1 Introdução

Este relatório documenta a aplicação prática de algoritmos de Aprendizado de Máquina em dois domínios: dados estruturados e não estruturados. O objetivo foi consolidar técnicas de pré-processamento e modelagem em cenários distintos.

O primeiro estudo de caso utiliza o Heart Disease Dataset para enfrentar desafios na área da saúde. Foram desenvolvidos modelos para classificação de doenças cardíacas e regressão de níveis de colesterol, com ênfase no tratamento de dados médicos incompletos e na precisão diagnóstica.

O segundo estudo foca no Processamento de Linguagem Natural (NLP) através do dataset AG News. O trabalho compara o impacto das técnicas de vetorização Bag of Words e TF-IDF na classificação de notícias em quatro tópicos, avaliando como a representação textual influencia a capacidade de generalização dos modelos.

1.1 Descrição dos Dados

A compreensão das variáveis e da estrutura dos dados foi fundamental para definir as estratégias de pré-processamento.

1.1.1 Dados Tabulares: Heart Disease

Para as tarefas de classificação e regressão, utilizou-se um conjunto de dados clínicos composto por 14 atributos. As variáveis preditoras incluem dados demográficos, como idade e sexo, e resultados de exames fisiológicos. Entre os principais atributos, destacam-se a pressão arterial em repouso, a frequência cardíaca máxima atingida em testes de esforço e a tipologia da dor no peito relatada pelo paciente.

A variável alvo para a classificação é binária, indicando a ausência ou presença de patologia cardíaca. Já para a regressão, a variável alvo é contínua, correspondendo ao nível de colesterol sérico. A análise exploratória revelou a presença de valores ausentes em colunas críticas, o que demandou uma estratégia de imputação cuidadosa.

1.1.2 Dados Textuais: AG News

O conjunto de dados textual é composto por títulos e descrições curtas de artigos jornalísticos. As classes são balanceadas, distribuídas igualmente entre os quatro tópicos de interesse. Uma característica importante deste dataset é a sobreposição de vocabulário entre certas categorias, como Negócios e Tecnologia, o que impõe um desafio adicional aos modelos lineares.

Para maximizar o contexto semântico disponível para os algoritmos, optou-se pela concatenação dos campos de título e descrição numa única variável textual. Esta abordagem permite que o modelo capture tanto as palavras-chave de alto impacto do título quanto o detalhamento fornecido no corpo da notícia.

1.2 Metodologia

A metodologia adotada priorizou a reprodutibilidade dos experimentos e a robustez das métricas de avaliação.

1.2.1 Pré-processamento de Dados

Nos dados tabulares, o tratamento consistiu na imputação de valores ausentes (mediana para numéricos e moda para categóricos), na aplicação de *One-Hot Encoding* para variáveis nominais e na padronização (*StandardScaler*). Este último passo, que garante média zero e desvio padrão unitário, foi fundamental para o desempenho dos modelos baseados em distância, como KNN e SVM.

No contexto de NLP, o *pipeline* incluiu a normalização textual (minúsculas, remoção de pontuação) e a filtragem de *stopwords*. A vetorização contrastou o método *Bag of Words* (frequência absoluta) com o *TF-IDF*, cuja ponderação penaliza termos genéricos para realçar palavras discriminantes dos tópicos.

1.2.2 Validação e Métricas

Para mitigar o risco de sobreajuste (overfitting), adotou-se a Validação Cruzada K-Fold Estratificada com 5 divisões. Esta técnica assegura que a proporção de classes em cada subconjunto de teste reflete a distribuição original dos dados.

As métricas de avaliação foram selecionadas conforme a natureza do problema. Para a classificação, priorizou-se o F1-Score devido à sua capacidade de balancear precisão e sensibilidade, além da análise da Matriz de Confusão. Para a regressão, utilizou-se o Erro Quadrático Médio (RMSE) e o coeficiente de determinação R2.

1.2.3 Parte I: Heart Disease

Na tarefa de classificação, o modelo K-Nearest Neighbors (KNN) apresentou o melhor desempenho geral. Após a otimização de hiperparâmetros, o modelo alcançou métricas excepcionais no conjunto de teste, com F1-Score de 1.00. Este resultado sugere que, após o escalonamento adequado das variáveis, as classes tornaram-se linearmente separáveis ou agrupadas de forma muito distinta no espaço de características. Embora resultados perfeitos exijam cautela quanto à generalização em novos dados, eles validam a eficácia do pré-processamento.

Para a regressão do nível de colesterol, o modelo Random Forest Regressor superou as abordagens lineares. O modelo obteve um R2 de 0.922 e um RMSE de 16.36. A capacidade das árvores de decisão em capturar relações não lineares provou-se essencial para modelar a complexidade dos indicadores biológicos.

A análise de explicabilidade revelou que o tipo de dor no peito (cp) e a talassemia (thal) foram as variáveis mais determinantes para as previsões, alinhando-se com o conhecimento médico estabelecido.

1.2.4 Parte II: Classificação de Texto (AG News)

A comparação entre as técnicas de vetorização demonstrou uma clara vantagem do TF-IDF sobre o Bag of Words. O uso do TF-IDF resultou num F1-Score médio de 0.876 na validação cruzada, superior aos resultados obtidos com contagem simples. Isso confirma que a penalização de termos genéricos ajuda o modelo a focar em palavras-chave específicas de cada editoria.

Entre os algoritmos testados, o Multinomial Naive Bayes combinado com TF-IDF foi o modelo mais eficaz, atingindo uma acurácia de 89.3% no conjunto de teste. Este modelo superou abordagens mais complexas, como Random Forest, demonstrando que métodos probabilísticos simples continuam sendo eficientes para classificação de texto quando os recursos computacionais são limitados.