PROJETO EBAC/SEMANTIX

Diagnóstico Precoce de Doenças Cardíacas

Utilizando Machine Learning

1. Coleta de dados

Para esse projeto foi usado dados públicos obtido no site data.gov , no seguinte url: https://catalog.data.gov/dataset/rates-and-trends-in-heart-disease-and-stroke-mortality-among-us-adults-35-by-county-a-2000-45659

Os dados foram baixados no formato .csv e tem os seguintes atributos:

Year: Ano dos dados, útil para analisar tendências temporais.

- LocationDesc: Descrição da localização (ex.: nome do condado ou cidade).
- DataSource: Fonte dos dados, pode ser útil para entender a origem dos dados.
- Class: Classe das doenças (doenças cardiovasculares).
- Topic: Tópico específico das doenças (ex.: doenças cardíacas).
- Data_Value: Valor dos dados (taxa de doenças cardíacas), a variável-alvo.
- Confidence_limit_Low: Limite inferior do intervalo de confiança, útil para entender a precisão dos dados.
- Confidence_limit_High: Limite superior do intervalo de confiança.
- StratificationCategory1: Categoria de estratificação (ex.: grupo etário).

- Stratification1: Estratificação específica (ex.: idades 35-64 anos).
- StratificationCategory2: Categoria de estratificação adicional (ex.: raça).
- Stratification2: Estratificação específica adicional (ex.: indígena americano/nativo do Alasca).
- StratificationCategory3: Categoria de estratificação adicional (ex.: sexo).
- Stratification3: Estratificação específica adicional (ex.: geral).
- LocationID: ID da localização, útil para indexação e agrupamento.
- GeographicLevel: Nível geográfico (ex.: condado, estado), pode ser derivado de LocationDesc.
- Data_Value_Unit: Unidade do valor dos dados (ex.: por 100.000), pode ser útil se houver diferentes unidades.
- Data_Value_Type: Tipo de valor dos dados (ex.: taxa ajustada por idade), pode ser útil para entender o tipo de medida.
- Data_Value_Footnote_Symbol: Símbolo de nota de rodapé do valor dos dados, pode ser descartado se não for relevante.
- Data_Value_Footnote: Nota de rodapé do valor dos dados, pode ser descartado se não for relevante.

2. Modelagem

Após tratamento dos dados, foi realizada a normalização dos dados para a modelagem.

Os dados foram separados em dados de treino e dados de teste usando train_test_split do sklearn.

O modelo escolhido foi o Random Forest Classifier com os parâmetros padrões.

Treinado o modelo, foi feita previsão e avaliado o modelo.

Para a avaliação foi usado o accuracy_score e classification_report, ambos da biblioteca sklearn.metrics.

Realizou-se também a visualização dos resultados através da matriz de confusão e a validação cruzada.

3. Conclusões

Com o modelo RandomForestClassifier foi obtido uma acurácia de 99,82%, além disso o modelo teve também precision, recall e f1-score altos, o que mostram que o modelo está equilibrado e é eficiente em identificar ambos os casos de presença e ausência de doenças cardíacas.

Pela matriz de confusão acima pode-se notar baixa taxa de falsos positivos (indica que há poucas previsões erradas quando elas não estão presentes) e falso negativos (indica que há poucas previsões erradas de ausência de doença cardíaca quando ela está presente.)

Esses resultados sugerem que o modelo é bem equilibrado e confiável para o diagnóstico precoce de doenças cardíacas. Contudo, é sempre prudente avaliar o modelo com dados adicionais ou através de validação cruzada para garantir que ele generalize bem para diferentes conjuntos de dados.

Pela validação cruzada obteve-se uma média de 0.8572, indicando que o modelo apresenta uma precisão de cerca de 85,72% ao ser testado em diferentes divisões dos dados, o que é um valor bem aceitável.

Ao final foi construído um gráfico de importância das variáveis usando Feature Importance, que é uma métrica utilizada em modelos de machine learning para avaliar a contribuição de cada variável (feature) na previsão do modelo. Em outras palavras, ela indica quão importante cada característica é para o modelo tomar uma decisão.