*INTERMATHS, VOL. 4, NO. 2 (**2023**)*, **54–74**
https://doi.org/10.22481/intermaths.v4i2.13422

Elementos Básicos de Aprendizagem de Máquina

Basic Elements of Machine Learning

João Socorro Pinheiro Ferreira [a]

[a] Universidade Federal do Amapá, Macapa - AP, Brasil

* Autor Correspondente: joaoferreira@unifap.br

Resumo: Neste trabalho realizamos a revisão de seis artigos científicos sobre a evasão escolar que utilizaram Aprendizagem de Máquina como metodologia para detectar as possíveis causas. A metodologia utilizada por esta pesquisa é a bibliográfica, pois-se investigou quais são os temas e elementos técnicos de cada um dos artigos, principalmente em relação à aprendizagem de máquina. Os resultados obtidos são sobre os principais elementos que estruturam os algoritmos de aprendizagem de máquina a partir dos seis artigos científicos estudados. O tema é muito abrangente e envolve diversas áreas e subáreas do conhecimento científico, como, por exemplo Álgebra Linear, Matrizes, Teoria da Computação, Computabilidade, Modelos de Computação, Linguagem Formais e Autômatos, Análise de Algoritmos e Complexidade Computacional. As Técnicas de Mineração de Dados (EDM) são as ações utilizadas para encontrar padrões em um grande volume de dados. Estes padrões podem ser explicativos, de modo a descrever as relações entre segmentos de dados, ou preditivos, os quais podem prever valores futuros baseados em dados anteriores. Ao final, o leitor terá uma visão ampla de como ocorre todo o processo metodológico de produção e construção de aprendizagem de máquina.

Palavras-chave: Aprendizagem de Máquina (ML); Inteligência Artificial (IA); Mineração de Dados Educacionais (EDM); Previsão de Abandono Escolar; Redes Neurais (RN).

Abstract: In this work, we reviewed six scientific articles on school dropouts that used machine learning as a methodology to point out the possible causes. The methodology used by this research is bibliographical because it investigated the themes and technical elements of each of the articles, especially in relation to machine learning. The results obtained are about the main elements that structure the machine learning algorithms from the six scientific articles studied. The theme is very comprehensive and involves several areas and subareas of scientific knowledge, such as linear algebra, matrices, theory of computation, computability, models of computation, formal language and automata, analysis of algorithms, and computational complexity. Data mining techniques (EDM) are the actions used to find patterns in a large volume of data. These patterns can be explanatory, to describe the relationships between data segments, or predictive, to predict future values based on previous data. At the end, the reader will have a broad view of how the entire methodological process of the production and construction of machine learning occurs.

keywords: Machine Learning (ML); Artificial Intelligence (AI); Educational Data Mining (EDM); School Dropout Prediction; Neural Networks (RN).

Submetido em: 14 de Agosto de 2023 Aprovado em: 22 de Dezembro de 2023 Publicado em: 30 de Dezembro de 2023
ISSN 2675-8318

1 Introdução

O objetivo deste texto é apresentar como se dá o processo de construção de um algoritmo de Aprendizagem de Máquina a partir da leitura de seis artigos científicos sobre o tema Evasão ou Abandono Escolar, em que os títulos estão identificados no próprio texto. Os artigos abordam o seguinte questionamento: Como é possível prever o abandono escolar a partir de dados levantados em um estabelecimento de ensino? A resposta que são apresentadas é sobre a utilização de Machine Learn (ML) ou Aprendizagem de Máquina.

Na AM há dois tipos de sistemas de aprendizado de máquina (ML): o primeiro, que trabalha com Aprendizagem Supervisionada e Não Supervisionada e o segundo, com Aprendizagem em Batch (lote) e On-line. Estes são os sistemas de se treinar e testar os algoritmos. A aprendizagem supervisionada e não supervisionada, podem ser subdivididas em quatro categorias: Aprendizagem Supervisionada, Aprendizagem Não Supervisionada, Aprendizagem Semisupervisionada e Aprendizagem por Reforço – aqui há alguma semelhança com Skinner, para treinar programas ou algoritmos robôs, pois há recompensa pelo desempenhoo. A Aprendizagem em Batch (Lote) e a On-line, são organizadas ou classificadas em três tipos conforme a seguir: Aprendizagem em Lote, Aprendizagem On-line e Aprendizado Baseada em Instância versus Aprendizado Baseada em Modelo.

Os principais desafios do aprendizado de máquina são: quantidade insuficiente de dados de treinamento, dados de treinamento não representativos, dados de baixa qualidade, recursos irrelevantes, sobreajustando os dados de treinamento e subajustando os dados de treinamento.

Este texto está organizado em quatro Seções. Na Seção 2, organizamos a descrição dos artigos à Tabela 1, com o título e as palavras-chave, com a finalidade de aprofundar sobre os fundamentos teóricos existentes em cada um dos artigos e verificar aquelas palavras-chave que são homônimas ou análogas, na busca de similaridade de métodos. Na Tabela 2, reunimos todos os elementos textuais de cada artigo, também para identificar semelhanças. Na Seção 3, organizamos na Tabela 3 os principais métodos, modelos e algoritmos utilizados para minerar dados de banco de dados para analisar evasão ou abandono escolar. Na Tabela 4, mostramos um exemplo de características utilizadas como variáveis para a mineração de dados (EDM) sobre evasão escolar, pesquisado no artigo [1]. Na Subseção 3.1 relatamos as principais métricas utilizadas nos seis artigos científicos, que tiveram a finalidade de classificar os diversos objetos de pesquisa que abordaram. As que se destacam são acurácia (ou precisão), especificidade, sensibilidade e a Área Sob a Curva (AUC). A Receiver Operating Characteristics (ROC) significa a curva de Características Operacionais do Receptor (ROC). Essas medidas são calculadas por meio dos valores da Matriz de Confusão*, de dimensão 2×2. Na Subseção 3.2, incluímos uma aplicação analítica do método *Linear Discriminant Analysis* (LDA) (Análise Discriminante Linear (LDA)), um dos métodos utilizados no artigo científico [1]. Na Seção 4, estão as Considerações Finais a respeito da compreensão dos seis artigos estudados e as possibilidades de produção de pesquisas futuras e publicação de textos científicos como forma de construir Aprendizagem de Máquina.

*No campo do Aprendizado de Máquina uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. Essa tabela de contingência 2×2 especial é também chamada de matriz de erro.

2 Revisão dos Artigos

Na Tabela 1, organizamos os seis artigos científicos que nos foi indicados para os estudos preliminares sobre *Machine Learning* (ML) (Aprendizagem de Máquina), apontando o título e as palavras-chave em língua portuguesa, para melhorar a compreensão dos mesmos, porém nas Referências, os títulos estão em Inglês. Vale ressaltar, que o sexto artigo não contém palavras-chave.

Tabela 1 – Relação de artigos científicos revisados e respectivas palavras-chave.

Artígo Científico	Título	Palavras-Chave
[1]	Previsão de Abandono do Aluno	Aprendizado de Máquina. Mineração de Dados Educacionais (EDM). Ferramentas de Suporte à Decisão
[2]	Os modelos de previsão de evasão universitária devem incluir atributos protegidos?	Previsão de abandono Análise preditiva Ensino superior Aprendizagem on-line Justiça algorítmica (Algoritmia)
[3]	Aplicação de regressão logística para prever a falha de alunos em disciplinas de um curso de graduação em Matemática	EDM. Regressão logística Análise de dados
[4]	Previsão de evasão precoce usando EDM: um estudo de caso com alunos do ensino médio	Predição de evasão Classificação, EDM, Programação genética baseada em gramática
[5]	Um estudo de inferência causal sobre os efeitos da Carga de Trabalho do Primeiro Ano sobre a Taxa de Evasão dos Graduandos	Abandono universitário Aprendizado de máquina Inferência causal Efeito médio do tratamento
[6]	Prevenindo a evasão de alunos no ensino a distância usando técnicas de aprendizado de máquina	

Fonte: elaborado pelo autor.

São dezenove palavras-chave, sendo que a frequência modal é sobre Mineração de Dados Educacionais (*Educational Data Mining* (EDM)), seguida por Aprendizado de Máquina (*Machine Learning* (ML)).

Considerando-se que estamos realizando um trabalho exploratório para compreender os termos e métodos de Ciências de Dados, organizamos na Tabela 2, os elementos presentes nos seis artigos científicos e percebemos que aqueles exigidos pela NBR 6028/2018, que normatiza Artigo em publicação periódica técnica e/ou científica, estão presentes em todos e surgem novos denominações, inerentes a ciência de dados.

Na Tabela 2, o 1 indica que o termo está presente e 0 ausente nos artigos científicos.

Observamos que há uma variedade de novos termos, sendo que alguns estão presentes em apenas um e outros são comuns a todos, como é o caso de Conjunto de Dados (*Dataset*).

No [1], encontra-se um estudo que dentre muitos problemas em aberto no processo

Tabela 2 – Elementos presentes nos artigos científicos

Elementos	Artigos					
	1	2	3	4	5	6
Introdução	1	1	1	1	1	1
Antecedentes	0	0	0	1	0	0
Metodologia de Ensino à Distância HOU e Descrição dos Dados	0	0	0	0	0	1
Trabalho Relatado (Descrição do Caso)	1	1	1	0	1	0
Previsão de abandono da faculdade	0	1	0	0	0	0
Imparcialidade algorítmica na educação	0	1	0	0	0	0
Metodologia	1	1	1	1	1	0
Algoritmos	0	0	0	1	0	0
Modelos	0	0	1	0	0	0
Conjunto de Dados	1	1	1	1	1	1
Meta de previsão e engenharia de recursos	0	1	0	0	0	0
Previsão de abandono	0	1	0	0	0	0
Avaliação de desempenho	0	1	0	0	0	0
Pré-processamento de Dados	1	0	1	0	0	0
Seleção de recursos e métricas de avaliação	1	0	1	0	0	0
Desenvolvimento e avaliação de modelos	0	0	1	0	0	0
Razão de probabilidade	0	0	1	0	0	0
Métricas de desempenho	0	0	1	0	0	0
Fluxograma de desenvolvimento de modelo	0	0	1	0	0	0
Experimento 1	0	0	0	1	0	0
Experimento 2	0	0	0	1	0	0
Experimento 3	0	0	0	1	0	0
Técnicas de aprendizado de máquina	0	0	0	0	0	1
Resultado Experimental	1	1	1	1	1	1
Seleção de Parâmetros e Escalonamento de Dados	1	0	1	0	0	0
Parâmetros do modelo e seleção de variáveis	1	0	1	0	0	0
Análise das Características.	1	0	0	0	0	0
Análise do Abandono por Escola Acadêmica	1	0	0	0	0	0
Desempenho Geral da Previsão	0	1	1	0	0	0
Desempenho do modelo	0	1	1	0	0	0
Imparcialidade de previsão	0	1	0	0	0	0
Modelos Descobertos	0	0	0	1	0	0
Trabalho e Discussão Relacionados	0	0	0	1	1	0
Experimentos e Resultados	0	0	0	0	0	1
Conclusão e Trabalho Futuro	1	1	1	1	1	1
Referências	1	1	1	1	1	1

Fonte: elaborado pelo autor.

0: Ausente; 1: Presente.

de ensino e aprendizagem, está à evasão escolar (abandono), pois é difícil de prever, tanto para o aluno quanto para as instituições, e predizer poderá ajudar a diminuir seus custos sociais e econômicos.

No [2], procura apresentar ferramentas à identificação precoce de evasão universitária pode fornecer um valor enorme para melhorar o sucesso do aluno e a eficácia institucional e a análise preditiva é cada vez mais usada para esse fim. No entanto, surgiram preocupações éticas sobre se a inclusão de atributos protegidos nesses modelos de previsão discrimina grupos de estudantes subrepresentados e exacerba as desigualdades existentes.

Em [3], os grandes índices de reprovação dos alunos é um problema muito frequente nos cursos de graduação, sendo ainda mais evidente nas ciências exatas. Apontar as razões desse problema é um tema primordial de pesquisa, embora não seja uma tarefa fácil.

Em [4], a previsão precoce da evasão escolar é um problema sério na educação, mas não é uma questão fácil de resolver. Por um lado, há muitos fatores que podem influenciar a retenção do aluno. Por outro lado, a abordagem de classificação tradicional usada para resolver esse problema normalmente deve ser implementada no final do curso para coletar o máximo de informações a fim de obter a maior precisão. Neste artigo, os autores propõem uma metodologia e um algoritmo de classificação específico para descobrir modelos de previsão compreensíveis da evasão escolar o mais rápido possível.

No artigo científico [5], os autores avaliaram o risco de abandono precoce na graduação usando métodos de inferência causal e focaram em grupos de alunos que têm um risco de abandono relativamente maior. Usamos um grande conjunto de dados composto por alunos de graduação admitidos em vários programas de estudo em oito faculdades/escolas de nossa universidade.

No artigo [6], a evasão estudantil ocorre com bastante frequência em universidades que oferecem Educação a Distância. O escopo desta pesquisa é estudar se o uso de técnicas de aprendizado de máquina pode ser útil para lidar com esse problema.

3 Metodologias e Métodos

Aqui estão alguns dos algoritmos de aprendizado supervisionado mais importantes:

- k-Nearest Neighbors (k-vizinhos mais próximos)
- Linear Regression (Regressão linear)
- Logistic Regression (Regressão Logística)
- Support Vectors Machine (Máquinas de Vetores de Suporte (SVMs))
- Decision Trees (Árvores de Decisão) e Random Foresta (Florestas Aleatórias)
- Neural Networks (Redes neurais)

Aqui estão alguns dos algoritmos de aprendizado não supervisionados mais importantes:

- Agrupamento
  - k-Means
  - Análise Hierárquica de Cluster (HCA)
  - Maximização da Expectativa
- Visualização e redução de dimensionalidade
  - Análise de Componentes Principais (PCA)
  - Núcleo PCA
  - Incorporação Localmente Linear (LLE)
  - Incorporação de vizinhos estocásticos distribuídos em t (t-SNE)
- Aprendizagem de regra de associação
  - A priori
  - Eclat

Na Tabela 3 estão relacionadas as algoritmos de aprendizado supervisionado e aprendizado não supervisionados utilizados em cada um dos seis artigos estudados.

As Metodologias utilizadas por cada um dos artigos científicos da Tabela 1 são diferenciadas, conforme sintetisamos na Tabela 2, por isso, faremos aqui um breve relato por artigo científico.

No artigo científico [1], em particular, os autores consideraram três métodos: Análise Discriminante Linear (LDA), Support Vector Machine (SVM) - [7] e Random Forest (RF), por serem os mais comumente modelos usados na literatura para resolver problemas semelhantes.

A LDA atua como um algoritmo de redução dimensional, tentando reduzir os dados complexidade, ou seja, projetando o espaço de recursos real em um espaço de dimensão inferior, ao tentar reter informações relevantes; além disso, não envolve configurações de parâmetros. O SVM é uma técnica bem estabelecida para classificação e regressão de dados. Ele encontra o melhor hiperplano de separação maximizando a margem no espaço de recursos. Os dados de treinamento que participam do processo de maximização são chamados de vetores de suporte. A RF constrói uma coleção de classificadores estruturados em árvore combinando-os aleatoriamente. Tem sido adotado na literatura para uma grande variedade de tarefas de regressão e predição [8].

A Tabela 4 mostra uma descrição detalhada das informações disponíveis no conjunto de dados do Artigo [1]. A primeira coluna lista o nome dos recursos, enquanto a segunda coluna descreve os possíveis valores ou intervalos. As duas primeiras funcionalidades representam dados pessoais dos alunos enquanto a terceira e a quarta são informações relacionadas ao ensino médio frequentado pelo aluno.

Tabela 3 – Algoritmos de aprendizado supervisionado mais importantes (abordados neste livro)

Artigo Científico	Título	Algoritmo de Aprendizagem
[1]	Previsão de Abandono do Aluno	Linear Discriminant Analysis (LDA) Support Vector Machine (SVM) Random Forest (RF)
[2]	Os modelos de previsão de evasão universitária devem incluir atributos protegidos?	Gradient Boosted Trees (GBT) Logistic Regression (LR)
[3]	Aplicação de regressão logística para prever a falha de alunos em disciplinas de um curso de graduação em matemática	Odds Ratio (OR) Ten-Fold Cross Validation Method
[4]	Previsão de evasão precoce usando mineração de dados (EDM): um estudo de caso com alunos do ensino médio	Genetic Programming (GP)
[5]	Um estudo de inferência causal sobre os efeitos da Carga de Trabalho do Primeiro Ano sobre a Taxa de Evasão dos Graduandos	Logistic Regression (LR) Multi-Layer Perceptron (MLP)
[6]	Prevenindo a evasão de alunos no ensino a distância usando técnicas de aprendizado de máquina	Decision Trees (DT) Neural Networks (NN) Naive Bayes algorithm (NBA) Instance-Based Learning Algorithms Logistic Regression (LR) Support Vector Machines (SVM)

Fonte: elaborado pelo autor.

Para LDA, RF e SVM, apenas mantevemos a escolha dos melhores parâmetros e monitoramos seu desempenho nos diferentes conjuntos de dados.

Considerando o conjunto básico de recursos, LDA e SVM obtêm o melhor desempenho com uma variação um pouco maior para os resultados do SVM. A introdução do recurso *Additional Learning Requirements* (ALR) melhora principalmente a precisão e a especificidade para LDA e SVM, mas diminui a sensibilidade. Pelo contrário, a introdução da funcionalidade ALR no RF ajuda a melhorar o desempenho final em todas as medidas, obtendo um desempenho superior face aos resultados de LDA e SVM em o conjunto básico de recursos.

No artigo [2], o método Gradient Boosted Trees (Árvores reforçadas com gradiente): Gradient Boosted Trees (também conhecido como GBT) é um algoritmo de ML baseado em árvore comumente usado que funciona tanto para regressão quanto para classificação de problemas de mineração de dados (EDM).

Para interpretação e avaliação dos modelos em [3], foram utilizados Odds Ratio (OR), Método dez-fold Cross Validation e as métricas: acurácia (Equação (1)), especificidade (Equação (2)), sensibilidade (Equação (4)) e área sob a curva ROC (AUC) (Equação (5)). A curva ROC é obtida representando a taxa de verdadeiros positivos (sensibilidade) no eixo x versus a taxa de falso positivo (especificidade) no eixo y para p_0 variando entre 0 e 1.

A AUC é a medida de desempenho frequentemente usada para modelos de ML

Tabela 4 – Recursos disponíveis para cada aluno no conjunto de dados original, juntamente com o intervalo de valores possíveis

Característica	Faixa de valor
Sexo do aluno	1, 2
Faixa etária do aluno	1 a 3
ID do ensino médio	1 a 10
Nota final do ensino médio	60 a 100
Requisitos Adicionais de Aprendizagem (ALR)	1, 2, 3
ID da escola acadêmica	1 a 11
Créditos do Curso (CC)	0 a 60
Desistência	0 , 1

Fonte: [1]

de classificação, onde a curva representa a curva Receiver Operating Characteristics (ROC) que pode incluir a totalidade do desempenho de previsão de um método de classificação para todos os limites de classificação ([9], 2020). A curva permite visualizar o comportamento dos modelos de classificação e denota o trade-off do classificador entre o número de previsões positivas corretas e incorretas ([9], 2020). Para ser mais específico, a curva ROC representa um gráfico bidimensional no qual a Taxa Positiva Verdadeira (TPR) denota o eixo y, enquanto a Taxa Positiva Falsa (FPR) representa o eixo x ([10] , 2019). Os valores de AUC variam entre 0,5 e 1,0. Quando o valor de AUC está próximo de 1, mostra que o modelo tem melhor desempenho, enquanto um valor menor que 0,5 indica desempenho ruim e a inclinação do ROC deve ser alta, pois representa alto TPR enquanto menos FPR ([10], 2019) .

No artigo [5], o estudo se concentra na modelagem de riscos de abandono e baixo desempenho usando dados disponíveis no momento em que os alunos se matriculam. O conjunto de recursos para nossos dois modelos consiste dados demográficos (sexo, idade e nacionalidade), tipo de acesso ao estudo, programa de estudo, número de créditos do primeiro ano e nota média de admissão. ML diferente algoritmos: regressão logística (LR), multi-layer perceptron (MLP) e decisão árvores são usadas para prever os riscos. Ambos os modelos de ML são treinados com alunos matriculados entre 2009 e 2015 (16.273 casos) e testados com alunos matriculados em 2016, 2017 e 2018 (6.823 casos). Devido a considerações de espaço e devido a a gravidade do abandono, focamos principalmente neste risco. Utilizando um método de seleção de recursos baseado em árvores de decisão (CART), verificamos que dentre os recursos disponíveis no momento da matrícula, os mais importantes na previsão do abandono risco são o número de créditos no primeiro ano (carga horária), série de ingresso, idade e tipo de acesso ao estudo.

No artigo [6], a fim de examinar o uso das técnicas de aprendizado no tema proposto, são aplicadas seis técnicas de aprendizado de máquina, a saber: Árvores de Decisão (*Decision Trees* (DT)) [11], Redes Neurais (*Neural Networks* (RN)) [12], Algoritmo Naive Bayes (ANB) [13] , Algoritmos de Aprendizagem Baseados em Instâncias (*Instance-Based Learning Algorithms* (IBLA)) [14], Regressão logística (*Logistic Regression* (LR)) [15] e Máquinas de vetores de suporte (*Support Vector Machines* (SVM)) [16] foram

usados para minerar dados educacionais. Posteriormente, foram realizadas a classificação e comparação entre as estas técnicas ou métodos de mineração de dados (EDM) para verificar a acurácia entre as mesmas.

3.1 Métricas de desempenho

Para avaliar os resultados nas métricas de classificação deste estudo, a matriz de confusão (ver Figura 1) e suas métricas de desempenho relacionadas, como exatidão (ACC) - Equação (1), precisão (PR) - Equação (3), recuperação (Rec) - Equação (4), área sob a curva de características operacionais do receptor (AUC) - Equação (5) e F1 Score (F1) - Equação (6), foram empregados ([17], 2020).

Diversas métricas de avaliação podem ser utilizadas para avaliar a qualidade dos classificadores tanto no processo de seleção da melhor configuração de hiperparâmetro quanto na classificando os diferentes modelos. A classificação produz valores de Verdadeiro Positivo (TP), Verdadeiro Negativo (TN), Falso Positivo (FP) e Falso Negativo (FN); no nosso caso, interpretamos um PF como a previsão de uma evasão que não ocorre, e um NF como um aluno que segundo a previsão do modelo continuará os estudos enquanto o fenômeno da evasão efetivamente ocorrer.

No caso de classificação binária, a precisão ou acurácia (ACC), a especificidade (SPEC) e a sensibilidade (SENS) são usados em vez de valores simples de TP, TN, FP e FN para melhorar a interpretabilidade dos resultados experimentais [18].

ACC é a razão entre a previsão correta sobre o número total de instâncias, medida pela Equação (1).

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

A SPEC ou True Negative Rate (TNR), é a proporção de TN para o número total de instâncias que possuem classe negativa real (Equação (2)). A "especificidade" é a razão entre os verdadeiros negativos (TN) e o total de negativos (TN + FP), ou seja, a especificidade mede a capacidade do modelo de classificar um *input* como negativo dado que ele realmente é negativo.

$$SPEC = \frac{TN}{TN + FP}. \tag{2}$$

e

$$PR = \frac{TP}{TN + FP}. \tag{3}$$

A SENS, também conhecida como recall ou True Positive Rate (TPR), é a proporção de TP para o número total de instâncias que possuem classe positiva real (Equação (4)).

$$SENS = \frac{TP}{TP + FP} = REC. \tag{4}$$

De acordo com [4], a AUC mostra o trade-off entre a taxa de TP e a taxa de FP e é calculada como [19]

$$AUC = \frac{1 + TP - FP}{2}. \tag{5}$$

e

$$\text{F1 SCORE} = \frac{2 \cdot REC \cdot PR}{REC + PR}. \tag{6}$$

O diagrama a seguir é conhecido como "Matriz de Confusão" e representa um resumo do desempenho de um modelo (Ver Figura 1).

Figura 1 – diagrama Matriz de Confusão : TP - Verdadeiro Positivo, FN - Falso Negativo, FP - Falso Positivo e TN - Verdadeiro Negativo

		Predicted class	
		Positive	Negative
True class	Positive	TP	FN
	Negative	FP	TN

Fonte: [3]

Finalmente, uma forma de avaliar o desempenho de um modelo de regressão logística é por meio da curva Receiver Operating Characteristic (ROC). [20], [21]

A curva ROC é obtida representando a taxa de verdadeiros positivos (sensibilidade) no eixo x versus a taxa de falso positivo (1-especificidade) no eixo y para p_0 variando entre 0 e 1.

A área sob a curva ROC (AUC) varia entre 0 e 1, essa métrica indica a capacidade do modelo de diferenciar corretamente entre casos de sucesso e fracasso [22]. Valores próximos a 1 indicam que o modelo apresenta bom desempenho (Tabela 5).

Tabela 5 – Níveis de poder de discriminação do modelo em função da AUC [22]

AUC = 0.5	Isso sugere nenhuma discriminação.
$0.7 \leq$ AUC < 0.8	Isso é considerado discriminação aceitável.
$0.8 \leq$ AUC < 0.9	Isso é considerado discriminação excelente.
AUC ≥ 0.9	Isso é considerado discriminação excepcional.

Fonte: [3].

$$\begin{cases} H_0 & : & \beta_i = 0 \\ H_1 & : & \beta_i \neq 0 \end{cases}, \ i = 0, 1, \cdots, n,$$

onde n é o número de variáveis independentes do modelo.

Portanto, quando o p-valor associado a uma variável é menor que o nível de significância α, rejeitamos a hipótese nula e concluímos que existe, de fato, uma associação entre a variável e a probabilidade de falha, ou seja, a variável independente é considerados significativos para o modelo.

3.1.1 Odds Ratio (OR): razão de probabilidade

Seja E um evento arbitrário e $p(E)$ sua respectiva probabilidade, as probabilidades de E são dadas por

$$Odds(E) = \frac{p(E)}{1 - p(E)}.$$

No modelo logístico, a chance de $E = S$, onde S é o evento de sucesso, depende do vetor de variáveis independentes $X = (x_1, x_2, \cdots, x_p)$, portanto, a chance de $E = S$ dado X (probabilidades (S/X)) são:

$$Odds(S/X) = \frac{p(S/X)}{1 - p(S/X)} = \frac{\dfrac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}}{1 - \dfrac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}} = \frac{\dfrac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}}{\dfrac{e^{-(\beta_0 + \sum \beta_i x_i)}}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}} = e^{(\beta_0 + \sum \beta_i x_i)},$$

onde $\sum \beta_i x_i$ é uma notação suprimida para

$$\sum_{i=1}^{p} \beta_i x_i.$$

Considerando dois vetores de variáveis independentes X_0 e X_1 , o Odds Ratio (OR) é dado por

$$OR(x_1, x_0) = \frac{odds(S/X_1)}{odds(S/X_0)} = \frac{e^{(\beta_0 + \sum \beta_i x_{1i})}}{e^{(\beta_0 + \sum \beta_i x_{0i})}} = e^{\sum \beta_i (x_{1i} - x_{0i})}.$$

O OR é a razão entre as chances de um determinado resultado ocorrer, considerando dois conjuntos possíveis de variáveis independentes X_0 e X_1. Quando a diferença entre X_0 e X_1 ocorre em uma única variável, o OR indica o quanto e como essa variável influencia na probabilidade de ocorrência do desfecho.

No escopo deste trabalho, o OR indica quanto cada variável preditora influencia a probabilidade de reprovação dos alunos.

No artigo [5], a propensão ao tratamento é estimada em cada cenário usando modelos de aprendizado de máquina (ML) e recursos de entrada, incluindo dados demográficos (sexo e nacionalidade), programas de estudo e nota média de admissão. Nos cenários 1 e 2, o tipo de acesso ao estudo também é adicionado como recurso e, no cenário 3, a idade é adicionada como recurso. Calculamos o efeito médio do tratamento (ATE) de cada tratamento na probabilidade de abandono usando vários métodos de inferência causal:

- O método de correspondência de pontuação de propensão [23], no qual os dados são classificados por pontuação de propensão e depois estratificados em listas (baldes) (cinco em nosso caso). Em nosso trabalho, obtemos o ATE subtraindo o abandono médio dos casos não tratados (controle) dos casos tratados em cada lista (balde).

- Ponderação de pontuação de propensão inversa (IPW) [24]: A ideia básica deste método está ponderando as medidas de resultado pelo inverso da probabilidade do indivíduo com um determinado conjunto de características sendo atribuído ao

tratamento de modo que sejam obtidas características basais semelhantes. Neste método, o tratamento efeito para o indivíduo i é obtido usando a seguinte equação:

$$TE_i = \frac{W_i Y_i}{p_i} - \frac{(1 - W_i)Y_i}{1 - p_i}, \tag{7}$$

W_i mostra tratamento (1 para casos tratados e 0 para casos de controle), p_i representa a probabilidade de receber tratamento (escore de propensão ao tratamento) e Y_i mostra abandono (1 se abandono e 0 se não abandono) para o indivíduo i.

– Augmented Inverse-Propensity Weighted (AIPW) [25]: Este método combina as propriedades do estimador baseado em regressão e do estimador IPW. Possui uma parte de aumento $(W_i - p_i)Y_i$ para o método IPW, em que Y_i é a probabilidade estimada de abandono usando todas as características aplicadas ao modelo de escore de propensão mais a variável de tratamento. Assim, este estimador pode levam a estimativas duplamente robustas que requerem apenas a propensão ou o modelo de resultado para serem especificados corretamente, mas não ambos. Podemos calcular o efeito do tratamento no indivíduo i como:

$$TE_i = \frac{W_i Y_i - (W_i - p_i)\hat{Y}_i}{p_i} - \frac{(1 - W_i)Y_i - (W_i - p_i)\hat{Y}_i}{1 - p_i}, \tag{8}$$

– Florestas causais do pacote EconML: Este método usa o Doublely Robust Florestas Ortogonais (DROrthoForest) que são uma combinação de florestas causais e aprendizado de máquina duplo para estimar não parametricamente o efeito do tratamento para cada indivíduo.

Em IPW, AIPW e DROrthoForest, obtemos o tratamento individual efeito TE_i, que é a diferença entre os resultados se a pessoa for tratada (tratamento) e não tratado (controle). Em outras palavras, esse efeito é a diferença de probabilidade de evasão quando o aluno é tratado e não tratado; um valor negativo indica um risco de abandono reduzido e um valor positivo indica um risco de abandono aumentado. O ATE resultante é a média sobre os efeitos individuais do tratamento.

3.2 Linear Discriminant Analysis (LDA)

Nesta Seção, iremos introduzir uma ideia sobre o método de análise LDA, com os fundamentos de Classificação, extraídos de [26], associada as ideias de [27] e [28], citados em [1].

O autor [27], aborda um exemplo de LDA para duas classes de populações pesquisadas, em que a função discriminante de Fisher é definida por

$$Y = \ell^T \cdot X,$$

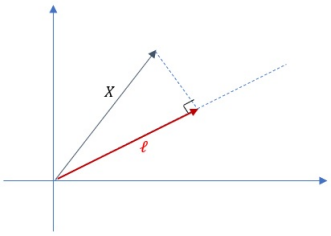
onde $\ell^T = (\ell_1, \ell_2, \cdots, \ell_p)$ e $X = (X_1, X_2, \cdots, X_p)$ são vetores constantes.

O objetivo da LDA é escolher o melhor vetor $\ell \in \mathbb{R}^p$ de forma que as funções Y's da população 1 (π_1) e Y's da população 2 (π_2), sejam o mais separados possíveis[†].

[†] Para um problema linearmente separável, o problema consiste em rotacionar os dados de maneira a maximizar a distância entre as classes e minimizar a distância intra-classe.

A metodologia utilizada para otimizar ℓ é utilizar-se da Projeção Ortogonal ℓ de X, conforme mostrado na Figura 2.

Figura 2 – Projeção ortogonal de X em ℓ.



Fonte: elaborada pelo autor.

Neste caso, a projeção ortogonal é definida por

$$\frac{X^T\ell}{\|\ell\|}\cdot\ell = \left(X^T\cdot\ell\right)\cdot\ell,$$

com $\|\ell\| = 1$, ou seja, ℓ seja ortonormal e $\left(X^T\cdot\ell\right)$ é o tamanho das projeções.

Suponha que $\|\ell\| = 1$, estamos buscando direção ℓ em que $\ell^T X = Y$ dos dois grupos sejam maximalmente separados.

Definimos a projeção $Y = \ell^T X$ de dois vetores X: um de pop1 (π_1) e outro de pop2 (π_2).

Qual a melhor direção de ℓ? Para encontrar a solução Fisher raciocinou assim inicialmente:

- Projete cada ponto X ao longo de ℓ gerando o escalar $Y = \ell^T X$.
- Calcule a média de: pop1 $= \overline{y}_1$ e a média de pop2 $= \overline{y}_2$.

Procure a direção ℓ em que $\|\overline{y}_1 - \overline{y}_2\|$ seja máxima.

Mas isso tem um problema: $\|\overline{y}_1 - \overline{y}_2\|$ pode ser grande mas as projeções não estão bem separadas.

Conclusão fundamental:

- A projeção dos dados aleatórios X em uma direção fixa ℓ produz a variável aleatória $Y = \ell^T X$, uni-dimensional.
- Para medir a separação entre as populações projetadas não devemos simplesmente olhar $\|\overline{y}_1 - \overline{y}_2\|$.
- A razão é que a projeção impacta não apenas as médias $\overline{y}_1$ e $\overline{y}_2$ mas impacta também a variabilidade dos dados.
- Assim, devemos considerar $\|\overline{y}_1 - \overline{y}_2\|$ relativamente ao desvio padrão[‡] s_y dos dados

[‡] O Desvio Padrão consiste em uma medida do nível de dispersão, isto é, ele indica quão uniforme está um conjunto de dados. É a raiz quadrada da variância.

projetados

$$s_y = \sqrt{\frac{\sum_{k=1}^{n}(y_k - \overline{y})}{n-1}}.$$

- Isto é, vamos procurar ℓ para maximizar $\dfrac{\|\overline{y}_1 - \overline{y}_2\|}{s_y}$.

Formulação do problema: Vamos supor que temos dados de duas populações ou duas classes: π_1 e π_2. Os dados são vetores aleatórios X de dimensão $p \times 1$ com densidades $f_1(x)$ e $f_2(x)$. Dados não precisam ser gaussianos. As classes possuem médias diferentes mas a mesma matriz de covariância (Equação (14)):

$$\begin{array}{rcll} \mathbb{E}(X| \in \pi_1) & = & \mu_1, & p \times 1 \\ \mathbb{E}(X| \in \pi_2) & = & \mu_2, & p \times 1 \\ \mathbb{V}(X| \in \pi_1) & = & \mathbb{V}(X| \in \pi_2) = \Sigma, & p \times p \end{array}$$

3.3 Função Discriminante Linear de Fisher

A função objetivo, deduzida por [27], tem as seguintes características: com um vetor $\ell \in \mathbb{R}^p$, reduzimos o vetor $p - dimensional$ X a um escalar unidimensional: $Y = \ell^T X$. Para π_1, o valor esperado de Y será $\mathbb{E}(Y| \in \pi_1) = m_1 = \ell^T \mu_1$ e para π_2, temos $\mathbb{E}(Y| \in \pi_2) = m_2 = \ell^T \mu_2$, isto é, a média das projeções é a projeção da média. A Variância[§] da projeção Y em torno de suas duas médias é a mesma: $\mathbb{V}(Y| \in \pi_1) = \mathbb{V}(Y| \in \pi_2) = \ell^T \Sigma \ell, 1 \times 1$. Veja que a variância de Y muda com ℓ.

Queremos encontrar ℓ que maximize a separação das duas populações. Como medir a separação?

Como discutimos, não queremos apenas maximizar $\|m_1 - m_2\|$. Devemos considerar também a dispersão (variância, desvio-padrão) de Y em torno de suas duas médias m_1 e m_2.

Queremos ℓ que maximize

$$\begin{array}{rcl} U & = & \dfrac{\|m_1 - m_2\|^2}{\mathbb{V}(Y)} \\ & = & \dfrac{\|\mathbb{E}(Y| \in \pi_1) - \mathbb{E}(Y| \in \pi_2)\|^2}{\mathbb{V}(Y)} \\ & = & \dfrac{\|\ell^T \mu_1 - \ell^T \mu_2\|^2}{\ell^T \Sigma \ell} \\ & = & \dfrac{\|\ell^T (\mu_1 - \mu_2)\|^2}{\ell^T \Sigma \ell} \\ & = & \dfrac{\ell^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \ell}{\ell^T \Sigma \ell}. \end{array}$$

Precisamos da última representação para usar um teorema de álgebra linear:

> **Teorema.** Seja v um vetor $p \times 1$ e Σ uma matriz $p \times p$ positiva definida e simétrica. Seja x um vetor $p \times 1$ não-nulo. Vamos considerar o comprimento

[§] Variância é uma medida unidimensional. É calculada de maneira independente pois não leva em consideração as outras dimensões.

(ao quadrado) de x usando a distância estatística (baseada em Σ) até a origem 0 e definida por $d^2 = x^T\Sigma x$. Assim, $\dfrac{x}{d} = \dfrac{x}{\sqrt{x^T\Sigma x}}$ é um vetor de comprimento 1 (ou norma-Σ igual a 1). Seja $\mathbf{B}$ o conjunto de vetores de norma-Σ igual a 1. O conjunto $\mathbf{B}$ é um $p-$dim elipsóide com eixos nas direções dos autovetores de Σ. Dentre todos os vetores w com norma$-\Sigma$ igual a 1, isto é, dentre todos os vetores $w = \mathbf{B}$, aquele vetor w que maximiza

$$\max_{w\in\mathbf{B}}(w^T v)^2 = \max_{x\neq 0}\frac{\|x^T v\|^2}{x^t\Sigma x}$$

é igual a

$$w = c\Sigma^{-1}v.$$

Aplicar o Teorema no LDA:
No problema do LDA precisamos encontrar um vetor ‘ que maximize

$$U = \frac{\|\ell^T(\mu_1 - \mu_2)\|^2}{\ell^T\Sigma\ell}.$$

Isto recai perfeitamente no caso do teorema anterior e a solução é

$$\ell = \Sigma^{-1}(\mu_1 - \mu_2).$$

Na prática, temos de estimar os parâmetros das duas distribuições com os dados da amostra. Por exemplo, μ_1 vira o vetor de médias aritméticas dos dados da amostra.

Queremos maximizar a separação $= \dfrac{\|\overline{y}_1 - \overline{y}_2\|}{s_y^2}$. S_y é o desvio padrão (DP) amostral com as observações no eixo da projeção

$$s_y^2 = \frac{1}{m_1 + m_2 - 2}\left(\sum_{j=1}^{m_1}(y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{m_2}(y_{2j} - \overline{y}_2)^2\right).$$

A combinação linear $Y = \ell^T X$ que maximiza a separação $\dfrac{\|\overline{y}_1 - \overline{y}_2\|}{s_y}$ é dada por

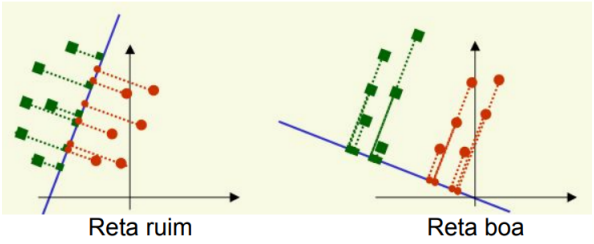
$$\ell^T = (\overline{x}_1 - \overline{x}_2)^T\left(s^2_{agrupado}\right)^{-1},$$

onde $s^2_{agrupado} = \dfrac{m_1}{m}s^2_{pop1} + \dfrac{m_2}{m}s^2_{pop2}$ e s^2_{pop2} a matriz de variância e covariância[¶] amostral da pop1 (um exemplo é a matriz (16) da população (π_1) da Tabela 6).

De acordo com [28], a LDA tenta encontrar uma transformação linear através da maximização da distância entre-classes e minimização da distância intra-classe. O método tenta encontrar a melhor direção de maneira que quando os dados são projetados em um plano, as classes possam ser separadas. A Figura 3 mostra duas situações em que o método LDA classifica duas classes com otimizações ruim e boa.

[¶] Covariância por sua vez, é uma medida bi-dimensional. Verifica a dispersão, mas levando em consideração duas variáveis aleatórias.

Figura 3 – Classificação de dois conjuntos de dados.



Fonte: [28].

A função discriminante linear de Fisher é uma combinação linear de características originais a qual se caracteriza por produzir separação máxima entre duas populações. Considerando que μ_i e $\sum$ são parâmetros conhecidos e respectivamente, os vetores de médias aritméticas e a matriz de covariâncias comum das populações π_i, $i = 1, \cdots, p$. Demonstra-se que a função linear do vetor aleatório $X = (X_1, X_2, \cdots, X_p)$ que produz separação máxima entre duas populações é dada por:

$$\begin{aligned} Y &= \ell^T \cdot X \\ &= [\mu_1 - \mu_2]^T \cdot \Sigma^{-1} \cdot X, \end{aligned} \tag{9}$$

onde $\ell^T = (L_1, L_2, \cdots, \cdots, L_p)$ é um vetor discriminante de constantes, X é vetor aleatório de características das populações, μ é o vetor de médias $p - variado$ e Σ é a matriz de covariâncias comum as populações π_1 e π_2.

O valor da função discriminante de Fisher para uma dada observação X_0 é:

$$Y(x_0) = [\mu_1 - \mu_2]^T \cdot \Sigma^{-1} \cdot x_0. \tag{10}$$

O ponto médio entre as duas médias populacionais univariadas μ_1 e μ_2 é:

$$\begin{aligned} m &= \frac{1}{2}[\mu_1 - \mu_2]^T \cdot \Sigma^{-1} \cdot [\mu_1 + \mu_2] \\ &= \frac{1}{2}[D(\mu_1) + D(\mu_2)]. \end{aligned} \tag{11}$$

A regra de classificação baseada na função discriminante de Fisher é:

(i) Alocar x_0 em π_1 se $D(x_0) = [\mu_1 - \mu_2]^T \cdot \Sigma^{-1} \cdot x_0 \geq m$;

(ii) alocar x_0 em π_2 se $D(x_0) = [\mu_1 - \mu_2]^T \cdot \Sigma^{-1} \cdot x_0 < m$.

Assumindo-se que as populações π_1 e π_2 têm a mesma matriz de covariâncias Σ podemos então estimar uma matriz comum de covariâncias S_c:

$$S_c = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] \cdot S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] \cdot S_2. \tag{12}$$

3.3.1 Função discriminante amostral

A função discriminante linear amostral de Fisher é obtida substituindo-se os parâmetros μ_1, μ_2 e Σ pelas respectivas quantidades amostrais $\overline{x}_1$, $\overline{x}_2$ e S_c na Equação (9):

$$\begin{aligned} Y(x) &= \hat{\ell}^T \cdot x \\ &= [\overline{x}_1 - \overline{x}_2]^T \cdot S_c^{-1} \cdot x, \end{aligned} \tag{13}$$

3.3.2 Aplicação

Como exemplo ilustrativo para obtenção da função discriminante linear amostral de Fisher, vamos considerar os dados de duas raças de insetos (Tabela 6), apresentados por HOEL (1966) e citado por REGAZZI (2000).

Tabela 6 – Número médio de cerdas primordiais (X_1) e número médio de cerdas distais (X_2) em duas raças de insetos

k	Raça A(π_1)		Raça B (π_2)	
	X_1	X_2	X_1	X_2
1	6.36	5.24	6.00	4.88
2	5.92	5,12	5.60	4.64
3	5.92	5,36	5.64	4.96
4	6.44	5,64	5.76	4.80
5	6.40	5,16	5.96	5.08
6	6.56	5,56	5.72	5.04
7	6.64	5,36	5.64	4.96
8	6.68	4,96	5.44	4.88
9	6.72	5,48	5.04	4.44
10	6.76	5,60	4.56	4.04
11	6.72	5,08	5.48	4.20
12	–	–	5.76	4.80
Somatório	71.12	58.56	66.60	56.72

3.3.3 Covariância

$$cov(X_i, X_i) = \frac{\sum_{k=1}^{n} (X_{k1} - \mu_1)(X_{k2} - \mu_2)}{n_i - 1}, \quad i = 1, 2.$$

3.3.4 Matriz de Covariância 2×2

$$S_i = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) \\ cov(X_2, X_1) & cov(X_2, X_2) \end{bmatrix}, \quad i = 1, 2. \tag{14}$$

Estimativa das médias das raças A e B

Para um dado conjunto de dados da Tabela 6, calcule os vetores médios de cada classe μ_1 e μ_2 (centróides) e o vetor médio geral, μ.

Com base nos dados apresentados na Tabela 6, temos:

As médias da classe **Raça A** são:

$$\overline{x}_{1A} = \frac{\sum_{k=1}^{n_1=11} X_{k1}}{n_1} = \frac{71.12}{11} = 6.46545$$

e

$$\overline{x}_{2A} = \frac{\sum_{k=1}^{n_1=11} X_{k1}}{n_1} = \frac{58.56}{11} = 5.32364$$

O vetor das médias da Raça A é:

$$\overline{x}_1 = \begin{bmatrix} \overline{x}_{1A} \\ \overline{x}_2 A \end{bmatrix} = \begin{bmatrix} 6.46545 \\ 5.32364 \end{bmatrix} \tag{15}$$

A matriz covariância da Raça A é:

$$S_1 = \begin{bmatrix} 0.09129 & 0.01126 \\ 0.01126 & 0.05263 \end{bmatrix}. \tag{16}$$

As médias da classe **Raça B** são:

$$\overline{x}_{2B} = \frac{\sum_{k=1}^{n_1=12} X_{k1}}{n_1} = \frac{66.60}{12} = 5.55000$$

e

$$\overline{x}_{2B} = \frac{\sum_{k=1}^{n_1=12} X_{k1}}{n_1} = \frac{56.72}{12} = 4.72667$$

O vetor das médias da Raça B é:

$$\overline{x}_2 = \begin{bmatrix} \overline{x}_{2B} \\ \overline{x}_2 B \end{bmatrix} = \begin{bmatrix} 5.5500 \\ 4.72667 \end{bmatrix} \tag{17}$$

A matriz covariância da Raça B é:

$$S_2 = \begin{bmatrix} 0.160327 & 0,107418 \\ 0,107418 & 0.111661 \end{bmatrix}. \tag{18}$$

Substituir $n_1 = 11$, $n_2 = 12$, as matrizes (16) e (18) na Equação (12):

$$S_c = \left[\frac{10}{21}\right] \cdot \begin{bmatrix} 0.09129 & 0.01126 \\ 0.01126 & 0.05263 \end{bmatrix} + \left[\frac{11}{21}\right] \cdot \begin{bmatrix} 0.160327 & 0,107418 \\ 0,107418 & 0.111661 \end{bmatrix}.$$

$$S_c = \begin{bmatrix} 0.12745 & 0.06163 \\ 0.06163 & 0.08355 \end{bmatrix}. \tag{19}$$

E a sua inversa é:

$$S_c^{-1} = \begin{bmatrix} 12.1960015 & -8.995964 \\ -8.995964 & 18.604583 \end{bmatrix}. \tag{20}$$

Obtenção da função discriminante linear amostral de Fisher

Substituir os vetores (15) e (17) e a matriz inversa (20) na Equação (13):

$$\begin{aligned} Y(x_1, x_2) &= \left[\begin{bmatrix} 6.46545 \\ 5.32364 \end{bmatrix} - \begin{bmatrix} 5.5500 \\ 4.72667 \end{bmatrix} \right] \cdot \begin{bmatrix} 12.1960015 & -8.995964 \\ -8.995964 & 18.604583 \end{bmatrix} \cdot x \\ &= \begin{bmatrix} 0.91545 \\ 0,59697 \end{bmatrix} \cdot \begin{bmatrix} 12.1960015 & -8.995964 \\ -8.995964 & 18.604583 \end{bmatrix} \cdot x \\ &= \begin{bmatrix} 0.91545 & 0,59697 \end{bmatrix}^T \cdot \begin{bmatrix} 12.1960015 & -8.995964 \\ -8.995964 & 18.604583 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} 5.794819 & 2,871023 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned} \tag{21}$$

Segue que:

$$Y(x_1, x_2) = 5.794819x_1 + 2.871023x_2.$$

4 Considerações Finais

No processo de leitura dos artigos científicos, tivemos a oportunidade de aprender novas definições inerentes a aprendizagem de máquina (*Machine Learning* (ML)) e a sua importância para detectar pontos relevantes, ou padrões, em um banco de dados sobre educação, através de uma técnica denominada de mineração de dados educacionais (Educational Data Mining (EDM)). Esta técnica ou método é estruturada por algoritmos para realizar a tarefa de lê e interpretar abundante dados numéricos inseridos na memória de computadores, o qual são os bancos de dados.

Existem na *internet* diversas plataformas que nos auxiliam, como é o caso do Oracle e da Weka - que são uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados (EDM). Eles contêm ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização.

Até agora você já sabe muito sobre Machine Learning. No entanto, passamos por tantos conceitos que você pode estar se sentindo um pouco perdido, então daremos um passo para trás e olhar para o quadro geral:

- Aprendizado de máquina é fazer com que as máquinas melhorem em alguma tarefa aprendendo com os dados, em vez de ter que codificar regras explicitamente.
- Existem muitos tipos diferentes de sistemas de ML: supervisionados ou não, em lote ou online, baseados em instâncias ou modelos, e assim por diante.
- Em um projeto de ML, você coleta dados em um conjunto de treinamento e alimenta o conjunto de treinamento para um algoritmo de aprendizado. Se o algoritmo for baseado em modelo, ele ajusta alguns parâmetros para ajustar o modelo ao conjunto de treinamento (ou seja, para fazer boas previsões no próprio conjunto de treinamento) e, com sorte, também será capaz de fazer boas previsões

em novos casos. Se o algoritmo for baseado em instâncias, ele apenas memoriza os exemplos e usa uma medida de similaridade para generalizar para novas instâncias.

- O sistema não funcionará bem se o seu conjunto de treinamento for muito pequeno ou se os dados não forem representativos, ruidosos ou poluídos com recursos irrelevantes (entrada de lixo, saída de lixo). Por fim, seu modelo não precisa ser nem muito simples (caso em que será subajustado) nem muito complexo (nesse caso, será superajustado).

Há apenas um último tópico importante a ser abordado: após treinar um modelo, você não quer apenas "esperar" que ele generalize para novos casos. Você deseja avaliá-lo e ajustá-lo, se necessário.

ORCID

João Socorro Pinheiro Ferreira https://orcid.org/0000-0002-3711-3602

Fontes de Financiamento

Não há.

Referências

1. DEL BONIFRO, Francesca *et al.* Student dropout prediction. *In*: **Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21**. Springer International Publishing, 2020. p. 129-140. https://doi.org/10.1007/978-3-030-52237-7_11
2. YU, Renzhe; LEE, Hansol; KIZILCEC, René F. Should college dropout prediction models include protected attributes?. *In*: **Proceedings of the eighth ACM conference on learning@ scale**. 2021. p. 91-100. https://doi.org/10.1145/3430895.3460139
3. COSTA, Stella F.; DINIZ, Michael M. Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course. **Education and Information Technologies**, v. 27, n. 9, p. 12381-12397, 2022. https://doi.org/10.1007/s10639-022-11117-1
4. MÁRQUEZ-VERA, Carlos et al. Early dropout prediction using data mining: a case study with high school students. **Expert Systems**, v. 33, n. 1, p. 107-124, 2016. https://doi.org/10.1111/exsy.12135
5. KARIMI-HAGHIGHI, Marzieh; CASTILLO, Carlos; HERNÁNDEZ-LEO, Davinia. A causal inference study on the effects of first year workload on the dropout rate of undergraduates. *In*: **International Conference on Artificial Intelligence in Education**. Cham: Springer International Publishing, 2022. p. 15-27. https://doi.org/10.1007/978-3-031-11644-5_2
6. KOTSIANTIS, Sotiris B.; PIERRAKEAS, C. J.; PINTELAS, Panayiotis E. Preventing student dropout in distance learning using machine learning techniques. *In*: **Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, September 2003. Proceedings, Part II 7**. Springer Berlin Heidelberg, 2003. p. 267-274.
7. CHANG, Chih-Chung; LIN, Chih-Jen. LIBSVM: a library for support vector machines. **ACM transactions on intelligent systems and technology (TIST)**, v. 2, n. 3, p. 1-27, 2011.

8. BREIMAN, Leo. Bagging predictors. **Machine learning**, v. 24, p. 123-140, 1996.
9. MUSCHELLI III, John. ROC and AUC with a binary predictor: a potentially misleading metric. **Journal of classification**, v. 37, n. 3, p. 696-708, 2020. https://doi.org/10.1007/s00357-019-09345-1
10. KOIZUMI, Yuma et al. SNIPER: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate. *In*: **ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2019. p. 915-919.
11. MURTHY, Sreerama K. Automatic construction of decision trees from data: A multi-disciplinary survey. **Data mining and knowledge discovery**, v. 2, p. 345-389, 1998.
12. MITCHELL, Tom M. Machine learning. 1997.
13. DOMINGOS, Pedro; PAZZANI, Michael. On the optimality of the simple Bayesian classifier under zero-one loss. **Machine learning**, v. 29, p. 103-130, 1997.
14. AHA, D. Lazy Learning. Kluwer Academic Publishers. 1997.
15. SCOTT LONG, John. Regression models for categorical and limited dependent variables. **Advanced quantitative techniques in the social sciences**, v. 7, 1997.
16. BURGES, Christopher J C. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, v. 2, n. 2, p. 121-167, 1998.
17. POWERS, David MW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061, 2020.
18. FREEMAN, Elizabeth A.; MOISEN, Gretchen G. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. **Ecological modelling**, v. 217, n. 1-2, p. 48-58, 2008.
19. LÓPEZ, Victoria et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. **Information sciences**, v. 250, p. 113-141, 2013.
20. BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern recognition**, v. 30, n. 7, p. 1145-1159, 1997.
21. SPACKMAN, Kent A. Signal detection theory: Valuable tools for evaluating inductive learning. *In*: **Proceedings of the sixth international workshop on Machine learning**. Morgan Kaufmann, 1989. p. 160-163.
22. HOSMER, D. W.; LEMESHOW, Stanley. John Wiley & Sons. **New York**, 2000.
23. ROSENBAUM, Paul R.; RUBIN, Donald B. The central role of the propensity score in observational studies for causal effects. **Biometrika**, v. 70, n. 1, p. 41-55, 1983.
24. BRAY, Bethany C. et al. Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later. **Prevention Science**, v. 20, p. 394-406, 2019.
25. GLYNN, Adam N.; QUINN, Kevin M. An introduction to the augmented inverse propensity weighted estimator. **Political analysis**, v. 18, n. 1, p. 36-56, 2010.
26. VARELLA, Carlos Alberto Alves. Análise multivariada aplicada as ciencias agrárias. **Seropédica: Universidade Federal Rural do Rio de Janeiro**, 2008.
27. ASSUNÇÃO, R. **Linear Discriminant Analysis**. Minas Gerais: DCC-UFMG, 2020.
28. MENOTTI, D. Classificação. Universidade Federal do Paraná (UFPR). Especialização em Engenharia Industrial 4.0. Paraná.