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

Submission: dd/Mmm/yyyy; 1st round notif.: dd/Mmm/yyyy; New version: dd/Mmm/yyyy; 2nd round notif.: dd/Mmm/yyyy;
Camera ready: dd/Mmm/yyyy; Edition review: dd/Mmm/yyyy; Available online: dd/Mmm/yyyy; Published: dd/Mmm/yyyy;

Modelagem Preditiva do Tempo de Permanência em Cursos de Computação: Uma Abordagem com Random Forest Comparando Regiões e o Cenário Nacional

Title: Predictive Modeling of Length of Stay in Computer Science Courses: A Random Forest Approach Comparing Regions and the National Scenario

Título: Modelado Predictivo de la Duración de la Estancia en los Cursos de Informática: un Enfoque de Random Forest que Compara Regiones y el Escenario Nacional

Caique Alves de Souza
Universidade de São Paulo
caiquealves@usp.br

Gabriel Almeida Correa
Universidade de São Paulo
correa_gabriel@usp.br

Gustavo Oliveira Souza
Universidade de São Paulo
gustavooliveira7@usp.br

William Jun Okinaka Suzuki
Universidade de São Paulo
willsuzuki@usp.br

Resumo

O tempo de permanência nos cursos de graduação, especialmente em áreas de alta demanda como a Computação, é um indicador fundamental da qualidade acadêmica da instituição, do bem-estar estudantil, de políticas de permanência e de características socioeconômicas pertinentes à região da instituição e ao próprio discente. A evasão, tanto definitiva quanto temporária, e a diplomação tardia representam custos significativos para alunos e instituições. Este artigo busca entender a capacidade de modelos de Machine Learning em prever o tempo de permanência de estudantes em cursos de Computação no Brasil. O objetivo principal é desenvolver modelos baseados nos algoritmos Random Forest e XGBoost, comparar a eficácia destes modelos entre si e comparar a eficácia destes modelos quando treinados com diferentes conjuntos de dados (conjunto nacional e conjuntos por região). A abordagem metodológica se baseia na análise dos dados disponibilizados pelo Exame Nacional de Desempenho dos Estudantes (ENADE), realizando uma comparação entre as métricas de um modelo preditivo treinado com dados de todo o Brasil e modelos treinados individualmente para cada região. Espera-se que os modelos alcancem alta performance preditiva, e que a análise comparativa entre os escopos nacional e regional forneça conhecimento valioso sobre como o tempo de permanência é afetado pelos fatores característicos de cada região do Brasil. Os resultados têm o potencial de auxiliar gestores educacionais na elaboração de estratégias de permanência para acompanhamento e retenção de estudantes.

Abstract

Length of stay in undergraduate programs, especially in high-demand fields such as Computer Science, is a fundamental indicator of an institution's academic quality, student well-being, retention policies, and socioeconomic characteristics relevant to the institution's region and the individual student. Dropout, both permanent and temporary, and delayed graduation represent significant costs for students and institutions. This article seeks to understand the ability of machine learning models to predict student retention times in Computer Science programs in Brazil. The main objective is to develop models based on the Random Forest and XGBoost algorithms, compare the effectiveness between algorithms, and compare their effectiveness when trained with different datasets (national dataset and datasets by region). The methodological approach is based on the analysis of data provided by the National Student

Cite as: Last name, Initials., ... & Last name, Initials. (Year). Article title in the original language. Revista Brasileira de Informática na Educação, vol, pp-pp. https://doi.org/10.5753/rbie.yyyy.id.

Performance Exam (ENADE), comparing the metrics of a predictive model trained with data from all over Brazil and models trained individually for each region. The models are expected to achieve high predictive performance, and the comparative analysis between the national and regional levels will provide valuable insights into how retention time is affected by factors specific to each region of Brazil. The results have the potential to assist educational administrators in developing retention strategies for student monitoring and retention.

Resumen

La duración de la permanencia en programas de grado, especialmente en áreas de alta demanda como Ciencias de la Computación, es un indicador fundamental de la calidad académica de una institución, el bienestar estudiantil, las políticas de retención y las características socioeconómicas relevantes para la región de la institución y para cada estudiante. Las tasas de deserción, tanto permanente como temporal, y el retraso en la graduación representan costos significativos para estudiantes e instituciones. Este artículo busca comprender la capacidad de los modelos de aprendizaje automático para predecir la retención estudiantil en programas de Ciencias de la Computación en Brasil. El objetivo principal es desarrollar modelos basados en los algoritmos Random Forest y XGBoost, comparar su efectividad y compararla al entrenarlos con diferentes conjuntos de datos (nacionales y regionales). El enfoque metodológico se basa en el análisis de datos proporcionados por el Examen Nacional de Desempeño Estudiantil (ENADE), comparando las métricas de un modelo predictivo entrenado con datos de todo Brasil y modelos entrenados individualmente para cada región. Se espera que los modelos alcancen un alto rendimiento predictivo, y el análisis comparativo entre los ámbitos nacional y regional proporcionará información valiosa sobre cómo la retención se ve afectada por factores específicos de cada región de Brasil. Los resultados tienen el potencial de ayudar a los administradores educativos a desarrollar estrategias de retención para el seguimiento y la retención de estudiantes.

1 Introdução

A permanência e graduação do estudante no ensino superior são temas centrais nas discussões sobre educação no Brasil. Indicadores como a taxa de evasão e o tempo médio para a conclusão do curso refletem não apenas as especificidades de um aluno, mas também a eficiência e a qualidade das instituições de ensino. Nos cursos da área de Computação, essa questão se torna ainda mais crítica devido à dinâmica do mercado de trabalho, que por vezes atrai os estudantes antes mesmo da conclusão de sua formação, e à complexidade técnica da área, que pode gerar dificuldades de adaptação.

A crescente disponibilidade de grandes volumes de dados educacionais, como os dados do Exame Nacional de Desempenho dos Estudantes (ENADE), abre novas oportunidades para a aplicação de técnicas de Machine Learning que busquem a criação de conhecimento sobre as especificidades do cenário educacional brasileiro. Diversos estudos têm utilizado esses dados para identificar fatores associados ao desempenho acadêmico, mas poucos se concentraram especificamente na predição do tempo de permanência até a diplomação por meio de algoritmos de Machine Learning. Uma lacuna notável na literatura é a comparação de modelos preditivos em diferentes contextos geográficos no Brasil.

Diante disto, este trabalho busca responder à seguinte pergunta: É possível construir modelos de grande poder preditivo para o tempo de permanência em cursos de Computação utilizando dados do ENADE, e qual a diferença de eficácia entre um modelo treinado com dados de escopo nacional e treinado com dados específicos de cada região?

O objetivo principal é, portanto, treinar e avaliar modelos de classificação baseados nos algoritmos Random Forest e XGBoost para prever o tempo de permanência de um aluno. A contribuição desta pesquisa reside na comparação direta entre a capacidade preditiva de um modelo geral (nacional) e modelos específicos(regionais), oferecendo um novo ângulo sobre a necessidade de contextualização regional em políticas de permanência estudantil.

2 Metodologia a ser usada

A metodologia a ser usada consiste na construção de dois modelos preditivos que serão treinados com dados de amplitude nacional e, posteriormente, com dados agrupados por estado. Após a extração das métricas de desempenho de cada modelo em cada cenário, os resultados serão comparados para definir se o uso dos modelos de amplitudade nacional é viável ou se é necessário construir modelos para cada estado.

3 Fontes de Dados

Os dados utilizados vêm dos microdados do ENADE 2023, disponibilizados publicamente pelo INEP. Os dados serão recortados por curso(serão considerados apenas os cursos de computação) e serão filtrados apenas os registros de alunos concluintes dos respectivos cursos. O objetivo será predizer o tempo de conclusão do curso com base nas respostas que o aluno deu ao questionário

do estudante, que inclui perguntas acerca dos hábitos, motivações e da condição socioeconômica do estudante. As perguntas estão disponíveis na tabela abaixo:

Tabela 1: Questionário do estudante.

Índice	**Pergunta**
1	Qual o seu estado civil?
2	Qual é a sua cor ou raça?
3	Qual a sua nacionalidade?
4	Até que etapa de escolarização seu pai concluiu?
5	Até que etapa de escolarização sua mãe concluiu?
6	Onde e com quem você mora atualmente?
7	Quantas pessoas da sua família moram com você? Considere seus pais, irmãos, cônjuge, filhos e outros parentes que moram na mesma casa com você.
8	Qual a renda total de sua família, incluindo seus rendimentos?
9	Qual alternativa a seguir melhor descreve sua situação financeira (incluindo bolsas)?
10	Qual alternativa a seguir melhor descreve sua situação de trabalho (exceto estágio ou bolsas)?
11	Que tipo de bolsa de estudos ou financiamento do curso você recebeu para custear todas ou a maior parte das mensalidades? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.
12	Ao longo da sua trajetória acadêmica, você recebeu algum tipo de auxílio de permanência? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.
13	Ao longo da sua trajetória acadêmica, você recebeu algum tipo de bolsa acadêmica? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração.
14	Durante o curso de graduação, você participou de programas e/ou atividades curriculares no exterior?
15	Seu ingresso no curso de graduação se deu por meio de políticas de ação afirmativa ou inclusão social?
16	Em que unidade da Federação você concluiu o ensino médio?
17	Em que tipo de escola você cursou o ensino médio?
18	Qual modalidade de ensino médio você concluiu?
19	Quem lhe deu maior incentivo para cursar a graduação?
20	Algum dos grupos abaixo foi determinante para você enfrentar dificuldades durante seu curso superior e concluí-lo?
21	Alguém em sua família concluiu um curso superior?
22	Excetuando-se os livros indicados na bibliografia do seu curso, quantos livros você leu neste ano?
23	Quantas horas por semana, aproximadamente, você dedicou aos estudos, excetuando as horas de aula?
24	Você teve oportunidade de aprendizado de idioma estrangeiro na Instituição?
25	Qual o principal motivo para você ter escolhido este curso?
26	Qual a principal razão para você ter escolhido a sua instituição de educação superior?

4 Cronograma

Data	Atividade
25/08 a 10/09	Organizar o DataSet (unificar dados do ENADE)
15/09 a 17/09	Modelagem e treinamento do algoritmo XGBoost
22/09 a 24/09	Modelagem e treinamento do algoritmo Random Forest
29/09 a 06/10	Finalização do vídeo e relatório parcial
29/09 a 15/10	Processamento e obtenção de resultados dos algoritmos
20/10 a 29/10	Análise de resultados
03/11 a 12/11	Finalização do artigo

5 Trabalhos Correlatos

5.1 Uso de técnicas de aprendizado de máquina para predição do tempo de graduação dos discentes de Engenharia da Computação na região Sudeste do Brasil

Autores: da Silva Macedo, B., & Saporetti, C. M. (2024)
Resumo: Este trabalho é relacionado pois usa técnicas de aprendizado de máquina para a predição do tempo de graduação utilizando a mesma base de dados, os microdados do ENADE.

5.2 Análise comparativa de modelos de aprendizagem automática para a previsão da permanência de estudantes de graduação presenciais na UFMT

Autores: de Lima, D. V., de Oliveira, A. C. S., & de Oliveira, E. C. S. (2024)
Resumo: O artigo é relevante pois utiliza árvores de decisão(mesma ferramenta a ser usada neste trabalho) para analisar a permanência de estudantes em cursos de graduação

5.3 Aprendizado de máquina para agrupamento e associação de dados do ensino superior público brasileiro

Autores: Rodrigues, E. M., Marques Gouveia, R. M., de Albuquerque Junior, G. A., & Moraes Batista, M. D. C. (2023)
Resumo: Este estudo utiliza algoritmos de clustering e regras de associação para analisar dados do ensino superior público brasileiro. A pesquisa identifica grupos de estudantes e cursos com características semelhantes, revelando padrões úteis para planejamento institucional e políticas educacionais.

5.4 Evasão ou permanência? Modelos preditivos para a gestão do Ensino Superior

Autores: Silva, F. C., Cabral, T. L. O., & Pacheco, A. S. V. (2020)
Resumo: O artigo analisa fatores relevantes para a evasão no ensino superior e, por isso, pode ser útil para comparar os modelos a serem desenvolvidos neste trabalho, tornando possível avaliar se os resultados obtidos estão associados aos fatores esperados ou se há over-fitting

Referências

da Silva Macedo, B., & Saporetti, C. M. (2024). Uso de técnicas de aprendizado de máquina para predição do tempo de graduação dos discentes de Engenharia da Computação na região Sudeste do Brasil. *Revista Brasileira de Computação Aplicada*, *16*(1), 26–37. [GS Search].

de Lima, D. V., de Oliveira, A. C. S., & de Oliveira, E. C. S. (2024). Análise comparativa de modelos de aprendizagem automática para a previsão da permanência de estudantes de graduação presenciais na UFMT. *Revista Meta: Avaliação*, *16*(52), 515–545. [GS Search].

Rodrigues, E. M., Marques Gouveia, R. M., de Albuquerque Junior, G. A., & Moraes Batista, M. D. C. (2023). Aprendizado de máquina para agrupamento e associação de dados do ensino superior público brasileiro. *Revista de Sistemas e Computação - RSC*, *13*(1). [GS Search].

Silva, F. C., Cabral, T. L. O., & Pacheco, A. S. V. (2020). Evasão ou permanência? Modelos preditivos para a gestão do Ensino Superior. *Arquivos Analíticos de Políticas Educativas*, *28*(149). https://doi.org/10.14507/epaa.28.5387 [GS Search].