Universidade Estadual de Londrina

GABRIEL TORRESIN DE OLIVEIRA GARDIN

PROJETO FINAL

LONDRINA
2021

GABRIEL TORRESIN DE OLIVEIRA GARDIN

PROJETO FINAL:
TREINAMENTO DE UM MODELO DE CLASSIFICAÇÃO

Relatório apresentado à Disciplina Fundamentos Matemáticos e Computacionais de Machine Learning do Departamento de Computação da Universidade Estadual de Londrina.

Docente: Ms. Edison Antonio Sahd Filho

Londrina
2021

OBJETIVOS

Utilizando os conceitos passados em aula realizar a análise exploratória de dados reais, aplicar técnicas de pré-processamento e por final gerar um modelo capaz de identificar se um paciente tem ou não diabetes, utilizando um classificador binário e o dataset Pima Indians Diabetes.

1 INTRODUÇÃO

Quando se discute aplicações de aprendizado de máquinas usos médicos para detecção de enfermidades é uma das primeiras aplicações que nos deparamos em artigos e livros.

Técnicas de Classificação são utilizadas para que utilizando uma série de atributos seja possível determinar se uma amostra pertence ou não a alguma classe, por exemplo: Utilizando fatores como sexo, idade, pressão sanguínea e nível de açúcar no sangue determinar se um paciente tem diabetes. Este é um exemplo de um classificador binário, que é um tipo de aprendizado de maquina supervisionado.

Existem diversos tipos de algoritmos de aprendizado de máquina capazes de serem usados para atividades de classificação, dentre eles podemos destacar: Regressão Logística, Máquinas de Vetores de Suporte, KNN, Árvores de decisão, Florestas aleatórias e Naive Bayes. Cada um destes algoritmos possuem caracterísitcas específicas além de vantagens e desvantagens, como tempo de treinamento e complexidade computacional.

ANÁLISE EXPLORATÓRIA E PRÉ PROCESSAMENTO DOS DADOS

Afim de conhecer melhor os dados e quais as limitações do dataset é necessário explorar os dados para obter informações úteis, como quantidades de dados ausentes, correlações entre diferentes atributos e outras métricas estatísticas como detecção de outliers, gráficos de disperção e avaliação de linearidade dos dados.

No dataset Pima Indians Diabetes não foi encontrado dados ausentes como NaN, porém, muitos dados foram imputados como 0, o que nos casos identificados, foi considerado como dados ausentes. Para imputar os dados ausentes, foi analisado a normalidade dos dados utilizando a curtose para definir se os atributos com dados ausentes seguem uma curva normal ou não, para então decidir qual técnica usar para imputação de dados.

Os valores de curtose foram todos próximos de zero, indicando que os parâmetros que possuem dados ausentes possuem distribuição normal, e por isto, foram imputados com a média dos dados.

As caracterísitcas também foram escalonadas utilizando a classe StandardScaler da biblioteca sklearn.

TREINAMENTO DOS MODELOS

Dois modelos foram treinados utilizando a biblioteca Scikit-Learn afim de determinar qual o melhor algoritmo para classificação do dataset Pima Indians Diabetes, a Regressão Logística e Florestas Aleatórias. O modelo de regressão logística obteve um desempenho um pouco pior do que o alcançado com as florestas aleatrórias, o que era esperado uma vez que o segundo algoritmo é mais potente.

Visto que o modelo RandomForest obteve um melhor resultado preliminar, foi utilizado a biblioteca GridSearchCV para encontrar os melhores valores de hiperparâmetros apenas para este modelo.

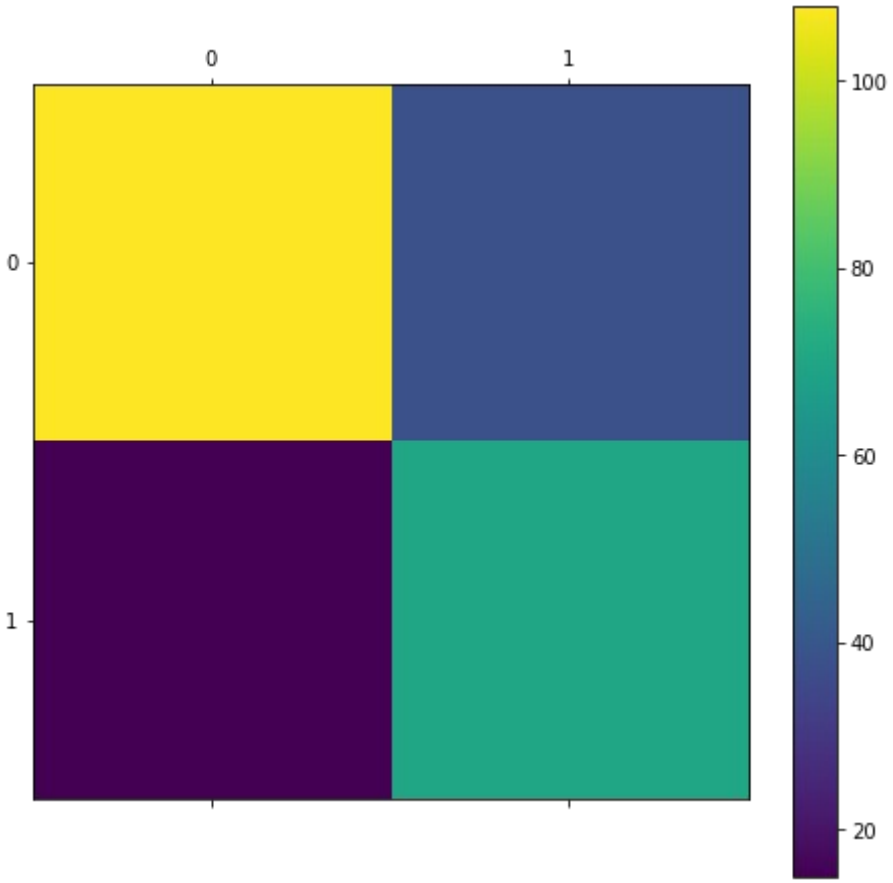
RESULTADOS

REGRESSÃO LOGÍSTICA

Métricas de avaliação obtidas com o modelo de Regressão Logística:

	precision	recall	f1-score	support
0	0.88	0.74	0.80	146
1	0.65	0.82	0.73	85
accuracy			0.77	231
macro avg	0.76	0.78	0.76	231
weighted avg	0.79	0.77	0.77	231

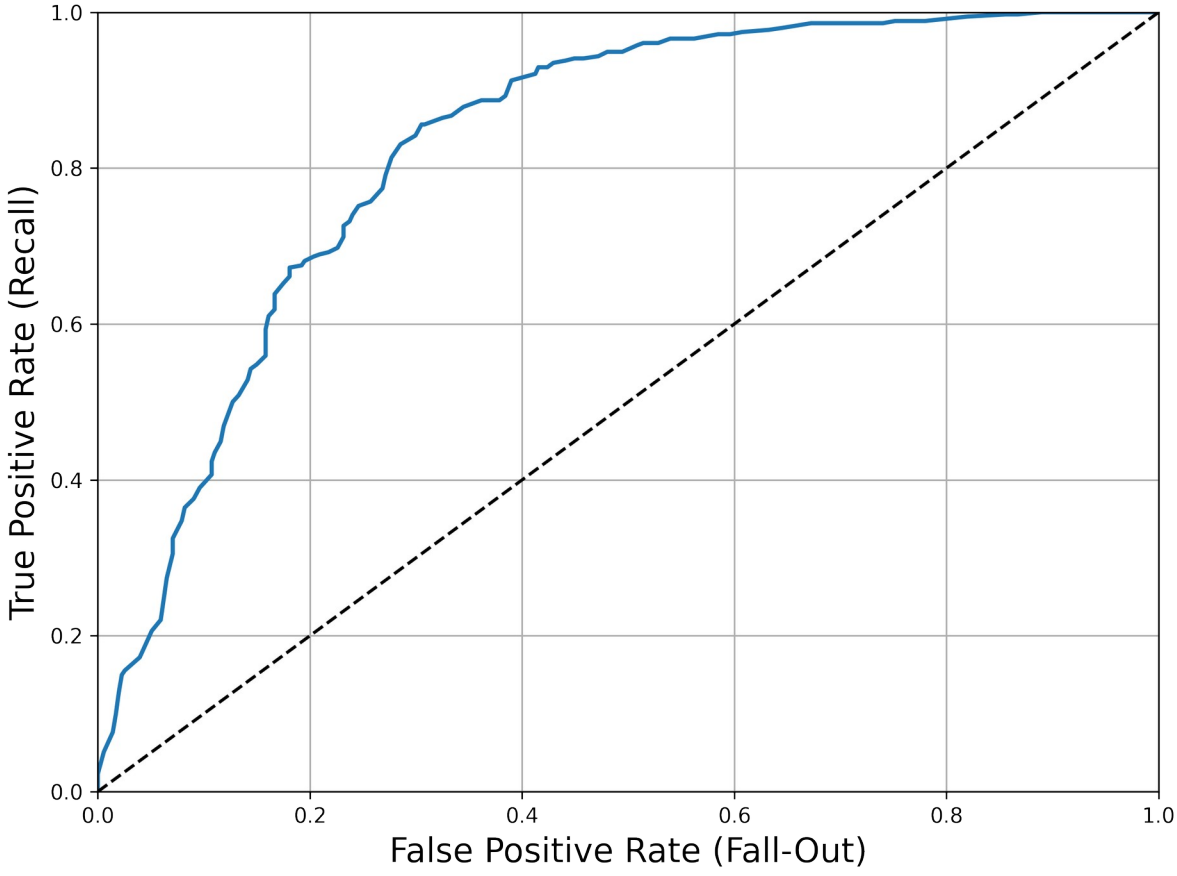
Matriz de confusão do modelo de regressão Logística:



Avaliando as métricas de avaliação e a matriz de confusão podemos notar que o modelo não se sai tão bem para classificação dos pacientes com diabetes, apresentando uma taxa de falso positivos elevada.

A Curva ROC nos ajuda a avaliar o Trade-off precisão/revocação, e uma maneira direta de comparar classificadores é comparar a área embaixo da Curva ROC, aonde, um classificador perfeito teria uma área igual a 1. A área obtida para o classificador de regressão logística foi igual a 0.829.

Curva ROC para o modelo de regressão Logística:

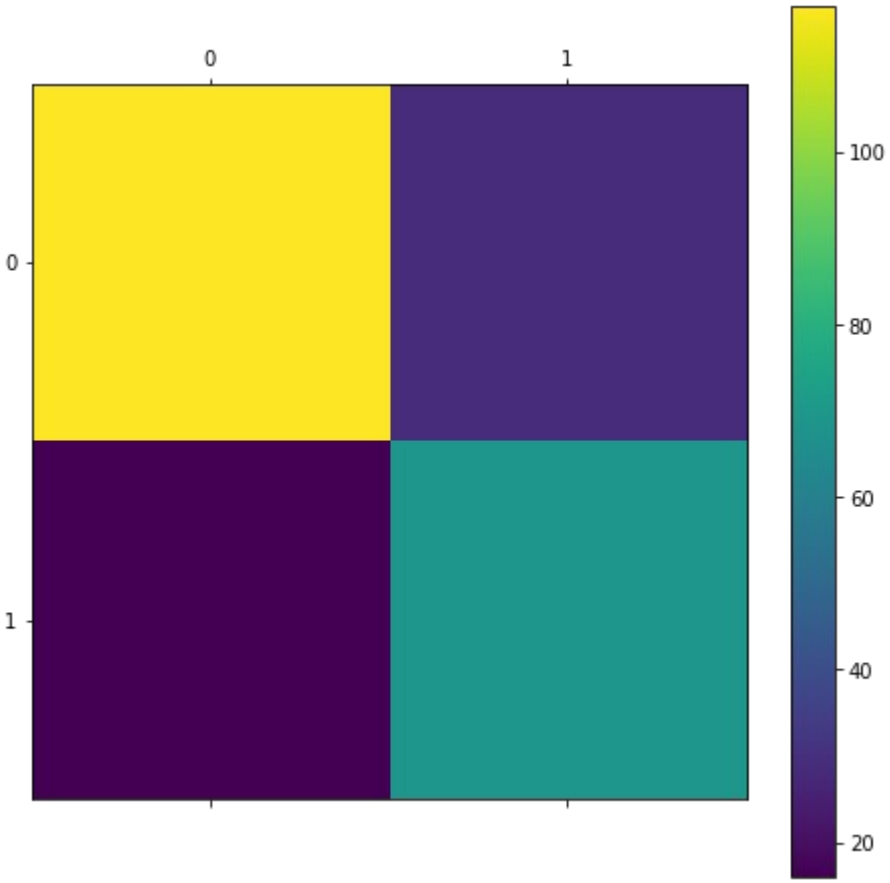


RANDOMFOREST

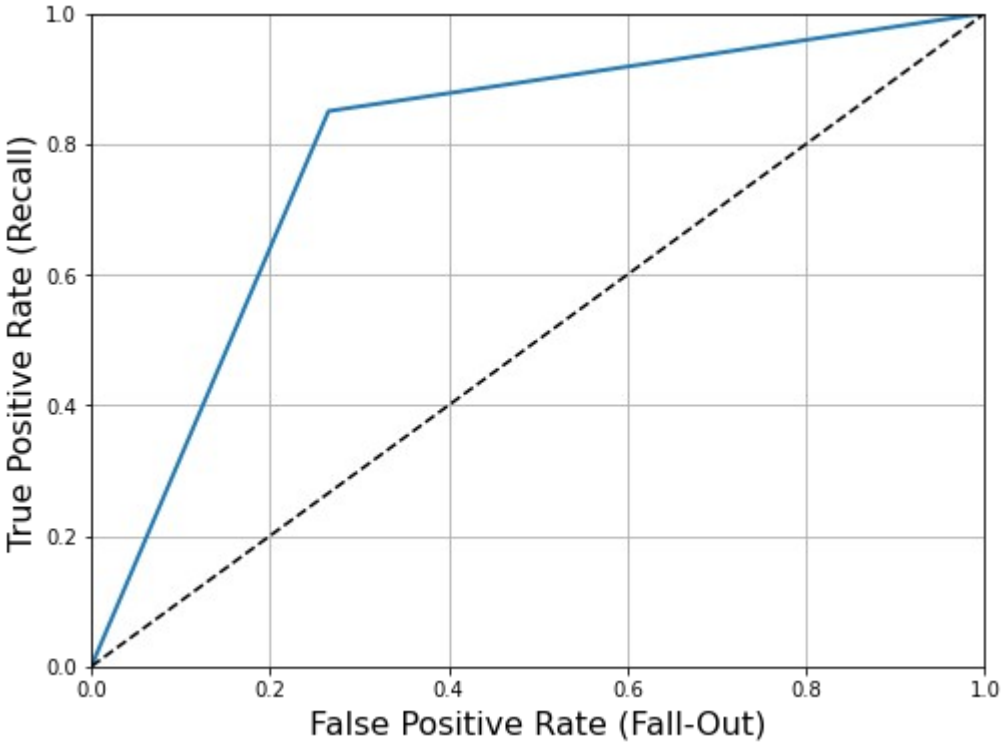
Métricas de avaliação obtidas com o modelo RandomForest

	precision	recall	f1-score	support
0	0.88	0.80	0.84	146
1	0.70	0.81	0.75	85
accuracy			0.81	231
macro avg	0.79	0.81	0.80	231
weighted avg	0.82	0.81	0.81	231

Matriz de confusão do modelo RandomForest



Curva ROC para o modelo de RandomForest:



Área = 0.793

CONCLUSÕES

Após analise dos resultados é possível concluir que o modelo obtido com o algoritmo RandomForest apresentou resultados um pouco melhores do que o modelo de Regressão Logística, com precisão igual a 0.79 e f1-score igual a 0.81. Ambos os modelos gerados tiveram um desempenho pior para classificar corretamente a classe de pacientes com diabetes, com uma taxa mais elevada de falso positivos. A precisão de classificação para pacientes com diabetes não ultrapassou 0.7 em nenhum dos modelos.

Como se trata de um modelo de uso médico, por mais que a precisão média não tenha sido muito elevada, a baixa taxa de falso negativos indicada pelas matrizes de confusão mostram que seria seguro utilizar o modelo como forma de pré análise dos pacientes, mas não como forma de diagnóstico final.