Disciplinarum Scientia. Série: Naturais e Tecnológicas, Santa Maria, v. 22, n. 2, p. 105-118, 2021.
Recebido em: 25.05.2021. Aprovado em: 18.03.2021.
ISSN 2176-462X
DOI: 10.37779/nt.v22i2.3889

CLASSIFICADOR PARA PREDIÇÃO DE ALUNO EVASOR DE GRADUAÇÃO EM UNIVERSIDADES PARTICULARES[1]

CLASSIFIER FOR PREDICTING DROPOUT OF GRADUATE STUDENT IN PRIVATE UNIVERSITIES

Mirkos Ortiz Martins[2] e Crhistopher Lenhard[3]

RESUMO

A evasão no ensino superior é um problema de difícil identificação, pois envolve um grande conjunto de características que a define. Esse trabalho teve como objetivo construir um sistema computacional, baseado em *machine learning*, para a identificação das características mais relevantes e consequente classificação dos alunos entre evasores e não evasores. Para isso foi utilizada a linguagem Python, juntamente com as bibliotecas Pandas, Numpy e Scikit-learn, resultando em uma implementação que comparou as arquiteturas árvore de decisão, floresta aleatória e árvores extra, demonstrando que a última obtém melhores resultados (92.84%) na acurácia da identificação do aluno evasor, em uma base de dados com informações no período de 10 anos na UFN.

Palavras-chave: árvore de decisão,ciência de dados, inteligência artificial.

ABSTRACT

Higher education dropout is a problem that is difficult to identify, because it involves a large set of characteristics that define it. This paper aimed to build a computational system, based on machine learning, for the identification of the most relevant characteristics and the consequent classification of students between evaders and non-evaders. For this, the Python language was used, together with the Pandas, Numpy and Scikit-learn libraries, resulting in an implementation that compared the decision tree, random forest and extra trees architectures, demonstrating that the latter obtains better results (92.84%) in accuracy of the identification of the evasive student, in a database with information from the 10-year period at UFN.

Keywords: *decision tree, data science, artificial intelligence.*

1 Trabalho Final de Graduação em Ciência da Computação - UFN

2 Professor Orientador. mirkos@ufn.edu.br

3 Aluno do curso de Ciência da Computação - UFN. crhistopher.lenhard@unifra.edu.br

INTRODUÇÃO

As instituições de Ensino Superior (IES), particulares e públicas enfrentam um problema comum, a evasão escolar. Estudos e análises são elaborados tendo o objetivo de diminuir a desistência de alunos evasores, e esses estudos se dividem em duas frentes: O desenvolvimento de campanhas para manter os alunos que possuem a probabilidade de evadir; E encontrar os padrões de perfil que alunos evasores possuem para melhorar foco das campanhas (TINTO, 1999).

Na avaliação da provável desistência de um aluno, é preciso considerar e encarar de forma mais individualizada o sucesso acadêmico, medido na forma de andamento do curso, objetivando atingir métricas mais concretas na interação IES - aluno e assim definir um formato mais dinâmico para o acompanhamento de sua vida acadêmica possibilitando a preempção de provável desistência do respectivo curso (SILVA *et al.*, 2007). Também é um complicador desse contexto de desistência, o perfil de cada aluno confrontado com as características particulares de cada curso, onde o casamento entre os diferentes atores nem sempre é possível de um sucesso (DEKKER *et al.*, 2009).

No contexto de solução desse problema, surge a implementação de uma ferramenta de *data science* e *machine learning* (GRUSS, 2009) para prever um provável evasor, antes que ele realmente desista do curso, assim auxiliando a gestão da IES o desenvolvimento de estratégias para retensão desses alunos. Um passo importante para a implantação de um sistema indicador de evasão é a identificação das regras que regem o comportamento de desistência do aluno, analisados sob a ótica de dados armazenados em informações acadêmicas.

A evasão discente é uma mazela presente em todas as IES, sendo ela privada ou pública. Essas IES buscam estratégias de manutenção do número de alunos matriculados, diminuição do fechamento de turmas ou mesmo turmas com baixo número de alunos. Com isso, um sistema de preempção (previsão) de evasão discente possui extrema importância para a manutenção da estrutura funcional e econômica da IES.

Nesse contexto, esse trabalho levantou uma série de regras de identificação de dados para serem utilizados como entrada em uma ferramenta de *data science*, estado da arte na computação científica, para a descoberta de padrões nos diversos dados em uma base de informações acadêmicas.

REVISÃO TEÓRICA

Nesta seção abordam-se os conceitos e tecnologias relacionadas ao desenvolvimento deste trabalho.

EVASÃO ESCOLAR

Um problema presente em todas as instituições de ensino é a da evasão escolar, que afeta desde o ensino fundamental até o superior e constroem um desperdício social, acadêmico e econômico. Nas IES a evasão é evidente internacionalmente, tanto no setor público quanto no privado. No setor público se têm gastos públicos sem obter a resposta esperada, já no setor privado se apresenta como uma perda de lucro (SILVA FILHO *et al.*, 2007).

No Brasil, o índice de evasão nas IES privadas aproximou-se de 53% e de 33% nas IES públicas de acordo com um estudo realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) no ano de 2006 (SAVIAN *et al.* 2018). No ano de 2010 segundo dados do Censo da Educação Superior, mostra que 53% dos alunos que ingressaram em IES privadas desistiram no decorrer do curso, nas IES públicas a evasão chegou a 47% nas municipais, 38% nas estaduais e 43% nas federais (OLIVEIRA *et al.*, 2019).

A procura das causas da evasão é o tema recorrente em muitos trabalhos e pesquisas educacionais, e de acordo com (TINTO, 1999), as IES que adotam um programa para obter uma redução na evasão escolar utilizam de dados ordinários da vida institucional dos alunos. Ainda para (TINTO, 1999), a evasão é algo a ser levado a sério, onde que a retenção dos alunos devem ser unânime e o primeiro ano de faculdade deve ser um ano de inclusão que "promova o ideal importante de que todas as pessoas possam e devem ter voz na construção do conhecimento".

INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é um campo de estudo da computação que abrange uma enorme variedade de subcampos, onde tenta não somente compreender entidades inteligentes, mas também contruí-las. Suas definições referem-se à processos de pensamento e raciocínio ou de comportamento que buscam pensar e agir como humanos ou racionalmente (NORVIG, 2013).

O aprendizado de máquina (ML, *Machine Learning*) é uma subárea da IA que estuda o desenvolvimento de técnicas para construção e aprendizado automático. Os algoritmos de ML possuem uma classificação de acordo com à linguagem de descrição, modo, paradigma e forma de aprendizado utilizado (MONARD e BARANAUSKAS, 2003).

O aprendizado indutivo é o topo da hierarquia, sendo ela a forma de inferência lógica um dos principais métodos para derivar conhecimento. O aprendizado indutivo se divide em supervisionado e não-supervisionado, no primeiro é dado ao algoritmo uma base de conhecimento onde se conhece o estado final e treina para determinar o estado final dos quais não se tem o conhecimento, separados posteriormente em classificação para os de saídas categóricas e como regressão os de saída numéricas. Já no aprendizado não-supervisionado o algoritmo tenta formar agrupamentos

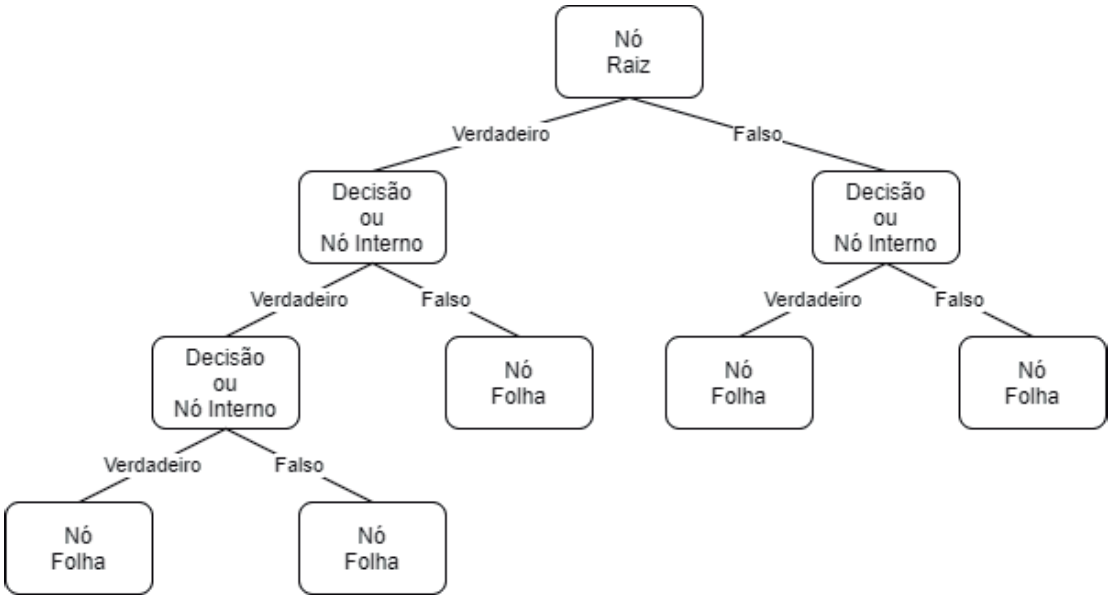
a partir de seu próprio conhecimento, essa determinação de agrupamentos devem ser validadas (MONARD e BARANAUSKAS, 2003).

A ciência de dados (*Data Science*) é uma área fortemente ligada ao ML pois é a arte de extrair conhecimento de dados desorganizados para detecção de padrões e tomada de decisões (GRUSS, 2016).

ÁRVORE DE DECISÃO

Árvore de decisão é um classificador estruturado no formato de uma árvore binária e apresenta possíveis caminhos de decisão e resultado para cada caminho. São muito recomendadas pois são fáceis de entender, interpretar e acompanhar a trajetória para uma previsão, além de poder trabalhar com atributos numéricos e categóricos (GRUSS, 2016).

Figura 1 - Modelo genérico de uma árvore de decisão.



Fonte: autor.

A Figura 1 apresenta um modelo genérico do funcionamento de uma árvore de decisão. Podemos observar que sua estrutura consiste em três tipos de nós. O *Nó Raiz*, presente no início da árvore, representa a aplicação de um teste que pode resultar em Verdadeiro ou Falso. O caso em teste passa pelos *Nós Internos*, que também os classificam através de testes até chegar em um *Nó Folha*, onde está a classificação encontrada pela árvore no final de seu ramo (NORVIG, 2013).

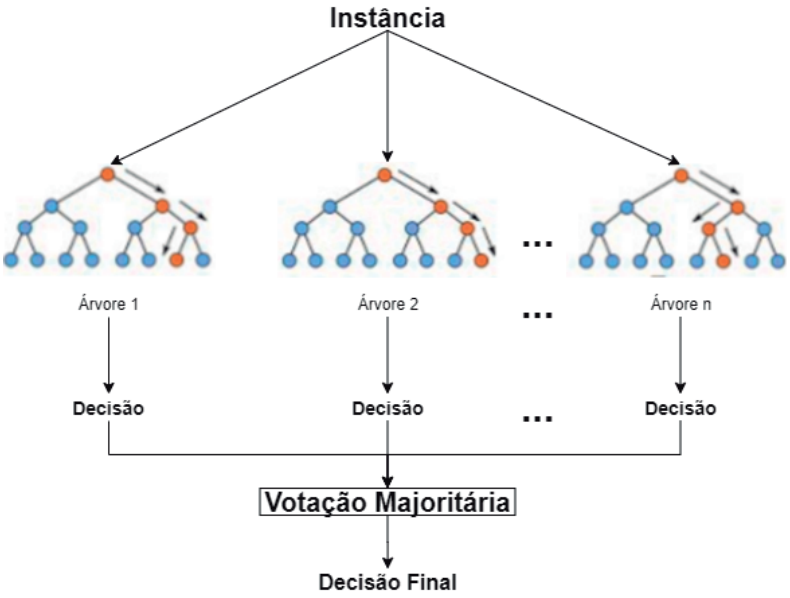
Para Norvig (2013), o formato da árvore de decisão atinge uma resultado agradável e conciso, mas podem ser ruins para alguns tipos de funções, como por exemplo, a função da maioria, que exige uma árvore extremamente grande para tomar uma decisão de verdadeiro, pois se e somente se as entradas possuírem mais da metade convergindo.

FLORESTA ALEATÓRIA

Quando uma árvore de decisão é construída ela é diferente de outra árvore que utilizou dados diferentes ou uma ordem diferente dos dados para ser criada, pois se ajustam com seus dados em seu treinamento. Uma floresta aleatória (*Random Forest*) permite com que possamos construir várias árvores de decisão e deixar com que decidam como classificar sua entrada, tornando um dos modelos mais versáteis disponíveis e assim diminuindo sua variância (GRUSS,2016).

A estrutura de uma floresta aleatória é formada por uma coleção de árvores de decisão, como podemos observar na Figura 2, onde são distribuídas cópias de um mesmo vetor para cada árvore da estrutura, e cada árvore retorna uma decisão que é passada para uma classificação final por votação da maioria. O crescimento do conjunto de árvores e a acurácia de suas decisões permitem melhorias significativas na precisão da floresta, fazendo com que seu erro se aproxime de um limite (BREIMAN, 2001).

Figura 2 - Modelo simplificado da lógica da floresta aleatória.



Fonte: Adaptado de (LORENZETT e TELÖCKEN, 2016).

ÁRVORES EXTRA

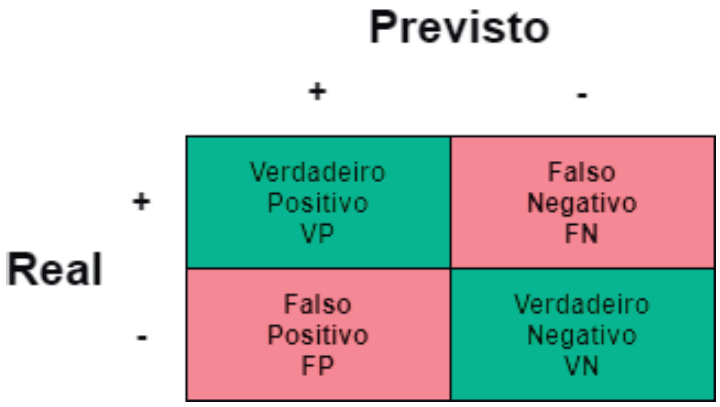
Árvores Extra ou Árvores Extremamente Randomizadas (*Extra Trees*), assim como a floresta aleatória, cria uma coleção de árvores de decisão e também toma a decisão com uma votação majoritária. Porém, diferencia-se da floresta aleatória ao não utilizar uma amostra inicial igual para todas as árvores, e a divisão nos cortes para os nós é realizado de forma aleatória, enquanto na floresta aleatória é realizado a divisão ideal. Essas diferenças fazem com que a árvores extra possuem uma redução no viés e uma menor variância em relação a floresta aleatória (GEURTS *et al.*, 2006).

VALIDAÇÃO DOS CLASSIFICADORES

Para avaliar o desempenho de um classificador binário é necessário separar uma porção dos dados para testes e validações. Métricas de validação buscam avaliar seu classificador, e existem alguns modelos que fazem isso.

Matriz de confusão - É utilizada para alcançar a quantidade de predições certas e erradas do modelo. A Figura 3 representa o modelo da matriz, onde que *verdadeiro positivo* (VP) para casos previstos corretamente, *falso positivo* (FP) para casos onde foi classificado como algo que não é. O *falso negativo* (FN) é uma classificação oposta ao FP e *verdadeiro negativo* (VN) para casos classificados certos também, porém com a classificação oposta de VP (AMIDI, 2020).

Figura 3 - Representação do modelo conceitual de uma matriz de confusão.

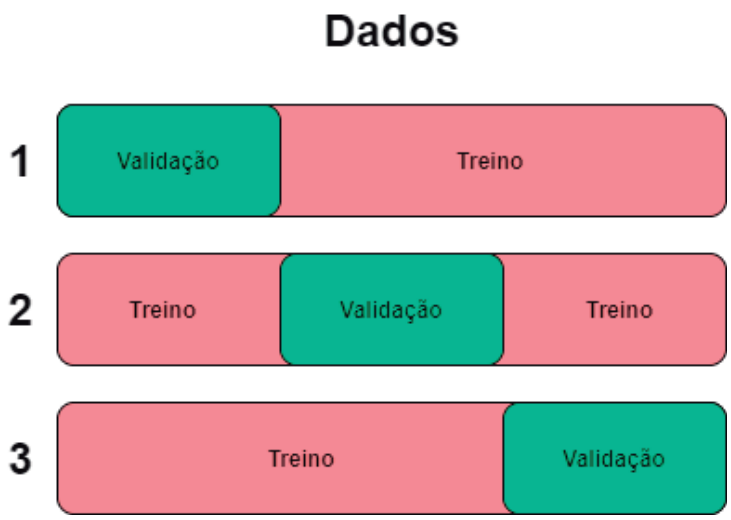


Fonte: Adaptado de (AMIDI, 2020).

ROC e AUC - Chamado de *Receiver Operating Characteristic* e *Area Under the Curve*, são métricas usadas para medição de performance de classificadores binários, onde é plotada a Sensibilidade pela Especificidade em um gráfico para visualizar a curva ROC. O AUC representa grau ou medida de separabilidade da curva, descrito de 0 à 1, onde 1 para um modelo que separa o conjunto de dados positivos dos negativos sem erro, e 0 o menor valor, no qual o modelo erra todas as classificações dos dados (AMIDI, 2020).

K-Fold cross validation - É um método de validação cruzada que busca dividir o conjunto de dados em *k* subconjuntos, onde uma das partes é usada para a validação e é repetido *k* vezes. Cada vez que um subconjunto é selecionado para ser a parte de validação, os outros *k - 1* subconjuntos são usados para o treinamento. O Treino é feito até todas as *k* partes terem sido utilizadas como validadores, como pode ser visto na Figura 4, e após isso, a média do erro ou o desvio padrão de todas as *k* tentativas é calculado (AMIDI, 2020).

Figura 4 - Representação da divisão feita na validação cruzada.



Fonte: Adaptado de (AMIDI, 2020).

TECNOLOGIAS UTILIZADAS

Foram utilizados nesse trabalho, as seguintes tecnologias:

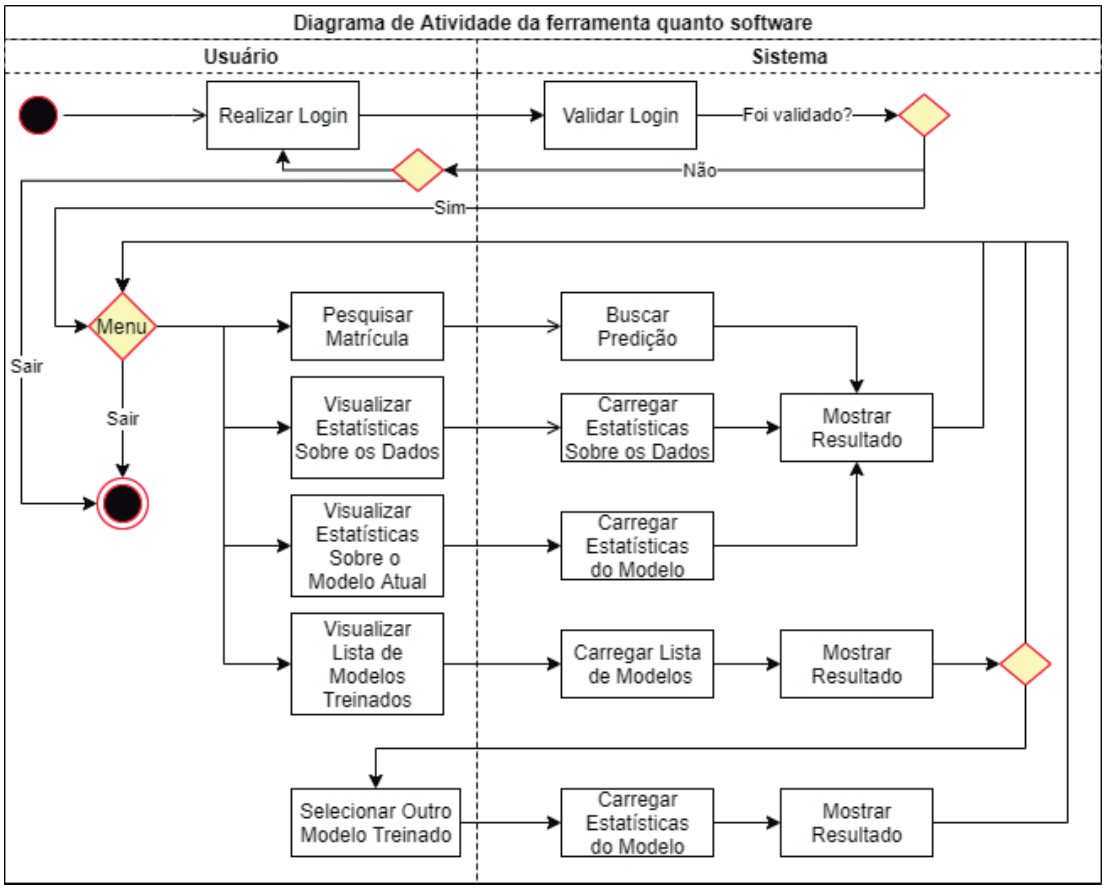
- Python: é uma linguagem orientada a objetos de alto nível, que apresenta tipagem dinâmica e forte, além de interpretada e interativa. É presente na área de \textit{data science} devido a simplicidade da escrita, pela quantidade de bibliotecas disponíveis para o tratamento de dados, além de bibliotecas para criação de classificadores (GRUSS, 2016).
- Scikit-learn: é uma biblioteca do Python que é utilizada para o aprendizado de máquina, pela eficiência para análise preditiva de dados, onde possui inúmeros algoritmos para o aprendizado supervisionado e não supervisionado (PEDREGOSA *et al.*, 2011).
- Pandas: é uma biblioteca do Python, que é utilizada para a análise e manipulação de dados em alta performance, pois contém o objeto *DataFrame* que possui indexação integrada rápida e eficiente (GRUSS, 2016).
- NumPy: é uma biblioteca do Python, que é utilizada principalmente para realizar operações em *Arrays* e matrizes multidimensionais, pois apresenta funções pré-compiladas, com grande capacidade de processamento numérico (GRUSS, 2016).
- Django: é um *framework* para desenvolvimento web de código aberto em Python, que utiliza o padrão *model-template-view* em seu funcionamento, focado para o desenvolvimento rápido, ágil e limpo do projeto (FORCIER *et al.*, 2008).
- Plotly: é uma plataforma colaborativa de gráficos e análises baseada na web. Permite a criação de gráficos iterativos com fácil implementação (INC., 2015).
- Seaborn: é uma biblioteca do Python, de visualização de dados baseada na biblioteca *matplotlib*. Fornece um alto nível para desenhar gráficos estatísticos (WASKOM *et al.*, 2017).

- Pickle: é um módulo da biblioteca padrão do Python, que realiza a escrita e leitura de arquivos binários que possuem a estrutura de um objeto Python. O que este módulo faz é serializar a estrutura em um fluxo de *bytes*, e o reconstrói a partir do arquivo binário dele (PYTHON SF 2020).
- Pep8: é uma proposta de aprimoramento do Python com a intenção de manter a consistência nos códigos. O PEP 8 possui propostas de guia de estilo para: Formatação do código; Comentários; *Docstrings*; Controle de versão; Nomes e identificadores; e Recomendações ao programar (VAN ROSSUM *et al.*, 2001).

METODOLOGIA

Este trabalho utiliza os classificadores árvore de decisão, floresta aleatória e árvores extra implementados com a biblioteca *Scikit-learn* e salvos com a biblioteca *Pickle*. A normalização é feita através da utilização das bibliotecas Pandas e NumPy. A validação é dada pelo K-Fold, ROC e AUC e matriz de confusão, com cálculo de sua acurácia, precisão, sensibilidade, especificidade e F1 score. A interface faz uso do *framework* Django, com gráficos criados pelas bibliotecas Plotly e Seaborn.

Figura 5 - Diagrama de atividade mostrando as fases de ações do projeto.



Fonte: Autor.

A Figura 5 apresenta um diagrama que possui o principal fluxo de atividades realizadas pelo software, onde o usuário primeiramente realiza o *login* no sistema. Após o *login*, o usuário pode realizar as ações de pesquisar um aluno pela matrícula, visualizar as estatísticas sobre os dados do banco, visualizar as estatísticas sobre o modelo em funcionamento e visualizar uma lista dos modelos treinados, onde pode selecionar um modelo para usar na classificação.

IDENTIFICAÇÃO DAS REGRAS E ANÁLISE DOS DADOS

O primeiro passo para criação de um classificador é a obtenção de dados para seu treinamento, e para isso é preciso primeiro identificar quais são os possíveis motivos que levam um aluno à evadir. Como metodologia para identificação das regras que podem influenciar na evasão, foi realizada uma pesquisa por trabalhos acadêmicos relacionados que tratam sobre o tema da evasão escolar e apresentam de alguma forma os possíveis motivos para ocorrerem.

A próxima etapa para a construção deste trabalho foi, a partir das regras identificadas para classificar um aluno evasor, fazer a análise dos dados da base acadêmica fornecida e preparar (normalizar) o conteúdo para o uso de um classificador.

O banco de dados utilizado neste projeto foi cedido pela Universidade Franciscana (UFN), contendo 413 tabelas somando 48 Gigabytes de informações referentes ao período de 2008 à 2018 de 93.5 mil alunos. Foram selecionadas 10 tabelas de interesse para a aplicação das regras. Os dados pessoais irrelevantes para essa pesquisa foram removidos para preservar a privacidade dos alunos.

Após a identificação dos motivos teóricos para evasão e durante a análise dos dados, foram definidas quais regras poderiam ser utilizadas através da disponibilidade de dados no banco, que são:

- Distância: Distância entre a casa do aluno e o campus onde estuda;
- Média de faltas: Média de faltas por disciplina cursada;
- Nº de parcelas pagas: Número de parcelas pagas pelo aluno dentro do prazo de validade + 5 dias de atraso;
- Média de dias em parcelas pagas atrasado: Média de dias de atraso do pagamento da parcela + 5 dias;
- Média das notas abaixo da média da turma: Média das notas do aluno que ficaram abaixo de 0.5 da média da turma;
- Média das notas na média da turma: Média das notas do aluno que ficaram ± 0.5 da média da turma;
- Média das notas acima da média da turma: Média das notas do aluno que ficaram acima da média da turma + 0.5;
- Razão das notas abaixo: Dado pelo Nº de disciplinas com a nota abaixo de 0.5 da média da turma dividido pelo Nº de disciplinas cursadas;

- Razão das notas na média: Dado pelo Nº de disciplinas com a nota ± 0.5 da média da turma dividido pelo Nº de disciplinas cursadas;
- Razão das notas acima: Dado pelo Nº de disciplinas com a nota acima da média da turma + 0.5 dividido pelo Nº de disciplinas cursadas;
- Nº de reprovações: Número de reprovações do aluno;
- Reprovações vs Semestre: Refere-se a regra apresentada em (LENHARD, MARTINS, 2019) que atribui um peso para a quantidade de reprovações de acordo com o semestre do aluno;
- Transferência interna anteriormente: Número de transferências internas já realizadas pelo aluno;
- Idade: Refere-se a idade do aluno;
- Gênero: Corresponde ao gênero do aluno;
- Forma de ingresso: Tipo de ingresso do aluno;
- Nº de cancelamentos: Número de cancelamentos já realizados pelo aluno;
- Estado Civil: Estado civil do aluno;
- Turno do curso: Turno do curso do aluno;
- Semestre: Semestre do aluno;
- Tempo cursado: Tempo cursado pelo aluno em anos.

A regra *Reprovações Vs Semestre* apresentada em Lenhard e Martins (2019), que atribui um peso para a quantidade de reprovações do aluno de acordo com o semestre dele, é dada por:

$$\alpha = \sum_{i=1}^{8}(9 - i) * disciplina\_reprovada_i \quad (1)$$

e adaptada para:

$$\alpha = \sum_{i=1}^{n\_semestres}(n\_semestres + 1 - i) * disciplina\_reprovada_i \quad (2)$$

onde *n_semestres* refere-se ao número de semestres do curso do aluno. Essa adaptação foi necessária, pois nem todos os cursos possuem a mesma quantidade de semestres.

RESULTADOS E DISCUSSÕES

Das regras idealizadas para identificação de perfis evasores pelos classificadores, as regras *Semestre, Tempo cursado* e *Número de parcelas pagas* foram desconsideradas, pois enviesavam como provável evasor todos os alunos que possuíam carga horária menor que a total para conclusão do curso, como por exemplo, qualquer aluno matriculado que está no meio de um curso seria um provável evasor.

A Tabela 1 mostra os resultados obtidos pelos classificadores para as quatro áreas do conhecimento individualmente e para todas as áreas em conjunto. É visível que a floresta aleatória e árvores extra possuem resultados melhores em todas as áreas que a árvore de decisão devido a sua modelagem. Para a área das Ciências Humanas e para a área das Ciências Sociais o modelo que obteve os melhores resultados foi a árvores extra com uma acurácia de 90.66% e 93.11% respectivamente.

Para a área das Ciências Tecnológicas e para área das Ciências da Saúde o modelo que obteve os melhores resultados foi a floresta aleatória com uma acurácia de 96.45% e 95.69% respectivamente.

Tabela 1 - Resultados obtidos pelos classificadores.

Área	Classificador	Métricas				
		Acurácia	**Precisão**	**Sensibilidade**	**Especificidade**	**F1 score**
Todas	Árv. Decisão	88.7%	86.85%	93.37%	83.12%	90.0%
	Flor. Aleatória	92.05%	96.35%	90.7%	94.28%	93.44%
	Árv. Extra	92.84%	94.21%	92.54%	93.2%	93.2%
Humanas	Árv. Decisão	86.13%	88.22%	81.61%	90.19%	84.78%
	Flor. Aleatória	90.38%	91.58%	94.62%	81.21%	93.08%
	Árv. Extra	90.66%	91.78%	94.82%	81.66%	93.28%
Sociais	Árv. Decisão	88.04%	89.78%	93.49%	75.48%	91.6%
	Flor. Aleatória	92.7%	94.48%	95.09%	87.19%	94.78%
	Árv. Extra	93.11%	95.63%	94.44%	90.05%	95.03%
Tecnológicas	Árv. Decisão	94.8%	97.33%	96.44%	86.22%	96.88%
	Flor. Aleatória	96.45%	98.99%	96.75%	94.88%	97.86%
	Árv. Extra	95.81%	99.06%	95.92%	95.28%	97.46%
Saúde	Árv. Decisão	89.68%	95.54%	87.09%	93.7%	91.12%
	Flor. Aleatória	95.69%	97.16%	95.48%	95.98%	96.31%
	Árv. Extra	91.67%	90.24	93.98%	89.21%	92.07%

Fonte: Autor.

Ao analisar as estruturas dos classificadores juntamente com valores estatísticos da maioria dos evasores, é possível determinar que o perfil mais aproximado a de um aluno evasor para todas as áreas é aquele que possui:

- *Razão_Notas_Abaixo* maior que 0.7
- *Razão_Notas_Média* menor que 0.1
- *Razão_Notas_Acima* menor que 0.2
- *Média_Das_Notas_Abaixo* menor que 1.27
- *Média_Das_Notas_Na_Média* igual a 0 ou entre 7.0 e 7.8
- *Média_Das_Notas_Acima* igual a 0 ou entre 7.5 e 8.5
- *Média_de_Faltas* igual a 0
- *Número_de_Disciplinas_Reprovadas* igual a 0 ou igual a 3
- Está na faixa etária de 19 a 22 anos

CONCLUSÃO

Os resultados obtidos mostram que a identificação de perfis evasores e seus padrões, utilizando os classificadores árvore de decisão, floresta aleatória e árvores extra, implementados pela biblioteca Scikit-learn, com acurácias superiores à 88% para todas as quatro áreas, foram satisfatórios.

Esses resultados apontam também que a utilização das bibliotecas Pandas e Numpy se mostraram eficientes para normalizar e modelar as regras a partir da base de dados acadêmica da UFN. É visível que os classificadores de floresta aleatória e árvores extra apresentam melhores resultados que os classificadores de árvore de decisão pelo fato de diminuírem a variância dos dados.

O requisito 'Mostrar Estrutura' mostra-se inviável sua implementação, pois os classificadores árvores extra e floresta aleatória são treinados com 100 árvores de decisão. Os requisitos 'Escolher Variáveis', 'Redefinir Modelo' e 'Carregar novos dados', que tratavam do treinamento de novos modelos não foram implementados, pois não era o foco principal deste trabalho a generalização dos classificadores.

Com isso, pode-se concluir que a utilização de aprendizado de máquina supervisionado juntamente com técnicas de ciência de dados e a utilização do banco de dados acadêmico da UFN, possui uma capacidade de predição média de 91.9% da evasão de alunos evasores, e que o classificador de árvores extra obteve os melhores resultados com 92.84%, considerando sua acurácia para as todas as áreas.

Existem diversos motivos que levam um aluno à evadir de um curso, porém nota-se que muitos destes motivos estão fora do alcance do banco de dados utilizado para o treinamento, e que possíveis pesquisas como questionários aplicados nos alunos poderiam auxiliar na predição destes motivos. Ao analisar os resultados, é possível concluir que a utilização das regras de forma correlacionada permite uma melhor predição do perfil evasor e que as regras Semestre; Tempo cursado; e Número de parcelas pagas, que foram desconsideradas ao enviesar os resultados podem ser reconsideradas se modelar os alunos não evasores de modo a retirar o sentido de tempo.

Ficam como sugestão para futuros trabalhos a implementação dos requisitos funcionais 'Escolher Variáveis', 'Redefinir Modelo' e 'Carregar Novos Dados' para a generalização dos classificadores, e da busca de novos dados. E também, a análise dos dados dos anos de 2019 e 2020 para verificação assertiva dos classificadores, além da análise do impacto causado pela pandemia do Covid-19 na evasão das instituições.

REFERÊNCIAS

AGARWAL, R. and Umphress, D. **Extreme programming for a single person team**. InProceedings of the 46th Annual Southeast Regional Conference on XX, pages 82-87, 2008.

AMIDI, A. and Amidi, S. **Machine learning tips and tricks cheatsheet**. https://stanford.io/3EhKsiJ. Visted on: 21 maio 2020.

BREIMAN, L. **Random forests. Machine learning**, 45(1):5-32, 2001.

DEKKER, G., PECHENIZKIY, M., and VLEESHOUWERS, J. **Predicting students drop out:** A case study. InComputers, Environment and Urban Systems, pages 41-50, 2009.

FORCIER, J., BISSEX, P., and CHUN, W. J. **Python web development with Django**. Addison--Wesley Professional, 2008.

FRITSCH, R., da ROCHA, C. S., and VITELLI, R. F. **A evasão nos cursos de graduação em uma instituição de ensino superior privada.** Revista Educação em Questão, 52(38), p. 81-108, 2015.

GEURTS, P., ERNST, D., and WEHENKEL, L. **Extremely randomized trees**. Machine learning, 63(1), p. 3-42, 2006.

GRUSS, J. **Data Science do zero**. Ed. Alta Books, 1. ed., 2016.

INC, P. T. **Collaborative data science**, 2015.

LENHARD, C. and MARTINS, M. O. Ia: **Descrição e aplicação de regras de evasão no curso de ciência da computação em IES**. Disciplinarum Scientia | Naturais e Tecnológicas, 20(2). p. 199-209, 2019.

LORENZETT, C. and TELÖCKEN, A. **Estudo comparativo entre os algoritmos demineração de dados random forest e j48 na tomada de decisão**. Simpósio de Pesquisa e Desenvolvimento em Computação (SPDC), 2(1), 2016.

MONARD, M. C. and BARANAUSKAS, J. A. **Conceitos sobre aprendizado de máquina.** Sistemas inteligentes-Fundamentos e aplicações, 1(1):32, 2003.

OLIVEIRA, C. H. M., SANTOS, F. R. T., LEITINHO, J. L., and FARIAS, L. G. A. T. **Busca dos fatores associados à evasão**. Revista Internacional de Educação Superior, 5:e019006-e019006, 2019.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL,M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU,D., BRUCHER, M., PERROT, M., and DUCHESNAY, E. **Scikit-learn:** Machine learningin Python. Journal of Machine Learning Research, 12. p. 2825-2830, 2011.

NORVIG, P. S. R. **Inteligência Artificial**. Elsevier, 3 ed., 2013.

PYTHON SOFTWARE FOUNDATION **Pickle - python object serialization**. https://docs.python.org/3/library/pickle.html. acessado em: 21 maio 2020.

TINTO, V. **Taking retention seriously:** Rethinking the first year of college. NACADA Journal, 19(2). p. 5-9, 1999.

VAN ROSSUM, G., WARSAW, B., and COGHLAN, N.**Pep 8: style guide for python code**. Python. org, 1565, 2001.

WASKOM, M., BOTVINNIK, O., O'KANE, D., HOBSON, P., LUKAUSKAS, S., GEMPERLINE, D. C., AUGSPURGER, T., HALCHENKO, Y., COLE, J. B., WARMENHOVEN, J., DE RUITER, J., PYE, C., HOYER, S., VANDERPLAS, J., VILLALBA, S., KUNTER, G., QUINTERO, E., BACHANT, P., MARTIN, M., MEYER, K., MILES, A., RAM, Y., YARKONI, T., WILLIAMS, M. L., EVANS, C., FITZGERALD, C., BRIAN, FONNESBECK, C., LEE, A., AND QALIEH, A. (2017). **mwaskom/ seaborn: v0.8.1** https://seaborn.pydata.org/. Acesso em: set. 2017.