VI Congresso sobre Tecnologias na Educação (Ctrl+E 2021)
Educação com tecnologia: pessoas, didática e currículo integrados
Pau dos Ferros – RN | Brasil | Online
24 a 27 de agosto de 2021

Técnicas de Machine Learning para Predição do Tempo de Permanência na Graduação no Âmbito do Ensino Superior Público Brasileiro

Ebony M. Rodrigues, Roberta M. M. Gouveia

Departamento de Estatística e Informática – DEINFO
Universidade Federal Rural de Pernambuco – UFRPE
R. Dom Manuel de Medeiros, s/n, Dois Irmãos – 52.171-900 – Recife – PE – Brasil

{ebony.marquesr,roberta.gouveia}@ufrpe.br

Abstract. *This article deals with the use of techniques of the Knowledge Discovery in Databases and Cross Industry Standard Process for Data Mining processes on educational databases made available by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (National Institute of Educational Studies and Research Anísio Teixeira) in order to enable the discovery of knowledge about the students' permanence time in undergraduate courses at Brazilian public higher education institutions. For this, Supervised Machine Learning methods were used to build models based on Decision Tree, Random Forest, XGBoost and Neural Network algorithms. XGBoost models stood out in all the experiments performed.*

Resumo. *Este artigo trata do uso de técnicas dos processos de Knowledge Discovery in Databases e Cross Industry Standard Process for Data Mining sobre bases de dados educacionais disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira visando possibilitar a descoberta de conhecimentos acerca do tempo de permanência de discentes em cursos de graduação de instituições de ensino superior públicas brasileiras. Para isso, métodos do Aprendizado de Máquina Supervisionado foram utilizados na construção de modelos baseados em algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural. Os modelos XGBoost destacaram-se em todos os experimentos executados.*

1. Introdução

O Exame Nacional de Desempenho dos Estudantes (ENADE) é um instrumento utilizado para avaliar o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral de concluintes de cursos de graduação no Brasil. Realizado com a observação de um ciclo que compreende três anos como período de avaliação, o exame é aplicado, a cada ano, para estudantes de cursos de áreas do conhecimento determinadas [INEP 2021c]. Além da prova propriamente dita, o ENADE possui um questionário que tem o objetivo de coletar informações que possibilitem caracterizar os perfis socioeconômicos dos estudantes, tal como os contextos de seus processos formativos [INEP 2021b].

O Censo da Educação Superior é uma ferramenta de pesquisa sobre as instituições de ensino superior (IES) brasileiras. O Censo tem o objetivo de subsidiar o Ministério da

Educação na concretização de suas atribuições ao permitir a compreensão do sistema brasileiro de educação superior. Além disso, o Censo contribui para os trabalhos de gestores do governo, de entidades públicas e privadas, de instituições de ensino, pesquisadores e especialistas brasileiros e estrangeiros e de organismos internacionais [INEP 2021a].

Este estudo tem como objetos dados das edições de 2016, 2017 e 2018 do ENADE [INEP 2020b] e do Censo da Educação Superior [INEP 2020a] – quanto ao Censo, dados constantes nas bases de IES, cursos de graduação e docentes –, empregados para possibilitar a descoberta de conhecimentos através do uso de técnicas de análise exploratória e de mineração de dados, considerando o Aprendizado de Máquina Supervisionado, a partir da observação dos métodos de *Knowledge Discovery in Databases* (KDD) e *Cross Industry Standard Process for Data Mining* (CRISP-DM). O ciclo de 2016 a 2018 do ENADE foi selecionado por ser o mais recente concluído até o início do trabalho, bem como os dados do Censo da Educação Superior no triênio foram usados por serem os mais recentes disponíveis até a data em questão.

A motivação deste trabalho está baseada no interesse em adquirir o respaldo científico necessário para evidenciar a possibilidade de descobrir conhecimentos, por meio da detecção de padrões e de regras significativas e não triviais, empregando os conjuntos de dados abertos educacionais observados, de maneira a permitir uma melhor compreensão acerca do tempo de permanência de discentes em cursos de graduação de instituições de ensino superior públicas federais e estaduais brasileiras.

De forma específica, este estudo interdisciplinar possui o objetivo de aplicar métodos do Aprendizado de Máquina Supervisionado para considerar os perfis socioeconômicos de concluintes de cursos de graduação brasileiros, de graus bacharelado e licenciatura, de IES federais e estaduais, abordando os distintos períodos percebidos entre o ano de início da graduação e o ano de realização do ENADE pelos estudantes, com a elaboração de quatro experimentos que envolvem a construção de 16 modelos de classificação baseados em algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural.

2. Ferramentas e método

Técnicas do processo de KDD, do CRISP-DM e de análise exploratória de dados constituem o método empregado neste estudo. Visando à concretização das atividades tratadas a seguir, a versão 3.8.3 da linguagem de programação Python foi usada, tal como as versões mais recentes das bibliotecas de Python Numpy, Seaborn, Matplotlib, Pandas, Pandas Profiling, Feature Engine, Scikit-Learn, Imbalanced-Learn e XGBoost disponíveis durante a execução do trabalho, além da ferramenta RapidMiner Studio[1], em sua versão 9.8, ativada com licença educacional.

Segundo Fayyad et al. (1996), o conhecimento é o produto final de uma descoberta baseada em dados. De acordo com os autores, o processo de KDD abrange um conjunto de cinco etapas que visam à descoberta de conhecimentos úteis a partir de dados, o que caracteriza um método não trivial de identificação de padrões novos e válidos. As etapas compreendem tarefas de seleção, pré-processamento, transformação e mineração de dados e interpretação e avaliação de resultados. O processo é não trivial porque é iterativo, já que cada etapa depende da finalização da etapa anterior, que pode ser repetida quantas vezes for preciso até que os dados estejam adequados para uso nas etapas subsequentes.

[1] A página oficial do RapidMiner Studio pode ser acessada em https://bit.ly/2W4tF1V.

De forma distinta da metodologia de KDD, que objetiva a descoberta de conhecimentos com base nos dados propriamente ditos, o CRISP-DM tem como objetos questões de negócio que baseiam a geração de modelos de aprendizado de máquina com a consideração de seis etapas, que tratam de entendimento do negócio, entendimento, preparação e modelagem de dados e avaliação e implantação de modelos [Shearer 2000]. Esse método possui etapas de modelagem e avaliação explícitas, o que é relevante para este estudo.

Muito utilizado para mineração de dados em aplicações que dependem do entendimento do negócio observado, o CRISP-DM costuma ser empregado, também, no contexto educacional, tendo, inclusive, recebido uma proposta de adaptação, desenvolvida por pesquisadores brasileiros, para basear a mineração na área [Ramos et al. 2020].

A análise exploratória compreende uma etapa fundamental em trabalhos com bases de dados que, em linhas gerais, trata de organizar, resumir, aplicar métodos estatísticos e visualizar os dados considerados. Apresentada em 1977, a análise exploratória destaca as principais características dos dados por meio de métodos visuais que, entre outras finalidades, possibilitam a visualização de tendências [Tukey 1977].

3. Seleção, pré-processamento e transformação de dados

As três primeiras etapas do processo de KDD tratam da execução de atividades de seleção, pré-processamento e transformação dos dados observados para a efetuação da descoberta de conhecimentos [Fayyad et al. 1996].

A primeira etapa, denominada seleção de dados, é crucial, pois consiste em selecionar um ou mais subconjuntos de dados de interesse (*target data*) para basear a mineração a partir das bases de dados consideradas [Fayyad et al. 1996].

Como exposto anteriormente, este estudo compreende dados das edições de 2016, 2017 e 2018 do ENADE. De acordo com os dados, houve 306.760 inscrições de estudantes de IES públicas federais e estaduais brasileiras no período, sendo que 84,9% (260.534) estão associadas à presenças válidas e 69,1% (211.950) designam presenças válidas de estudantes de cursos de graduação de graus bacharelado e licenciatura, cuja diferença entre o ano de início da graduação e o ano de realização do ENADE é de 2 anos ou mais, que não apresentavam inconsistências após a realização das tarefas de identificação e tratamento abordadas a seguir. O subconjunto de 211.950 registros do ENADE em questão foi considerado para a execução dos experimentos de aprendizado de máquina neste trabalho.

As bases de dados das edições de 2016, 2017 e 2018 do ENADE contêm, respectivamente, 141, 150 e 137 atributos, enquanto a base de dados de 2018 das Instituições de Ensino Superior do Censo da Educação Superior contém 48 atributos. Tendo em mente o objetivo da primeira etapa do processo de KDD, 30 atributos[2] do ENADE, constantes nas bases das três edições, foram selecionados, a partir da observação dos atributos considerados mais relevantes pelo RapidMiner Studio, para compor o subconjunto de dados usado para embasar os experimentos executados após operações de pré-processamento e transformação de dados. Os atributos de IES, da base do Censo, foram utilizados na análise exploratória que fundamentou a definição dos experimentos em questão.

O RapidMiner Studio foi empregado neste trabalho pois possui um módulo nativo para seleção automática de atributos de um conjunto de dados de entrada, tendo em vista

[2]Os atributos originais do ENADE selecionados podem ser vistos em https://bit.ly/3k2yQaV.

os seus graus de correlação e estabilidade. As análises realizadas pela ferramenta tratam da identificação de atributos que contêm variados valores distintos, que têm diversos valores idênticos e que espelham de forma muito próxima o atributo ou variável classe do experimento considerado, quando o aprendizado de máquina supervisionado é observado [Mierswa et al. 2006].

Dentre os 30 atributos do ENADE considerados, há informações sobre cada participante, como sua faixa etária no ato de realização do exame, sexo, ano de conclusão do ensino médio, ano de início da graduação etc.; a Instituição de Ensino Superior vinculada ao participante, a exemplo da categoria administrativa, organização acadêmica, localização, questões de infraestrutura etc.; o curso de graduação do participante, a exemplo de sua área, modalidade, turno etc.; e, por fim, as respostas do participante para o questionário socioeconômico, acerca de seu estado civil no momento de realização do exame, cor/raça, renda familiar, motivo de escolha do curso, se o seu ingresso ocorreu por meio de políticas de ação afirmativa ou inclusão social etc.

A segunda etapa do processo de KDD, denominada pré-processamento de dados, trata de executar processamentos gerais sobre os atributos, bem como de analisar os dados selecionados na primeira etapa buscando identificar dados ausentes e outras inconsistências [Fayyad et al. 1996]. Os dados inconsistentes devem ser corrigidos ou removidos de maneira a possibilitar a sequência do estudo e não comprometer a qualidade dos modelos de aprendizado de máquina a serem construídos.

As operações executadas na segunda etapa do estudo consistem na padronização da estrutura de determinados atributos — como acontece, por exemplo, com os atributos que tratam do turno do curso de graduação do estudante — que apresentam divergências nos conjuntos das três edições utilizadas. Algumas alterações gerais também foram realizadas sobre todos os dados, envolvendo, por exemplo, a tarefa de renomear os atributos para tornar seus nomes intuitivos. Além disso, um conjunto de análises ou verificações específicas para a identificação de inconsistências foi executado sobre os dados, bem como os seus respectivos tratamentos. Registros que apresentam inconsistências quanto ao ano de conclusão do ensino médio, ano de início da graduação, ano de realização do ENADE ou idade do estudante foram corrigidos ou removidos, a depender da verificação em questão. Os registros que contêm dados ausentes foram removidos.

Uma das verificações das quais o parágrafo anterior trata compreende a identificação de registros onde a idade do estudante é igual ou menor ao período de diferença entre o ano de conclusão do ensino médio e o ano de início da graduação. Quando um registro apresenta tal inconsistência, o ano de conclusão do ensino médio é corrigido com o uso da média da variável, observando dados de estudantes de faixa etária semelhante.

A terceira etapa do método de KDD, denominada transformação de dados, designa a análise dos dados percebidos ao término da etapa anterior para subsequente reorganização e formatação, observando os objetivos do trabalho [Fayyad et al. 1996]. Nesta etapa, novas informações podem ser adicionadas à base através da transformação ou criação de atributos. Atributos originais podem ser alterados com o uso de técnicas de discretização, *undersampling*, binarização e estratificação de dados, e também podem ser substituídos por atributos criados por meio de operações que envolvam atributos originais.

A partir de alguns dos 30 atributos do ENADE selecionados na primeira etapa do

processo e de atributos constantes nas bases de IES, de docentes e de cursos de graduação do Censo da Educação Superior, três atributos foram criados, três foram removidos e vários foram transformados. Os atributos criados têm a diferença entre o ano de conclusão do ensino médio e o ano de início da graduação, a diferença entre o ano de início da graduação e o ano de realização do ENADE e, por fim, o grau acadêmico do curso de graduação. Os atributos que continham o ano de realização do ENADE, o ano de conclusão do ensino médio e o ano de início da graduação, que basearam a criação de dois atributos, foram removidos. Um dos atributos transformados possuía a área de formação geral do curso e passou a ter a grande área do conhecimento do curso, com base na Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica (Cine Brasil) [GOV.BR 2021].

Alguns atributos numéricos constantes na base, que têm como categorias valores contínuos, foram discretizados para equilibrar as frequências de suas observações. A discretização trata de apresentar os dados de maneira alternativa através de intervalos ou faixas de valores. Com isso, atributos originalmente contínuos tornam-se discretos, ou seja, passam a ter quantidades específicas e limitadas de categorias [Feature-Engine 2020]. A biblioteca de Python Feature Engine — em especial, sua classe *EqualFrequencyDiscretiser* — foi empregada na discretização de dados neste trabalho.

A técnica de *undersampling*, que trata da remoção de determinada quantidade de observações das categorias predominantes dos atributos, de maneira que todas as categorias consideradas passem a ter a mesma quantidade de observações, foi executada sobre a variável dependente dos experimentos de classificação de dados, apresentada em breve [Bilogur 2018]. A classe *RandomUnderSampler* da biblioteca Imbalanced-Learn foi utilizada nesta tarefa.

Considerando que alguns dos modelos de aprendizado de máquina supervisionado construídos não conseguem lidar com dados categóricos de maneira direta, os 30 atributos observados foram *binarizados* por meio de variáveis *dummy*, sendo, então, representados de forma numérica [Scikit-Learn 2021d]. Essa operação foi realizada com o emprego da classe *OneHotEncoder* da biblioteca Scikit-Learn.

A estratificação de dados envolve a separação de um conjunto de dados em conjuntos menores, de forma que cada conjunto menor seja uma representação justa do conjunto total. A observação da operação é importante para que a mineração seja feita de maneira correta, utilizando as devidas proporções de dados nos treinamentos e testes dos modelos [Scikit-Learn 2021c]. A estratificação foi realizada sobre a variável dependente do cenário de mineração, com o uso da classe *StratifiedKFold* da biblioteca Scikit-Learn.

4. Experimentos de Aprendizado de Máquina

A quarta etapa do processo de KDD, denominada mineração de dados, abrange a execução de métodos que podem basear a descoberta de conhecimentos [Fayyad et al. 1996]. Nesta etapa, os dados anteriormente tratados são empregados para a identificação de padrões e correlações, objetivando a geração de informações úteis.

Este trabalho tratou da execução de metas preditivas através da construção e avaliação de modelos de classificação baseados em algoritmos de *Decision Tree* – Árvore de Decisão –, *Random Forest* – Floresta Aleatória –, *eXtreme Gradient Boosting* (XGBoost)

– Aumento de Gradiente Extremo, em tradução livre – e *Neural Network* – Rede Neural. Os modelos foram construídos por meio da linguagem de programação Python e das bibliotecas de Python Scikit-Learn e XGBoost.

O método de Árvore de Decisão designa um fluxograma em forma de árvore onde pode-se classificar dados ao seguir as ligações entre os nós da estrutura por meio de decisões sobre valores [Quinlan 1986]. Neste estudo, os modelos construídos com base no método consideram o classificador *DecisionTreeClassifier* da biblioteca Scikit-Learn e os seus parâmetros padrões.

O método denominado Floresta Aleatória usa um conjunto de árvores de decisão em seu processo de classificação. A floresta, estrutura que resulta da combinação de várias árvores, busca aprimorar os resultados obtidos na classificação. O principal hiperparâmetro do método é o número de árvores empregadas [Breiman 2001]. Os modelos de Floresta Aleatória construídos neste trabalho baseiam-se no classificador *RandomForestClassifier* da biblioteca Scikit-Learn e observam os seus parâmetros padrões, com o número de árvores igual a 100.

O método XGBoost caracteriza um sistema de aumento de árvore (*tree boosting*) escalonável de ponta a ponta que, empregando menos recursos do que os outros sistemas, com a otimização de hardware e software, promete retornar resultados bastante superiores [Chen and Guestrin 2016]. Os modelos de XGBoost construídos neste estudo baseiam-se no classificador *XGBClassifier* da biblioteca XGBoost com os seus parâmetros padrões.

Uma rede neural perceptron multicamadas compreende um conjunto de unidades de entrada e saída conectadas por sinapses com pesos associados. O modelo é capaz de aprender com dados de entrada para classificar dados futuros [Popescu et al. 2009]. Para a criação de redes neurais, o classificador *MLPClassifier* da biblioteca Scikit-Learn, com a observação dos parâmetros padrões, foi utilizado.

Os modelos construídos neste estudo foram treinados e avaliados a partir do método de validação cruzada (*cross-validation*) de 5 grupos. Esse procedimento de reamostragem é utilizado na avaliação de modelos com base em uma amostra de dados limitada, que é empregada para estimar o desempenho do modelo quando usado para prever sobre dados não observados durante o treinamento. O único parâmetro do método é a quantidade de grupos em que a amostra deve ser dividida [Scikit-Learn 2021a].

O cenário de mineração de dados considerado neste estudo tem o atributo ou variável que armazena a diferença entre o ano de início da graduação e o ano de realização do ENADE como dependente. 30 variáveis[3], incluindo a variável dependente, resultantes das etapas anteriores à mineração, compõem o conjunto de dados que baseia este cenário. Além da identificação do melhor algoritmo para predição do tempo de graduação entre os quatro algoritmos observados, este cenário objetiva prever, de forma indireta, a retenção estudantil no âmbito do ensino superior público brasileiro.

Quatro experimentos, que distinguem-se quanto às categorias possíveis para a variável dependente e quanto ao balanceamento dos dados dessa variável, foram executados a partir do cenário exposto e são apresentados a seguir. Em cada um dos experimentos, quatro modelos de aprendizado de máquina supervisionado foram construídos, através de

[3]Os atributos percebidos após a etapa de transformação podem ser vistos em https://bit.ly/3iZ5wCK.

algoritmos de Árvore de Decisão, Floresta Aleatória, XGBoost e Rede Neural, para serem avaliados com base em métricas tratadas na próxima seção.

Sobre as categorias da variável dependente, os experimentos 1 e 2 observam duas classes – 'Entre 2 e 4 anos' e '5 anos ou mais' –, enquanto os experimentos 3 e 4 consideram três – 'Entre 2 e 4 anos', 'Entre 5 e 7 anos' e '8 anos ou mais' –, como exposto na Tabela 1. Sobre os experimentos 1 e 2, cada um deles emprega um conjunto de dados distinto. O Experimento 1 usa um conjunto desbalanceado, com todos os dados possíveis. O Experimento 2, por sua vez, observa um conjunto de dados balanceado, que possui os dados balanceados por classe a partir da classe com a menor quantidade de observações. Para os experimentos 3 e 4, a situação quanto aos conjuntos de dados empregados é análoga. Dessa maneira, os modelos construídos nos experimentos 1 e 3 foram treinados e testados com o conjunto de dados desbalanceado, enquanto os modelos dos experimentos 2 e 4 empregaram o conjunto de dados balanceado nessas tarefas.

Tabela 1. Quantidades de observações dos conjuntos com dados desbalanceados e balanceados empregados nos experimentos 1, 2, 3 e 4.

Classe	Qtde. de observações do conjunto desbalanceado	Qtde. de observações o conjunto balanceado
1: Entre 2 e 4 anos 0: 5 anos ou mais	118.646 93.304 **(Experimento 1)**	93.304 93.304 **(Experimento 2)**
1: Entre 2 e 4 anos 2: Entre 5 e 7 anos 0: 8 anos ou mais	118.646 84.104 9.200 **(Experimento 3)**	9.200 9.200 9.200 **(Experimento 4)**

Sabe-se que a abordagem tida como correta para a implementação de experimentos de classificação envolve o balanceamento de classes da variável dependente na fase de treino e a utilização de todos os dados possíveis na fase de teste. Apesar disso, decidiu-se executar os experimentos da forma detalhada acima para que fosse possível visualizar as consequências da observação das distintas abordagens de treino e teste tratadas.

5. Interpretação e avaliação de resultados

Bem como as etapas anteriores, a última etapa do processo de KDD é bastante importante, pois compreende as tarefas de interpretação e avaliação dos resultados apresentados pelos modelos construídos observando as regras empregadas [Fayyad et al. 1996]. Métricas de avaliação são instrumentos usados nesta etapa, que possibilitam mensurar e comparar os resultados obtidos pelos modelos. Quatro métricas foram consideradas neste estudo e são tratadas a seguir: acurácia, precisão, *recall* e *f1-score*.

A acurácia caracteriza a média de predições identificadas de forma correta. É justo dizer que essa medida é sensível a desbalanceamentos do conjunto de dados empregado, o que pode induzir a conclusões equivocadas sobre o desempenho do modelo de aprendizado considerado [GoogleDevelopers 2021a]. A métrica de precisão, por sua vez, associa verdadeiros e falsos positivos e retorna a proporção de identificações positivas corretas. Já a métrica de *recall* relaciona os verdadeiros e falsos negativos [GoogleDevelopers 2021b]. Por fim, a *f1-score*, ao contrário da acurácia, considera tanto os falsos positivos quanto os

falsos negativos e é indicada na tentativa de não enviesar a avaliação para as classes mais populosas quando há distribuição desigual de classes [Scikit-Learn 2021b].

A Figura 1 expõe os resultados[4] apresentados pelos modelos construídos nos quatro experimentos executados. As tabelas 2 e 3 contêm detalhes do aprendizado dos modelos XGBoost dos experimentos 1 e 4. A Figura 2 exibe a matriz de confusão do modelo XGBoost do Experimento 1, enquanto a Figura 3 apresenta o gráfico de radar desse experimento. As categorias da variável dependente são representadas na matriz de confusão de acordo com os valores expostos na Tabela 1.

Considerando as métricas antes tratadas, os modelos baseados em XGBoost retornaram os melhores resultados em todos os quatro experimentos. Nota-se que, observando o emprego de duas classes no aprendizado, o modelo XGBoost do Experimento 1 teve o melhor desempenho, enquanto o modelo XGBoost do Experimento 4 apresentou o melhor desempenho quando considerou-se o uso de três classes. Esses modelos foram treinados e testados, a partir do método de validação cruzada, em 3 minutos e 30 segundos e 1 minuto e 18 segundos, respectivamente. A média macro foi considerada no cálculo dos resultados. É justo destacar que os modelos de Rede Neural e Floresta Aleatória tiveram resultados próximos aos dos modelos XGBoost nos experimentos 1 e 4.

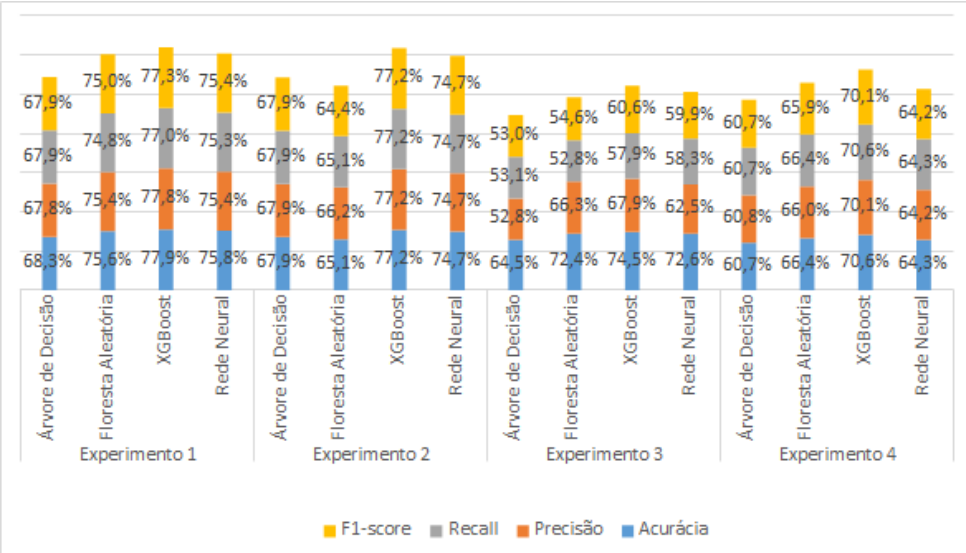

Figura 1. Resultados de classificação dos modelos construídos.

Tabela 2. Detalhes do aprendizado do modelo XGBoost - Experimento 1.

Classe	Precisão	*Recall*	*F1-score*
Entre 2 e 4 anos	78,13%	83,92%	80,92%
5 anos ou mais	77,43%	70,12%	73,59%

Tabela 3. Detalhes do aprendizado do modelo XGBoost - Experimento 4.

Classe	Precisão	*Recall*	*F1-score*
Entre 2 e 4 anos	71,51%	73,54%	72,51%
Entre 5 e 7 anos	65,34%	53,61%	58,90%
8 anos ou mais	73,56%	84,68%	78,73%

6. Conclusão

Nota-se que, no âmbito dos experimentos 1 e 2, os modelos XGBoost conseguiram compreender de maneira satisfatória as características dos estudantes que perceberam um pe-

[4]Os detalhes do aprendizado, as matrizes de confusão e os períodos de treino e teste de todos os 16 modelos, tal como os gráficos de radar dos quatro experimentos, podem ser vistos em https://bit.ly/3g9mH2L.

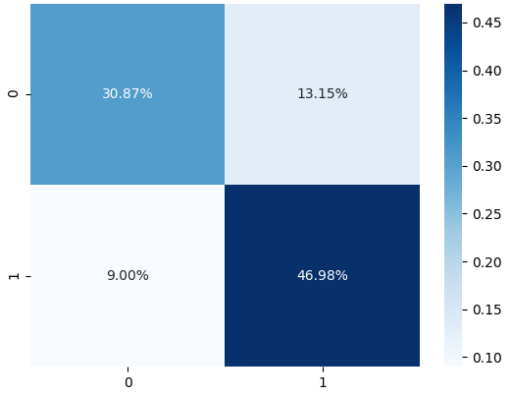

Figura 2. Matriz de confusão do modelo XGBoost - Experimento 1.

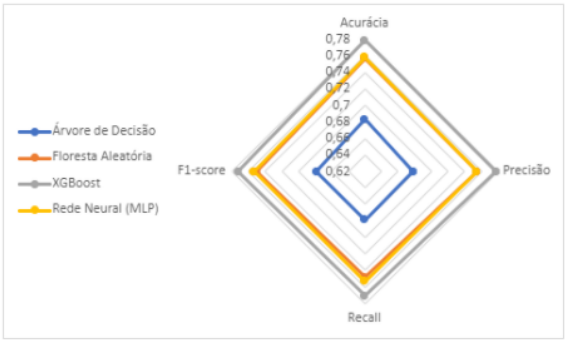

Figura 3. Gráfico de radar - Experimento 1.

ríodo de cerca de 2 a 4 anos, assim como de 5 anos ou mais, entre o início da graduação e o ano de realização do ENADE. No contexto dos experimentos 3 e 4, os modelos XGBoost compreenderam de forma adequada as características dos estudantes que levaram cerca de 2 a 4 anos, assim como compreenderam bem o perfil dos estudantes que levaram 8 anos ou mais, entre os momentos em questão.

As observações expostas indicam a possibilidade da descoberta de conhecimentos por meio dos dados, ferramentas e método, destacando o algoritmo XGBoost, empregados neste estudo. Diante disso, é justo considerar a importância e estimular a continuidade de trabalhos como este, que objetivam a reflexão e uma melhor compreensão sobre a permanência de estudantes no ensino superior brasileiro e que podem possibilitar a elaboração de propostas de intervenção institucionais e políticas públicas que reduzam os índices de retenção e evasão em universidades públicas de todo o Brasil.

Referências

Bilogur, A. (2018). Undersampling and oversampling imbalanced data. `https://bit.ly/3vG48Zw`. Acesso em 10 de janeiro de 2021.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37.

Feature-Engine (2020). EqualFrequencyDiscretiser. `https://bit.ly/34wskBO`. Acesso em 20 de novembro de 2020.

GoogleDevelopers (2021a). Classification: Accuracy. `https://bit.ly/2RaOt5G`. Acesso em 15 de fevereiro de 2021.

GoogleDevelopers (2021b). Classification: Precision and Recall. `https://bit.ly/3uBs9zJ`. Acesso em 15 de fevereiro de 2021.

GOV.BR (2021). Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica (Cine Brasil). https://bit.ly/3fInU1g. Acesso em 12 de janeiro de 2021.

INEP (2020a). Censo da Educação Superior: Microdados. https://bit.ly/3fAh2mf. Acesso em 13 de agosto de 2020.

INEP (2020b). ENADE: Microdados. https://bit.ly/3g21SFB. Acesso em 13 de agosto de 2020.

INEP (2021a). Censo da Educação Superior. https://bit.ly/34A6oWo. Acesso em 12 de janeiro de 2021.

INEP (2021b). ENADE: questionário do estudante. https://bit.ly/3p8z8PB. Acesso em 13 de janeiro de 2021.

INEP (2021c). Exame Nacional de Desempenho dos Estudantes. https://bit.ly/3i2yL7H. Acesso em 12 de janeiro de 2021.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 935–940, New York, NY, USA. Association for Computing Machinery.

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., and Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Trans. Cir. and Sys.*, 8(7):579–588.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.*, 1(1):81–106.

Ramos, J., Rodrigues, R., Silva, J., and Oliveira, P. (2020). 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*, pages 1092–1101, Porto Alegre, RS, Brasil. SBC.

Scikit-Learn (2021a). Cross-validation: evaluating estimator performance. https://bit.ly/3fVPbfv. Acesso em 8 de fevereiro de 2021.

Scikit-Learn (2021b). sklearn.metrics.f1-score. https://bit.ly/3yTCp9K. Acesso em 15 de fevereiro de 2021.

Scikit-Learn (2021c). sklearn.model-selection.StratifiedKFold. https://bit.ly/3g1Uhqo. Acesso em 12 de fevereiro de 2021.

Scikit-Learn (2021d). sklearn.preprocessing.OneHotEncoder. https://bit.ly/2SHi4Ei. Acesso em 10 de fevereiro de 2021.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4).

Tukey, J. W. (1977). *Exploratory Data Analysis*. Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, Mass.