Entrega da parte inicial da Documentação do projeto

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

O conjunto de dados "Câncer de Mama" é composto por duas colunas: "Número de Identificação" e "Diagnóstico". Abaixo, descrevo cada uma das variáveis/colunas e os tipos de dados armazenados em cada uma:

Número de Identificação:

Descrição: Esta coluna contém número de identificação do exame realizado para o câncer de mama. O conjunto contém diversas identificações realizadas.

Tipo de Dados: Número (String)

Representação: Cada entrada nesta coluna representa uma identificação de câncer de mama encontrado.

Diagnóstico:

Descrição: O diagnóstico do câncer relacionado a um número de identificação. Os valores possíveis são “benigno” ou “maligno”.

Tipo de Dados: Texto (String)

Representação: Cada entrada nesta coluna é uma categoria que indica se o câncer de mama é maligno ou benigno. Esses valores são atribuídos ao número de Identificação.

Raio (média):

Descrição: Distância média do centro aos pontos na periferia.

Tipo de Dados: Número(int)

Representação: Valores numéricos que indicam a extensão média do tumor.

Textura:

Descrição: Desvio padrão dos valores em escala de cinza.

Tipo de Dados: Texto(String)

Representação:Valores que descrevem a variação na textura da imagem do tecido mamário.

Perímetro:

Descrição: Perímetro da massa ou tumor.

Tipo de Dados: Número(int)

Representação: Valores numéricos indicando o comprimento total da borda do tumor.

Área:

Descrição: Área da massa ou tumor.

Tipo de Dados: Número(int)

Representação:Valores numéricos indicando a extensão da área ocupada pelo tumor.

Suavidade:

Descrição: Variação local nos comprimentos dos raios.

Tipo de Dados: Texto (String)

Representação: Valores textuais que descrevem a suavidade ou irregularidade da superfície do tumor.

Compacidade:

Descrição: Indica a compactação do tumor.

Tipo de Dados: Texto(String)

Representação:Valores textuais que descrevem o quão compacta é a forma da massa ou tumor.

Cavidade:

Descrição:Severidade de porções côncavas na borda.

Tipo de Dados: Texto (String)

Representação: Valores textuais indicando a presença e gravidade de cavidades na borda.

Pontos côncavos:

Descrição: Número de porções côncavas na borda.

Tipo de Dados: Texto(String)

Representação:Valores textuais indicando a quantidade de áreas côncavas na borda da massa.

Simetria:

Descrição: Avalia a simetria da massa ou tumor.

Tipo de Dados: Número(int)

Representação: Valores numéricos indicando o grau de simetria na forma do tumor.

Dimensão fractal:

Descrição: Aproximação de "contorno costeiro" - 1.

Tipo de Dados: Número(int)

Representação: Valores numéricos que descrevem a complexidade da forma do tumor, com base em conceitos fracta.

Em resumo, esse conjunto de dados contém informações de análises de identificações de câncer de mama contendo a classificação do tipo do câncer, sendo benigno ou maligno. Com esses dados podem ser utilizados para desenvolver modelos de aprendizado de máquina para classificar de forma automática se possui ou não câncer de mama facilitando sua identificação e melhorando a rapidez de um início de tratamento para este câncer.

Variável Target: Como a Variável Target é a Variável de interesse do banco de dados, nossa variável seria o Diagnóstico onde ele teria somente 2 variáveis sendo M para Maligno e B para Benigno com isso podemos fazer diversas análises para estudos como porcentagem de câncer mais agressivo ou menos agressivo, porcentagem de câncer de mama entre as mulheres e também utilizando os diagnósticos temos como calcular a porcentagem de risco por idade, onde é possível identificar idades onde tem um risco maior e definir motivos para isso.

A variável target em um conjunto de dados é aquela que você está tentando prever ou entender. A variável target é o diagnóstico, que tem duas categorias: M (Maligno) e B (Benigno). Quando se prepara os dados para aplicação de técnicas de Inteligência Artificial (IA),

Codificação da Variável Target:

Modelos de IA geralmente requerem que as variáveis categóricas sejam convertidas em valores numéricos. É possível usar uma técnica chamada codificação one-hot para transformar a variável target em dois vetores binários, um para Maligno e outro para Benigno.

Tratamento de Dados Ausentes:

Verificação se há valores ausentes na variável target. Algoritmos de IA muitas vezes não lidam bem com dados ausentes, então é importante garantir que todas as amostras tenham um valor na variável target.

Balanceamento de Classes:

Verificar se as classes (Maligno e Benigno) estão balanceadas. Se houver um desequilíbrio significativo, isso pode afetar o desempenho do modelo, e técnicas como oversampling (aumentar o número de amostras da classe minoritária) ou undersampling (reduzir o número de amostras da classe majoritária) podem ser aplicadas.

Normalização/Padronização:

Dependendo do algoritmo escolhido, pode ser necessário normalizar e padronizar as características do conjunto de dados. Isso ajuda a garantir que todas as características tenham a mesma escala, o que pode melhorar o desempenho de certos algoritmos.

Engenharia de Recursos (Feature Engineering):

Avaliar se há oportunidades para criar novas características (features) derivadas das existentes que podem ser mais informativas para o modelo. Por exemplo, criar uma nova variável representando a idade do paciente, se isso não estiver diretamente disponível.

Transformação de Variáveis Contínuas:

Se houver variáveis contínuas no conjunto de dados, pode ser útil aplicar transformações como logaritmo ou normalização para garantir que a distribuição seja mais adequada para o modelo.

Tratamento de Outliers:

Identifique e trate outliers nas variáveis relevantes, pois eles podem afetar negativamente o desempenho de alguns modelos.

Validação do Modelo:

Prepare conjuntos de treino, validação e teste. Isso é essencial para avaliar o desempenho do modelo em dados não vistos.

Seleção de Modelo e Avaliação:

Escolher um modelo apropriado para o seu problema (classificação, neste caso) e avalie seu desempenho usando métricas relevantes, como precisão, recall, F1-score, área sob a curva ROC, etc.

Entrega 16/11/2023

Escolhi fazer o primeiro método de aprendizado de máquinas por árvore de decisões.

O modelo de árvore de decisão alcança uma determinada acurácia no conjunto de teste.

A matriz de confusão e o relatório de classificação fornecem informações detalhadas sobre o desempenho do modelo em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos assim dando se o câncer é benigno ou maligno.

A visualização da árvore de decisão ajuda na interpretação das regras de decisão aprendidas pelo modelo .