Uso do Teorema de Bayes para Classificação de Alunos com Alto Risco de Evasão Escolar: Um Estudo de Caso

Ênio Henrique Nunes Ribeiro
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
enhr@cin.ufpe.br

Filipe Maciel Leicht
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
fml2@cin.ufpe.br

Matheus Augusto Monte Silva
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mams4@cin.ufpe.br

Thiago José Grangeiro Costa
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
tjgc@cin.ufpe.br

Victória Xavier Queiroz
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
vxq@cin.ufpe.br

Resumo — O objetivo desse artigo é utilizar modelos de aprendizado de máquina para classificar estudantes universitários com maior probabilidade de evasão ou sucesso acadêmico com base no teorema Naïve Bayes. Para isso será usada a Base de dados "Predict Students' dropout and academic success", disponível no Repositório Online para Aprendizado de máquina UCI

Keywords—Desempenho Acadêmico, Evasão Acadêmica, Aprendizado de máquina, Naïve Bayes

I. Introdução

A evasão escolar é um fenômeno que afeta diretamente o desenvolvimento econômico e social de um país. A alta demanda por vagas no ensino superior tem gerado expansão e aprimoramentos estruturais nas instituições gerando cada vez mais oportunidades de ingresso em cursos superiores para estudantes de todas as idades e origens. Entretanto, a alta taxa de evasão escolar compromete o investimento público e privado no setor da educação como também o desenvolvimento intelectual e profissional dos indivíduos, prejudicando assim não apenas o futuro do indivíduo, mas também o do país.

II. Objetivos

Esse projeto tem como objetivo principal a utilização de técnicas de *Machine Learning* para a classificação de alunos com maior risco de evasão escolar baseado em informações prévias como método de admissão, curso, notas anteriores dentre outras informações pessoais.

III. Justificativa

Com a identificação dos alunos com maior risco de evasão é possível utilizar medidas preventivas como a criação de aulas especiais, aconselhamento acadêmico, grupos de apoio, auxílio psicológico como também financeiro e entre outros.

Dessa maneira, se utilizadas corretamente, essas medidas podem diminuir a taxa de evasão ao integrar mais uma vez os estudantes em risco com a comunidade acadêmica, proporcionando assim uma vida acadêmica mais saudável e significativa tanto para os indivíduos quanto para o ambiente educacional como um todo.

IV. Metodologia

A evasão escolar é um problema significativo enfrentado por instituições de ensino em todo o mundo. A identificação precoce de alunos em risco de evasão é fundamental para que sejam tomadas medidas preventivas e de intervenção.

Nesse contexto, o Teorema de Bayes, uma técnica estatística de inferência probabilística, pode ser aplicado para auxiliar na classificação de alunos com alto risco de evasão escolar.

Este artigo apresenta um estudo detalhado sobre o uso do Teorema de Bayes como ferramenta de classificação e sua aplicação em um caso de estudo específico.

A. Dataset

Para realizar o estudo, foi obtido um abrangente conjunto de dados [1], criado por uma instituição de ensino superior, que inclui informações sobre:

1. A matrícula do estudante:
    a. Trajetória acadêmica.
    b. Dados demográficos.
    c. Fatores socioeconômicos.
2. O desempenho dos alunos ao final do primeiro e segundo semestres de seus respectivos cursos.

O conjunto de dados utilizado é composto por 4424 instâncias, descrito a partir de 36 atributos diferentes e será utilizado para construir modelos de classificação que preveem a evasão dos alunos e o sucesso acadêmico.

O problema é formulado como uma tarefa de classificação em três categorias (evasão, matriculado e graduado) ao final da duração normal do curso, na qual percebe-se um desequilíbrio acentuado em relação a uma das classes [2].

B. *Processamento do dataset*

Os conjuntos de dados coletados será submetido a uma série de etapas de pré-processamento, incluindo:

1. *Filtro de Instancias:* Nesta etapa, teremos que remover ou preencher os valores faltantes, dado que existe uma abundância de estudantes apenas matriculados, os quais nem concluíram e nem abandonaram seus cursos. Essas instâncias não são interessantes para o estudo em questão.
2. *Seleção de Atributos:* Nem todos os atributos contidos no conjunto de dados serão utilizados como *inputs* para os modelos. Essa etapa consiste em selecionar um conjunto de atributos preciso para o modelo de aprendizado. Esse conjunto de atributos precisa ter uma dimensão razoável. Dessa forma, não iremos incluir atributos muito específicos, pois isso prejudicaria o desempenho dos modelos.
3. *Engenharia de atributos:* Esta etapa consiste em transformação dos atributos para permitir que sejam utilizadas pelos algoritmos de aprendizado e, até mesmo, melhorar o desempenho deles. Precisaremos, por exemplo, transformar dados nominais em numéricos através de um *encoding*.

C. *Construção, treinamento e avaliação do modelo*

O Teorema de Bayes é uma ferramenta matemática fundamental na teoria das probabilidades, que nos permite atualizar as probabilidades de um evento com base em novas informações ou evidências. Ele é amplamente utilizado em diversas áreas, incluindo a ciência de dados e o aprendizado de máquina [4].

Suponha que temos dois eventos, A e B. A probabilidade condicional de A ocorrer, dado que B ocorreu, é denotada por P(A|B) e é calculada usando a fórmula do Teorema de Bayes:

$$P(A|B) = \frac{(P(B|A) \times P(A))}{P(B)}$$

onde:

- P(A|B): Probabilidade condicional de A ocorrer dado que B ocorreu.
- P(B|A): Probabilidade condicional de B ocorrer dado que A ocorreu.
- P(A): Probabilidade de A ocorrer (antes de considerarmos a evidência de B).
- P(B): Probabilidade de B ocorrer (antes de considerarmos a evidência de A).

Nesta seção, iremos apresentar o classificador Ingênuo de Bayes [3], que será utilizado no desenvolvimento do projeto e o teorema de Bayes, que é a base para o classificador.

1. *Classificador Ingênuo de Bayes:* é um algoritmo de aprendizado de máquina amplamente utilizado para problemas de classificação. Ele é baseado no Teorema de Bayes, já apresentado anteriormente.

O algoritmo se baseia na suposição "ingênua" de que os atributos (ou *features*) são independentes entre si, ou seja, a presença ou ausência de um atributo não afeta a presença ou ausência de outros atributos. Essa simplificação permite que o algoritmo seja computacionalmente eficiente e adequado para conjuntos de dados com alta dimensionalidade.

Durante a fase de treinamento, o classificador estima as probabilidades a priori de cada classe no conjunto de treinamento. Em seguida, calcula a probabilidade condicional de cada atributo dado cada classe. Esse processo permite criar um modelo probabilístico para cada classe com base nos atributos observados nos dados de treinamento.

Após o treinamento, o classificador pode ser usado para classificar novas instâncias. Dado um conjunto de atributos de uma instância desconhecida, o algoritmo calcula a probabilidade de pertencer a cada classe com base nos atributos fornecidos. A instância é então atribuída à classe com a maior probabilidade.

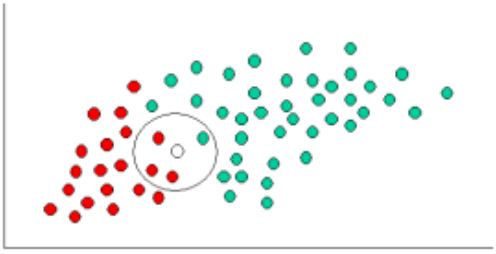

Figura 1: demonstração gráfica do Naïve Bayes[5]

V. CRONOGRAMA DE ATIVIDADES

Tabela 1

Data*	Atividade Planejada	Detalhes
25/07/2023	Seleção do *dataset*	Em grupo (Discord)
27/07/2023	Escrita da proposta	Em grupo (Discord)
01/08/2023	**Entrega da proposta**	-
03/08/2023	Desenvolvimento do projeto	Colab/Python
08/08/2023	Desenvolvimento do projeto	Colab/Pyton
15/08/2023	Finalização do projeto	Em grupo
17/08/2023	Planejamento para apresentação	Em grupo
22/08/2023	Escrita do relatório	Em grupo (Discord)
24/08/2023	Elaboração dos slides	Em grupo (Discord)
29/08/2023	Gravação da apresentação	Em grupo (Google Meet)
31/08/2023	**Entrega do projeto final**	-

* As datas podem mudar em caso de imprevistos.

A Tabela 1 mostra o planejamento detalhado para a execução do projeto, desde a concepção da proposta até a entrega final. O cronograma inclui reuniões em grupo para definir as tarefas e discutir a direção do projeto, bem como atividades individuais, nas quais cada membro trabalhará em datas de sua conveniência, porém seguindo as sugestões de datas estipuladas no cronograma.

A elaboração da proposta ocorreu em dois dias. No primeiro dia, foi realizado o processo de seleção do dataset, discussão dos objetivos e metodologia do projeto, além da divisão de responsabilidades. Posteriormente, nos primeiros e segundos dias seguintes, as seções da proposta foram elaboradas individualmente ou em equipe.

Após a entrega da proposta, cada membro receberá tarefas específicas para o desenvolvimento do projeto, seguindo a metodologia definida. O projeto utilizará o Google Colaboratory, e a previsão para sua conclusão é em 15 de agosto. Nessa etapa, haverá uma nova divisão de tarefas para a elaboração do relatório e dos slides de apresentação.

Estima-se que aproximadamente uma semana antes da entrega do projeto final, a gravação da apresentação final será realizada. Por fim, todos os artefatos, incluindo o relatório, os slides de apresentação e demais documentos pertinentes, serão entregues no dia 31 de agosto.

VI. Referências

[1] Predict students' dropout and academic success | UCI Machine Learning Repository [https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success]

[2] Data | Free Full-Text | Predicting Student Dropout and Academic Success [https://www.mdpi.com/2306-5729/7/11/146]

[3] What is Naïve Bayes | IBM [https://www.ibm.com/topics/naive-bayes]

[4] Bayes Theorem - an overview | ScienceDirect Topics [https://www.sciencedirect.com/topics/engineering/bayes-theorem]

[5] Understanding Naive Bayes [https://stats.stackexchange.com/questions/21822/understanding-naive-bayes]