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, 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. **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:
  - **VP** - 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;
  - **VN** - Verdadeiro Negativo: Valores que são negativos no conjunto de testes e também negativos na predição.

- Precisão (P): 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;
- Acurácia (ACC): É a proporção dada como: $ACC = \frac{VP+VN}{VP+FN+FP+VN}$;
- Recall (R): É a proporção dada como: $R = \frac{VP}{VP+FN}$;
- F-Score (F): Trade-off entre recall e Precisão e dado como: $F = \frac{2\times P\times R}{P+R}$;

Com esses modelos e métricas, analisou-se o conjunto de dados, em uma proporção de 70% para treino e 30% para teste e econtrou-se os seguintes resultados:

RESULTADOS

Modelo	Acurácia (ACC)
LR	0.774603
LDA	0.785516
KNN	0.766865
CART	0.681746
NB	0.790278

Tabela 1: Comparação inicial entre modelos

A partir de agora, a fim de facilitar a escrita e compreensão, denominar-se-á como:

- LDA - Análise discriminante;
- LR - Regressão Logística;
- KNN - K-ésimo Vizinho mais Próximo;
- CART - Árvore de decisão;
- NB - Naive Bayes;

Em primeira mão, analisou-se os modelos com os conjuntos de treinamento para verificar qual deles aparesentaria a *priori* a melhor precisão. Por se tratar de uma comparação entre os modelos, utilizou-se a métrica de acurácia (ACC) para fazer essa avaliação. O Resultado obtido é dado na Tabela 3.

É possível observar pela validação cruzada (Kfold) que o modelo NB apresenta a maior acurácia, com um valor de 0.790278 e o CART o menor com um valor de 0.681746. Verificar-se-á com a execução dos modelos se esse resultado é de fato o encontrado.

1) LR - REGRESSÃO LOGÍSTICA

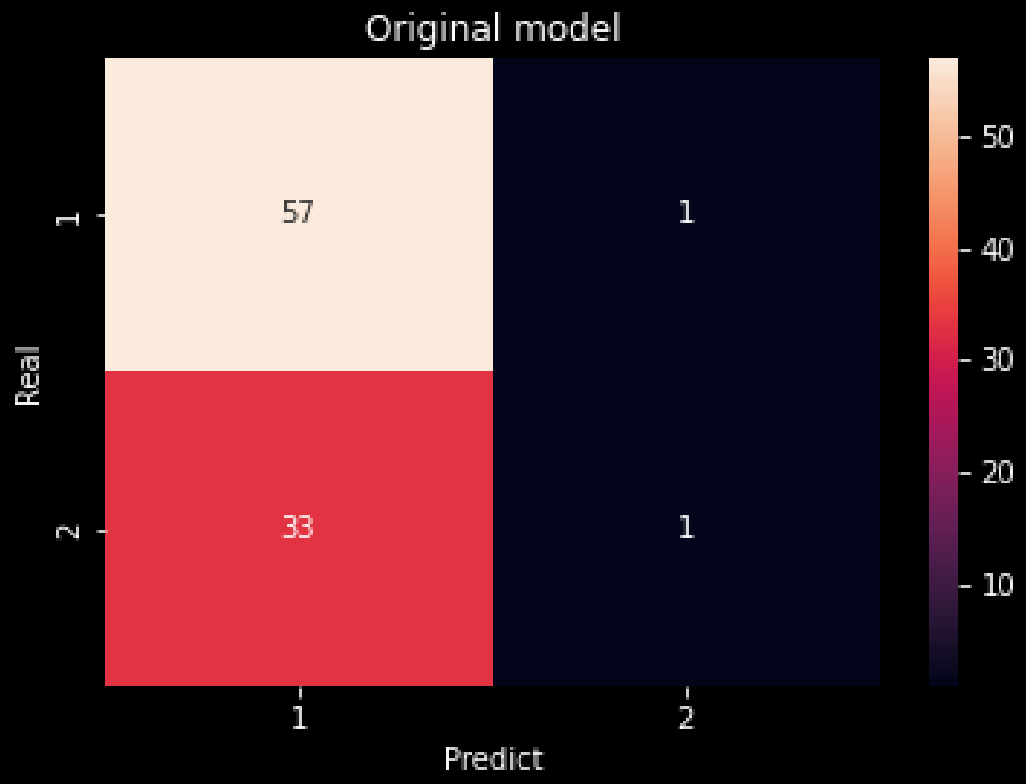

Figura 1: Matriz de confusão da Regressão Logística

Para o caso da regressão Logística simples, sem nenhum parâmetro de regularização obteve um MCC bastante baixo (4%), muito próximo da aleatoriedade (Tabela 2). Pela matriz de confusão, o modelo indica também que há maior probabilidade de resultados 1, indicando uma sobrevida maior que 5 anos do paciente.

Esse modelo obteve uma acurácia muito aquém da validação cruzada feita inicialmente, o que indica uma possibilidade de utilizaçãod e parâmetros de regularização numa tentativa

Métrica	Valor
ACC	0.63043
MCC	0.04028
R	0.98275
P	0.63333
F	0.77027

Tabela 2: LR métricas

de aumentar esse resultado.

2) LDA - ANÁLISE DISCRIMINANTE

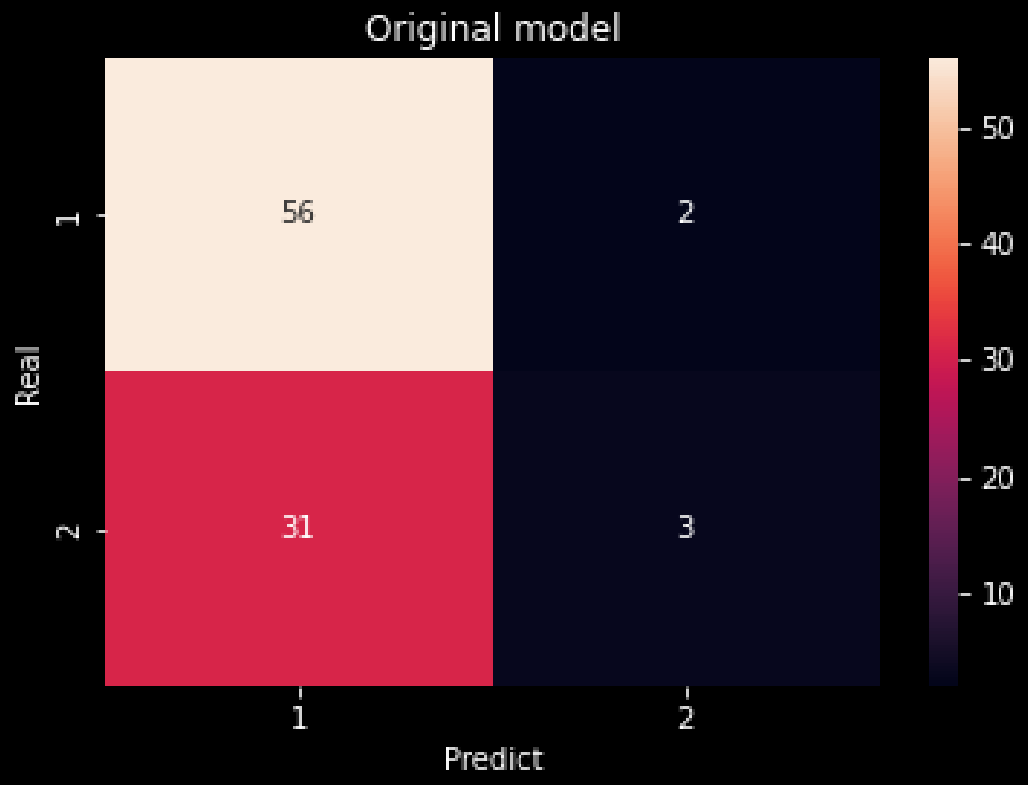

Figura 2: Matriz de confusão da Análise discriminante

No caso da análise discriminante, quase não houve diferença em relação à regressão Logística. A acurácia se manteve muito parecida, não valendo a pena preteri-la. Há de se

Métrica	Valor
ACC	0.64130
MCC	0.11444
R	0.96551
P	0.64367
F	0.77241

Tabela 3: LDA métricas

observar a situação do Recall (R). O modelo apresenta um valor menor que na regressão logística, sendo o fato prejudicial, já que criar a expectativa de sobrevida maior que 5 anos é uma situação ruim. No caso, falsos negativos são mais prejudiciais que os falso positivos.

3) CART - ÁRVORE DE DECISÃO

Métrica	Original	Melhorado
ACC	0.608696	0.630435
MCC	0.119594	0.160538
R	0.758621	0.793103
P	0.666667	0.676471
F	0.709677	0.730159

Tabela 4: CART métricas comparativas

A árvore de decisão apresentou uma acurácia (ACC) bastante pequena inicialmente. Então, aplicou-se alguns parâmetros de regularização para melhorá-la e ainda assim, não se obteve um resultado satisfatório, sendo ele o menor resultado de todos os modelos (63%).

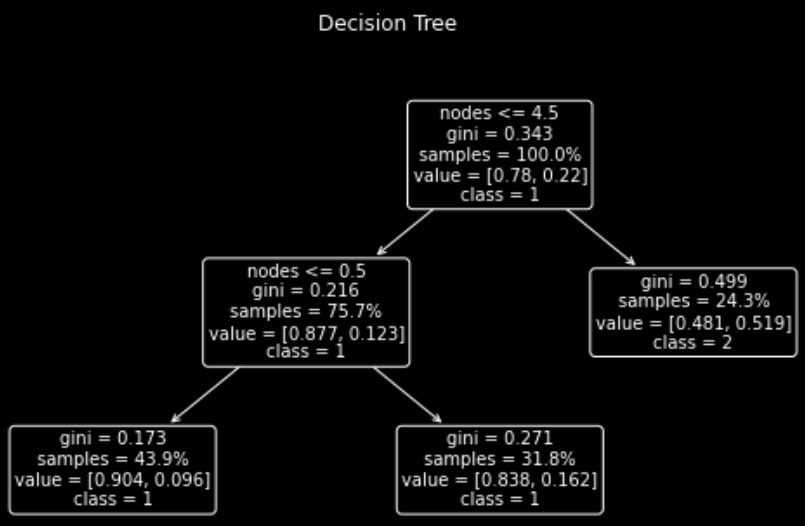

Figura 3: Árvore de decisão

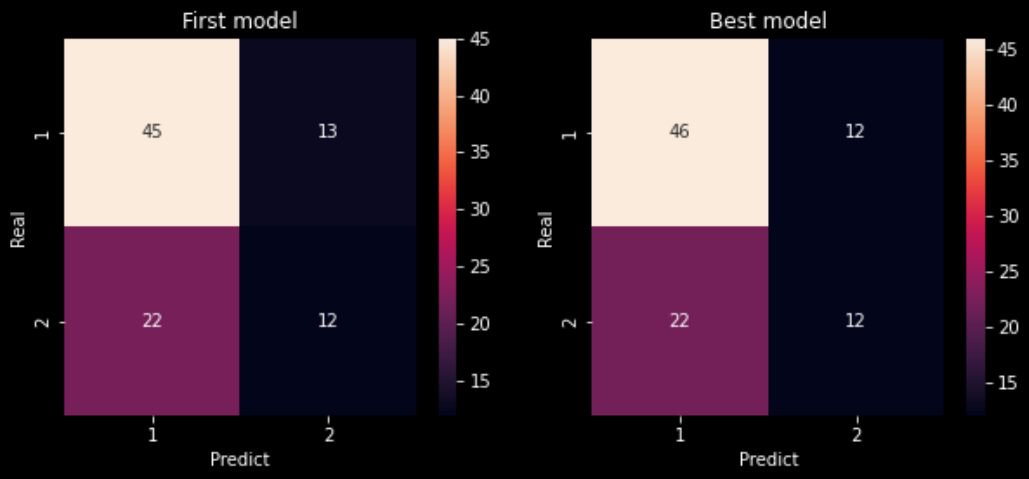

Figura 4: Matriz de confusão comparativa da Árvore de decisão

4) KNN

Para o KNN, inicialmente obteve-se uma acurácia de 64%, o que se encontrou muito parecido com os resultados obtidos anteriormente nas outras análises, apesar de ser um dos maiores ainda assim. Mudou-se então os parâmetros de distância, sendo então testados com as distâncias Euclidiana, de Manhattan e Minkowski e com valores de k variando de 1 a 23. Para uma ditância euclidiana e um k=21, encontrou-se uma acurácia mais elevada

correspondente a mais que 70%. Tem-se um MCC também mais elevado entre todos os outros modelos, indicando um dado mais próximo do real, além de um R mais elevado. Observa-se qeu o valor de F também se eleva com o aumento de P, o que indica uma maior quantidade de acertos positivos.

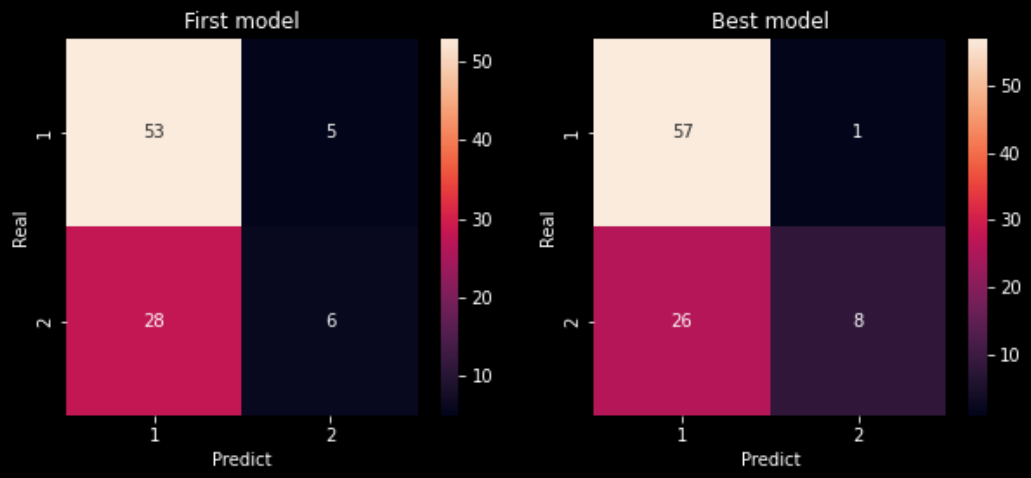

Figura 5: Matriz de confusão comparativa do modelo KNN

Métrica	Original	Melhorado
ACC	0.641304	0.706522
MCC	0.134285	0.354287
R	0.913793	0.982759
P	0.654321	0.686747
F	0.762590	0.808511

Tabela 5: KNN métricas originais e ajustadas

5) NB - NAIVE BAYES

Inicialmente, o modelo NB apresentou o maior valor de precisão com a avaliação cruzada do conjunto de treinamento. Todavia, após a análise com o conjunto de testes, verificou-se que os resultados encontrados não apresentam nenhuma tendência de melhora com ajuste de parâmetros e também com todos os valores inferiores ao KNN ajustado, por exemplo. A matriz de confusão continua a mostrar uma prevalecência de valores 1 em ambos os casos.

Métrica	Original	Melhorado
ACC	0.652174	0.652174
MCC	0.162579	0.162579
R	0.965517	0.965517
P	0.651163	0.651163
F	0.777778	0.777778

Tabela 6: NB métricas originais e ajustadas

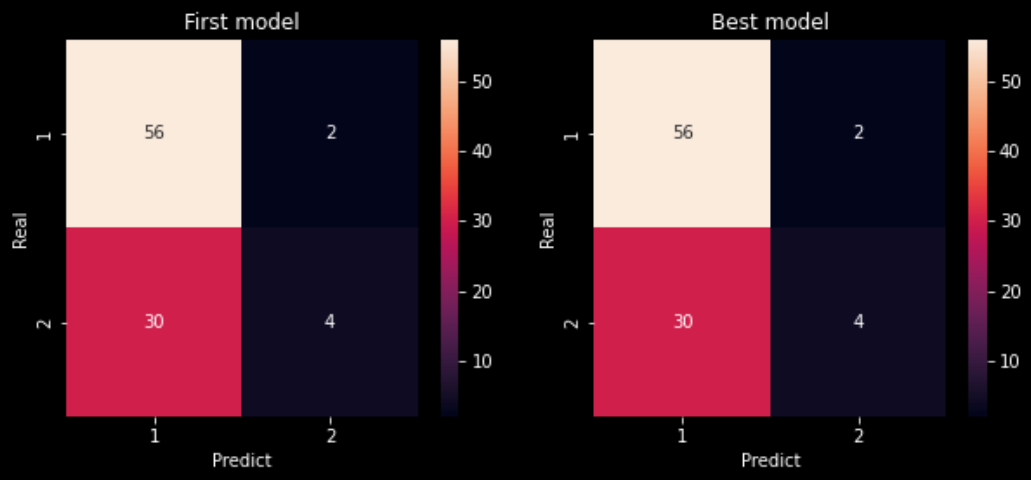

Figura 6: Matriz de confusão comparativa do modelo Naive Bayes

CONCLUSÃO

Ao analisar os cinco modelos propostos, observa-se que todos eles apresentam uma acurácia acima de 50%. Num caso como tal, por se tratar de um problema que envolve a saúde e a condição de vida de pacientes com cancer, quanto menor a quatidade de erros e maior a acurácia, melhor.

Isto posto, Ao observar que o modelo KNN não somente apresentou maior acurácia (ACC), como também melhor predição de VP e todas as outras métricas melhores que os outros, pode-se ecolhê-lo como o modelo mais indicado para a análise.

Observa-se também que a análise inicial com os conjuntos de treinamento não foram contempladas, já que o modelo NB era o mais indicado a primeira vista.