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

# Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior*

**Cássio S. Carvalho[1], Júlio C. B. Mattos[1], Marilton S. Aguiar[1]**

[1]Programa de Pós-Graduação em Computação – Universidade Federal de Pelotas (UFPel)
Rua Gomes Carneiro, 1 – 96.010-610 – Pelotas – RS – Brasil

{cassio.carvalho,julius,marilton}@inf.ufpel.edu.br

***Abstract.*** *This paper investigates model interpretability aspects in the educational data mining context, specifically for the school dropout problem in higher education. Prediction models are trained and then explained with LIME. Explanations are analyzed by nonsupervised learning, and a prediction method based on central explanations is proposed. Combining a prediction model and prediction by explanations allows us to analyze performance aspects and quality. An interpretability metric is presented. Results indicate that models with similar performance can present different characteristics regarding interpretability metrics.*

***Resumo.*** *Este trabalho propõe investigar aspectos de interpretabilidade de modelos no contexto da mineração de dados educacionais, especificamente para o problema da evasão no ensino superior. Modelos de predição são treinados e então explicados com LIME. Explicações são analisadas utilizando aprendizado não supervisionado, e é proposto um método de predição baseado em explicações centrais. O uso combinado das predições do modelo com as predições pelas explicações permite analisar aspectos de desempenho e qualidade das explicações. Apresenta-se uma métrica de interpretabilidade. Resultados indicam que modelos com performance de desempenho similar podem apresentar diferentes características quanto a métricas de interpretabilidade.*

## 1. Introdução

A mineração de dados educacionais (MDE) tem por objetivo produzir conhecimento útil a partir de dados educacionais digitalizados, destacando-se neste contexto como uma das principais formas de análise de *big data* educacional. Sendo uma área interdisciplinar, aplica técnicas de aprendizado de máquina (AM), estatística, mineração de dados (MD), psicologia educacional, psicologia cognitiva, e outras teorias e métodos de análise de dados educacionais [Xiao et al. 2022].

Os métodos utilizados na MDE são comuns aos da área de mineração de dados em geral, sendo que existem múltiplos métodos para cada uma das várias aplicações possíveis. Dentre os métodos mais utilizados, destacam-se o agrupamento, a mineração de associações, a regressão e a classificação [Romero and Ventura 2010,

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Bakhshinategh et al. 2018]. As aplicações no contexto da MDE também são diversas e podem estar relacionadas a diferentes métodos. A predição de desempenho e de características do aluno, por exemplo, normalmente utiliza os métodos de classificação e regressão, no entanto outras técnicas como agrupamento também já foram utilizadas [Bakhshinategh et al. 2018].

Em razão da grande disponibilidade de dados heterogêneos e poder computacional, algoritmos de AM alcançam desempenhos preditivos cada vez melhores. A maioria dos modelos gerados, entretanto, apresentam maior complexidade e limitações como opacidade ou falta de transparência [Carvalho et al. 2019]. Talvez o principal objetivo da previsão de performance de alunos seja identificar os fatores que realmente impactam nessa performance. Entretanto, poucos estudos deram atenção para esse aspecto utilizando teorias de aprendizagem para explicar os princípios do modelo, o processo de predição e porque diferentes características afetam os resultados. Poucos pesquisadores buscam melhorar a interpretabilidade do modelo de previsão no contexto de MDE [Xiao et al. 2022].

Percebe-se que o grau de interpretabilidade de um modelo está relacionado a facilidade de compreender suas decisões e/ou predições. Ao mesmo tempo, um modelo é mais interpretável que outro se suas decisões são mais facilmente compreendidas do que as decisões de outro modelo [Molnar 2022]. Interpretabilidade e explicabilidade são conceitos muito próximos, intimamente relacionados. Alguns autores costumam diferenciá-los. Quando os humanos são capazes de entender o que um modelo fez, então este modelo é interpretável. Ao considerar que os humanos são capazes de explicar e descobrir por que e quais atributos têm uma influência significativa no resultado, então diz-se que o modelo é explicável [Burkart and Huber 2021]. No contexto deste trabalho, assim como em [Molnar 2022], os termos interpretabilidade e explicabilidade serão usados de forma intercambiável.

Este trabalho é um esforço para investigar aspectos de interpretabilidade de modelos no contexto da MDE, mais especificamente em relação ao problema da evasão de estudantes no ensino superior. Nesse sentido, o presente trabalho busca responder as seguintes questões de pesquisa:

Q1 – É possível usar explicações para predizer novas amostras?
Q2 – É possível avaliar a qualidade individual de uma explicação?
Q3 – É possível usar a interpretabilidade para melhoria da predição de um modelo?
Q4 – É possível identificar uma métrica relacionada a interpretabilidade?

O artigo está organizado da seguinte forma. A Seção 2 apresenta trabalhos com investigações similares ao presente estudo. A Seção 3 traz detalhes do *dataset* utilizado bem como todos os procedimentos metodológicos aplicados. A Seção 4 apresenta e discute os resultados obtidos nos experimentos. Por fim, a Seção 5 conclui sobre os resultados obtidos, apresentando as respostas para as questões de pesquisa propostas.

## 2. Trabalhos Relacionados

O estudo de [Alwarthan et al. 2022] busca identificar alunos em risco de reprovação em curso chamado de ano preparatório do estudante (*humanities track*), oferecido pela Universidade Imam Abdulrahman bin Faisal (IAU). O *dataset* foi coletado a partir de várias fontes, incluindo sistemas de gestão acadêmica, sistemas de gerenciamento de aprendizagem, entre outros. Emprega técnicas como LIME

[Ribeiro et al. 2016], SHAP [Lundberg and Lee 2017] e *Global Surrogate Model* para explicar modelos de predição, destacando as razões para reprovação do aluno. O trabalho de [Vultureanu-Albisi and Badica 2021] propõe identificar as técnicas de classificação mais apropriadas para melhorar a predição da performance de estudantes, interpretando os resultados com o algoritmo LIME. Uma das suas questões de pesquisa foi "até que ponto as explicações interpretáveis ajudam na predição de desempenho dos alunos?". O *dataset* contém informações como notas, características demográficas, sociais, comportamentais e relacionadas a escola. A interpretabilidade foi abordada com o LIME, escolhendo-se algumas instâncias para checar se as explicações melhoram a transparência de classificação dos modelos.

O estudo de [Kumar and Sharma 2020] propõe predizer o desempenho de alunos estrangeiros em uma universidade do norte da Índia. Foram aplicados algoritmos como Regressão Logística, Naïve Bayes, CART e *Random Forest* (RF) em um *dataset* contendo 399 alunos, de graduação e pós-graduação. Foram utilizados dados como média cumulativa e relacionados ao país de origem (região e economia). O método LIME foi utilizado para gerar explicações e realizar breve análise sobre algumas amostras. Em [Pei and Xing 2022], uma nova abordagem para identificação de alunos em risco, com foco na interpretabilidade, é introduzida. O estudo foi realizado em nível universitário, compreendendo os tópicos de Ciências Sociais e STEM (*Science, technology, engineering, and mathematics*). O *dataset* possui informações dos cursos, dos estudantes e suas interações com ambiente virtal de aprendizagem. A interpretabilidade é explorada no escopo local, utilizando LIME e escolhendo alguns alunos em risco.

Poucos estudos na MDE têm iniciativa no sentido de medir a interpretabilidade de modelos. [Alharbi 2022] propôs medir a interpretabilidade a partir de uma métrica de complexidade, limitando o número de regras por classe, justificado pelo fato de que humanos podem lidar com no máximo $7\pm2$ entidades cognitivas ao mesmo tempo [Miller 1956]. O estudo de [Al-Jallad et al. 2019] utilizou modelos interpretáveis e propôs uma medida de interpretabilidade baseada no tamanho e na complexidade das regras. São apresentadas as medições de interpretabilidade de cada modelo e uma análise dos efeitos da etapa de pré-processamento nessas medições. Nesse sentido, esse trabalho diferencia-se por analisar um conjunto de explicações e propor uma métrica de interpretabilidade para modelos de predição.

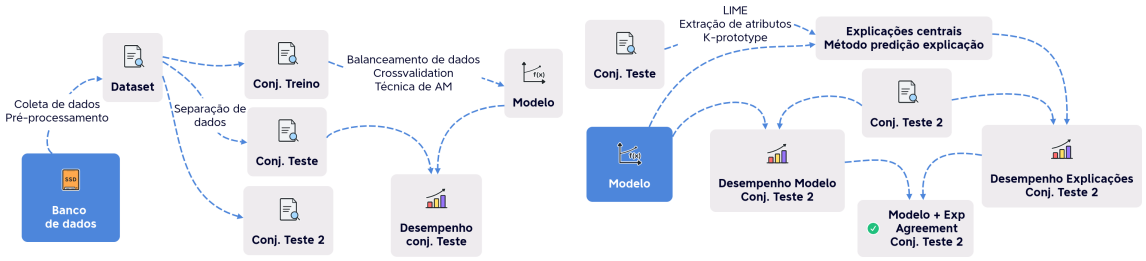
## 3. Materiais e Métodos

Os experimentos foram implementados utilizando a linguagem de programação Python no ambiente online Google Colaboratory[1]. As bibliotecas NumPy[2] e Pandas[3] foram utilizadas para algebra linear e processamento de dados, respectivamente. Classificadores, métricas e outras técnicas relacionadas ao AM foram importadas a partir do scikit-learn[4] ou de pacote específico como é o caso do XGBoost[5]. Destaca-se ainda a utilização do SMOTE[6] para balanceamento de dados, LIME[7] para interpretabilidade de modelos e k-prototype[8] para clusterização de dados contendo variáveis numéricas e categóricas.

Os experimentos descritos nesse trabalho estão organizados em duas etapas. Na primeira etapa, técnicas de AM são utilizadas para evoluir modelos no contexto da

[1] https://colab.research.google.com/ [2] https://numpy.org [3] https://pandas.pydata.org
[4] https://scikit-learn.org/ [5] https://xgboost.readthedocs.io/ [6] https://imbalanced-learn.org/
[7] https://github.com/marcotcr/lime [8] https://github.com/nicodv/kmodes

predição de evasão de alunos no ensino superior, conforme ilustrado na Figura 1a. Para compreender essa etapa, a Seção 3.1 detalha o *dataset* utilizado, enquanto a Seção 3.2 apresenta a separação e balanceamento de dados e, posteriormente, a Seção 3.3 descreve como os modelos foram treinados. A segunda etapa dos experimentos consiste em explicar um conjunto de amostras e agrupá-las, conforme ilustrado na Figura 1b, com intuito de responder às questões de pesquisa apresentadas na introdução deste trabalho (Seção 1). Em linha com essa etapa, são descritas as explicações e a extração de atributos na Seção 3.4, já a clusterização de explicações é apresentada na Seção 3.5 e, por fim, a predição pelas explicações na Seção 3.6.




**(a) Modelos de predição.** **(b) Explicação das amostras.**

**Figura 1. Fluxo das etapas da abordagem.**

### 3.1. Dataset

O *dataset* utilizado nos experimentos foi gerado especificamente para esta pesquisa por meio de consulta a base de dados do COBALTO[9] (Sistema Integrado de Gestão) da Universidade Federal de Pelotas (UFPel). Os dados obtidos podem ser categorizados principalmente em dois grupos: demográficos e de desempenho acadêmico. Vale ressaltar que muitos dos atributos relacionados ao desempenho acadêmico não estão disponíveis de forma imediata no sistema acadêmico. Tais atributos foram derivados com intenção de enriquecer o conjunto de treino no contexto a que essa pesquisa se propõe. A Tabela 1 apresenta os 44 atributos gerados e suas respectivas descrições, sendo que o *target* é o atributo alvo para o problema de predição. Cada linha do conjunto de dados é referente a um aluno de graduação, com um total de 14.452 instâncias. A geração do *dataset* respeitou algumas condições, a saber:

- Alunos de qualquer curso de graduação da Universidade que tiveram seu ingresso entre os anos de 2010 e 2015, exceto alunos do curso de Direito (curso anual);
- Alunos que tiveram registro de saída até 2019 (inclusive). Condição para desconsiderar alunos formados ou evadidos no período da pandemia de COVID-19;
- Foram obtidos 14.452 alunos, sendo que os tipos de saídas identificadas foram Abandono (3.379), Cancelamento (1.088), Desligado (13), Desligado PEC-G(1), Formado (8.938), Jubilado (3), Reopção (847), Reopção compulsória (1) e Transferido (182). Todas as saídas diferentes de "Formado" foram agrupadas como sendo da classe "Evadido". A classe alvo resultou em 8.938 alunos formados (*target*=0) e 5.514 evadidos (*target*=1);
- Foram considerados os registros dos 3 primeiros semestres de cada aluno; e;
- Foram descartados alunos com trancamento geral de matrícula em um dos seus 3 primeiros semestres. Foram descartados alunos com zero matrículas no 3º semestre, zero matrículas no 2º e 3º semestres, zero matrículas no 1º, 2º e 3º semestre.

**Tabela 1. Relação de atributos.**

| # | Atributo | Descrição |
|---|---|---|
| 1 | idade_ingresso | Idade de ingresso |
| 2 | tempo_entre_medio_grad | Tempo entre ensino médio e graduação |
| 3 | sexo_code | Sexo |
| 4 | cor_code | Cor |
| 5 | estado_civil_code | Estados civil |
| 6 | codigo_sisu_code | Código da cota de ingresso |
| 7 | flg_publica_code | Flag escola pública |
| 8 | nivel_superior_code | Flag possui nível superior |
| 9 | turno_code | Turno do curso |
| 10 | area_fundamental_code | Área fundamental do curso |
| 11,12,13 | nr_disciplinas_{1,2,3} | Nr. disciplinas 1º, 2º e 3º semestre |
| 14,15,16 | nr_creditos_{1,2,3} | Nr. créditos 1º, 2º e 3º semestre |
| 17,18,19 | media_semestre_{1,2,3} | Média do 1º, 2º e 3º semestre |
| 20,21,22 | nr_creditos_aprovado_{1,2,3} | Créditos aprovados 1º, 2º e 3º semestre |
| 23,24,25 | nr_creditos_dispensa_{1,2,3} | Créditos dispensados 1º, 2º e 3º semestre |
| 26,27,28 | nr_creditos_reprovado_{1,2,3} | Créditos reprovados 1º, 2º e 3º semestre |
| 29,30,31 | nr_creditos_infrequente_{1,2,3} | Créditos infrequente 1º, 2º e 3º semestre |
| 32,33,34 | nr_creditos_trancado_{1,2,3} | Créditos trancados 1º, 2º e 3º semestre |
| 35,36,37 | nr_creditos_semestre_referencia_{1,2,3} | Créditos a serem cursados no 1º, 2º e 3º semestre do currículo |
| 38,39,40 | nr_creditos_semestre_referencia_historico_{1,2,3} | Nr créditos matriculados e que pertencem ao 1º, 2º e 3º semestre do curso |
| 41,42,43 | relacao_apr_semestre_{1,2,3} | Relação entre créditos aprovados e créditos do currículo, no 1º, 2º e 3º |
| 44 | target | Atributo alvo. Formado (0) ou Evadido (1) |

Durante o pré-processamento dos dados foi realizada a "inputação" de valores faltantes. Os atributos categóricos e dois atributos numéricos (idade e tempo_entre_medio_grad) foram preenchidos com os valores mais frequentes no *dataset*. Os demais atributos numéricos foram preenchidos com zero. Foram criados 3 atributos derivados (*relacao_apr_semestre_1*, *relacao_apr_semestre_2*, *relacao_apr_semestre_3*), já descritos na Tabela 1.

### 3.2. Separação e Balanceamento de Dados

Os dados foram separados de forma estratificada em 3 conjuntos: treino (60%), teste (20%) e teste 2 (20%). A Tabela 2 apresenta a quantidade de amostras por classe em cada um dos conjuntos obtidos. A utilização de cada um dos conjuntos é descrita ao longo desta Seção. Como observado, o número de alunos formados é maior que o número de alunos evadidos, sendo recomendado um balanceados dos dados. Esse problema foi tratado por meio da biblioteca SMOTE, aplicando a técnica de *oversampling* na classe minoritária do conjunto de treino.

**Tabela 2. Instâncias por classe antes do balanceamento no conjunto de treino.**

| Tipo da saída | Alunos | Treino | Teste | Teste 2 |
|---|---|---|---|---|
| Formado | 8938 | 5363 | 1787 | 1788 |
| Evadido | 5514 | 3308 | 1103 | 1103 |

### 3.3. Modelos

Esta Seção apresenta a metodologia utilizada ao treinar os modelos de predição. Todos os experimentos consideraram as seguintes técnicas de AM para classificação: *Decision Tree*[10] (DT), *Random Forest*[11] (RF), XGBoost[12] (XGB), *Artificial neural network*[13]

[9] https://cobalto.ufpel.edu.br [10] DecisionTreeClassifier() [11] RandomForestClassifier()
[12] https://xgboost.readthedocs.io [13] MLPClassifier()

(ANN). A seguir são descritos os passos realizados ao treinar os modelos para cada uma das técnicas de AM:

- Separação dos dados em dados de treino, teste e teste 2 (conforme Seção 3.2);
- Balanceamento dos dados de treino (*oversampling* da classe minoritária);
- Avaliação da metodologia[14] (técnica de AM + hiperparâmetros *default*) utilizando validação cruzada estratificada (*n_splits*=10 e *shuffle=True*). A métrica utilizada nesta avaliação é a acurácia balanceada;
- Obtenção do modelo final a partir do treino da metodologia utilizando o conjunto completo de dados de treino;
- Aplicação do modelo para prever a evasão no conjunto de teste. Avaliação das métricas *precision*, *recall* e *f1-score*.

### 3.4. Explicações e Extração de Atributos

O LIME [Ribeiro et al. 2016] é um método de interpretabilidade agnóstico e de escopo local. Agnóstico porquê pode ser aplicado a qualquer classe de modelos e local pelo fato de explicar uma amostra específica. A explicação entregue pelo LIME consiste de uma lista de tuplas, onde cada tupla possui uma expressão e um peso associado. Visando analisar um conjunto de explicações, propõe-se um método para extração de atributos a partir de uma explicação. Essa extração de atributos transforma uma explicação do LIME em uma nova amostra. Cada tupla na explicação gera 3 atributos na nova amostra. Sendo assim, se uma explicação possui $N$ tuplas, então a amostra resultante terá um total de $N * 3$ atributos. Para compreensão, considere a tupla de exemplo ("3 $<=$ *nr_creditos_aprovado_3* $<=$ 10", 0.05). Neste caso serão gerados os atributos *nr_creditos_aprovado_3_lim_inf*, *nr_creditos_aprovado_3_lim_sup* e *nr_creditos_aprovado_3_weight*, com os valores 3, 10 e 0.05, respectivamente. O nome de cada atributo começará com o nome do atributo original presente na expressão, concatenado com termos que referenciam o limite inferior da expressão (*_lim_inf*), limite superior (*_lim_sup*) e o peso associado (*_weight*).

### 3.5. Clusterização de Explicações

Cada uma das instâncias do conjunto de teste é explicada com o LIME e submetida ao método de extração de atributos. Como resultado, o conjunto de teste é mapeado para um novo conjunto onde cada instância gerada representa a explicação para a amostra original no contexto de um determinado modelo. Portanto, obtém-se um conjunto de amostras de explicações para cada um dos modelos treinados. A partir desse ponto, o conjunto de explicações passa a ser um novo conjunto de dados a ser analisado, porém sem rótulos. A clusterização das explicações inicia com a normalização das amostras, identificação do número ótimo de *cluster* de forma automatizada com o método de Elbow e, por último, o agrupamento propriamente dito com a biblioteca k-prototype. O *centroid* de cada *cluster* é calculado e submetido ao processo inverso da extração de atributos, resultando numa explicação (no formato original de lista de tuplas) que pode ser entendida como uma explicação central do *cluster*. Uma vez que existe um mapeamento do conjunto de teste para o conjunto de explicações, sabe-se também a qual classe (Formado ou Evadido) uma explicação está associada. O que permite calcular os *centroids* por classe dentro de cada *cluster*, os quais também são transformados para explicações no formato de lista de tuplas.

[14] Nesse contexto, utiliza-se o termo metodologia para a combinação de uma técnica de AM com uma determinada configuração de hiperparâmetros dessa técnica

### 3.6. Predição pelas Explicações

De posse das explicações centrais (do *cluster* e do *cluster* por classe), propõe-se predizer novas amostras de alunos sem a utilização do modelo de AM treinado originalmente. A ideia proposta é simples. Inserem-se os dados acadêmicos de um aluno nas expressões obtidas para cada explicação central. Calcula-se o número de expressões satisfeitas, sempre ponderadas pelo peso. Isso permitirá descobrir em qual *cluster* o aluno melhor se encaixa. Cumprida esta etapa, pegam-se as explicações centrais por classe deste *cluster* e repete-se o processo inserindo novamente os dados acadêmicos do aluno. A explicação central por classe que melhor for satisfeita, definirá a classe de resposta.

## 4. Experimentos e Resultados

### 4.1. Predição de evasão

As técnicas de DT, RF, XGB e ANN foram utilizadas para obter modelos de predição de evasão no ensino superior. Todos os modelos foram obtidos conforme procedimentos descritos na Seção 3.3. A Tabela 3 apresenta a acurácia balanceada de cada metodologia (Técnica de AM + hiperparâmetros *default*) para os *Folds* (divisões) durante a validação cruzada. Apresenta-se o desvio padrão dessas métricas, bem como o *mean score*. Os resultados validam as metodologias, uma vez que o desempenho nos diferentes *Folds* são bem próximos, com desvio padrão menor ou igual a 0,01. A Tabela 4 apresenta o desempenho desses modelos no conjunto de teste. As técnicas RF, XGB e ANN apresentaram desempenho acima de 80% para todas as métricas quando considerada a classe Formado. Para a classe Evadido, destacaram-se XGB e RF com *precision* de 81% e 79%, respectivamente.

**Tabela 3. Desempenho de cada metodologia na validação cruzada.**

| Metodologia | F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | *MeanScore* | $\sigma$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DT Default | 0,808 | 0,801 | 0,798 | 0,792 | 0,803 | 0,804 | 0,785 | 0,810 | 0,814 | 0,807 | 0,802 | 0,008 |
| RF Default | 0,852 | 0,859 | 0,850 | 0,856 | 0,862 | 0,865 | 0,868 | 0,856 | 0,871 | 0,871 | 0,861 | 0,007 |
| XGB Default | 0,862 | 0,870 | 0,854 | 0,858 | 0,862 | 0,877 | 0,871 | 0,861 | 0,875 | 0,876 | 0,867 | 0,007 |
| ANN Default | 0,813 | 0,824 | 0,823 | 0,818 | 0,801 | 0,835 | 0,834 | 0,814 | 0,833 | 0,812 | 0,821 | 0,010 |

**Tabela 4. Desempenho dos modelos de predição no conjunto de teste.**

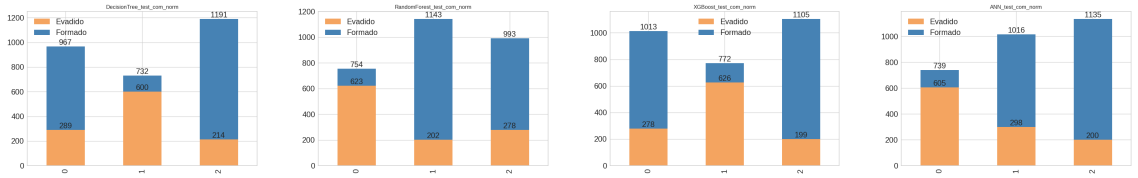
| Classe | Técnica | *precision* | *recall* | *f1-score* | suporte |
|---|---|---|---|---|---|
| Formado | DT | 0,80 | 0,78 | 0,79 | 1787 |
| | RF | 0,84 | 0,89 | 0,86 | |
| | XGB | 0,84 | 0,89 | 0,87 | |
| | ANN | 0,83 | 0,87 | 0,85 | |
| Evadido | DT | 0,66 | 0,68 | 0,67 | 1103 |
| | RF | 0,79 | 0,72 | 0,75 | |
| | XGB | 0,81 | 0,73 | 0,77 | |
| | ANN | 0,78 | 0,71 | 0,74 | |
| | | | | | 2890 |

### 4.2. Explicabilidade

Nesta Seção são apresentados os resultados referentes a explicabilidade dos modelos. Para cada modelo obtido na Seção 4.1, realizou-se o mapeamento do conjunto de teste para um conjunto de amostras que representam explicações para o modelo. Foram aplicados os métodos de extração de atributos e clusterização conforme descritos nas Seções

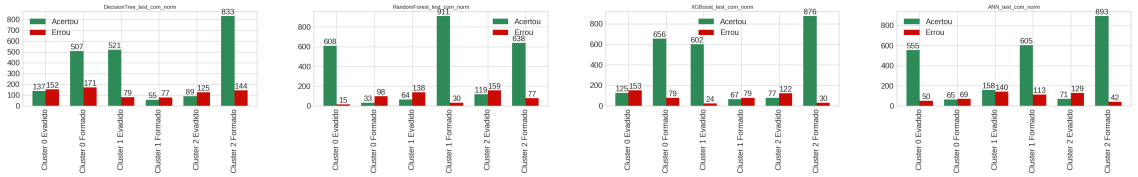
3.4 e 3.5 e, como resultado, cada modelo gerou o total de 3 *clusters*. O mapeamento das explicações com as amostras originais de alunos permite identificar a quantidade de explicações por classe (Formado e Evadido) dentro de cada *cluster*, conforme detalhado nas Figuras 2a, 2b, 2c e 2d. Nota-se que alguns *clusters* explicam melhor alunos Evadidos, enquanto outros, Formados. Como exemplo, destaca-se o *cluster* 0 do modelo *Random Forest* (Figura 2b). Esse *cluster* agrupou 754 explicações, das quais 623 são referentes a alunos que evadiram. Para todas as técnicas obteve-se 2 *clusters* com predomínio de alunos Formados e 1 *cluster* com predomínio de Evadidos.

De forma complementar, também é possível associar as explicações ao desempenho do modelo original, ou seja, destacando os acertos e erros do modelo no escopo de cada *cluster*. As Figuras 3a, 3b, 3c e 3d separam cada *cluster* em dois, exibindo o número de acertos e erros para a classe Evadido e para a classe Formado. Ao observar o *cluster* 2 do XGBoost (3c), por exemplo, identifica-se que o *cluster* 2 agrupou majoritariamente explicações da classe Formado. Ao mesmo tempo, dentre as explicações para Formado, o modelo original acertou a maioria das predições. Por outro lado, as explicações para alunos Evadidos no *cluster* 2, estão associadas a muitos erros do modelo original. Dando continuidade, calculou-se os *centroids* de cada *cluster*, bem como o *centroid* por classe dentro no *cluster*. Esses *centroids* são convertidos para o formato de lista de tuplas. A Figura 4 apresenta a explicação central do *cluster* 0 para o modelo obtido pela técnica RF.




**(a) DT.** **(b) RF.** **(c) XGB.** **(d) ANN.**

**Figura 2. Instâncias por classe em cada *cluster*.**




**(a) DT.** **(b) RF.** **(c) XGB.** **(d) ANN.**

**Figura 3. Acertos e erros para cada classe e *cluster*.**

Com base nas explicações centrais, aplicou-se o método de predição pelas explicações conforme descrito na Seção 3.6. Enquanto as explicações centrais foram geradas a partir do conjunto de teste, os resultados das métricas foram gerados utilizando o conjunto de teste 2. As Tabelas 5, 6, 7 e 8 apresentam os desempenhos das métricas para essa etapa dos experimentos. Para cada técnica são apresentados os desempenhos do modelo, do método de predição pelas explicações e, por último, o desempenho combinado do modelos com as explicações. As métricas são calculadas por classe, portanto, a Tabela está organizada em duas regiões (Formado e Evadido). Importante detalhar o significado da coluna suporte. Conforme a separação dos dados (Seção 3.2), o conjunto teste 2 possui 2.891 amostras (1.788 formados e 1.103 evadidos). Sendo assim, as linhas referentes a

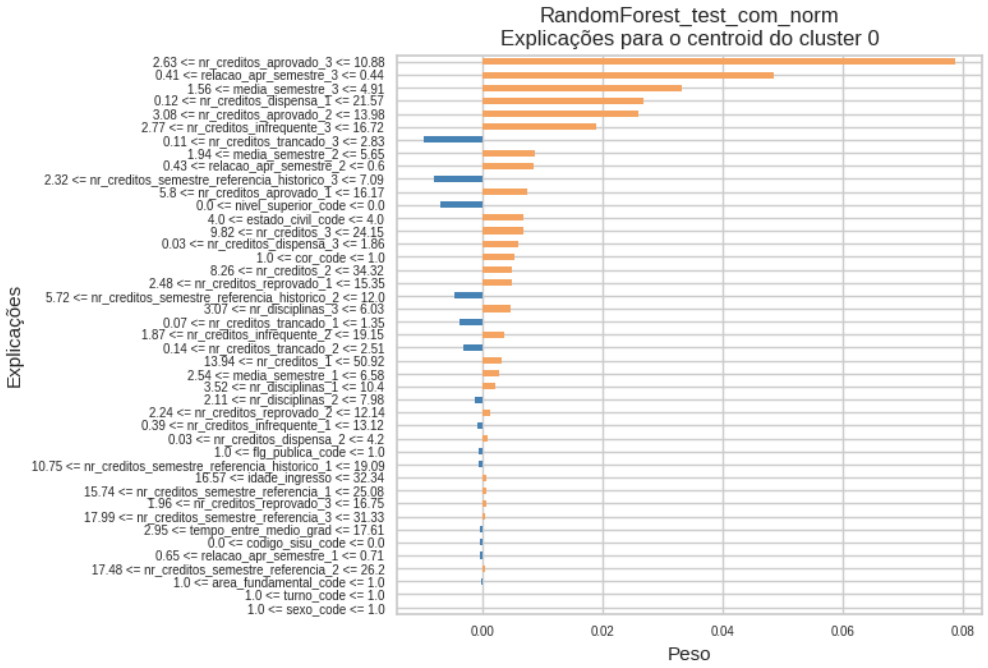



**Figura 4. Explicação central para o *cluster* 0 da técnica *Random Forest*.**

predição do modelo e a predição pelo método das explicações, sempre apresentarão os suportes de 1.788 para classe 0 (Formado) e 1.103 para classe 1 (Evadido). Isso pode ser identificado, por exemplo, nas linhas das técnicas RF e Exp da Tabela 5. Por outro lado, ainda nessa tabela, a técnica RF+Exp refere-se a combinação das duas técnicas. Nesse caso o suporte será menor pois é referente apenas às amostras que as duas técnicas concordam na predição. O percentual de concordância é aqui chamado de *Agreement*. Na tabela 5 é possível identificar que o *Agreement* final para a técnica RF+Exp é de 87,1%, sendo de 86,2% para classe Formado e 86,7% para classe Evadido.

Primeiro ponto a observar é que o método de predição pelas explicações apresentou resultados interessantes embora, na maioria dos casos, inferior ao do modelo original. Por outro lado, quando a resposta do modelo é igual a resposta do método de predição pelas explicações (*Agreement*), obtém-se um novo suporte no qual o desempenho é igual ou superior ao do modelo original. Outra característica identificada é que o *Agreement* é diferente para diferentes técnicas. Conforme as Tabelas 5, 6, 7 e 8, obteve-se novos suportes finais de 1.967 para DT+Exp, 2.509 para RF+Exp, 2.274 para XGBoost+Exp e 1.890 para ANN+Exp. Quando um modelo está respondendo igual ao método de predição pela explicação, considera-se que sua predição é coerente com as explicações centrais obtidas pela clusterização. Por fim, quanto maior o *agreement*, maior a coerência do modelo com as explicações centrais. Ao analisar os resultados obtidos, é possível observar que os modelos das técnicas RF e XGBoost apresentaram desempenho muito similares para as métricas (*precision*, *recall* e *f1-score*) no conjunto de teste 2. Por outro lado, nota-se que o *agreement* do RF+Exp é de 86,7% contra 78,6% do XGBoost+Exp. A técnica ANN+Exp, por outro lado, ficou com *agreement* de 65,3%. Esses números indicam que modelos com desempenho similar em termos de predição podem apresentar características diferentes em relação a sua interpretabilidade.

**Tabela 5. DT - Teste 2.**

| Classe | Técnica | precision | recall | f1-score | suporte | *Agreement* |
|---|---|---|---|---|---|---|
| Formado | DT | 0,82 | 0,79 | 0,80 | 1788 | |
| | Exp | 0,68 | 0,93 | 0,79 | 1788 | |
| | DT + Exp | 0,84 | 0,95 | 0,89 | 1448 | 80,9% |
| Evadido | DT | 0,68 | 0,71 | 0,70 | 1103 | |
| | Exp | 0,73 | 0,29 | 0,42 | 1103 | |
| | DT + Exp | 0,78 | 0,51 | 0,61 | 519 | 47,0% |
| | | | | Novo suporte | 1967 | 68,0% |

**Tabela 6. RF - Teste 2.**

| Classe | Técnica | precision | recall | f1-score | support | *Agreement* |
|---|---|---|---|---|---|---|
| Formado | RF | 0,85 | 0,90 | 0,88 | 1788 | |
| | Exp | 0,82 | 0,82 | 0,82 | 1788 | |
| | RF+Exp | 0,86 | 0,92 | 0,89 | 1558 | 87,1% |
| Evadido | RF | 0,83 | 0,74 | 0,78 | 1103 | |
| | Exp | 0,71 | 0,71 | 0,71 | 1103 | |
| | RF+Exp | 0,85 | 0,76 | 0,80 | 951 | 86,2% |
| | | | | Novo suporte | 2509 | 86,7% |

**Tabela 7. XGB - Teste 2.**

| Classe | Técnica | precision | recall | f1-score | support | *Agreement* |
|---|---|---|---|---|---|---|
| Formado | XGB | 0,86 | 0,90 | 0,88 | 1788 | |
| | Exp | 0,81 | 0,74 | 0,77 | 1788 | |
| | XGB + Exp | 0,87 | 0,91 | 0,89 | 1375 | 76,9% |
| Evadido | XGB | 0,82 | 0,76 | 0,79 | 1103 | |
| | Exp | 0,63 | 0,72 | 0,67 | 1103 | |
| | XGB + Exp | 0,86 | 0,79 | 0,82 | 899 | 81,5 |
| | | | | Novo suporte | 2274 | 78,6% |

**Tabela 8. ANN - Teste 2.**

| Classe | Técnica | precision | recall | f1-score | support | *Agreement* |
|---|---|---|---|---|---|---|
| Formado | ANN | 0,84 | 0,89 | 0,86 | 1788 | |
| | Exp | 0,83 | 0,50 | 0,63 | 1788 | |
| | ANN + Exp | 0,87 | 0,84 | 0,85 | 1025 | 57,3% |
| Evadido | ANN | 0,80 | 0,72 | 0,75 | 1103 | |
| | Exp | 0,51 | 0,84 | 0,63 | 1103 | |
| | ANN + Exp | 0,82 | 0,85 | 0,83 | 865 | 78,4% |
| | | | | Novo suporte | 1890 | 65,3% |

## 5. Conclusões

Este trabalho propôs investigar aspectos de interpretabilidade de modelos de predição no contexto da evasão de alunos no ensino superior. Um conjunto de dados demográficos e de desempenho acadêmico foi elaborado especificamente para esta pesquisa, contendo 14.452 instâncias de alunos de diferentes cursos de graduação. As técnicas DT, RF, XGB e ANN foram utilizadas para obter os modelos de predição. A clusterização das explicações permitiu identificar que alguns *clusters* estão relacionados predominantemente a explicações da classe Formado, enquanto outros, a explicações da classe Evadido. Identificou-se também que o desempenho dos modelos é diferente no contexto de cada *cluster*. O *centroid* de cada *cluster* permitiu gerar explicações centrais, que representam explicações chave. O uso dessas explicações centrais demonstrou a viabilidade de predição de novas amostras por meio da clusterização de explicações, respondendo a questão de pesquisa Q1. A concordância entre modelo e predição pelas explicações (*agreement*) permitiu identificar regiões em que o modelo apresenta melhor desempenho para as métricas *precision*, *recall* e *f1-score*, respondendo a questão Q3. Ao mesmo tempo, a concordância entre modelo e predição pela explicação, contribui para avaliar a explicação de uma determinada amostra. Respondendo a questão Q2, portanto, quando o modelo faz uma predição similar a predição com base nas explicações, significa que o modelo está sendo coerente com suas explicações centrais. Por fim, com base nos experimentos apresentados, acredita-se que é possível distinguir os modelos em relação a similaridade de suas predições com as predições pelo método das explicações. Como métrica possível, e resposta para questão Q4, apresentou-se o percentual de concordância (*agreement*) do modelo com as predições pelas explicações. Quanto maior o *agreement* maior a coerência entre modelo e explicações. Destaca-se, por fim, a contribuição deste trabalho em ir além do simples uso de métodos explicativos, buscando analisar um conjunto de explicações com o objetivo de comparar modelos de predição sob aspectos de interpretabilidade. Como trabalho futuro, pretende-se avaliar a relação entre ajuste fino de hiperparâmetros de técnicas de AM e o seu impacto nas métricas de interpretabilidade. O material gerado pelo estudo pode ser consultado no GitHub[15].

[15] https://github.com/cassiocarvalho/clustering-lime-explanations

## Referências


Al-Jallad, N., Ning, X., Khairalla, M., and Al-Qaness, M. (2019). Rule mining models for predicting dropout/ stopout and switcher at college using satisfaction and ses features. *International Journal of Management in Education*, 13(2):97–118.

Alharbi, B. (2022). Back to basics: An interpretable multi-class grade prediction framework. *Arabian Journal for Science and Engineering*, 47(2):2171–2186.

Alwarthan, S., Aslam, N., and Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, 10:107649–107668.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., and Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23(1):537–553.

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Kumar, P. and Sharma, M. (2020). Predicting academic performance of international students using machine learning techniques and human interpretable explanations using lime—case study of an indian university. *Advances in Intelligent Systems and Computing*, 1087:289–303.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition.

Pei, B. and Xing, W. (2022). An interpretable pipeline for identifying at-risk students. *Journal of Educational Computing Research*, 60(2):380–405.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Vultureanu-Albisi, A. and Badica, C. (2021). Improving students' performance by interpretable explanations using ensemble tree-based approaches. In *Proceedings of SACI 2021 - IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, pages 215–220.

Xiao, W., Ji, P., and Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, 4(5):e12482.