Aplicação de Modelos Supervisionados para Previsão de Sobrevivência no Titanic com scikit-learn

Maria Luiza de Assis Coelho, RGM: 11231103347
Universidade de Mogi das Cruzes (UMC)
Engenharia de Software – Disciplina: Inteligência Artificial

Introdução

A Inteligência Artificial tem se consolidado como ferramenta essencial em aplicações críticas, como diagnóstico médico e análise preditiva. Este trabalho propõe a aplicação de modelos de classificação supervisionada para prever a sobrevivência de passageiros do Titanic, utilizando o conjunto de dados público disponível via seaborn. A tarefa consiste em identificar padrões que indiquem maior probabilidade de sobrevivência com base em atributos demográficos e socioeconômicos.

Fundamentação Teórica

Foram utilizados três algoritmos de classificação supervisionada:

- **XGBoost**: modelo baseado em árvores de decisão com boosting, reconhecido por sua alta performance.
- **SVM com kernel RBF**: eficaz em problemas com fronteiras não lineares.
- **Random Forest**: conjunto de árvores de decisão que melhora a generalização e reduz overfitting.

Cada modelo foi avaliado quanto à capacidade de discriminação entre classes, interpretabilidade e custo computacional.

Metodologia Computacional

Pré-processamento

- Imputação de valores ausentes com mediana (Age) e moda (Embarked).
- Escalonamento de variáveis numéricas com StandardScaler.
- Codificação de variáveis categóricas com OneHotEncoder.

Hiperparâmetros Utilizados

- RandomizedSearchCV foi aplicado com StratifiedKFold para otimização dos hiperparâmetros.

- O Random Forest tunado apresentou os melhores resultados com:
  - n_estimators=300
  - min_samples_split=10
  - max_depth=None

Engenharia de Atributos

Não foi aplicada redução de dimensionalidade, pois o conjunto de atributos é pequeno e interpretável.

Validação Cruzada

A validação foi realizada com 5 folds estratificados, garantindo equilíbrio entre classes. A métrica principal foi **ROC-AUC**, complementada por **acurácia**, **F1-score** e **matriz de confusão**.

Resultados

Modelo	Acurácia	F1 - score	ROC - AUC	Hiperparâmetros Tunados
XGBoost (padrão)	0.827	0.770	0.838	-
SVM (padrão)	0.810	0.707	0.824	-
Random Forest (padrão)	0.810	0.707	0.824	-
Randon Forest (Tunado)	0.810	0.730	0.838	n_estimators=300, min_samples_split=10, max_depth=None

Discussão

O modelo Random Forest tunado apresentou o melhor equilíbrio entre desempenho e interpretabilidade. O ajuste de hiperparâmetros foi decisivo para melhorar o F1-score. O XGBoost também teve excelente desempenho, mas com

maior custo computacional. O SVM mostrou-se menos eficaz, com maior número de falsos negativos.

Conclusões

A aplicação de técnicas de classificação supervisionada demonstrou ser eficaz na previsão de sobrevivência no Titanic. O uso de validação cruzada e ajuste de hiperparâmetros contribuiu para a robustez dos modelos. O projeto reforça a importância da IA em contextos críticos e reprodutíveis.

Referências

- PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.
- CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD, 2016.
- SEABORN Documentation. Disponível em: https://seaborn.pydata.org/
- KAGGLE Titanic Dataset. Disponível em: https://www.kaggle.com/c/titanic

Link para GitHub: https://github.com/maluassiscoelho-sketch/titanic-classification