UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

ALUNOS
André Miyazawa
Erick Henrique Barbosa Albuquerque
Felipe Mateos Castro de Souza
Matheus Silva Lopes da Costa

# Planejamento de Estudo Exploratório para Predição de Evasão Estudantil no Ensino Superior

São Paulo
2024

**Sumário**

## 1 Questão de pesquisa geral

A questão geral de pesquisa se concentra em estudar possíveis padrões no desempenho acadêmico e fatores socieconomicos na probabilidade de um aluno abandonar a graduação.

## 2 Definição e coleta dos dados necessários

Dada a questão de pesquisa geral, procuramos por conjuntos de dados que reunissem informações como desempenho acadêmico e fatores socioeconomicos de estudantes. A partir disso, optamos por restringir nossa busca à plataforma Kaggle, uma plataforma de ciência de dados e inteligência artificial amplamente reconhecida por hospedar competições com prêmios, além de oferecer recursos para compartilhamento e análise de conjuntos de dados. Os dados específicos que escolhemos estão disponíveis em: https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention, esse conjunto de dados foi selecionado por possuir a avaliação máxima de usabilidade do site (10.0).

O conjunto de dados completo consiste em registros de 4.424 estudantes. Para esta análise, optamos por utilizar uma amostra aleatória simples, composta de 10% desse total, o que equivale a 442 instâncias. É importante destacar que essa amostra será utilizada exclusivamente para a fase de análise exploratória e será descartada em análises subsequentes que incluirão a aplicação de um modelo de aprendizado de máquina.

## 3 Descrição do conjunto de dados

Este conjunto de dados contém informações de uma instituição de ensino superior sobre várias variáveis relacionadas a estudantes de graduação, incluindo dados demográficos, fatores socioeconômicos e desempenho acadêmico, para investigar o impacto desses fatores na evasão estudantil.

- **Marital status** (*texto, qualitativa nominal*): O estado civil do estudante
  - 1-Solteiro
  - 2-Casado
  - 3-Viúvo
  - 4-Divorciado
  - 5-União estável
  - 6-Legalmente separado.
- **Application mode** (*texto, qualitativa nominal*): O método de aplicação utilizado pelo estudante.
  - 1-1ª fase-contingente geral
  - 2-Portaria Nº 612/93
  - 3-1ª fase-contingente especial (Ilha dos Açores)
  - 4-Portadores de outros cursos superiores
  - 5-Portaria Nº 854-B/99
  - 6-Estudante internacional (bacharelado)
  - 7-1ª fase-contingente especial (Ilha da Madeira)
  - 8-2ª fase-contingente geral
  - 9-3ª fase-contingente geral
  - 10-Portaria Nº 533-A/99, item b2) (Plano Diferente)
  - 11-Portaria Nº 533-A/99, item b3) (Outra Instituição)
  - 12-Maior de 23 anos
  - 13-Transferência
  - 14-Mudança de curso
  - 15-Portadores de diploma de especialização tecnológica
  - 16-Mudança de instituição/curso
  - 17-Portadores de diploma de ciclo curto

  - 18-Mudança de instituição/curso (Internacional)

- **Application order** (*texto, qualitativa ordinal*): A ordem em que o estudante se candidatou.
- **Course** (*texto, qualitativa nominal*): O curso realizado pelo estudante.
  - 1-Tecnologias de Produção de Biocombustíveis
  - 2-Design de Animação e Multimídia
  - 3-Serviço Social (frequência noturna)
  - 4-Agronomia
  - 5-Design de Comunicação
  - 6-Enfermagem Veterinária
  - 7-Engenharia Informática
  - 8-Equinicultura
  - 9-Gestão
  - 10-Serviço Social
  - 11-Turismo
  - 12-Enfermagem
  - 13-Higiene Oral
  - 14-Gestão de Publicidade e Marketing
  - 15-Jornalismo e Comunicação
  - 16-Educação Básica
  - 17-Gestão (frequência noturna)
- **Daytime/evening attendance** (*texto, qualitativa nominal*): Se o estudante frequenta as aulas durante o dia ou à noite.
  - 1-Diurno
  - 0-Noturno
- **Previous qualification** (*texto, qualitativa nominal*): A qualificação obtida pelo estudante antes de ingressar no ensino superior.
  - 1-Educação secundária
  - 2-Educação superior-bacharelado
  - 3-Educação superior-licenciatura
  - 4-Educação superior-mestrado

  - 5-Educação superior-doutorado
  - 6-Frequência de educação superior
  - 7-12º ano de escolaridade-não concluído
  - 8-11º ano de escolaridade-não concluído
  - 9-Outro-11º ano de escolaridade
  - 10-10º ano de escolaridade
  - 11-10º ano de escolaridade-não concluído
  - 12-Educação básica 3º ciclo (9º/10º/11º ano) ou equivalente
  - 13-Educação básica 2º ciclo (6º/7º/8º ano) ou equivalente
  - 14-Curso de especialização tecnológica
  - 15-Educação superior-licenciatura (1º ciclo)
  - 16-Curso superior técnico profissional
  - 17-Educação superior-mestrado (2º ciclo)

- **Nacionality** (*texto, qualitativa nominal*): A nacionalidade do estudante.
  - 1-Portuguesa
  - 2-Alemã
  - 3-Espanhola
  - 4-Italiana
  - 5-Holandesa
  - 6-Inglesa
  - 7-Lituana
  - 8-Angolana
  - 9-Cabo-verdiana
  - 10-Guineense
  - 11-Moçambicana
  - 12-Santomense
  - 13-Turca
  - 14-Brasileira
  - 15-Romena
  - 16-Moldava (República da Moldávia)
  - 17-Mexicana
  - 18-Ucraniana

  - 19-Russa
  - 20-Cubana
  - 21-Colombiana.

- **Mother's qualification** (*texto, qualitativa nominal*): A qualificação da mãe do estudante.
  - 1-Ensino Secundário-12º Ano de Escolaridade ou Equivalente
  - 2- Ensino Superior - bacharelado
  - 3-Técnica Superior
  - 4-Ensino Superior-mestrado
  - 5-Ensino Superior-doutorado
  - 6-Frequência do Ensino Superior
  - 7-12º ano de escolaridade - não concluído
  - 8º-11º ano de escolaridade - não concluído
  - 9-7º ano (antigo)
  - 10-Outros-11º Ano de Escolaridade
  - 11-2º ano do ensino médio complementar
  - 12-10º ano de escolaridade
  - 13-Curso de comércio geral
  - 14-Ensino Básico, 3º Ciclo (9º/10º/11º Ano) ou Equivalente
  - 15-Curso Complementar de Ensino Médio
  - 16-Curso técnico-profissional
  - 17-Curso Complementar de Ensino Médio - não concluído
  - 18-7º ano de escolaridade
  - 19-2º ciclo do curso secundário geral
  - 20-9º ano de escolaridade - não concluído
  - 21-8º ano de escolaridade
  - 22-Curso Geral de Administração e Comércio
  - 23-Contabilidade e Administração Complementares
  - 24-Desconhecido
  - 25-Não sabe ler nem escrever
  - 26-Consegue ler sem ter o 4º ano de escolaridade
  - 27-Ensino básico 1º ciclo (4º/5º ano) ou equivalente

  - 28-Ensino Básico 2º Ciclo (6º/7º/8º Ano) ou equivalente
  - 29-Curso de especialização tecnológica
  - 30-Licenciatura superior (1º ciclo)
  - 31-Curso superior especializado
  - 32-Curso técnico superior profissional
  - 33-Ensino Superior-mestrado (2º ciclo)
  - 34-Ensino Superior-doutoramento (3º ciclo).

- **Father's qualification** (*texto, qualitativa nominal*): A qualificação do pai do estudante.
  - 1-Ensino Secundário-12º Ano de Escolaridade ou Equivalente
  - 2- Ensino Superior - bacharelado
  - 3-Técnica Superior
  - 4-Ensino Superior-mestrado
  - 5-Ensino Superior-doutorado
  - 6-Frequência do Ensino Superior
  - 7-12º ano de escolaridade - não concluído
  - 8º-11º ano de escolaridade - não concluído
  - 9-7º ano (antigo)
  - 10-Outros-11º Ano de Escolaridade
  - 11-2º ano do ensino médio complementar
  - 12-10º ano de escolaridade
  - 13-Curso de comércio geral
  - 14-Ensino Básico, 3º Ciclo (9º/10º/11º Ano) ou Equivalente
  - 15-Curso Complementar de Ensino Médio
  - 16-Curso técnico-profissional
  - 17-Curso Complementar de Ensino Médio - não concluído
  - 18-7º ano de escolaridade
  - 19-2º ciclo do curso secundário geral
  - 20-9º ano de escolaridade - não concluído
  - 21-8º ano de escolaridade
  - 22-Curso Geral de Administração e Comércio
  - 23-Contabilidade e Administração Complementares

  - 24-Desconhecido
  - 25-Não sabe ler nem escrever
  - 26-Consegue ler sem ter o 4º ano de escolaridade
  - 27-Ensino básico 1º ciclo (4º/5º ano) ou equivalente
  - 28-Ensino Básico 2º Ciclo (6º/7º/8º Ano) ou equivalente
  - 29-Curso de especialização tecnológica
  - 30-Licenciatura superior (1º ciclo)
  - 31-Curso superior especializado
  - 32-Curso técnico superior profissional
  - 33-Ensino Superior-mestrado (2º ciclo)
  - 34-Ensino Superior-doutoramento (3º ciclo).

- **Mother's occupation** (*texto, qualitativa nominal*): A ocupação da mãe do estudante.
  - 1-Estudantes
  - 2-Representantes do Poder Legislativo e dos Órgãos Executivos, Diretores, Diretores e Gerentes Executivos
  - 3-Especialistas em Atividades Intelectuais e Científicas,
  - 4-Técnicos e Profissões de Nível Intermediário,
  - 5-Pessoal administrativo
  - 6-Serviços pessoais, trabalhadores de segurança e proteção e vendedores
  - 7-Agricultores e Trabalhadores Qualificados na Agricultura, Pesca e Silvicultura
  - 8-Trabalhadores Qualificados na Indústria, Construção e Artesãos
  - 9-Operadores de Instalação e Máquinas e Trabalhadores de Montagem
  - 10-Trabalhadores Não Qualificados, 11-Profissões das Forças Armadas
  - 12-Outra Situação
  - 13-(em branco)
  - 14-Oficiais das Forças Armadas
  - 15-Sargentos das Forças Armadas
  - 16-Demais militares das Forças Armadas
  - 17-Diretores de serviços administrativos e comerciais
  - 18-Diretores de hotelaria, restauração, comércio e outros serviços
  - 19-Especialistas em ciências físicas, matemática, engenharia e técnicas afins

- 20-Profissionais de saúde
- 21-Professores
- 22-Especialistas em finanças, contabilidade, organização administrativa e relações públicas e comerciais
- 23-Técnicos e profissões de ciência e engenharia de nível intermediário
- 24-Técnicos e profissionais de nível intermediário de saúde
- 25-Técnicos de nível intermediário dos serviços jurídicos, sociais, desportivos, culturais e similares
- 26-Técnicos de tecnologia da informação e comunicação
- 27-Trabalhadores de escritório, secretárias em geral e operadores de processamento de dados
- 28-Operadores de dados, contabilidade, estatística, serviços financeiros e registradores
- 29-Outro pessoal de apoio administrativo
- 30-Trabalhadores de serviços pessoais
- 31-Vendedores
- 32-Trabalhadores de cuidados pessoais e similares
- 33-Pessoal dos serviços de proteção e segurança
- 34-Agricultores orientados para o mercado e trabalhadores qualificados na produção agrícola e animal
- 35-Agricultores, criadores de gado, pescadores, caçadores e coletores, e subsistência
- 36-Trabalhadores qualificados da construção civil e similares, exceto eletricistas
- 37-Trabalhadores qualificados em metalurgia, metalomecânica e similares
- 38-Trabalhadores qualificados em eletricidade e eletrônica
- 39-Trabalhadores em processamento de alimentos, marcenaria e vestuário e outras indústrias e artesanato, 40-Operadores fixos de instalações e máquinas
- 41-Trabalhadores de montagem
- 42-Motoristas de veículos e operadores de equipamentos móveis
- 43-Trabalhadores não qualificados na agricultura, produção animal, pesca e silvicultura
- 44-Trabalhadores não qualificados na indústria extrativa, construção, manufatura e transporte

  - 45-Auxiliares de preparação de refeições
  - 46-Vendedores ambulantes (exceto alimentos) e prestadores de serviços ambulantes.

- **Father's occupation** (*texto, qualitativa nominal*): A ocupação do pai do estudante.
  - 1-Estudantes
  - 2-Representantes do Poder Legislativo e dos Órgãos Executivos, Diretores, Diretores e Gerentes Executivos
  - 3-Especialistas em Atividades Intelectuais e Científicas,
  - 4-Técnicos e Profissões de Nível Intermediário,
  - 5-Pessoal administrativo
  - 6-Serviços pessoais, trabalhadores de segurança e proteção e vendedores
  - 7-Agricultores e Trabalhadores Qualificados na Agricultura, Pesca e Silvicultura
  - 8-Trabalhadores Qualificados na Indústria, Construção e Artesãos
  - 9-Operadores de Instalação e Máquinas e Trabalhadores de Montagem
  - 10-Trabalhadores Não Qualificados, 11-Profissões das Forças Armadas
  - 12-Outra Situação
  - 13-(em branco)
  - 14-Oficiais das Forças Armadas
  - 15-Sargentos das Forças Armadas
  - 16-Demais militares das Forças Armadas
  - 17-Diretores de serviços administrativos e comerciais
  - 18-Diretores de hotelaria, restauração, comércio e outros serviços
  - 19-Especialistas em ciências físicas, matemática, engenharia e técnicas afins
  - 20-Profissionais de saúde
  - 21-Professores
  - 22-Especialistas em finanças, contabilidade, organização administrativa e relações públicas e comerciais
  - 23-Técnicos e profissões de ciência e engenharia de nível intermediário
  - 24-Técnicos e profissionais de nível intermediário de saúde
  - 25-Técnicos de nível intermediário dos serviços jurídicos, sociais, desportivos, culturais e similares
  - 26-Técnicos de tecnologia da informação e comunicação

  - 27-Trabalhadores de escritório, secretárias em geral e operadores de processamento de dados
  - 28-Operadores de dados, contabilidade, estatística, serviços financeiros e registradores
  - 29-Outro pessoal de apoio administrativo
  - 30-Trabalhadores de serviços pessoais
  - 31-Vendedores
  - 32-Trabalhadores de cuidados pessoais e similares
  - 33-Pessoal dos serviços de proteção e segurança
  - 34-Agricultores orientados para o mercado e trabalhadores qualificados na produção agrícola e animal
  - 35-Agricultores, criadores de gado, pescadores, caçadores e coletores, e subsistência
  - 36-Trabalhadores qualificados da construção civil e similares, exceto eletricistas
  - 37-Trabalhadores qualificados em metalurgia, metalomecânica e similares
  - 38-Trabalhadores qualificados em eletricidade e eletrônica
  - 39-Trabalhadores em processamento de alimentos, marcenaria e vestuário e outras indústrias e artesanato, 40-Operadores fixos de instalações e máquinas
  - 41-Trabalhadores de montagem
  - 42-Motoristas de veículos e operadores de equipamentos móveis
  - 43-Trabalhadores não qualificados na agricultura, produção animal, pesca e silvicultura
  - 44-Trabalhadores não qualificados na indústria extrativa, construção, manufatura e transporte
  - 45-Auxiliares de preparação de refeições
  - 46-Vendedores ambulantes (exceto alimentos) e prestadores de serviços ambulantes.

- **Displaced** (*texto, qualitativa nominal*): Se o estudante é uma pessoa deslocada.
  - 1-Sim
  - 0-Não

- **Educational special needs** (*texto, qualitativa nominal*): Se o estudante tem necessidades educacionais especiais.

  - 1-Sim
  - 0-Não

- **Debtor** (*texto, qualitativa nominal*): Se o estudante é um devedor.
  - 1-Sim
  - 0-Não
- **Tuition fees up to date** (*texto, qualitativa nominal*): Se as taxas de matrícula do estudante estão em dia.
  - 1-Sim
  - 0-Não
- **Gender** (*texto, qualitativa nominal*): O gênero do estudante.
  - 1-Masculino
  - 0-Feminino.
- **Scholarship holder** (*texto, qualitativa nominal*): Se o estudante é bolsista.
  - 1-Sim
  - 0-Não
- **Age at enrollment** (*discreto, quantitativa razão*): A idade do estudante no momento da matrícula.
- **International** (*texto, qualitativa nominal*): Se o estudante é um estudante internacional.
  - 1-Sim
  - 0-Não
- **Curricular units 1st sem (credited)** (*Inteiro, quantitativa razão*): O número de unidades curriculares creditadas pelo estudante no primeiro semestre.
- **Curricular units 1st sem (enrolled)** (*Inteiro, quantitativa razão*): O número de unidades curriculares matriculadas pelo estudante no primeiro semestre.
- **Curricular units 1st sem (evaluations)** (*Inteiro, quantitativa razão*): O número de unidades curriculares avaliadas pelo estudante no primeiro semestre.
- **Curricular units 1st sem (approved)** (*Inteiro, quantitativa razão*): O número de unidades curriculares aprovadas pelo estudante no primeiro semestre.
- **Curricular units 1st sem (grade)** (*Real, quantitativa razão*): A nota obtida pelo estudante nas unidades curriculares do primeiro semestre.

- **Curricular units 1st sem (without evaluations)** (*Inteiro, quantitativa razão*): O número de unidades curriculares do primeiro semestre sem avaliações.
- **Curricular units 2nd sem (credited)** (*Inteiro, quantitativa razão*): O número de unidades curriculares creditadas pelo estudante no segundo semestre.
- **Curricular units 2nd sem (enrolled)** (*Inteiro, quantitativa razão*): O número de unidades curriculares matriculadas pelo estudante no segundo semestre.
- **Curricular units 2nd sem (evaluations)** (*Inteiro, quantitativa razão*): O número de unidades curriculares avaliadas pelo estudante no segundo semestre.
- **Curricular units 2nd sem (approved)** (*Inteiro, quantitativa razão*): O número de unidades curriculares aprovadas pelo estudante no segundo semestre.
- **Curricular units 2nd sem (grade)** (*Real, quantitativa razão*): A nota obtida pelo estudante nas unidades curriculares do segundo semestre.
- **Curricular units 2nd sem (without evaluations)** (*Inteiro, quantitativa razão*): O número de unidades curriculares do segundo semestre sem avaliações.
- **Unemployment rate** (*Real, quantitativa razão*): A taxa de desemprego.
- **Inflation rate** (*Real, quantitativa razão*): A taxa de inflação.
- **GDP** (*Real, quantitativa razão*): O Produto Interno Bruto.
- **Target** (*texto, qualitativa nominal*): O objetivo. Nesse caso é a situação que o estudante se encontra após um período que corresponde a duração ideal do seu curso, sendo os possíveis valores:
  - Inscrito
  - Graduado
  - Desistente

Será realizada uma análise para investigar se técnicas de normalização, com o objetivo de uniformizar a escala de todas as características numéricas, serão benéficas para esses dados. Além disso, examinaremos se há desbalanceamento de classes no conjunto de dados. Se identificado desbalanceamento, avaliaremos se técnicas de balanceamento, como oversampling ou undersampling, podem melhorar a capacidade de predição.

Será verificado se existem dados nulos no conjunto de dados. Se forem identificados dados nulos, será realizada uma análise para determinar se sua presença é provavelmente devido a um fenômeno sistemático ou ao acaso. Se os dados nulos forem encontrados, as observações contendo esses dados poderão removidas ou preenchidas usando técnicas

de aprendizado de máquina, especialmente considerando que não há disponibilidade de muitos dados.

Por fim, a análise exploratória futura dos dados, incluindo distribuições, gráficos, etc., será conduzida exclusivamente na amostra que contém 10% dos dados. Essa abordagem visa preservar a integridade dos dados restantes, representando 90% do conjunto total, para análises futuras que dependam de uma quantidade maior de dados, como a aplicação de algoritmos de aprendizado de máquina.

## 4 Questões de pesquisa refinadas

Foram formuladas algumas questões específicas a serem investigadas no conjunto de dados:

- Será que a ocupação dos pais é mais impactante do que a sua formação para probabilidade de um aluno se graduar?
- Será que a ocupação e/ou formação do pai é mais impactante do que a ocupação e/ou formação da mãe para probabilidade de um aluno se graduar?
- Qual o atributo do conjunto de dados mais impactante na probabilidade de um estudante abandonar a graduação?

## 5 Cronograma

Nas duas primeiras semanas, os dados do conjunto "Predict students' dropout and academic success"serão minuciosamente analisados, com ênfase na compreensão dos atributos disponíveis. Uma amostra representativa de 10% dos dados será selecionada aleatoriamente para uma análise inicial, permitindo uma exploração inicial de sua estrutura e distribuição. Além disso, haverá uma verificação da qualidade dos dados, com identificação de variáveis relevantes e compreensão das características. O refinamento das questões de pesquisa ocorrerá com base nos insights obtidos dessa análise da amostra.

Na terceira semana, o foco estará na modelagem e avaliação dos dados. Utilizando os dados ainda não explorados do conjunto, eles serão divididos em conjuntos de treinamento e teste. Modelos de aprendizado de máquina serão treinados e validados com base nos dados de treinamento, e sua performance será avaliada utilizando métricas apropriadas, como acurácia, precisão, recall e F1-score.

Finalmente, na quarta semana, ocorrerá a análise e conclusões finais do projeto. Os resultados obtidos pelos modelos serão detalhadamente analisados, com o objetivo de identificar padrões ou insights relevantes na predição da situação do estudante. Haverá uma comparação dos resultados com as expectativas iniciais e com a literatura existente, seguida pela documentação dos principais achados e conclusões do projeto. Por fim, o relatório final será preparado, seguindo o formato especificado no enunciado do trabalho.

- Semana 1: Análise exploratória.
- Semana 2: Pré-processamento.
- Semana 3: Treinamento do modelos de aprendizado de máquina e avaliação.
- Semana 4: Comparação dos resultados e escrita do relatório.