Universidade Federal da Bahia
Programa de Pós-Graduação em Ciência de Dados e Big Data
Carlos Magno Santos Ribeiro de Brito

Prof Dr. Paulo Henrique Ferreira da Silva

Trabalho Individual de Machine Learning

INTRODUÇÃO

O câncer é um problema de saúde que se caracteriza por uma mortalidade anual bastante elevada em todo o planeta e que é evidentemente um risco para a saúde humana.

Contudo, o tratamento e diagnóstico vem sendo cada vez mais eficiente e facilitado por uso de tecnologias. Uma dessas tecnologias é a *Machine Learning*, que é uma técnica de aprendizado supervisionado que permite ao usuário aplicar uma série de regras para identificar padrões e entender o comportamento de um conjunto de dados.

Nesse contexto, a taxa de sobrevida é uma estimativa utilizada que indica a probabilidade de que um paciente seja recuperado após um diagnóstico a partir de um histórico de dados de outros pacientes que também tiveram a mesma doença, com características semelhantes.

Então, com esse trabalho se discute o mais adequado modelo de classificação para prever a sobrevida de paciente com câncer de mama após 5 anos. Utiliza-se o conjunto de dados que contém casos de um estudo realizado entre 1958 e 1970 no Hospitaln Billings da Universidade de Chicago, acerca da sobrevivência de 306 pacientes que se submeteram a cirurgia para câncer de mama.

PRINCIPAIS CARACTERÍSTICAS

Para o tratamento desses dados foi utilizada a linguagem python, que é uma linguagem de programação de alto nível, com uma sintaxe simples e flexível.

Além disso, conforme dito na introdução, tem-se um conjunto de dados onde estão disponíveis dados de 306 pacientes submetidos à cirurgia de câncer de mama. Esse *dataset* contém as seguintes informações:

- *Idade* - Idade do paciente no momento da cirurgia variando de 30 a 83 anos;
- *Ano* - Ano em que o paciente foi submetido à cirurgia variando entre 1958 e 1969;
- *Número de nódulos* - Número de nódulos encontrados na mama variando entre 0 e 52;
- *Status de sobrevivência* - Com status igual a 1, o paciente sobreviveu 5 anos ou mais; com status igual a 2, o paciente morreu dentro de 5 anos;

Com essas quatro variáveis se analisou o dataset e foi possível identificar qual melhor modelo melhor se adequa na classificação desses resultados.

MODELOS E MÉTRICAS UTILIZADOS

Para o presente trabalho comparou-se os resultados obtivos com os modelos de predição: Análise discriminante, Regressão Logística, Árvore de Decisão e Random Forest, Naive Bayes e KNN.

1. **Análise discriminante** - Essa técnica compara diferença entre grupos e classifica o dado analisado no grupo que tenha as características mais semelhantes a ele. Todas as variáveis utilziadas são categóricas, mesmo podendo ter mais de 2 categorias. No Python ele é implementado com pacote de uso livre da comunidade *Sklearn*, usando o método `LinearDiscriminantAnalysis()`;

2. **Regressão Logística** - Para essa técnica é utilizado o método `LogisticRegression()` do pacote *scikit-learn*. Nessa regressão, tem-se um modelo para prever valores que serão assumidos por uma varável categórica por meio de variáveis contínuas independentes;

3. **Árvore de Decisão** - Para essa técnica é utilizado o pacote *scikit-learn* e o método `DecisionTreeClassifier()`. A Árvore de Decisão também utiliza do aprendizado supervisionado para classificar e prever os dados e utiliza de uma árvore com ramificações variáveis de acordo com a quantidade de atributos e seus valores. Essa técnica busca identificar os atributos que fornecem a maioria das informações, removendo os raros, para melhorar o modelo;

4. **Naive Bayes** - Para essa técnica é utilizado o método `GaussianNB()` do pacote *scikit-learn*. Naive Bayes é um modelo de aprendizado supervisionado onde cada variável de análise recebe um peso em cada uma das classes, esses pesos serão somados e a classe com maior peso será a que classificará o novo objeto;

5. **Random Forest** - Para essa técnica é utilizado o método `RandomForestClassifier()` do pacote *scikit-learn*. O Random Forest utiliza o modelo de árvore de decisão a partir de suconjuntos de atributos aleatoriamente selecionados e em seguida efetua a classificação da variável de interesse de acordo com a árvore que possui a melhor lógica e vantagens para tomar a decisão;

6. **KNN** - Para essa técnica é utilizado o método `KNeighborsClassifier()` do pacote *scikit-learn*. O KNN é um modelo de aprendizado supervisionado que utiliza a distância euclidiana entre os pontos de treinamento para classificar um novo objeto.

Além desses métodos, as seguintes métricas de avaliação foram utilizadas: Matriz de confusão, Precisão, Coeficiente de Correlação de Matthews, Recall, F-Score e Accuracy.

- Matriz de confusão:
  - **TP** - Verdadeiro Positivo: Valores que são positivos no conjunto de testes e também positivos na predição;
  - **FN** - Falso Negativo: Valores que são positivos no conjunto de testes e negativos na predição;
  - **FP** - Falso Positivo: Valores que são negativos no conjunto de testes e positivos na predição;
  - **TN** - Verdadeiro Negativo: Valores que são negativos no conjunto de testes e também negativos na predição.
- Precisão: É a proporção dada como: $P = \frac{VP}{VP+FP}$;
- Coeficiente de Correlação de Matthews (MCC) - Interpreta aleatoriedade da relação. Sendo próximo de **1**, a classificação é perfeita. Próximo de **-1** a classificação é inversa e se próximo de **0** a classificação é aleatória;
- Recall: É a proporção dada como: $P = \frac{VP}{VP+FN}$;
- F-Score: Trade-off entre recall e Precisão e dado como: $F = \frac{2 \times P \times R}{P+R}$;