PROJETO EBAC SEMANTIX

Diagnóstico Precoce de Doenças Cardíacas (Machine Learning)

1. **Coleta de dados**

Para o projeto, os dados foram retirados do seguinte dataset público: 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 retirados e manuseados de um arquivo csv, contendo 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 o tratamento adequado dos dados, foi feita a normalização (ou padronização) para que a modelagem ocorra da melhor forma. Foi feito:

- Separação dos dados em treino e teste usando o train_test_split do skylearn;
- Utilizando o modelo Random Forest Classifier com os parâmetros padrões;
- Treinamento, avaliando o modelo (com accuracy_store e classification_report) e fazendo previsões;
- Visualização dos resultados através da matriz de confusão e validação cruzada.

3. Conclusão

Com o modelo Random Forest Classifier foi obtido uma acurácia de *99,82%.* O modelo possui precision, recall e f1-score altos, sinal que ele possui equilíbrio e é eficiente em identificar os casos de presença e ausência de doenças cardíacas.

Pela matriz de confusão, pode se notar baixa taxa de falsos positivos e falsos negativos, indicando que há pouca presença de previsões erradas presentes e não presentes.

Os resultados acima sugerem que o modelo está eficiente, equilibrado e confiável para o diagnóstico precoce de doenças cardíacas. Porém, de qualquer forma, é recomendável avaliar o modelo com dados adicionais ou através da validação cruzada para garantir sua adaptação para novos conjuntos de dados.

Assim, com a validação cruzada, obteve-se a média de 0.8572 (85,72%) ao ser testado em diferentes divisões de dados. Ao final do estudo em Machine Learning, foi construído um gráfico de importância das variáveis utilizando o Feature Importance, métrica usada em modelos de Machine Learning para avaliar a contribuição de cada variável na previsão do modelo, indicando quão importante é cada característica para a previsão do modelo.