CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Predição de Evasão por meio de um Instrumento Sistemático de Avaliação Institucional

Ronei dos Santos Oliveira, Francisco Petrônio Alencar de Medeiros, Kamila Alves

Programa de Pós-Graduação em Tecnologia da Informação - Instituto Federal da Paraíba (IFPB) João Pessoa – PB – Brasil

{ronei.santos, petronio, alves.kamila}@ifpb.edu.br

Abstract. *This research investigated the feasibility of predicting dropout at Federal University of Paraiba through a compulsory institutional assessment instrument administered to students. Classification algorithms were applied and evaluated using performance metrics, revealing an accuracy of 87.97%, precision of 91.72%, recall of 91.67%, and an F1 score of 91.57% in identifying students prone to dropout. Approximately 59% of active students admitted to the University since 2017 showed a likelihood of abandoning their courses in the predictive model tests, approaching the consolidated dropout rates at the University between 2010 and 2018, which ranged from 51 to 60%*

Resumo. *Esta pesquisa investigou a viabilidade da predição de evasão na Universidade Federal da Paraíba (UFPB) por meio de um instrumento de avaliação institucional aplicado juntos aos discentes de forma compulsória. Algoritmos de classificação foram aplicados e 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 propensos a evadir. Cerca de 59% dos alunos ativos da Universidade admitidos a partir de 2017 demonstram probabilidade de abandonar seus cursos nos testes do modelo preditivo, aproximando-se dos valores consolidados de evasão da Universidade entre 2010 e 2018, que alcançaram de 51 a 60%.*

1. Introdução

Para uma compreensão mais clara da evasão no contexto da educação superior, uma comissão formada pelo MEC definiu diferentes categorias de evasão, a saber: evasão do curso, da instituição e evasão do sistema. Este trabalho se concentrou 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. Apesar das várias observações e concepções sobre a evasão escolar, compreender as suas razões continua sendo um desafio para a maioria das instituições [ANDIFES et al., 1996].

A avaliação 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 UFPB realiza, a cada semestre, uma avaliação compulsória, sistemática e anônima dos cursos de graduação por meio de um instrumento respondido por todos os mais de 39 mil estudantes matriculados no período de matrícula. Essa avaliação fornece uma ampla gama de dados que podem ser explorados para compreender as especificidades da instituição e é aplicado desde o ano de 2017.

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 [Colpo 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, não foi encontrado nenhum que explore o uso de dados de avaliação institucional sistemática de cursos de graduação. É mais comum encontrar o uso de dados como fatores pessoais, acadêmicos, econômicos, sociais e institucionais [Namoun & Alshanqiti, 2020].

A motivação para esse estudo reside na possibilidade de utilizar a MDE sobre o vasto volume de dados gerados pelas respostas do instrumento de avaliação sistemático da Universidade, oferecendo a detecção precoce dos alunos com maior propensão à evasão e com isso possibilitar que os stakeholders, como a alta gestão, coordenadores de cursos e professores, tomem decisões para mitigar a evasão desses alunos [Rodrigues et al., 2013].

2. Trabalhos Relacionados

O estudo conduzido por Rafiq et al. (2021) sobre evasão escolar na Universidade de Bangladesh assume um papel de destaque, oferecendo discernimentos valiosos, apesar de seu enfoque distinto deste trabalho. Os dados foram coletados da base universitária e de pesquisas estudantis, resultando na seleção de 17 variáveis abrangendo aspectos acadêmicos, socioeconômicos e pessoais. A construção do modelo de prognóstico se baseou em métodos de classificação supervisionada, notadamente os algoritmos One-vs-Rest e Floresta Aleatória (*Random Forest* – RF), implementados via biblioteca Scikit-learn em Python. A performance superior do RF, com índices de precisão de 0,9745 e AUC de 0,98, reforça sua relevância para a concepção do modelo preditivo na instituição de interesse.

Lottering et al. (2020) aplicaram estratégias de Mineração de Dados Educacionais (MDE) e Aprendizado de Máquina (AM) para a detecção de alunos com risco de evasão em um curso de graduação em universidade de tecnologia na África do Sul. O estudo empregou informações de desempenho acadêmico ao longo dos anos letivos e promoveu a seleção de 19 atributos. Diante do desequilíbrio dos dados, foi adotada uma subamostragem e, posteriormente, algoritmos de AM de classificação supervisionada, como Naive Bayes, Árvore de Decisão (Decision Tree – DT), SVM, Nearest Neighbor e RF, foram aplicados. O SVM se destacou, alcançando F-Measure de 99,32%. O estudo oferece contribuições importantes ao campo da modelagem preditiva de evasão escolar, por sua avaliação comparativa de algoritmos e enfoque em mitigar o desbalanceamento de dados.

A investigação de Santos et al. (2021) é notável pelo emprego de técnicas de MDE, especificamente classificação, na análise da previsibilidade da evasão e graduação dos alunos, utilizando apenas dados de desempenho acadêmico. A pesquisa, voltada a uma instituição brasileira de ensino superior, construiu dez modelos correspondentes a diferentes semestres acadêmicos, usando o algoritmo DT. Resultados satisfatórios, com acurácias entre 79,31% e 98,25%, revelam que a predição se torna mais precisa nos semestres posteriores, em virtude do aumento de atributos analisados. A intersecção entre o desempenho acadêmico e a probabilidade de evasão ou conclusão ao longo dos semestres é abordada, enriquecendo o escopo do modelo preditivo. Manrique et al. (2019) exploram três abordagens distintas para a predição da evasão em uma instituição de ensino superior no Brasil. As abordagens abrangem modelos globais, específicos de curso e análise temporal dos dados do aluno, utilizando diferentes algoritmos de AM. Os resultados apontam a eficácia do modelo global em conjunto com o RF, salientando sua relevância na implementação do modelo preditivo de evasão escolar.

Este trabalho se diferencia dos trabalhos relacionados principalmente em relação ao contexto dos dados utilizados. A pesquisa de predição de evasão escolar com base em dados de um instrumento sistemático de autoavaliação e avaliação de cursos de graduação pode trazer avanços práticos significativos ao proporcionar uma compreensão mais profunda dos fatores e padrões que levam à evasão na Universidade,

bem como ser replicado para outras instituições. 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.

3. Métodos

Esta pesquisa utilizou a metodologia CRISP-EDM (CRoss-Industry Standard Process for Educational Data Mining), uma adaptação da metodologia CRISP-DM para o contexto educacional [Ramos et al., 2020].

3.1. Instrumento de Avaliação e Conjunto de Dados

Na primeira etapa do processo houve uma compreensão do contexto dos dados educacionais e sua relação com a evasão. Os dados foram obtidos por meio da avaliação dos cursos de graduação, realizada pelos alunos da instituição a cada início de semestre. Essa avaliação é compulsória e necessária para a matrícula no semestre seguinte. O instrumento de avaliação possui quatro dimensões: a discente, onde os alunos avaliam seu próprio desempenho e motivação em cada disciplina; a disciplina, onde avaliam a importância e dificuldade das disciplinas; a docente, onde avaliam aspectos relacionados aos professores; e a dimensão curso, onde indicam a probabilidade de recomendar o curso e expressam possibilidade de desistir. Essa avaliação foi adotada a partir do primeiro semestre de 2017 para todos os cursos de graduação da Universidade.

Os dados analisados consideraram somente alunos matriculados a partir desse semestre, visando evitar enviesamentos de dados que prejudicariam a eficácia do modelo. Esse ajuste foi realizado ao longo do processo, com refinamentos contínuos no modelo. O conjunto dos dados oriundos do instrumento de avaliação discente é apresentado na Tabela 1.

Tabela 1: 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
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

Foram analisados 1.156.891 registros das avaliações dos cursos de graduação na modalidade presencial. A variável STATUS_DISCENTE é a variável preditiva ou alvo e assume um dos valores: ATIVO, TRANCADO, CONCLUIDO, ATIVO – FORMANDO e ATIVO – CONCLUINTE e CANCELADO. 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.

3.2. Preparação dos dados – Filtragem, remoção de *outliers* e transformações

Realizou-se uma filtragem dos dados de acordo com o ano de matrícula dos alunos, selecionando apenas as avaliações daqueles que ingressaram a partir do primeiro semestre de 2017, garantindo a representação completa de seus ciclos acadêmicos até a extração dos dados. A exclusão das matrículas anteriores a 2017 se deu devido à ausência das avaliações para disciplinas cursadas antes da implementação da avaliação sistemática. Isso resultou em um novo conjunto de dados brutos com 683.634 registros. Além disso, uma filtragem adicional foi realizada na variável alvo "STATUS_DISCENTE", considerando apenas os valores que indicam conclusão ou evasão, excluindo os estados ativos e trancados que não permitem a determinação final. Com essa seleção, o conjunto de dados brutos foi reduzido para 128.235 registros.

Algumas variáveis demográficas foram descartadas por não se encaixarem no contexto da pesquisa, bem como foram removidos *outliers* gerados no processo de migração dos dados entre servidores, visto que as perguntas do instrumento são fechadas, o que impossibilitaria que o estudante respondesse fora do intervalo esperado para a variável. Finalizado a etapa de remoção de *outliers*, o conjunto de dados ficou com 120.341 registros. A variável alvo STATUS_DISCENTE foi transformada em valores numéricos com status 1 para CANCELADO e 0 para alunos que concluíram ou são concluintes. Por fim, obteve-se um registro único para cada aluno com a média que representa toda sua vida acadêmica, passando a ter um registro atômico para realizar a predição de evasão. As transformações resultaram em um subconjunto final com 6.138 registros, que representa a média dos 120.341 registros.

3.3. Modelagem – Seleção de atributos, separação e balanceamento dos dados

Utilizou-se os algoritmos de classificação Máquina de Vetores de Suporte (*Support Vector Machine* – SVM), Floresta Aleatória (*Random Forest* – RF) e Árvore de Decisão (Decision Tree – DT) com base na RSL proposta por proposta por dos Santos et al., (2021). As Support Vector Machines (SVMs) são poderosas e versáteis no aprendizado de máquina, capazes de lidar com classificações lineares ou não lineares, regressão e detecção de outliers [Gerón, 2019]. Da mesma forma, Decision Trees (DTs) também são algoritmos versáteis que realizam classificação e regressão, e se destacam pela interpretação intuitiva de suas regras de predição [Louppe, 2014]. Esses modelos são amplamente adotados na previsão de evasão escolar [Pereira & Zambrano, 2017; Sukhbaatar et al., 2018]. Random Forest (RF), por outro lado, é um modelo baseado em árvores de decisão adequado para conjuntos de dados de alta dimensão e amplamente utilizado em tarefas de classificação, regressão, estudo de importância de variáveis, seleção de variáveis e detecção de outliers [Verikas et al., 2011].

Durante o processo de seleção de atributos, foram utilizados os métodos *Filter*, *Wrapper* e *Embedded* durante 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. O atributo de maior importância corresponde a própria manifestação do aluno em sair do curso (4.2.1), no entanto, apenas essa informação seria insuficiente para prever a evasão escolar, pois (18) dezoito atributos apresentaram graus elevados de importância na classificação, em destaque para média final, autoavaliação em relação ao comprometimento e motivação na disciplina (1.1.1) e faltas. A definição dos pesos/importâncias é determinante para que o algoritmo consiga treinar o modelo de AM de forma mais eficiente e eficaz. Na fase de avaliação dos resultados da metodologia é realizada a validação do modelo treinado e constatado que a seleção de atributos refletiu em um bom desempenho do modelo classificador na predição da evasão [Oliveira et al., 2022].

Durante a classificação, um desafio foi o desequilíbrio nos dados de treinamento (4296 registros), com mais registros de alunos evadidos (3.026) do que alunos concluídos (1.270). Muitos algoritmos de Aprendizado de Máquina presumem equilíbrio nos conjuntos de treinamento, o que nem sempre ocorre na realidade, resultando em desequilíbrio de classes, prejudicando o desempenho [Batista et al., 2004]. Neste trabalho, avaliamos diferentes métodos de subamostragem e sobreamostragem para balancear a distribuição de classes nos dados de treinamento. 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 sobreamostragem, foi gerada uma sobreamostra 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 era detectar os alunos com maior risco de evasão.

4. Resultados e Discussões

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. Como se pode observar na Tabela 2, 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.

Tabela 2: Métricas com dados desbalanceados

	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

Com o algoritmo Floresta Aleatória, 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 foram divididos em 10 subconjuntos. Em seguida, o método *holdout* foi repetido k vezes, de tal forma que, a cada vez, um dos k subconjuntos fosse usado como set de validação e os outros subconjuntos k-1 fossem colocados juntos para formar um set de treinamento. A média dos k resultados das avaliações realizadas, conforme Tabela 3, mostra que o algoritmo Floresta Aleatória obteve os melhores resultados em todas as métricas.

Tabela 3: 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, foi necessário atentar-se para a especialização do modelo quanto a classe majoritária (CANCELADO). 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 do modelo com baixo número de FN, poderia 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 foi a mesma aplicada para os demais modelos e é a métrica utilizada para a tomada de decisão quanto ao modelo a ser proposto.

Para dados balanceados com subamostragem aleatória, 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. O modelo que apresentou o melhor resultado utilizando a técnica *Holdout* para um conjunto de dados de treinamento (70%) e teste (30%) aleatório foi o Floresta Aleatória, conforme mostra a Tabela 4.

Tabela 4: 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 5 apresenta a média das avaliações.

Tabela 5: 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

Ao diminuir a quantidade de exemplos da classe majoritária, ocorreu uma piora no desempenho do modelo em relação à quantidade de FN. Isso pode ter ocorrido devido à perda de dados relevantes para o treinamento do algoritmo, o que resultou na dificuldade do algoritmo em classificar adequadamente a sobreposição de classes. Analisando as métricas obtidas, verifica-se praticamente um diferença 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.

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 em relação aos outros modelos. Analisando os resultados apresentados na Tabela 6, observa-se um equilíbrio nas métricas *Recall* e Precisão, refletindo uma equiparação das métricas quanto a classificação das sobreposições de classes que reflete em um modelo mais assertivo na predição da evasão escolar.

Tabela 6: 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 7 apresenta a média das avaliações.

Tabela 7: 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. Com base nos resultados das avaliações dos modelos nas diferentes configurações, o modelo Floresta Aleatória com os dados de treinamento balanceados pela técnica de sobreamostragem

SMOTE foi o modelo escolhido por esta pesquisa como o proposto para tentar mitigar o fenômeno da evasão na UFPB.

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 avaliações 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 a proporção de ingressantes e concluintes no estado da Paraíba, Mapa do Ensino Superior no Brasil – 11ª Edição (2021), temos uma porcentagem compatível com o cenário atual no estado. Além disso, quando comparado com as taxas consolidadas de evasão da UFPB entre os anos de 2010 a 2018, que alcançaram valores entre 51 e 60%, também mostra uma aproximação com os resultados alcançados no modelo de predição, tornando o modelo desenvolvido, portanto, um instrumento poderoso para individualizar ações no sentido de mitigar a evasão na Universidade.

5. Conclusão

É possível afirmar que o modelo proposto pode ser utilizado para implementação de soluções educacionais que auxiliem os *stakeholders* na tomada de decisão que resultem em intervenções para melhoria do processo educacional e mitigação da evasão, 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. A implementação efetiva dessas soluções extrapola o escopo desta pesquisa, pois ela precisa ser validada e proposta pela alta gestão da instituição, processo iniciado a partir da apresentação dos resultados à Pró Reitoria de Ensino da Universidade.

Foram implementadas e validadas diferentes abordagens de classificação para predição de evasão tendo em vista a interpretação de dados da avaliação dos cursos de graduação da UFPB. Com a finalidade de obter esse objetivo, foi utilizado a metodologia CRISP-EDM para guiar o trabalho de mineração de dados. Cada uma das 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*.

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, 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. Por fim, o desenvolvimento de *dashboards* interativos que apresentem as predições de evasão escolar para cada semestre permitirá aos *stakeholders* visualizarem tendências e padrões, auxiliando-os na elaboração de políticas e melhoria de processos com base nessas informações.

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.

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.

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.

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. Revista Brasileira de Informática na Educação, 19(02), 03.

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.

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.

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.

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.

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.

Joshi, A. V. (2020). Machine Learning and Artificial Intelligence. Springer.

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.

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

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 – 11ª Edição. (2021). Instituto Semesp. Recuperado de https://www.semesp.org.br/wp-content/uploads/2021/06/Mapa-do-Ensino-Superior-Completo.pdf

Oliveira, I. S., Medeiros, F. P. A., & Andrade, F. G. (2022). Seleção de Atributos para Classificadores de Evasão Escolar com Dados da Plataforma Nilo Peçanha. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil* (pp. 30-39). SBC. DOI: https://doi.org/10.5753/wapla.2022.226769

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.

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.

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.

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.

Rodrigues, R. L., Medeiros, F. P., & Gomes, A. S. (2013). Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In *Brazilian symposium on computers in education (Simpósio Brasileiro de Informática na Educação)* (Vol. 24, No. 1, p. 607).

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.

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.

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 Educação em Computação (pp. 319-333). SBC. DOI: https://doi.org/10.5753/wei.2019.6639.

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.

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.