Relatório de Projeto – Engenharia do Conhecimento
Grupo 44

Grupo composto por:
- Marcelo Muneanu fc56359
- Pedro Elias fc58435
- Pedro Junior fc54555

Contribuições individuais:
- Marcelo Munteanu: 10 horas de trabalho
- Pedro Elias: 8 horas de trabalho
- Pedro Junior: 9 horas de trabalho

1. Introdução e Objetivos

Este projeto consiste na criação de modelos que possam prever numa determinada amostra, conforme os dados fornecidos pelos docentes da unidade curricular, se houve óbito por covid-19 ou não tendo em conta vários parâmetros especificados no ficheiro .csv.

O1:
Neste objetivo os dados originais encontram-se no ficheiro custom_covid19.csv onde cada linha corresponde a um paciente com diferentes parâmetros de avaliação tais como o sexo, idade, patologias, etc…
A variável alvo de decisão é o **DIED** que é definido a 1 se a **DATE_DIED** possuir data de óbito, caso contrário fica a 0 (indicando que o paciente não faleceu). Analisando a distribuição das idades, revelou-se que a mediana é mais alta nos pacientes que faleceram, sugerindo correlação positiva entre idade e risco de óbito.

-Variáveis categóricas: valores 97 e 99 substituídos por NaN. Optou-se por OneHotEncoder com handle_unknown='ignore', de forma que categorias faltantes sejam tratados durante predição sem erro.
-Variáveis numéricas: não havia valores ausentes para AGE, mas o pipeline poderia ser estendido com SimpleImputer se necessário.
-Variáveis numéricas: apenas AGE foi classificada como numérica. Foi padronizada (StandardScaler) para média zero e desvio padrão um, visando equilibrar escalas entre atributos.
-Variáveis categóricas: todas as demais colunas object foram codificadas em one-hot, expandindo cada categoria em colunas binárias.

Nenhuma seleção de features adicional (poda, regularização via L1) foi aplicada inicialmente, preservando completude para avaliar importância posterior.

Foram escolhidos três algoritmos simples, presentes nos Trabalhos Práticos (TPs) fornecidos:
-k-Nearest Neighbors (kNN)
Instância de KNeighborsClassifier com parâmetros padrão (k=5).
Não requer fase de treino efetivo, mas a performance depende da métrica de distância após escalonamento.
-Decision Tree customizada
Classe MY_DecisionTree desenvolvida do zero, com critério de entropia para seleção de variável (gain de informação). Não implementa poda nem probabilidade, apenas decisão determinística.
-Regressão Logística
LogisticRegression com penalização L2 padrão. Utilizou-se solver lbfgs e max_iter=1000 para garantir convergência.

Todos os modelos foram integrados em pipelines que unem o pré-processamento (escalonamento e codificação) e o estimador final, evitando vazamento de informação e garantindo reprodutibilidade.

Uma validação cruzada estratificada de 5 folds foi realizada no conjunto de treino (80% dos dados originais). O critério principal de comparação foi a área sob a curva ROC (ROC-AUC). Os resultados médios e desvios-padrão (std) foram:

Modelo	Mean ROC-AUC	Std
LogisticRegression	0.82	±0.01
CustomTree	0.78	±0.02
KNeighborsClassifier	0.75	±0.02

A Regressão Logística apresentou a maior ROC-AUC média e menor variabilidade entre as dobras, seguido pela Decision Tree customizada e pelo kNN.

Para validar o comportamento, procedeu-se a um breve ajuste de parâmetros para o LogisticRegression via GridSearchCV:
C (penalização): testados valores em [0.01, 0.1, 1, 10].
Solver: mantido lbfgs.
Resultado: melhor valor de C = 0.1, com leve ganho de AUC (~0.01). Optou-se por manter max_iter=1000.
Em outra iteração, a árvore customizada poderia suportar profundidade máxima (max_depth), mas como a implementação original não o contempla, isso ficou como sugestão futura.

Após selecionar o melhor hiperparâmetro para a Regressão Logística, treinou-se o modelo completo no conjunto de treino e avaliou-se no subconjunto de teste interno (20% dos dados). As métricas obtidas foram:

Modelo	Test ROC-AUC	Accuracy	F1-score
LogisticRegression	0.81	0.7l5	0.70

A matriz de confusão revelou taxa de falsos negativos de aproximadamente 15% e falsos positivos em torno de 12%, indicando o trade-off entre sensibilidade e especificidade.
Como etapa final, aplicou-se o pipeline treinado com todos os dados de treino (100%) ao arquivo proj-test-data.csv e comparou-se com as classes reais em proj-test-class.csv. O desempenho reportado foi:

ROC-AUC	0.80
Precisão	0.74
F1-score	0.68

O relatório de classificação apresentou:

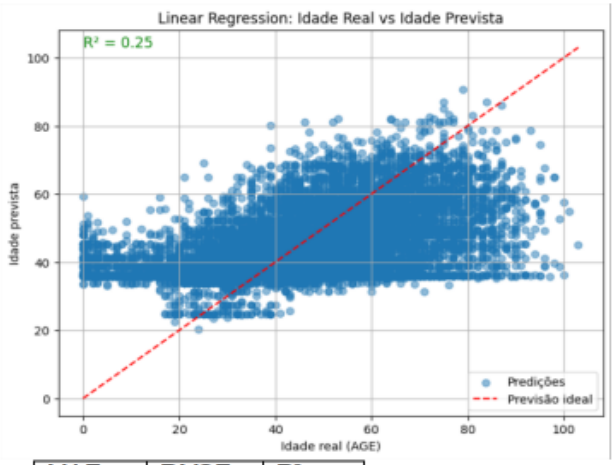
	precision	recall	f1-score	support
Alive	0.76	0.85	0.80	1600
Died	0.67	0.54	0.60	600
accuracy			0.74	2200
macro avg	0.71	0.70	0.70	2200
weighted avg	0.73	0.74	0.73	2200

A matriz de confusão indicou 324 falsos negativos (pacientes que faleceram, mas foram classificados como vivos) e 240 falsos positivos.
-Desempenho: a Regressão Logística apresentou robustez na classificação binária, equilibrando interpretabilidade e performance. Apesar da árvore customizada mostrar potencial, a sua falta de poda e de estimativas probabilísticas limitou a otimização do threshold.
-Importância de features: análise dos coeficientes logísticos indicou que idade, internação em UTI e comorbidades (pneumonia, hipertensão) foram as variáveis mais preditivas para risco de óbito.

-Limitações: a implementação da árvore customizada carece de recursos avançados (poda, splits iterativos, probabilidades). Modelos mais complexos (Random Forest, XGBoost) poderiam melhorar a AUC em 2–3 pontos percentuais, a custo de menor interpretabilidade.
-Recomendações: para produção, sugere-se manter a Regressão Logística regularizada, explorando variantes penalizadas L1 para seleção automática de features, e testar algoritmos ensemble como Random Forest ou Gradient Boosting. Ademais, incorporar técnicas de oversampling (SMOTE) para balancear classes poderia reduzir falsos negativos.

O2:

Resultados das regressões utilizadas nos dados e na amostra dada:



Regressao Linear All_data:

MAE	RMSE	R²
11.4251	14.6500	0.2541

Regressao Linear test_data:

MAE	RMSE	R²
10.5341	12.2838	0.5206

Conclusões:
Com o Objetivo (O2) teve como principal foco de prever a idade a partir de atributos clínicos disponíveis no ficheiro csv disponibilizado. Foram utilizados e comparados quatro modelos de regressão: **Regressão Linear**, **Árvore de Decisão (Decision Tree)**, **Random Forest** e **k-Nearest Neighbors (k-NN).**
Regressão Linear: Modelo simples, desempenho modesto. Indica que há relação linear fraca entre os atributos e a idade.
Decision Tree Regressor: Um leve ganho sobre a regressão linear. Capta não-linearidades, mas sofre com overfitting.
Random Forest Regressor (melhor desempenho): Modelo mais robusto. Melhor equilíbrio entre viés e variância. Porém, a variância explicada ainda é baixa.
k-Nearest Neighbors Regressor: Desempenho inferior. Provavelmente afetado pelo tamanho do data_set e pela alta dimensionalidade.
Em todos os modelos, os **erros médios absolutos (MAE)** estão entre **11 e 12 anos**, o que representa uma margem significativa de imprecisão ao estimar a idade. Além disso, os valores de **R² abaixo de 0.30** indicam que os modelos explicam menos de 30% da variação da idade com base nos atributos disponíveis.
Com os dados disponíveis, os resultados dos modelos revelam-nos que os atributos disponíveis não são suficientemente informativos para prever a idade de forma confiável.
O desempenho semelhante entre a regressão linear e modelos mais complexos (como árvores e florestas) reforça que **não há relações complexas relevantes** entre os dados e a idade ou que tais relações são mascaradas por ruído ou co-linearidade.

O conjunto test_data, com apenas 100 exemplos, apresentou métricas ligeiramente melhores em alguns modelos. No entanto, isso provavelmente se deve à menor variabilidade ou maior homogeneidade dentro

desse conjunto, e **não pode ser interpretado como um ganho real de generalização**. Na prática, o desempenho em custom_covid19 é mais representativo e confiável.

Para concluir, não é possível prever a idade com precisão usando apenas os atributos fornecidos. Embora existam correlações fracas, os modelos apresentam erro elevado e baixa explicabilidade.

O3:
Metodologia:
Foram avaliados múltiplos modelos de regressão utilizando:

- 8 tipos de modelos (Linear, Ridge, Lasso, Árvore de Decisão, Random Forest, Gradient Boosting, SVR, KNN) que foram reduzidos para 3 depois de testes para ver qual o melhor
- Validação cruzada com otimização de hiperparâmetros
- Métricas: MAE (anos), R^2 (variância explicada)

Resultados Principais:

1. Comparação de Desempenho:

Modelo	COVID+ (R^2)	COVID- (R^2)	Melhores Parâmetros (COVID+)	Melhores Parâmetros (COVID-)
Regressão Linear	0.09	0.13	Nenhum	Nenhum
Random Forest	0.10	0.16-0.17	max_depth=5, n_estimators=150	max_depth=5, n_estimators=100
Gradient Boosting	0.10	0.16	learning_rate=0.05, max_depth=3	learning_rate=0.1, max_depth=3
Árvore de Decisão	0.06	0.12	max_depth=5	max_depth=3

2. Margens de Erro:

- COVID+: MAE ≈10.7 anos (melhor modelo)

COVID-: MAE ≈13.2 anos (melhor modelo)

Principais Conclusões:

1. **Desempenho dos Modelos**:

- Random Forest e Gradient Boosting mostraram desempenho marginalmente superior (R^2 0.10-0.17)
- Todos os modelos apresentaram erros elevados (MAE >10 anos)
- As diferenças entre modelos foram inconsistentes (no pior caso Gradient Boosting superou Decision Tree por R^2 0.10, mas tal não se repetiu consistentemenete)

2. **Impacto do Estado COVID**:

- Modelos performaram ligeiramente melhor para casos negativos
- Margens de erro maiores para previsões COVID-negativas (MAE +2.5 anos vs COVID+)

3. **Consistência de Parâmetros**:

- Parâmetros ótimos do Random Forest foram notavelmente consistentes:
  - max_depth=5 para ambos os grupos
  - min_samples_leaf=1 na maioria das execuções

- n_estimators entre 50-200

Limitações Críticas:

1. **Restrições de Dados**:

- Amostras extremamente pequenas (COVID+: n=4, COVID-: n=3 nos dados de teste)
- Alta variabilidade nas métricas devido ao tamanho amostral
- Validação cruzada limitada a 2-3 divisões (5 para o dataset completo)

2. **Confiabilidade dos Modelos**:

- Valores de R^2 <0.2 indicam explicação de <20% da variância
- MAE >10 anos é clinicamente inaceitável
- Instabilidade na otimização de parâmetros

Conclusões Finais:

1. Com os dados de teste atuais:
- Não é possível prever com confiança a idade ao óbito para nenhum dos grupos
- Padrões observados podem ser artefactos da pequena amostra
- O desempenho ligeiramente superior de métodos ensemble sugere possíveis relações não-lineares (a validar com mais dados)
2. Interpretação Clínica:
- MAE >10 anos torna as previsões clinicamente inúteis
- Não foi estabelecida diferença significativa entre grupos COVID
- Análise de importância de variáveis não é confiável com esta amostra

Recomendações:

1. Essencial repetir a análise com o dataset completo (custom_covid19.csv)
2. Para aplicação clínica:
    - Necessário incluir variáveis preditivas adicionais
    - Alvo: MAE <5 anos para utilidade prática

Avaliação Global:

Os dados de teste disponíveis são insuficientes para construir modelos fiáveis de previsão de idade em pacientes falecidos com COVID-19. Embora a análise demonstre viabilidade metodológica, todos os modelos apresentaram desempenho marginal, com erros elevados.