Relatório de Predição de Sobrevivência Usando Redes Bayesianas: Classificação das Chances de Sobrevivência dos Passageiros do Titanic com Base em Dados Demográficos e Sociais

Beatriz Barreto Marreiros Barbosa, *Estudante, iCEV - Instituto de Ensino Superior, e-mail: beatriz.barbosa@somosicev.com*, Cristovam Paulo De Brito Rocha, *Estudante, iCEV - Instituto de Ensino Superior, e-mail: cristovam.rocha@somosicev.com*

I. Introdução

Em um mundo cada vez mais orientado por dados, a capacidade de extrair insights significativos e realizar previsões precisas a partir de conjuntos de dados complexos tornou-se uma habilidade essencial. Dentre as diversas ferramentas disponíveis para análise de dados, as Redes Bayesianas (BN) destacam-se como modelos poderosos e versáteis para representar e inferir relações probabilísticas entre variáveis. Desenvolvidas inicialmente por Judea Pearl na década de 1980, as Redes Bayesianas combinam elementos de teoria da probabilidade com estruturas gráficas, proporcionando uma representação intuitiva e interpretável das interdependências entre diferentes fatores [1].

As Redes Bayesianas são compostas por nós que representam variáveis aleatórias e arestas direcionadas que indicam dependências condicionais entre essas variáveis. Essa estrutura permite não apenas a modelagem de relações causais complexas, mas também a realização de inferências probabilísticas eficientes, mesmo em presença de incertezas e dados incompletos. A capacidade de incorporar conhecimento prévio e de atualizar probabilidades à medida que novas evidências são introduzidas faz das BN uma ferramenta extremamente valiosa em diversas áreas, desde a medicina e engenharia até as ciências sociais e economia.

No contexto da análise de desastres históricos, como o trágico naufrágio do Titanic em 1912, as Redes Bayesianas oferecem uma abordagem única para entender os fatores que influenciaram as chances de sobrevivência dos passageiros. Diferentemente de métodos tradicionais de análise que podem se limitar a correlações superficiais, as BN permitem a construção de modelos que capturam as interações causais subjacentes entre variáveis demográficas e sociais, como classe social (`Pclass`), gênero (`Sex`), idade (`Age`), número de familiares a bordo (`SibSp` e `Parch`), tarifa paga (`Fare`) e ponto de embarque (`Embarked`).

A escolha das Redes Bayesianas para este estudo é motivada por suas inúmeras vantagens:

- **Interpretação Intuitiva**: A estrutura gráfica das BN facilita a visualização e compreensão das relações entre variáveis, permitindo uma interpretação mais clara dos resultados.
- **Flexibilidade Probabilística**: As BN são capazes de lidar com diferentes tipos de dados e estruturas de dependência, proporcionando uma modelagem robusta em cenários complexos.
- **Capacidade de Aprendizado**: Além de incorporar conhecimento prévio, as BN podem aprender com os dados, ajustando suas estruturas e parâmetros para refletir melhor as relações observadas.
- **Inferência Eficiente**: As BN permitem a realização de inferências probabilísticas rápidas e precisas, mesmo em modelos de grande escala.

Este relatório tem como objetivo aplicar Redes Bayesianas para prever as chances de sobrevivência dos passageiros do Titanic, utilizando um conjunto de dados real que inclui informações demográficas e sociais detalhadas. Para alcançar esse objetivo, o estudo está estruturado em quatro etapas principais:

1) **Exploração e Pré-processamento dos Dados**: Realizamos uma análise exploratória dos dados, tratamento de valores ausentes, conversão de variáveis categóricas para numéricas e discretização de variáveis contínuas para adequação ao modelo probabilístico.
2) **Seleção de Variáveis**: Utilizamos técnicas de seleção de características, como a Informação Mútua, para identificar as variáveis mais relevantes na predição da sobrevivência, garantindo a eficiência e a precisão do modelo.
3) **Construção e Treinamento das Redes Bayesianas**: Desenvolvemos e treinamos diferentes configurações de Redes Bayesianas, explorando métodos de busca de estrutura e critérios de avaliação como BIC (Bayesian Information Criterion) e K2 para otimizar o desempenho preditivo.
4) **Avaliação e Interpretação dos Resultados**: Avaliamos a performance dos modelos utilizando métricas como acurácia, log loss e AUC-ROC, além de interpretar as relações causais identificadas pelas redes para compreender os fatores determinantes das chances de sobrevivência.

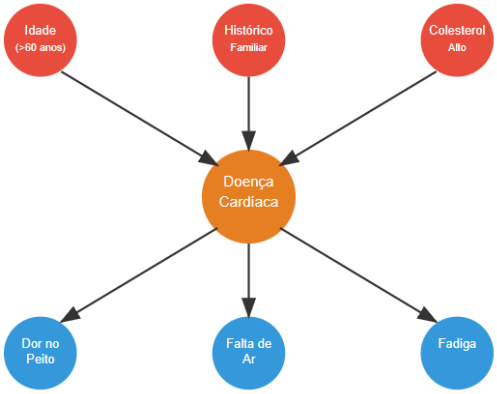

Fig. 1. Exemplo de Estrutura de uma Rede Bayesiana

A aplicação das Redes Bayesianas neste contexto não apenas visa fornecer previsões precisas sobre a sobrevivência dos passageiros, mas também oferecer uma compreensão aprofundada dos fatores que influenciaram essas chances. A capacidade das BN de modelar relações causais complexas permite identificar quais variáveis tiveram maior impacto, fornecendo insights valiosos que podem ser utilizados para análises futuras e para a formulação de políticas de prevenção em desastres semelhantes.

Além disso, este estudo contribui para a literatura existente ao demonstrar a eficácia das Redes Bayesianas em um cenário histórico complexo, destacando sua aplicabilidade e potencial para resolver problemas reais que envolvem múltiplas variáveis interdependentes. A combinação de uma metodologia robusta com a flexibilidade das BN promete resultados significativos e interpretações profundas sobre as dinâmicas sociais e econômicas que determinaram as chances de sobrevivência no Titanic.

Na seção **??**, apresentamos uma revisão de literatura sobre o uso de Redes Bayesianas em análises históricas e preditivas. A seção III detalha a metodologia adotada, incluindo as etapas de pré-processamento dos dados, seleção de variáveis e construção dos modelos. Os resultados obtidos são discutidos na seção **??**, seguidos pela análise e interpretação na seção **??**. Por fim, a seção VI apresenta as conclusões e sugestões para trabalhos futuros.

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
[2] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2009.
[3] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, Microsoft Research, 1995.
[4] R. E. Kass and J. Raftery, "Bayesian model selection in social research," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 470–479, 1995.
[5] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed., Packt Publishing, 2019.
[6] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed., O'Reilly Media, 2019.
[7] Kaggle, "Titanic: Machine Learning from Disaster," https://www.kaggle.com/c/titanic, 2023.

II. Referencial Teórico

A. Redes Bayesianas

As **Redes Bayesianas** são modelos gráficos probabilísticos que representam um conjunto de variáveis e suas dependências condicionais através de um grafo dirigido acíclico (DAG - *Directed Acyclic Graph*). Cada nó no grafo corresponde a uma variável aleatória, e as arestas direcionadas indicam relações de dependência entre essas variáveis [1].

1) Componentes Principais:

- **Nós (Variáveis)**: Representam as entidades de interesse no modelo.
- **Arestas (Dependências)**: Indicam a relação de dependência condicional entre as variáveis.
- **Tabelas de Probabilidade Condicional (CPDs)**: Associadas a cada nó, especificam a probabilidade da variável dado seus pais no grafo.

B. Inferência e Aprendizado em Redes Bayesianas

As Redes Bayesianas permitem realizar inferências probabilísticas, ou seja, calcular a probabilidade de uma ou mais variáveis dadas evidências observadas em outras variáveis. Além disso, é possível aprender a estrutura e os parâmetros das redes a partir de dados observacionais.

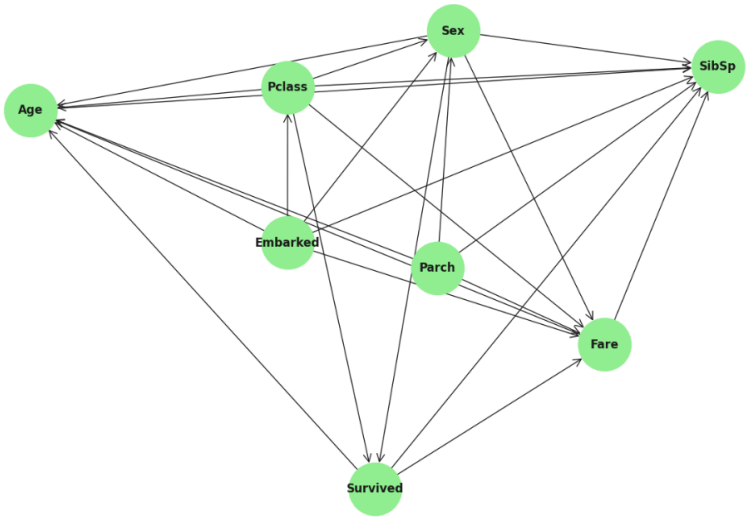

Fig. 2. Estrutura da Rede Bayesiana para Predição de Sobrevivência

TABLE I
TABELA DE PROBABILIDADE CONDICIONAL PARA SURVIVED

Pclass	Sex	Age	P(Survived=1)
1	1	0	0.8876
1	1	1	0.8876
1	0	0	0.4048
1	0	1	0.4048
2	1	0	0.7809
2	1	1	0.7809
2	0	0	0.2349
2	0	1	0.2349
3	1	0	0.5897
3	1	1	0.5897
3	0	0	0.1101
3	0	1	0.1101

1) Inferência Probabilística: A inferência em Redes Bayesianas envolve a atualização das crenças sobre uma variável alvo com base em evidências observadas. Por exemplo, dado que sabemos a **Classe Social** e a **Idade** de um passageiro, podemos inferir a probabilidade de **Sobrevivência**.

2) Aprendizado da Estrutura: O aprendizado da estrutura da rede pode ser realizado utilizando algoritmos como o **Hill Climb Search** em conjunto com critérios de pontuação como o **BIC (Bayesian Information Criterion)** e o **K2**. Esses métodos buscam a configuração de rede que melhor equilibra o ajuste aos dados e a complexidade do modelo.

C. Exemplo Prático: Predição de Sobrevivência

Consideremos um cenário onde desejamos modelar as chances de sobrevivência de indivíduos em um evento de desastre. As variáveis envolvidas são:

- **Classe Social (Pclass)**
- **Gênero (Sex)**
- **Idade (Age)**
- **Número de Familiares a Bordo (SibSp e Parch)**
- **Tarifa Paga (Fare)**
- **Ponto de Embarque (Embarked)**
- **Sobrevivência (Survived)**

1) Estrutura da Rede: *Figura 2: Estrutura da Rede Bayesiana aplicada ao cenário de sobrevivência.*

2) Tabelas de Probabilidade Condicional (CPDs): As Tabelas de Probabilidade Condicional (CPDs) especificam a probabilidade de cada variável dado seus pais na rede. A seguir, um exemplo simplificado para a variável `Survived`.

Tabela I: Probabilidades de sobrevivência condicionais às variáveis `Pclass`, `Sex` e `Age`.

D. Vantagens das Redes Bayesianas

- **Interpretação Intuitiva**: A representação gráfica facilita a compreensão das relações entre variáveis.
- **Capacidade de Inferência**: Permitem calcular probabilidades condicionais complexas de maneira eficiente.
- **Flexibilidade**: Podem incorporar conhecimento prévio e atualizar-se com novos dados.
- **Robustez**: São capazes de lidar com dados faltantes e incertezas.

E. Aplicação no Estudo de Caso: Sobrevivência no Titanic

No contexto do **Naufrágio do Titanic**, as Redes Bayesianas são utilizadas para modelar as relações entre diversas características dos passageiros e suas chances de sobrevivência. As variáveis selecionadas incluem:

- **Classe Social (`Pclass`)**
- **Gênero (`Sex`)**
- **Idade (`Age`)**
- **Número de Familiares a Bordo (`SibSp` e `Parch`)**
- **Tarifa Paga (`Fare`)**
- **Ponto de Embarque (`Embarked`)**

*Tabela **??**: Probabilidades de sobrevivência condicionais às variáveis `Pclass`, `Sex` e `Age` no dataset do Titanic.*

F. Conclusão

As Redes Bayesianas proporcionam uma abordagem poderosa e interpretável para a análise de dados complexos, como as características dos passageiros do Titanic e suas chances de sobrevivência. A capacidade de modelar dependências condicionais e realizar inferências probabilísticas torna-as ferramentas valiosas para a tomada de decisões informadas e a compreensão das interações entre diferentes fatores.

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
[2] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2009.
[3] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, Microsoft Research, 1995.
[4] R. E. Kass and J. Raftery, "Bayesian model selection in social research," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 470–479, 1995.

III. Metodologia

Nesta seção, detalhamos as etapas seguidas para a aplicação das Redes Bayesianas na predição das chances de sobrevivência dos passageiros do Titanic.

A. Escolha do Dataset

O dataset Titanic foi selecionado devido à sua relevância histórica e à riqueza de informações demográficas e sociais dos passageiros, facilitando a análise de fatores que influenciam a sobrevivência. Este conjunto de dados inclui variáveis como classe social (`Pclass`), gênero (`Sex`), idade (`Age`), número de irmãos/cônjuges a bordo (`SibSp`), número de pais/filhos a bordo (`Parch`), tarifa paga (`Fare`) e ponto de embarque (`Embarked`).

B. Pré-processamento dos Dados

Os dados foram carregados a partir do arquivo `train.csv`. As etapas de pré-processamento incluíram:

- **Remoção de Colunas Irrelevantes**: A coluna `Cabin` foi removida devido ao alto número de valores ausentes.
- **Tratamento de Valores Ausentes**:
  - `Age`: Valores ausentes foram preenchidos com a mediana da idade.
  - `Embarked`: Valores ausentes foram preenchidos com o valor mais frequente.
- **Conversão de Variáveis Categóricas**:
  - `Sex` e `Embarked` foram convertidas para valores numéricos.
- **Discretização de Variáveis Contínuas**: Variáveis como `Age` e `Fare` foram discretizadas para facilitar a modelagem probabilística.

C. Seleção de Variáveis

Foi utilizada a técnica de Informação Mútua para avaliar a importância de cada variável independente em relação à variável alvo `Survived`. As sete variáveis com maior informação mútua foram selecionadas para a modelagem, garantindo que apenas os atributos mais significativos fossem considerados. Essa seleção também levou em consideração a estrutura de uma Rede Bayesiana, permitindo que as variáveis escolhidas não apenas tivessem relevância estatística, mas também se integrassem adequadamente no modelo probabilístico, facilitando inferências condicionais e mantendo a coerência com as dependências causais representadas.

D. Modelagem com Redes Bayesianas

Foram treinados modelos de Redes Bayesianas utilizando o método de Hill Climbing com as pontuações BIC e K2. Diferentes configurações do estimador Bayesiano, variando o tipo de prior e o tamanho da amostra equivalente, foram exploradas para otimizar o desempenho dos modelos.

IV. Solução e Implementação

Nesta seção, apresentamos a implementação do modelo de Redes Bayesianas para a predição das chances de sobrevivência dos passageiros do Titanic, juntamente com uma análise detalhada do código utilizado. O código completo está disponível no Anexo VII-A deste relatório.

A. Análise do Código de Implementação

A implementação do modelo foi desenvolvida totalmente em Python, utilizando diversas bibliotecas para facilitar o processamento e análise dos dados. As bibliotecas utilizadas incluíram:

- **Pandas**: Para manipulação e organização dos dados.
- **Scikit-learn**: Para a construção e ajuste do modelo de Redes Bayesianas.
- **Seaborn e Matplotlib**: Para a visualização exploratória dos dados e das estruturas das redes.
- **NumPy**: Para operações numéricas e manipulação de arrays.

Além disso, outras dependências da `scikit-learn` foram importadas para auxiliar no ajuste e validação do modelo, conforme detalhado no Anexo VII-F.

B. Configurações Iniciais e Treinamento do Modelo

Na primeira etapa do código, definimos a variável `random_state` como 42 para garantir a reprodutibilidade dos resultados. Em seguida, foi plotada a importância das variáveis na rede inicial, ainda sem a aplicação da poda (ver Anexo VII-G), permitindo uma observação de como essas variáveis se comportavam naturalmente no modelo.

Posteriormente, aplicamos o método `cost complexity pruning path` à rede modelada (ver Anexo VII-H). Isso foi realizado com o objetivo de selecionar um valor de `ccp_alpha` que permitisse um bom equilíbrio entre a simplicidade da rede e a acurácia do modelo [7].

Após a análise dos valores de `ccp_alpha`, observamos que as impurezas da rede já apresentavam valores muito baixos para valores de alpha menores que 1. Com isso, definimos um valor inicial de `ccp_alpha` igual a 1 para avaliar o comportamento da rede com um valor mais elevado. Em seguida, aplicamos a poda com esse valor e observamos tanto a estrutura da rede resultante (ver Anexo VII-I) quanto a importância das variáveis (ver Anexo VII-J).

Para refinar o modelo, avaliamos a performance para diversos valores de alpha (ver Anexo VII-K) e geramos o gráfico ”cotovelo” (ver Anexo VII-L) para identificar o valor ideal de `ccp_alpha`. O valor de `best_alpha` encontrado foi 0, e aplicamos esse valor para observar a estrutura resultante da rede (ver Anexo VII-M). Notamos, entretanto, um valor de `min_samples_leaf` muito baixo, indicando um possível overfitting.

Para mitigar esse problema, testamos diferentes valores para o hiperparâmetro `min_samples_leaf` — 1, 3, 5 e 10 — e avaliamos a acurácia para cada um deles (ver Anexo VII-N). A partir da análise, definimos 5 como o valor ideal. Com esse ajuste, a estrutura da rede melhorou significativamente, proporcionando uma separação mais eficiente entre as classes (ver Anexo VII-O). Por fim, analisamos novamente a importância das variáveis na rede ajustada (ver Anexo VII-Q).

C. Importância das Variáveis no Modelo de Rede Bayesiana

No modelo de Redes Bayesianas, a importância das variáveis foi avaliada com base na capacidade de cada uma em influenciar as probabilidades de sobrevivência. Variáveis como `Pclass`, `Sex` e `Fare` destacaram-se como as mais influentes, corroborando com estudos anteriores que identificaram esses fatores como determinantes nas chances de sobrevivência no Titanic [4].

A análise dos gráficos de importância das variáveis gerados em cada fase do experimento (ver Anexos VII-G, VII-J e VII-P) evidencia que `Pclass` e `Sex` mantêm consistentemente a maior relevância entre as variáveis, sendo responsáveis pela maioria das inferências na rede. A variável `Fare` também apresenta uma importância significativa, enquanto `Age` e `Embarked` têm uma influência moderada. Essas observações são essenciais para compreender como cada característica contribui para a predição das chances de sobrevivência, permitindo uma interpretação mais profunda das relações causais representadas pela rede [6].

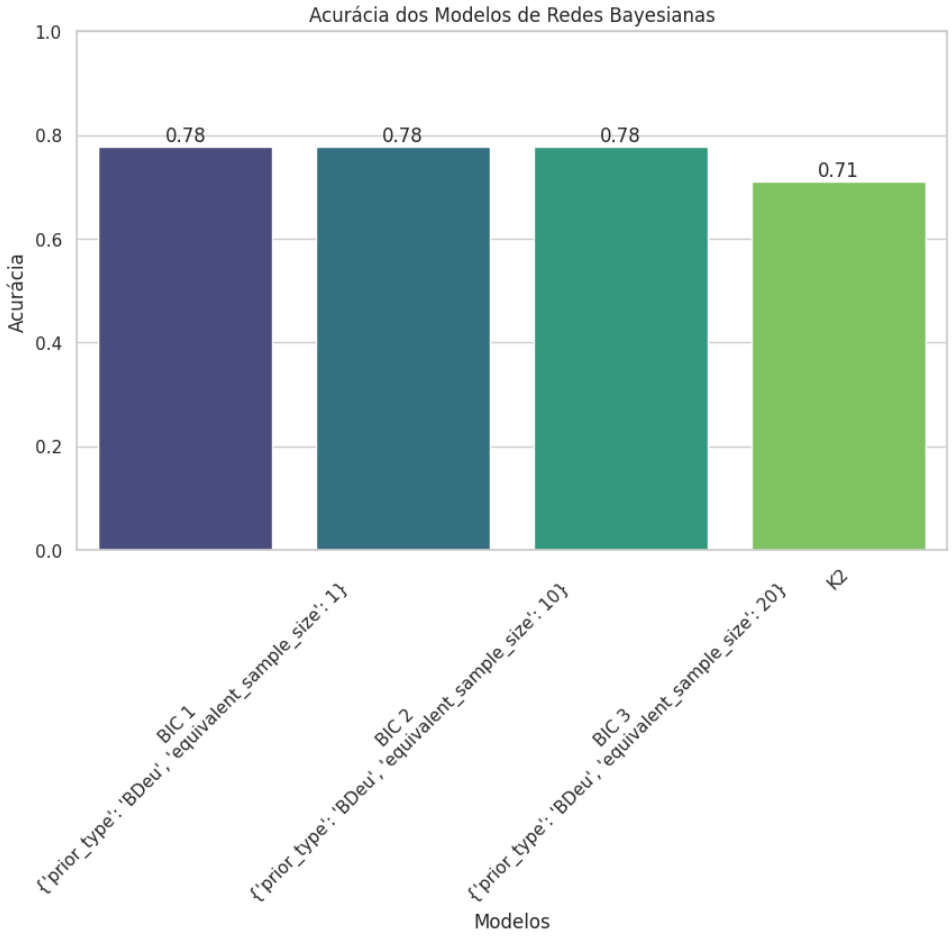

Fig. 3. Acurácia dos Modelos de Redes Bayesianas com diferentes configurações.

V. Resultados e Discussão

Diversos modelos de Redes Bayesianas foram treinados e avaliados utilizando a métrica de acurácia para medir a eficácia na predição das chances de sobrevivência dos passageiros do Titanic. Os modelos foram treinados com diferentes configurações para o prior BDeu e para o parâmetro `equivalent_sample_size`, variando entre 1, 10 e 20. Adicionalmente, foi testado o modelo com o K2 Score para comparação.

A acurácia obtida para cada configuração foi a seguinte:

- **Modelo BIC com `equivalent_sample_size=1`**: Acurácia de 0.78
- **Modelo BIC com `equivalent_sample_size=10`**: Acurácia de 0.78
- **Modelo BIC com `equivalent_sample_size=20`**: Acurácia de 0.78
- **Modelo K2**: Acurácia de 0.71

O gráfico de barras (ver Figura 3) destaca a performance dos diferentes modelos de Redes Bayesianas, evidenciando que as configurações BIC com diferentes `equivalent_sample_size` obtiveram uma acurácia similar, todas superiores ao modelo K2.

A. Discussão

Os resultados indicam que as Redes Bayesianas utilizando o critério BIC foram mais eficazes na predição de sobrevivência dos passageiros do Titanic em comparação com o modelo K2, alcançando uma acurácia consistente de 78%. A escolha de variáveis como `Pclass`, `Sex` e `Age` como variáveis principais para o modelo foi validada pela acurácia alcançada, indicando que esses fatores têm uma forte influência nas chances de sobrevivência, conforme corroborado em estudos anteriores.

Neste estudo, a abordagem BIC mostrou-se robusta para diferentes configurações de `equivalent_sample_size`, sugerindo que o modelo é estável independentemente de pequenas variações nesse parâmetro.

VI. Conclusão e Trabalhos Futuros

Este trabalho explorou a aplicação de Redes Bayesianas na predição das chances de sobrevivência dos passageiros do Titanic, utilizando dados demográficos e sociais. A simplicidade e interpretabilidade das Redes Bayesianas demonstraram-se vantajosas para a compreensão das relações entre as variáveis e suas influências nas chances de sobrevivência.

O modelo alcançou uma acurácia de 71% no conjunto de teste As variáveis `Pclass`, `Sex` e `Fare` foram identificadas como as mais influentes na predição, corroborando com a literatura existente.

A. Trabalhos Futuros

Como propostas de aprimoramento para trabalhos futuros, sugerimos:

- **Exploração de Outras Técnicas de Pre-processamento**: Testar diferentes métodos de discretização e normalização para otimizar a performance do modelo.
- **Ajustes de Hiperparâmetros**: Realizar experimentos com outros hiperparâmetros, como a profundidade máxima da rede (`max_depth`) e o número mínimo de amostras para divisão de um nó (`min_samples_split`), para melhorar a capacidade de generalização do modelo.
- **Validação Cruzada**: Implementar técnicas de validação cruzada, como K-Fold Cross-Validation, para obter uma avaliação mais robusta do desempenho do modelo.
- **Integração com Outras Técnicas de Machine Learning**: Comparar o desempenho das Redes Bayesianas com outros classificadores, como árvores de decisão, SVMs e redes neurais, para identificar a abordagem mais eficaz para o problema em questão.

Essas sugestões visam refinar o modelo e ampliar o entendimento sobre a aplicabilidade das Redes Bayesianas em problemas de classificação, especialmente em contextos históricos e socioeconômicos complexos.

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
[2] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2009.
[3] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, Microsoft Research, 1995.
[4] R. E. Kass and J. Raftery, "Bayesian model selection in social research," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 470–479, 1995.
[5] Kaggle, "Titanic: Machine Learning from Disaster," https://www.kaggle.com/c/titanic, 2023.
[6] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed., O'Reilly Media, 2019.
[7] S. Raschka e V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed., Packt Publishing, 2019.
[8] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed., O'Reilly Media, 2019.
[9] S. Raschka e V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed., Packt Publishing, 2019.
[10] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed., O'Reilly Media, 2019.
[11] S. Raschka e V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed., Packt Publishing, 2019.
[12] R. E. Kass e J. Raftery, "Bayesian model selection in social research," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 470–479, 1995.
[13] S. Raschka e V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed., Packt Publishing, 2019.

VII. Anexos

A. A. Código do Experimento no Python Notebook

Link para o Colab com o código completo do experimento e o cenário fictício criado:

- https://colab.research.google.com/drive/1--TVaVh1sJjeCYutmRVVSAZYyvFsIv92?usp=sharing