UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS

RELATÓRIO

DISCIPLINA: CIÊNCIA DOS DADOS
DOCENTE: RENAN GOMES VIEIRA

CARLOS EDUARDO RODRIGUES PITA – 509630

FRANCISCO RIAN RIBEIRO MEDEIRO – 552915

MARLON MELO MOURA – 55324

Crateús/CE
11/01/2026

RELATÓRIO INSUFICIÊNCIA CARDÍACA

CONTEXTUALIZAÇÃO

O objetivo central deste relatório é aplicar um fluxo completo de Machine Learning para prever a mortalidade em pacientes com insuficiência cardíaca. O foco metodológico consiste na **avaliação comparativa de diferentes algoritmos de classificação** para identificar o modelo mais promissor para este problema.

DESCRIÇÃO DOS DADOS

O conjunto de dados contém registros médicos de 299 pacientes e foi obtido através do Kaggle, sendo eles descritos por: Age (Idade), Anaemia (Diminuição dos glóbulos vermelhos ou da hemoglobina), Creatinine_phoshokinase (Nível da enzima CPK (Creatina Fosfoquinase) no sangue), Diabetes (Se o paciente tem diabetes), Ejection_fraction (Porcentagem de sangue que sai do coração a cada contração) (porcentagem), High_blood_pressur (Se o paciente tiver hipertensão), Platelets (Indicativo de plaquetas), Serum_Creatinine (Nível de creatinina sérica no sangue), Serum_Sodium (Nível de sódio sérico no sangue), Sex (Sexo do indivíduo), Smoking (Indicativo de fumante/não fumante), Time (Tempo de observação) e Death_Event (Se o paciente faleceu ou sobreviveu durante o período de acompanhamento).

JUSTIFICATIVAS

Para garantir a robustez do projeto, isolamos as variáveis preditoras e o alvo (DEATH_EVENT). Além disso, aplicamos o *StandardScaler* nas colunas numéricas para realizar a padronização, garantindo que todas as variáveis estejam na mesma escala para não enviesar o modelo.

RESULTADOS

Após a avaliação comparativa das abordagens testadas, o modelo selecionado como o mais robusto para a solução do problema foi a Árvore de Decisão (Decision Tree), otimizada e integrada a um pipeline de balanceamento de classes (SMOTE).

A seleção deste modelo baseou-se em uma análise crítica das métricas de desempenho, priorizando a capacidade de detecção da classe minoritária (Classe 1: Óbito) sem sacrificar a precisão global.

CONCLUSÕES E LIMITAÇÕES

Logo a seguir, detalha-se a justificativa técnica frente aos demais algoritmos avaliados:

1. Embora modelos como Random Forest, SVC e Logistic Regression tenham alcançado uma acurácia global competitiva (82%), eles falharam gravemente na métrica mais importante para este domínio médico: o Recall (Revocação) da classe positiva.

- Esses modelos apresentaram um Recall entre 0.55 e 0.59 para a classe 1. Isso significa que eles falharam em identificar aproximadamente 40-45% dos pacientes que realmente vieram a óbito (Falsos Negativos), tornando-os clinicamente inseguros.

2. Ao compararmos a Decision Tree com o Gradient Boosting Classifier (que obteve o maior Recall de 0.79), a Árvore de Decisão mostrou-se superior no equilíbrio geral:

- O Gradient Boosting teve um Recall alto, mas sua Precisão caiu para 0.64, indicando um alto número de Falsos Positivos (alarmes falsos).

- A Decision Tree alcançou um Recall robusto de 0.76 (detectando a grande maioria dos óbitos) mantendo uma Precisão superior de 0.71.

- A Árvore de Decisão apresentou o maior F1-Score para a classe de interesse (0.73) dentre todos os modelos testados.

Este resultado comprova que este foi o algoritmo que melhor conseguiu balancear a sensibilidade necessária para não negligenciar pacientes graves (minimizar Falsos Negativos) com a especificidade necessária para evitar diagnósticos alarmistas incorretos (Falsos Positivos).

Como limitações houve a tentativa de usar os dados do datatran, mas pela quantidade exorbitante de dados, sendo ainda muito difíceis de prever, optamos por utilizar o dataset de insuficiência cardíaca, uma vez que os modelos para o datatran não tinham bons resultados (sempre abaixo de 50% f1-score e de acurácia).

RELATÓRIO ADVERTISING

CONTEXTUALIZAÇÃO

O objetivo desse relatório é analisar o impacto de diferentes canais de comunicação (TV, Rádio e Jornal) no volume de vendas de um produto. O problema é classificado como uma regressão, pois buscamos prever a variável contínua ‘Sales’ e identificar a eficiência marginal de cada dólar investido. Este tipo de análise, conhecido como *Marketing Mix Modeling*, é crucial para a tomada de decisão estratégica e realocação de orçamentos de marketing.

DESCRIÇÃO DOS DADOS

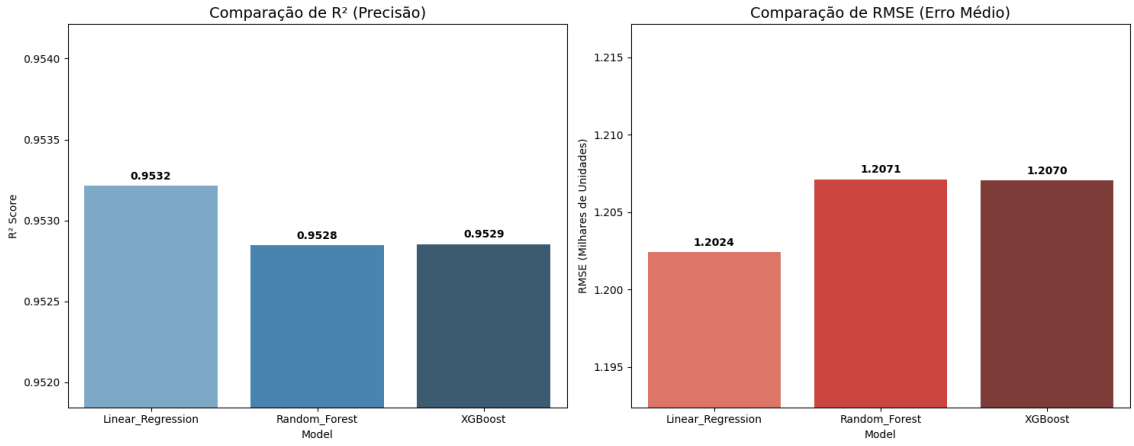
A base de dados ‘advertising.csv’ contém 200 registros de investimentos publicitários, sendo descritos por: TV / Radio / Newspaper (Variáveis independentes mensuradas em milhares de dólares) e Sales (Variável dependente mensurada em milhares de unidades vendidas).

JUSTIFICATIVAS

Para garantir a robustez do projeto, adotamos além das práticas já abordadas no documento de explicação do trabalho, as a aplicação de Polinômios de 2º Grau, permitindo ao modelo capturar interações entre mídias (ex: TV e Rádio juntos) e efeitos de curvatura (saturação). Sobre a escolha de modelos, a regressão linear atua como o nosso "baseline". É um modelo altamente interpretável que permite entender a relação direta entre cada dólar investido e o retorno em vendas. Já o Random Forest foi escolhido por ser conhecido por sua robustez contra *outliers* e por não assumir que os dados seguem uma distribuição linear. Por último o XGBoost (eXtreme Gradient Boosting) que diferente do Random Forest, ele utiliza *Boosting*, onde cada nova árvore tenta corrigir os erros das árvores anteriores, foi incluído para verificar se o ganho de precisão compensaria a perda de interpretabilidade em relação ao modelo linear.

RESULTADOS

O modelo de Regressão Linear Polinomial apresentou o melhor desempenho geral, sugerindo que a relação entre as mídias e as vendas segue uma estrutura matemática bem definida de interações.



CONCLUSÕES E LIMITAÇÕES

A Regressão Linear superou modelos de florestas, provando que, para este volume de dados, a captura explícita de interações (sinergia) é mais eficaz que a busca por padrões não-lineares arbitrários. Sobre as limitações podemos abordar que os dados não consideram datas festivas ou meses do ano, que podem influenciar drasticamente as vendas, preços dos concorrentes, promoções no ponto de venda e fatores macroeconômicos não foram incluídos no modelo e por fim uma base de 200 linhas é pequena para modelos de *Deep Learning* ou árvores muito profundas, o que justifica a performance superior da regressão linear.

RELATÓRIO REVIEWS DE E-COMMERCE (B2W)

CONTEXTUALIZAÇÃO

Este relatório descreve o desenvolvimento de um pipeline de Processamento de Linguagem Natural (NLP) aplicado a um conjunto de dados reais de avaliações de e-commerce. O objetivo principal foi explorar técnicas de manipulação de texto não estruturado para extrair insights sobre o comportamento e a satisfação dos consumidores, além de preparar os dados para potenciais modelos de classificação.

DESCRIÇÃO DOS DADOS

Foi utilizado o corpus **B2W-Reviews01**, que contém avaliações de usuários sobre produtos adquiridos online. O dataset possui atributos como o texto da revisão, a nota atribuída (estrelas), data e metadados do produto.

- Dataset: B2W-Reviews01

Alvo (Target): ‘sentimento’ (Inferido a partir da coluna ‘overall_rating’: Positivo se nota > 3, negativo caso contrário).

JUSTIFICATIVAS

Para a tarefa de classificação textual, a conversão de dados não estruturados em vetores numéricos foi realizada através de duas abordagens distintas: Bag of Words (BoW) e TF-IDF (Term Frequency-Inverse Document Frequency). A escolha dos hiperparâmetros para cada vetorizador foi pautada na necessidade de equilibrar a captura de contexto semântico com a redução de ruído e dimensionalidade.

RESULTADOS

Abaixo apresentamos o modelo vencedor e seus hiperparâmetros ideais. A métrica final é calculada sobre o conjunto de teste (dados nunca vistos pelo modelo).

Pontos de Análise:

1. BoW vs TF-IDF: Observe se a penalização de termos comuns feita pelo TF-IDF trouxe ganho de performance.

2. N-grams: Verifique nos ‘best_params_’ se o modelo preferiu usar Bigramas ‘(1, 2)’. Em português, isso geralmente ajuda a capturar negações (ex: "não gostei").

```
=== MODELO VENCEDOR: TF-IDF ===
Melhores Hiperparâmetros: {'clf__alpha': 1.0, 'tfidf__min_df': 2, 'tfidf__ngram_range': (1, 2), 'tfidf__sublinear_tf': False}

Relatório de Classificação:
              precision    recall  f1-score   support

    Negativo       0.87      0.78      0.82     14935
    Positivo       0.87      0.93      0.90     23795

    accuracy                           0.87     38730
   macro avg       0.87      0.85      0.86     38730
weighted avg       0.87      0.87      0.87     38730
```

CONCLUSÕES E LIMITAÇÕES

A aplicação das técnicas de NLP permitiu transformar dados textuais brutos em informações estruturadas. O pré-processamento mostrou-se essencial para reduzir o ruído nos dados, permitindo uma distinção clara entre o vocabulário utilizado em experiências positivas e negativas de compra.