Inteligência Artificial

Gabriel Rodrigues da Silva Costa	822137024
Lucas Ribeiro Pedroso	823149292
Fabrício Peres dos Santos	822160071
Thiago Duarte Reis	822141527

Projeto A3

Universidade São Judas Tadeu

Orientador: EVANDRO CATELANI FERRAZ

O conjunto de dados "Câncer de Mama" é composto por doze colunas, tendo "Diagnóstico" como variável target. Abaixo, descrevo cada uma das variáveis/colunas e os tipos de dados armazenados em cada uma:

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

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.

O 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 dependente 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.

Variável Preditora: As variáveis preditoras são independentes, ou seja, a que iremos utilizar para treinar o modelo de regressão(resposta), prever uma variável de resultado.

Quando se prepara os dados para aplicação de técnicas de Inteligência Artificial (IA),

Codificação da variável alvo: Muitos algoritmos de IA requerem que a variável alvo seja numérica. Normalmente, é usada uma codificação binária (0 para benigno, 1 para maligno).

Padronização ou Normalização: Em alguns algoritmos, normalizar ou padronizar os dados pode ajudar no desempenho do modelo, especialmente se os dados estiverem em diferentes escalas. Por exemplo, aplicar a normalização min-max ou padronização Z-score às características numéricas.

Lidar com valores ausentes: Verificar se existem valores ausentes e decidir sobre a imputação de dados (preenchendo-os com a média, mediana, ou usando métodos mais avançados, como a imputação por K-vizinhos mais próximos).

Redução de dimensionalidade: Dependendo da complexidade do modelo e do número de características, técnicas como Análise de Componentes Principais (PCA) podem ser úteis para reduzir a dimensionalidade e preservar a variância dos dados.

Tratamento de outliers: Identificar e tratar outliers, se existirem, para evitar que eles afetem negativamente o desempenho do modelo.

Seleção de características: Algumas técnicas de IA podem se beneficiar da seleção de características, removendo aquelas que têm baixa importância ou não contribuem significativamente para o modelo.

Primeiro método

O primeiro método de aprendizado de máquinas por árvore de decisões.

O código começa carregando dados de um arquivo CSV que contém informações sobre tumores. Ele remove colunas não essenciais e converte a variável alvo ('diagnosis') para um formato numérico ('M' para 1 e 'B' para 0) para permitir a modelagem.

Divide os dados em conjuntos de treino e teste para avaliar o desempenho do modelo.

Em seguida, utiliza um algoritmo de busca em grade (GridSearchCV) para encontrar os melhores hiperparâmetros para o modelo de árvore de decisão, como a profundidade máxima da árvore e os critérios para divisão e tamanho mínimo das amostras nas folhas.

Após a busca em grade, ajusta o modelo com os melhores hiperparâmetros encontrados e realiza previsões nos dados de teste.

Avalia o desempenho do modelo através da acurácia, matriz de confusão e relatório de classificação, fornecendo uma visão abrangente da capacidade do modelo em prever tumores como benignos ou malignos.

Por fim, exibe visualmente a árvore de decisão resultante, permitindo uma interpretação mais intuitiva das decisões tomadas pelo modelo.

Este código representa um fluxo típico de construção, ajuste e avaliação de um modelo de classificação usando a árvore de decisão, com o objetivo de identificar tumores malignos e benignos com base em suas características.

Segundo método

Esse código é um exemplo de como usar uma Rede Neural Artificial (RNA) para classificar dados. Vamos passar por isso passo a passo:

Importações:

Importa as bibliotecas necessárias, incluindo numpy, pandas e ferramentas do scikit-learn para redes neurais, pré-processamento de dados e avaliação de modelos.

Leitura e Preparação dos Dados:

Carrega um conjunto de dados a partir de um arquivo CSV.

Remove uma coluna ('Unnamed: 32') que parece ser sem nome ou irrelevante.

Divide os dados em variáveis independentes (X) e dependente (y).

Pré-processamento dos Dados:

Identifica colunas numéricas e categóricas nos dados.

Cria transformadores para lidar com valores ausentes, substituindo valores numéricos pela média e valores categóricos pelo valor mais frequente.

Usa esses transformadores para criar um pré-processador que aplica essas transformações aos dados.

Divisão em Conjuntos de Treino e Teste:

Separa os dados em conjuntos de treino e teste para poder avaliar o desempenho do modelo.

Escalonamento de Dados Numéricos:

Padroniza os dados numéricos (escalonamento) para ter média zero e variância unitária, o que pode ajudar na convergência do modelo.

Configuração e Treinamento da Rede Neural:

Define um grid de hiperparâmetros para a rede neural (número de neurônios, taxa de aprendizado, etc.).

Utiliza GridSearchCV para encontrar a combinação mais adequada de hiperparâmetros por meio de validação cruzada nos dados de treino.

Imprime os melhores parâmetros encontrados.

Avalia o modelo utilizando um relatório de classificação, matriz de confusão e a acurácia do modelo nos dados de teste.

Calcula e plota a curva ROC (Receiver Operating Characteristic) para avaliar o desempenho do modelo em diferentes limiares de classificação.

Esse código é uma boa demonstração de como construir e otimizar um modelo de rede neural para classificação, além de realizar uma análise detalhada do desempenho do modelo usando métricas comuns de avaliação.

Terceiro método

A escolha do terceiro método algoritmo KNN (K-Nearest Neighbors) para classificação. Vou resumir cada etapa para ajudar no entendimento:

Importação de Bibliotecas:

Importa as bibliotecas necessárias para trabalhar com os dados, modelagem e avaliação do algoritmo.

Carregamento e Preparação dos Dados:

Carrega um conjunto de dados do arquivo 'data.csv'.

Remove colunas ('id', 'Unnamed: 32') que não são relevantes para o modelo.

Mapeia a coluna 'diagnosis' para valores binários (M: 1, B: 0).

Divide os dados em variáveis preditoras (X) e alvo/target (y).

Pré-processamento:

Normaliza os dados usando a classe Normalizer do scikit-learn.

Realiza a redução de dimensionalidade com PCA (Principal Component Analysis) para 10 componentes.

Divisão dos Dados em Treino e Teste:

Separa os dados em conjuntos de treino e teste.

Definição dos Parâmetros e Busca pelo Melhor Modelo:

Define uma grade de parâmetros para o modelo KNN (número de vizinhos).

Utiliza o GridSearchCV para encontrar os melhores parâmetros do modelo com validação cruzada (5 folds).

Treinamento do Melhor Modelo:

Treina o melhor modelo encontrado com base nos melhores parâmetros e dados de treino.

Avaliação do Modelo:

Avalia a precisão do modelo final utilizando todo o conjunto de dados.

Conclusão

Neste projeto, utilizamos três métodos de aplicação de IA em um conjunto de dados: Câncer de Mama. Ao realizar os três métodos , podemos concluir que todos foram precisos e eficazes(com sucesso de 95%) , mas a Árvore de Decisão é capaz de visualizar melhor os caminhos de decisões IA operando no código e assim, ajudando no entendimento. Mas, os métodos KNN(K-Nearest Neighbors) e Redes Neurais também exerceram seu papel.