**Objetivo do Trabalho**

O objetivo deste trabalho é analisar dados de pacientes com cancro do fígado e usar diferentes algoritmos de aprendizagem automática para prever se um paciente irá sobreviver ou não.

**Passos do Trabalho**

1. **Carregar e Limpar os Dados**:
   * **Carregamento dos Dados**: O trabalho começa por carregar um ficheiro de dados que contém informações sobre pacientes com cancro do fígado.
   * **Limpeza dos Dados**: Os dados podem ter informações em falta (por exemplo, dados não preenchidos). O trabalho preenche estas lacunas com valores apropriados para garantir que os dados estejam completos.
2. **Transformar os Dados**:
   * **Identificação de Tipos de Dados**: As colunas do ficheiro são identificadas como numéricas (como idade, nível de hemoglobina) ou categóricas (como sintomas, hábitos de consumo de álcool).
   * **Conversão de Dados**: Os dados categóricos são convertidos em números, pois os algoritmos de aprendizagem automática funcionam melhor com números. Isto é feito transformando cada categoria numa coluna separada com valores binários (0 ou 1).
3. **Selecionar as Variáveis Mais Importantes**:
   * **Matriz de Correlação**: Calcula a correlação entre cada variável e a sobrevivência do paciente. A correlação é uma medida que indica a força e a direção do relacionamento entre duas variáveis.
   * **Selecionar Variáveis**: Escolhe as variáveis mais importantes que têm uma forte correlação com a sobrevivência do paciente.
4. **Dividir os Dados**:
   * **Variáveis Independentes e Dependentes**: Divide os dados em variáveis independentes (fatores que podem influenciar a sobrevivência, como idade, níveis de certos compostos no sangue) e a variável dependente (se o paciente sobreviveu ou não).
   * **Conjuntos de Treino e Teste**: Divide os dados em dois conjuntos: um para treinar os modelos (conjunto de treino) e outro para testar os modelos (conjunto de teste).
5. **Treinar e Avaliar os Modelos**:
   * **Modelos de Aprendizagem Automática**: Utiliza diferentes modelos para prever a sobrevivência dos pacientes. Os modelos utilizados são:
     + **Árvore de Decisão (Decision Tree)**: Um modelo que toma decisões baseadas numa série de perguntas sim/não.
     + **K-Nearest Neighbors (KNN)**: Um modelo que classifica um paciente com base nos pacientes mais semelhantes a ele.
     + **Random Forest**: Um modelo que combina várias árvores de decisão para melhorar a precisão.
     + **Gradient Boosting**: Um modelo que corrige erros de previsões anteriores para melhorar a precisão.
     + **Multi-Layer Perceptron (MLP)**: Uma rede neural simples que imita a forma como o cérebro humano processa informações.
     + **Regressão Logística (Logistic Regression)**: Um modelo que prevê a probabilidade de um evento ocorrer (neste caso, a sobrevivência do paciente).
   * **Avaliação dos Modelos**: Cada modelo é testado com o conjunto de teste para verificar a sua precisão e outras métricas de desempenho.
6. **Validação Cruzada**:
   * **Cross-Validation (Validação Cruzada)**: Uma técnica usada para avaliar a fiabilidade dos modelos, dividindo os dados várias vezes em diferentes partes de treino e teste.

**Resultado**

No final, o trabalho gera relatórios que mostram o desempenho de cada modelo. Esses relatórios incluem métricas como precisão (quantas previsões foram corretas), recall (quantas das situações reais de sobrevivência foram corretamente identificadas) e F1-score (uma média ponderada de precisão e recall).

**Em Resumo**

Este trabalho pega nos dados de pacientes com cancro do fígado, limpa e transforma esses dados, seleciona os fatores mais importantes, divide os dados para treino e teste, treina vários modelos de aprendizagem automática para prever a sobrevivência dos pacientes, e finalmente avalia a precisão desses modelos para identificar qual deles funciona melhor para este conjunto de dados.

Parte superior do formulário

Parte inferior do formulário

### Análise do Notebook

Vamos começar analisando o conteúdo do notebook:

#### 1. Exploração de Dados (Data Exploration)

* **Análise Exploratória dos Dados (EDA)**: Presente.
* **Visualizações**: Presente (histogramas e gráficos de barras).

#### 2. Pré-processamento dos Dados (Data Preprocessing)

* **Imputação de Valores em Falta**: Presente.
* **Transformação dos Dados**: Presente (codificação one-hot).
* **Escalonamento dos Dados**: Não mencionado explicitamente. (pode ser necessário verificar ou adicionar).
* **Engenharia de Características**: Não mencionado explicitamente. (pode ser necessário verificar ou adicionar).

#### 3. Modelagem de Dados (Data Modeling)

* **Definição do Conjunto Alvo**: Presente.
* **Divisão dos Dados**: Presente.
* **Seleção e Parametrização dos Algoritmos**: Presente.
* **Algoritmos**: Decision Trees, KNN, Random Forest, Gradient Boosting, MLP, Logistic Regression.
* **Treinamento dos Modelos**: Presente.

#### 4. Avaliação dos Dados (Data Evaluation)

* **Comparação dos Resultados de Classificação**: Presente (precisão, recall, F1-score).
* **Visualização dos Resultados**: Presente (tabelas e gráficos de desempenho).

#### 5. Interpretação dos Resultados (Interpretation of Results)

* **Extração de Insights**: Não mencionado explicitamente. (pode ser necessário adicionar uma seção para discutir os resultados).
* **Recomendações**: Não mencionado explicitamente. (pode ser necessário adicionar recomendações para análises futuras).

#### Elementos Extras (Extra Elements)

* **Técnicas para Imputação de Dados em Falta**: Não mencionado explicitamente. (pode ser necessário verificar ou adicionar).
* **Técnicas para Balanceamento de Dados**: Não mencionado explicitamente. (pode ser necessário verificar ou adicionar).
* **Métodos de Particionamento Adicionais**: Validação cruzada presente.
* **Algoritmos e Hiperparâmetros Adicionais**: Presente.
* **Desdobramento da Solução**: Não mencionado explicitamente. (pode ser necessário verificar ou adicionar

Temos de por todas as ultimas células a funcionar, é IMPORTANTE

URGENTE: OS NOSSOS ALGORITMOS EXEMPLO: DECISION TREE NÃO ESTAO A FUNCIONAR COM BASE EM TODOS OS DADOS, SO ESTAO A USAR A CLASS LIVES E TEMOS DE POR A USAR TODOS

Neste momento as células do knn e da decision tree não funcionam porque eu tive a tentar por com os dados todos e não consegui