruzada.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.80448055	0.86616020	0.85489494	0.86028856
Floresta Aleatória	**0.87827977**	**0.90075786**	**0.93139125**	**0.91542677**
Máquinas de Vetores de Suporte	0.83071188	0.88979089	0.86647439	0.87783455

Apesar dos resultados satisfatórios, no qual foi possível obter um *Recall* de 93,13% em média, devemos nos atentar para a especialização do modelo quanto a classe majoritária (CANCELADO). Como podemos observar, o modelo apresenta uma quantidade significativa de FP relativo à classe minoritária, dessa forma se tivéssemos um maior número de caso de testes para a classe minoritária (CONCLUIDO), a tendência é que a métrica Precisão diminuísse proporcionalmente, diminuindo também a *F-Measure*. Diante do contexto, apesar de termos um modelo com baixo número de FN, poderíamos ter outro problema, que é ter um número significativo de alunos sendo classificados como com potencial de evasão escolar, gerando uma perda significativa de recursos da instituição ao concentrar esforços dos *stakeholders* na mitigação de evasão escolar em casos FP. A avaliação por validação cruzada aplicada nesse modelo é a mesma aplicada para os demais modelos e é a métrica utilizada para a tomada de decisão quanto ao modelo a ser proposto.

4.1.2 Dados balanceados: subamostragem aleatória

Nesta tarefa de classificação, foi utilizado os dados de treinamento balanceados através do método de subamostragem aleatória, ou seja, um subconjunto foi selecionado aleatoriamente a partir da classe de maior frequência (CANCELADO). Dessa forma, a classe majoritária passou a ter 1.270 exemplos, igualmente a classe minoritária. A Tabela 6 traz a avaliação dos modelos desenvolvidos com dados de treinamento balanceados pelo método de subamostragem aleatória, o modelo que apresentou o melhor resultado, utilizando a técnica *Holdout* para um conjunto de dados de treinamento (70%) e teste (30%) aleatório (Figura 5), foi o Floresta Aleatória.

Tabela 7: Métricas com dados balanceados por subamostragem aleatória utilizando um conjunto aleatório.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.76981541	0.89059674	0.76475155	0.82289055
Floresta Aleatória	**0.87133550**	**0.93974895**	**0.87189440**	**0.90455094**
Máquinas de Vetores de Suporte	0.82138979	0.93159315	0.80357142	0.86285952

Para uma melhor avaliação dos modelos, foi aplicada a técnica de validação cruzada para os modelos com os dados de treinamento balanceados por subamostragem aleatória, a Tabela 7 apresenta a média das avaliações.

Tabela 8: Média dos resultados das métricas com dados de treinamento balanceados por subamostragem aleatória utilizando a técnica de validação cruzada.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.79339979	0.90131798	0.78628351	0.84267739
Floresta Aleatória	**0.86068648**	**0.93532036**	**0.86417354**	**0.89483536**
Máquinas de Vetores de Suporte	0.80870923	0.93459041	0.77954960	0.85006802

Analisando as métricas obtidas, verifica-se praticamente um inversão de resultados entre as métricas *Recall* e Precisão, esse fenômeno pode ser explicado pelo mesmo motivo da avaliação com os dados de treinamento desbalanceados, ou seja, devido a perda de dados na classe majoritária, o algoritmo de AM passou a ser mais tendencioso a classificar as sobreposições de classes como sendo da classe minoritária, o contrário do caso anterior que devido ao *overfiting* da classe majoritária classificava as sobreposições tendiam a ser classificadas como a classe majoritária.

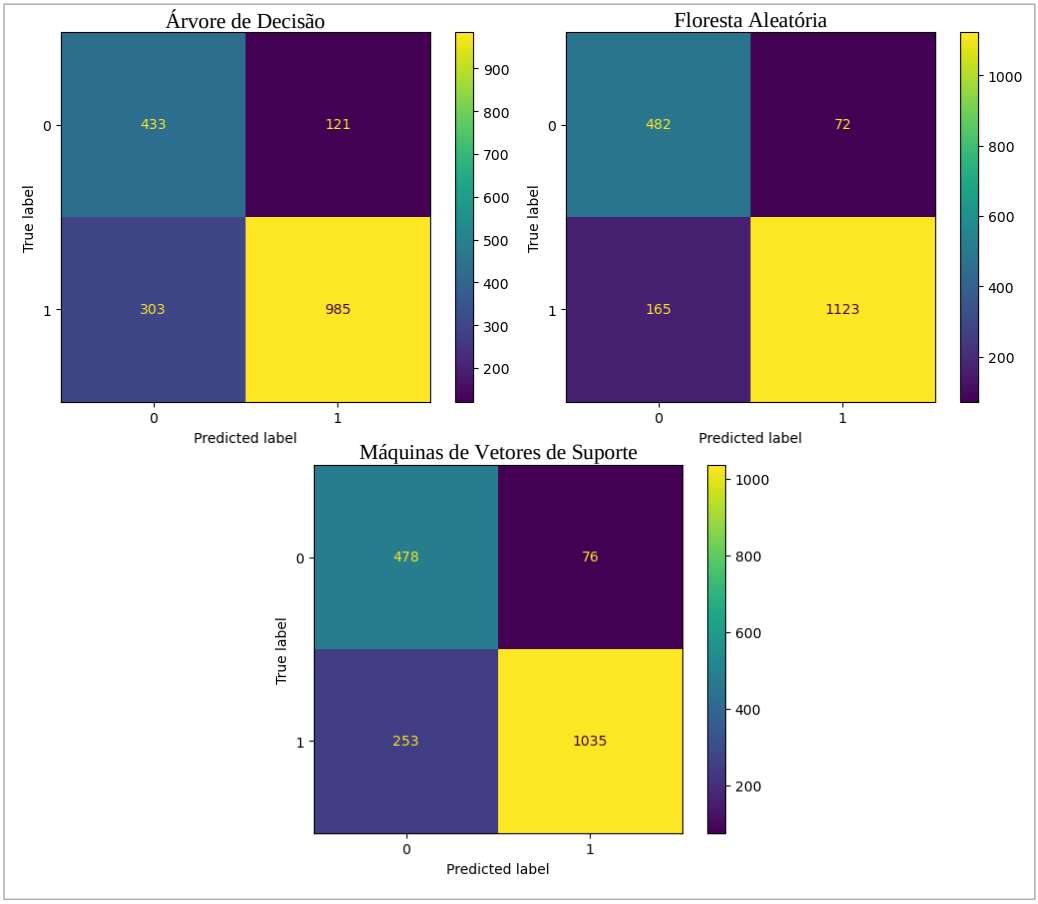

Figura 5: Matrizes de confusão dos modelos balanceados por subamostragem aleatória utilizando um conjunto de dados aleatório.

Outros métodos de subamostragem como o *Edited Nearest Neighbours* (ENN) foram testados, no entanto, apresentou desempenho abaixo da subamostragem aleatória e ocorreu os mesmos problemas de classificação quanto as sobreposições de classe. Diante desse contexto, optou-se por verificar os dados balanceados através dos métodos de sobreamostragem.

4.1.3 Dados balanceados: sobreamostragem SMOTE

Utilizando a técnica de sobreamostragem SMOTE (*Synthetic Minority Oversampling Technique*), que gera novos exemplos da classe minoritária através de interpolação entre os pontos mais próximos, fez com o modelo obtivesse uma melhor performance quanto aos outros modelos. Na matriz de confusão apresentada na Figura 6, é possível observar que o número de FP teve uma melhora significativa com relação ao modelo desbalanceado e o número de FN teve uma piora relativa, tornando esse modelo, com o algoritmo Floresta Aleatória, o mais equilibrado entre os modelos analisados.

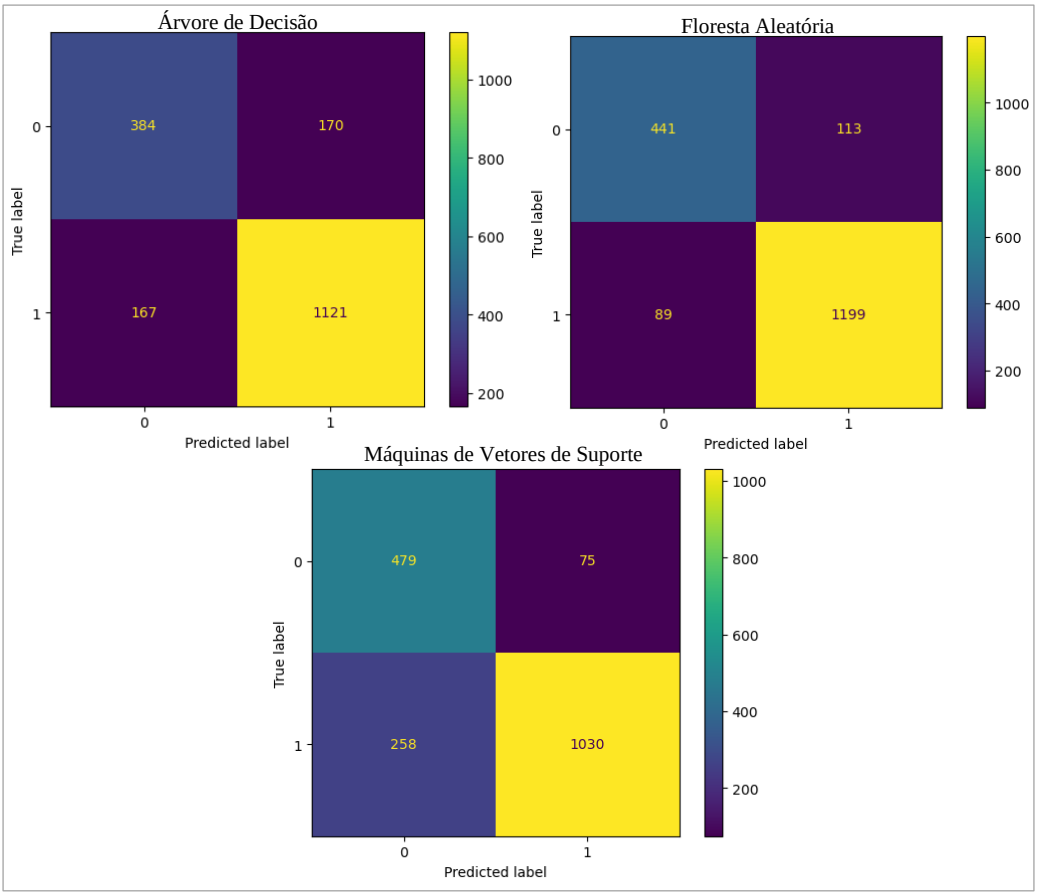

Figura 6: Matrizes de confusão dos modelos balanceados por sobreamostragem SMOTE utilizando um conjunto de dados aleatório.

Ao analisarmos os resultados obtidos (Tabela 8), observamos que esse equilíbrio se reflete nas métricas *Recall* e Precisão, refletindo em uma equiparação das métricas quanto a classificação das sobreposições de classes que reflete em modelo mais assertivo na predição da evasão escolar.

Tabela 9: Métricas com dados balanceados por sobreamostragem SMOTE utilizando um conjunto de dados aleatório.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.81704668	0.86831913	0.87034161	0.86932919
Floresta Aleatória	**0.89033659**	0.91387195	**0.93090062**	**0.92230769**
Máquinas de Vetores de Suporte	0.81921824	**0.93212669**	0.79968944	0.86084412

Para uma melhor avaliação dos modelos, foi aplicada a técnica de validação cruzada para os modelos com os dados de treinamento balanceados por sobreamostragem SMOTE, a Tabela 9 apresenta a média das avaliações.

Tabela 10: Média dos resultados das métricas com dados de treinamento balanceados por sobreamostragem SMOTE utilizando a técnica de validação cruzada.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.80578029	0.88209109	0.84863098	0.86641259
Floresta Aleatória	**0.87974371**	0.91724143	**0.91679234**	**0.91574044**
Máquinas de Vetores de Suporte	0.80968616	**0.93308219**	0.78580819	0.85346256

Também foram utilizados os métodos de sobreamostragem aleatória e ADASYN, no entanto, o método SMOTE foi o que apresentou os melhores resultados. O modelo Floresta Aleatória com os dados de treinamento balanceados pela técnica de sobreamostragem SMOTE também foi o modelo escolhido por esta pesquisa como o proposto para tentar mitigar o fenômeno

da evasão escolar na UFPB. A seguir na Tabela 10 é possível ver um quadro resumo com todos os modelos desenvolvidos e analisados.

Tabela 11: Quadro resumo com todos os modelos desenvolvidos e analisados com base na técnica de validação cruzada.

	Acurácia	Precisão	*Recall*	*F-Measure*
Árvore de Decisão Desbalanceado	0.80448055	0.86616020	0.85489494	0.86028856
Árvore de Decisão Balanceado por Subamostragem	0.79339979	0.90131798	0.78628351	0.84267739
Árvore de Decisão Balanceado por Sobreamostragem SMOTE	0.80578029	0.88209109	0.84863098	0.86641259
Floresta Aleatória Desbalanceado	0.87827977	0.90075786	**0.93139125**	0.91542677
Floresta Aleatória Balanceado por Subamostragem	0.86068648	**0.93532036**	0.86417354	0.89483536
Floresta Aleatória Balanceado por Sobreamostragem SMOTE	**0.87974371**	0.91724143	0.91679234	**0.91574044**
Máquinas de Vetores de Suporte Desbalanceado	0.83071188	0.88979089	0.86647439	0.87783455
Máquinas de vetores de suporte Balanceado por Subamostragem	0.80870923	0.93459041	0.77954960	0.85006802
Máquinas de vetores de suporte Balanceado por Sobreamostragem SMOTE	0.80968616	0.93308219	0.78580819	0.85346256

O modelo proposto foi aplicado para a base de dados de alunos ativos. Foram utilizados os 555.399 registros referentes aos status não finalizadores que não são utilizados na modelagem. Após a aplicação das mesmas etapas realizadas na modelagem, foi obtido 525.694 registros tratados. Foi realizada a predição de evasão sobre os dados da média dos registros tratados, resultando em 22.560 médias de autoavaliações para cada matrícula distinta, utilizando o modelo com o algoritmo Floresta Aleatória com dados de treinamento balanceados por sobreamostragem SMOTE.

Das 22.560 matrículas ativas desde o ano de 2017, o modelo classificou 13.310 como CANCELADO e 9.250 como CONCLUIDO, ou seja, para o modelo preditor, 59% dos alunos têm potencial de evasão escolar. Se compararmos com os indicadores de trajetória em cursos presenciais do ensino superior no estado da Paraíba (Figura 7), presente no Mapa do Ensino Superior no Brasil – 13ª Edição (2023), temos uma porcentagem compatível com o cenário atual no estado.

Figura 7: Indicadores de trajetória em cursos presenciais na Paraíba.

Com base nesses dados, é possível afirmar que o modelo proposto tem potencial para ser utilizado em implementações de soluções educacionais que auxiliem os *stakeholders* na tomada de decisão que resulte em intervenções para melhoria do processo educacional e mitigação da evasão escolar, antecipando problemas antes que se tornem irreversíveis. Ao identificar estudantes em risco de evasão com base em indicadores do modelo de predição desenvolvido, as instituições podem intervir prontamente e oferecer apoio personalizado. Isso pode incluir tutorias adicionais, aconselhamento acadêmico ou a implementação de programas de orientação e suporte emocional. No entanto, a implementação efetiva dessas soluções extrapola o escopo desta pesquisa, pois ela precisa ser validada e proposta pela alta gestão da instituição.

5 Conclusão e Trabalhos futuros

No presente trabalho foram apresentadas e implementadas diferentes abordagens de classificação para predição de evasão escolar tendo em vista a interpretação de dados da autoavaliação dos cursos de graduação da UFPB. Com a finalidade de obter esse objetivo, foi utilizada a metodologia CRISP-EDM para guiar o trabalho de mineração de dados. Cada uma dessas abordagens se diferencia da outra tanto no balanceamento dos dados de treinamento quanto nos algoritmos de AM que as compõem. Para validar as diferentes abordagens de classificação, foram utilizadas as métricas Acurácia, Precisão, *Recall* e *F-Measure*. A partir dos resultados dessas métricas, foi proposto um método de predição baseado no algoritmo Floresta Aleatória com o balanceamento dos dados utilizando a técnica de sobreamostragem SMOTE, no qual obteve 87,97% de Acurácia, 91,72% de Precisão, 91,67 de *Recall* e 91,57 de *F-Measure*. Além disso, o método proposto foi aplicado nos dados das autoavaliações dos alunos ativos e o resultado apresentou compatibilidade com os índices de evasão escolar atuais no estado, quando consideramos a relação de ingressante e concluinte em um mesmo ano.

As principais conclusões e considerações relacionadas aos experimentos foram que é possível prever de forma genérica, com taxa de acerto satisfatória, a evasão escolar na instituição a partir dos dados de autoavaliação dos cursos de graduação. No entanto, o modelo não obteve bons resultados para os alunos que ingressaram antes de 2017, possivelmente devido à falta de inclusão de disciplinas cursadas anteriormente no instrumento de autoavaliação. Observou-se que as questões diretamente relacionadas à evasão escolar foram as mais importantes para a classificação, destacando-se a questão 4.2.1, em que o aluno expressa sua intenção de sair do curso, e a questão 1.1.1, em que o aluno avalia seu desempenho pessoal na disciplina. De acordo com o modelo preditor, dos 22.560 alunos ativos que ingressaram na instituição a partir de 2017, 59% têm probabilidade de evadir do curso, refletindo a situação atual apontada pelo Mapa do Ensino Superior no Brasil.

A partir do trabalho desenvolvido, surgem novos desafios e oportunidades que direcionam a necessidade de investigações futuras. Para isso, é importante explorar a possibilidade de identificar se as disciplinas específicas têm um impacto significativo na evasão escolar, o que permitiria a criação de modelos mais precisos e direcionados. Além disso, é importante investigar se modelos específicos por curso podem superar o desempenho do modelo genérico proposto inicialmente. Outro aspecto a ser considerado é a incorporação de dados socioeconômicos, acadêmicos, culturais e outros provenientes de estudos anteriores, a fim de enriquecer o modelo preditor e obter resultados mais assertivos. Uma maneira de facilitar o processamento dos dados e a aplicação do modelo é criar uma API (*Application Programming Interface*), permitindo a automatização do processo e a integração com outros sistemas. Além disso, é fundamental coletar e incorporar dados atualizados de semestres posteriores ao modelo, garantindo sua melhoria contínua e precisão ao longo do tempo. Para apresentar insights relevantes e em tempo real aos stakeholders, é importante integrar a API desenvolvida com o SIGAA da instituição de ensino.

Isso fornecerá informações valiosas para tomar decisões estratégicas e implementar ações específicas para mitigar a evasão escolar. Por fim, o desenvolvimento de dashboards interativos que apresentem as predições de evasão escolar para cada semestre permitirá aos stakeholders visualizar tendências e padrões, auxiliando-os na elaboração de políticas e melhoria de processos com base nessas informações.

Esses trabalhos futuros visam aprimorar o modelo de predição de evasão escolar, aumentando sua precisão, incorporando dados relevantes, automatizando o processamento de dados, integrando-o com sistemas existentes e fornecendo insights acionáveis para combater a evasão escolar de forma eficaz.

Referências

Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: A Systematic Literature. Indian Journal of Science and Technology, 12(4), 1-12. [GS Search]

ANDIFES, A., ABRUEM, A., & SESu/MEC, S. (1996). Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. Avaliação: Revista Da Avaliação Da Educação Superior, 1(2). Recuperado de https://periodicos.uniso.br/avaliacao/article/view/739. [GS Search]

Baggi, C. A. D. S., & Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: Revista da Avaliação da Educação Superior (Campinas), 16(02), 355-374. [GS Search]

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. Revista Brasileira de informática na educação, 19(02), 03. [GS Search]

Batista, G. E., Prati, R. C., & Monard, M. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29. DOI: https://doi.org/10.1145/1007730.1007735. [GS Search]

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34, 483-519. DOI: https://doi.org/10.1007/s10115-012-0487-8. [GS Search]

Costa, F. J., Dias, J. J. L. Avaliação da formação superior pelo discente: proposta de um instrumento. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 25, n. 2, p. 275–296, maio 2020. DOI: https://doi.org/10.1590/S1414-4077/S1414-40772020000200003. [GS Search]

dos Santos, V. H. B., Saraiva, D. V., & de Oliveira, C. T. (2021). Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In Anais do XXXII Simpósio Brasileiro de Informática na Educação (pp. 1196-1210). SBC. DOI: https://doi.org/10.5753/sbie.2021.218167. [GS Search]

Gamba, E., & Righetti, S. (2022). Em crise, universidades federais participam de mais da metade da produção científica. Folha de São Paulo. Recuperado de https://www1.folha.uol.com.br/educacao/2022/12/em-crise-universidades-federais-participam-de-mais-da-metade-da-producao-cientifica.shtml

Géron, A. (2019). Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow. Alta Books. [GS Search]

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. [GS Search]

Joshi, A. V. (2020). Machine Learning and Artificial Intelligence. Springer. [GS Search]

Lottering, R., Hans, R., & Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. In 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-8). IEEE. DOI: https://doi.org/10.1109/icABCD49160.2020.9183874. [GS Search]

Louppe, G. (2014). Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502. DOI: https://doi.org/10.48550/arXiv.1407.7502. [GS Search]

Lousrhania, L. (2021). Universidades públicas lideram ranking brasileiro de patentes. Rádio Agência Nacional. Recuperado de https://agenciabrasil.ebc.com.br/radioagencia-nacional/pesquisa-e-inovacao/audio/2021-07/universidades-publicas-lideram-ranking-brasileiro-de-patentes

Manrique, R., Nunes, B. P., Marino, O., Casanova, M. A., & Nurmikko-Fuller, T. (2019). An analysis of student representation, representative features and classification algorithms to predict degree dropout. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (pp. 401-410). DOI: https://doi.org/10.1145/3303772.3303800. [GS Search]

Mapa do Ensino Superior no Brasil – 13ª Edição. Instituto Semesp, 2023. Recuperado de https://www.semesp.org.br/wp-content/uploads/2023/06/mapa-do-ensino-superior-no-brasil-2023.pdf

Pereira, R. T., & Zambrano, J. C. (2017). Application of decision trees for detection of student dropout profiles. In 2017 16th IEEE international conference on machine learning and applications (ICMLA) (pp. 528-531). IEEE. DOI: https://doi.org/10.1109/ICMLA.2017.0-107. [GS Search]

Prestes, E. M. D. T., & Fialho, M. G. D. (2018). Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. Ensaio: Avaliação e Políticas Públicas em Educação, 26, 869-889. DOI: https://doi.org/10.1590/S0104-40362018002601104. [GS Search]

Rafiq, M. A., Rabbi, A. M., & Ahammad, R. (2021, June). A data science approach to Predict the University Students at risk of semester dropout: Bangladeshi University Perspective. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1350-1354). IEEE. DOI: https://doi.org/10.1109/ICOEI51242.2021.9453067. [GS Search]

Ramos, J. L. C., Rodrigues, R. L., Silva, J. C. S., & de Oliveira, P. L. S. (2020, November). CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In Anais do XXXI Simpósio Brasileiro de Informática na Educação (pp. 1092-1101). SBC. DOI: https://doi.org/10.5753/cbie.sbie.2020.1092. [GS Search]

Saccaro, A., França, M. T. A., & Jacinto, P. D. A. (2019). Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas. Estudos Econômicos (São Paulo), 49, 337-373. DOI: https://doi.org/10.1590/0101-41614925amp. [GS Search]

Santos, C. H. D., de Lima Martins, S., & Plastino, A. (2021). É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?. In Anais do XXXII Simpósio Brasileiro de Informática

na Educação (pp. 792-802). SBC. DOI: https://doi.org/10.5753/sbie.2021.218105. [GS Search]

Saraiva, D., Pereira, S., Gallindo, E., Braga, R., & Oliveira, C. (2019, July). Uma proposta para prediçao de risco de evasao de estudantes em um curso técnico em informática. In Anais do XXVII Workshop sobre Educaçao em Computaçao (pp. 319-333). SBC. DOI: https://doi.org/10.5753/wei.2019.6639. [GS Search]

Sukhbaatar, O., Ogata, K., & Usagawa, T. (2018). Mining educational data to predict academic dropouts: a case study in blended learning course. In TENCON 2018-2018 IEEE region 10 conference (pp. 2205-2208). IEEE. DOI: https://doi.org/10.1109/TENCON.2018.8650138. [GS Search]

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. Pattern recognition, 44(2), 330-349. DOI: https://doi.org/10.1016/j.patcog.2010.08.011. [GS Search]