Predição de Evasão e Desempenho Estudantil com Machine Learning

Uma análise comparativa e implementação manual de Árvore de Decisão

Disciplina: Machine Learning

Instituição: Fatec Mauá

Apresentação: Kaike Medeiros Dourado e Sérgio Henrique Basilio Reis

A Relevância Crítica da Predição de Evasão

Problema Global

A evasão afeta a eficácia e o desempenho de instituições de ensino superior em todo o mundo.

Identificação de Risco

Modelos preditivos localizam alunos em potencial risco, permitindo a implementação de medidas preventivas.

Precisão Analítica

ML oferece maior exatidão na análise e interpretação de grandes volumes de dados acadêmicos e comportamentais.

Intervir antes que o abandono se materialize transforma dados em oportunidades de sucesso educacional.

Estudo de Caso 1: Previsão de Evasão

Predicting Student Dropout Using Machine Learning Algorithms (2024)

Este artigo de **Suleyman A. Sulak** e **Nigmet Koklu** foca na aplicação de modelos de ML para prever e, consequentemente, reduzir as taxas de evasão em instituições de ensino.

Objetivo Central

Aplicar ML para prever o estado final do estudante (evasão, matriculado, graduado).

Fonte de Dados

Dataset público do Kaggle, compreendendo **4.424 registros**.

Variáveis

O estudo utilizou **37 variáveis**, incluindo dados demográficos, desempenho e informações curriculares.

3 Classes de Saída

Evasão (Dropout) • Ativo (Enrolled) • Concluído (Graduate)

Metodologia: Algoritmos e Avaliação

Decision Tree (DT)

Focado na interpretabilidade e na separação hierárquica das classes.

Random Forest (RF)

Usa múltiplos DTs para maior robustez e menor risco de overfitting.

Artificial Neural Network (ANN)

Modelo complexo capaz de capturar relações não lineares nos dados.

Métricas de Desempenho

O desempenho dos modelos foi avaliado usando métricas padrão em classificação, garantindo uma avaliação completa da qualidade preditiva:

Acurácia

Proporção de previsões corretas.

Precisão

Relevância dos resultados positivos.

Recall

Capacidade de identificar corretamente todos os positivos (alunos em risco).

F-Score

Média harmônica entre Precisão e Recall.

Resultados do Estudo de Evasão

Decision Tree (DT)

70.1%

Focado na interpretabilidade e na separação hierárquica das classes.

Desempenho inicial, base para comparação.

Random Forest (RF)

75.5%

Usa múltiplos DTs para maior robustez e menor risco de overfitting.

Melhoria significativa devido ao ensemble learning.

Artificial Neural Network (ANN)

77.3%

Modelo complexo capaz de capturar relações não lineares nos dados.

Melhor resultado, destacando sua capacidade com complexidade de dados.

Conclusão: Os autores concluíram que a ANN é mais apta para a tarefa. Contudo, enfatizaram que a evasão é multifatorial e que variáveis emocionais e sociais não capturadas no dataset inicial possuem forte influência e merecem atenção em futuras análises.

Estudo de Caso 2: Revisão de Desempenho

Estudos Analisados

82

Early Prediction of Student Learning Performance Through Data Mining (2021)

López-Zambrano, Lara Torralbo e Romero

Objetivo

Mapear e sintetizar os métodos e resultados mais eficazes na predição de performance estudantil.

Metodologia

Análise de 82 estudos de ponta publicados até o ano de 2020.

Foco

Identificar padrões e a eficiência de diferentes algoritmos no contexto educacional.

Panorama Metodológico e Algorítmico (Artigo 2)

Técnicas Predominantes

Classificação

Algoritmos supervisionados para categorizar estados estudantis.

Regressão

Modelos para prever valores contínuos de desempenho.

Outras Técnicas

Clustering, análise de padrões e mineração de regras.

Algoritmos Mais Comuns

- **Decision Tree (J48)**
- **Random Forest**
- **Support Vector Machine (SVM)**
- **Naive Bayes**
- **Logistic Regression**

Acurácias alcançadas: Estudos utilizando Random Forest e Regressão Linear demonstraram acurácias de até **96%** em diferentes ambientes educacionais (E-learning, B-learning e Presencial).

Comparativo: Síntese e Convergência de Resultados

Ambos os estudos, um experimental e outro de revisão, confirmam o valor do Machine Learning para o futuro da educação preditiva.

01

Potencial Comprovado

Os dois artigos atestam o poder do ML para prever resultados cruciais (evasão e desempenho).

02

Melhores Modelos

Redes Neurais (ANN) e Random Forest (RF) consistentemente demonstram maior poder preditivo em datasets complexos.

03

Foco em Comportamento

O segundo estudo reforça a importância das variáveis comportamentais e de engajamento em ambientes digitais.

04

Metodologias

Artigo 1: Experimento prático; Artigo 2: Revisão sistemática e mapeamento de técnicas.

Conclusões e Perspectivas Futuras

O Machine Learning é inquestionavelmente uma ferramenta eficaz, mas o futuro da predição na educação exige uma abordagem mais holística.

Eficácia de ML

Algoritmos como ANN e Random Forest são ideais para lidar com a alta dimensionalidade e as relações não lineares dos dados acadêmicos.

Exploração Holística

É crucial incluir variáveis latentes, como fatores emocionais, motivação e engajamento social, para refinar a precisão preditiva.

Sistemas de Alerta

A integração de modelos preditivos em sistemas de alerta precoce pode permitir intervenções oportunas, reduzindo drasticamente a evasão.

O futuro da retenção passa pela intervenção orientada por dados.

Síntese da Análise Comparativa

A análise dos dois estudos (experimental e revisão sistemática) confirma que o **Machine Learning é uma ferramenta eficaz** para prever evasão e desempenho estudantil.

Modelos como **Random Forest** e **Redes Neurais Artificiais** demonstram capacidade superior na captura de padrões complexos em dados educacionais.

Próximos Passos

Exploraremos a implementação manual de uma **Árvore de Decisão (CART)**, demonstrando como esses algoritmos funcionam internamente e como o **Índice de Gini** guia o processo de aprendizado.

O Desafio da Implementação Manual: Entendendo o Core do CART

A implementação manual do algoritmo **CART (Classification and Regression Trees)**, utilizando o **Índice de Gini**, demonstra o entendimento profundo da mecânica de Machine Learning, indo além do uso de bibliotecas de alto nível.

Objetivo

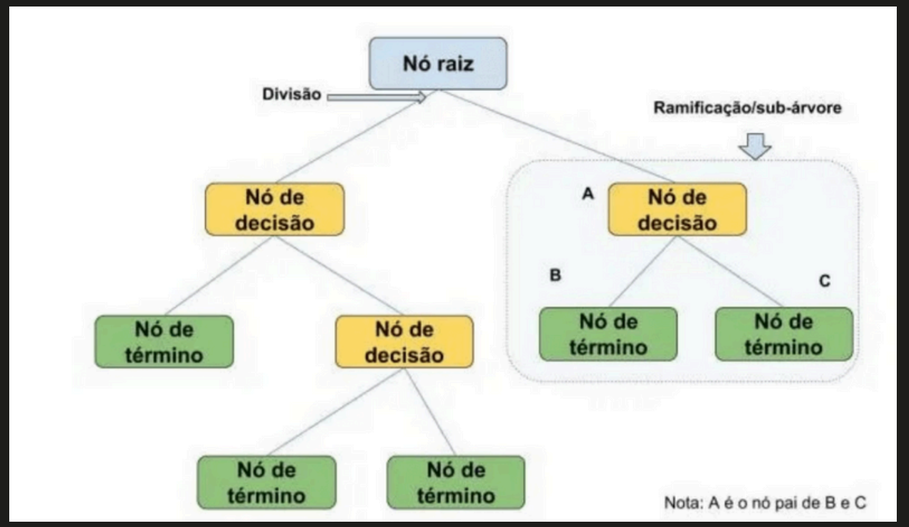
Replicar a lógica de construção da Árvore de Decisão para classificar o estado final do estudante (Evasão, Ativo, Graduado).

Metodologia

- **Algoritmo:** CART (Árvore de Decisão Binária).
- **Critério de Divisão:** Índice de Gini (medida de impureza).
- **Parâmetros:** Profundidade Máxima = 5; Tamanho Mínimo do Nó = 10.

Por que a Implementação Manual?

- **Transparência:** Revela como o modelo toma decisões em cada nó.
- **Controle:** Permite otimizar cada etapa do processo de divisão.
- **Aprendizado:** Consolida o conhecimento teórico sobre a teoria de árvores.



O Coração do Algoritmo: Índice de Gini e Divisão

1. Cálculo do Índice de Gini

Mede a **impureza** de um nó.

$$Gini(D) = 1 - \Sigma(p_i)^2$$

onde p_i = proporção da classe i

O Gini é minimizado quando o nó é "puro" (todas as amostras pertencem à mesma classe).

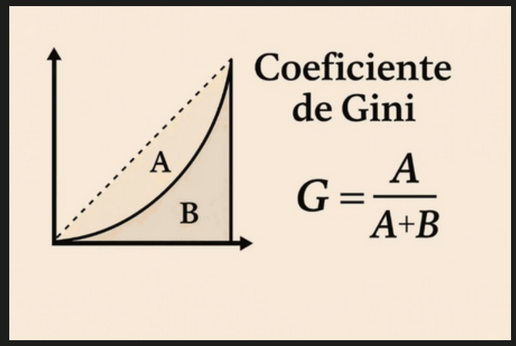
2. Encontrando a Melhor Divisão

A função get_best_split itera sobre:

- **Todos os atributos** (colunas) do dataset.
- **Todos os valores** únicos como pontos de corte.

Calcula o Gini ponderado para cada divisão.

A divisão escolhida resulta no **menor Índice de Gini** (maior ganho de informação).



3. Construção Recursiva

A árvore é construída recursivamente, parando quando:

- Profundidade máxima é atingida.
- Tamanho mínimo do nó é alcançado.
- Nó é puro (uma única classe).

Nós terminais (to_terminal) são criados com a **classe majoritária** do grupo.

Interpretação da Árvore Gerada: Fatores Chave de Evasão

Nó Raiz (Fator Mais Importante)

Variável:

Curricular units 2nd sem (approved)

Divisão:

[Curricular units 2nd sem (approved) < 4.00]

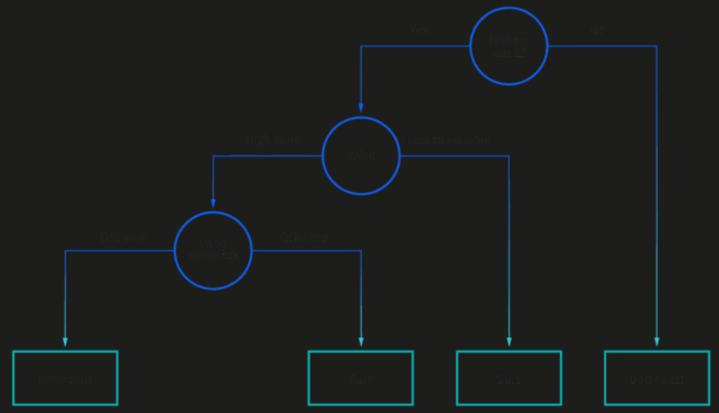
Interpretação

Ramo Esquerdo (< 4 UCs):

Alunos que aprovam **menos de 4 UCs** no 2º semestre tendem a seguir um caminho de **Evasão** ou permanência **Ativa** (com risco).

Ramo Direito (≥ 4 UCs):

Alunos que aprovam **4 ou mais UCs** tendem a seguir um caminho de **Graduação**.



Outros Fatores Relevantes

Tuition fees up to date

Fator administrativo/financeiro crucial.

Admission grade

Indicador inicial de aptidão.

Curricular units 1st sem (enrolled)

Engajamento inicial no programa.

Conclusão da Implementação

Resultado da Implementação

A árvore de profundidade 5 e tamanho mínimo de nó 10 alcançou uma **precisão de 75,45%** no conjunto de treinamento, validando a eficácia do algoritmo CART com o Índice de Gini como critério de divisão.

A implementação manual confirmou que o algoritmo consegue capturar padrões complexos nos dados educacionais.

Validação das Variáveis

A análise da árvore gerada **confirma que o desempenho no segundo semestre** é o principal preditor de sucesso ou evasão, reforçando os achados da literatura.

Variáveis críticas identificadas:

- Unidades Curriculares Aprovadas (2º Semestre)
- Situação de Propinas
- Nota de Admissão

Valor Agregado

A capacidade de construir o modelo do zero fornece uma **base sólida** para a compreensão e customização de algoritmos de Machine Learning em contextos educacionais.

Implementações manuais permitem:

- Otimização específica para domínios
- Transparência total do processo
- Controle fino sobre parâmetros

A implementação manual do CART valida a importância das variáveis de desempenho acadêmico e demonstra a eficácia do Índice de Gini como critério de divisão na predição de evasão estudantil.