Projeto de Ciência de Dados

Hector S. Pinheiro[1], **Jacques Wolbeck S. de Melo G. Amorim**[2]

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)
Caixa Postal 15.064 – 52.072-970 – Maceió – AL – Brazil

hsp@ic.ufal.br, jwsmga@ic.ufal.br

Abstract. *This study seeks to analyze the relationships of the parameters present in the data set, in order to determine whether these relationships can influence or not, in the occurrence of stroke cases and from that to apply supervised machine learning techniques, to create an efficient predictive model in the classification of cases of the disease.*

Resumo. *Este estudo busca analisar as relações dos parâmetros presentes no conjunto de dados, a fim de, determinar se essas relações podem influir ou não, na ocorrência de casos de AVC (acidente vascular cerebral) e partir disso aplicar técnicas de aprendizado de máquina supervisionado, para criar um modelo preditivo eficiente na classificação de casos da doença.*

1. Aplicação

Atraves de técnicas de visualização de dados procuramos entender as relações dos atributos com a ocorrencia de acidente vascular cerebral (stroke). Após a visualização submetemos os dados aos algoritmos de aprendizagem de máquina supervisionada: regressão logistica e árvore de decisão, com o objetivo de construir um modelo que prevê casos de AVC, avaliando as suas respectivas ácuracias e curva ROC.

2. Experimentos

2.1. Base de Dados

O conjunto de dados para este estudo foi extraído dos repositórios de dados Kaggle, ele é usado com intuito de prever se um paciente tem probabilidade de desenvolver AVC. Os dados estão distribuídos em 12 colunas, sendo estas: id, gender, age, hypertension, heart disease, ever married, work type, residence type, avg glucoselevel, bmi, smoking status, stroke.

- A coluna de “id” foi descartada, pois não houve necessidade do seu uso.
- Na coluna “bmi” havia 281 dados faltantes, sendo estes substituídos pela média dos valores deste atributo.
- Após a análise dos gráficos foi constatado que o atributo ”residence type” pouco influenciava na ocorrência de ”stroke”, sendo também removido.
- Para facilitar o uso dos algoritmos de aprendizagem, os dados das colunas ”work type”, ”gender”, ”smoking status” e ”ever married” foram transformados de categóricos para numéricos.
- Foi realizado a padronização dos dados de treino e teste, pois eles estavam com diferentes escalas. Isso tem como objetivo fazer os dados se parecerem mais normalmente distriuídos, o que melhora o desempenho dos algoritmos de aprendizagem.

2.2. Estatística descritiva e inferência

Foram utilizados gráficos de dispersão e histogramas para analisar a distribuição dos dados no conjunto e extrair alguma informação relevante, um gráfico com mapa de calor também foi utilizado, com o próposito de verificar a correlação das variáveis e determinar se há uma multicolinearidade forte entre elas.

2.3. Métodos avaliados

Os algoritmos de aprendizagem supervisonada (classificação) utilizados foram a regressão logística e árvore de decisão. A regressão logística é uma técnica que permite estimar a probabilidade associada à ocorrênica de uma determinado evento, ela é indicada para problemas em que a variável dependente(stroke) é de característica binária e as variáveis independentes podem ser categóricas ou numéricas, já a árvore de decisão foi utilizada apenas como exemplo de outro algoritmo atuando neste problema, ela também pode ser utilizada para a previsão de dados.

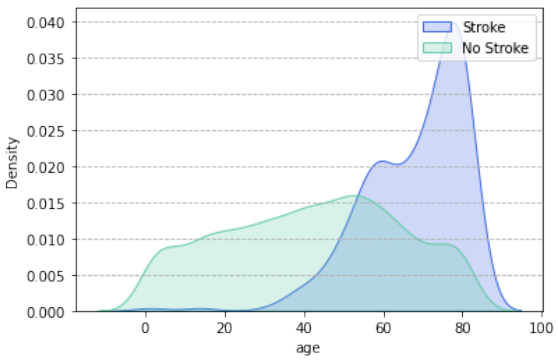
2.4. Métricas de avaliação

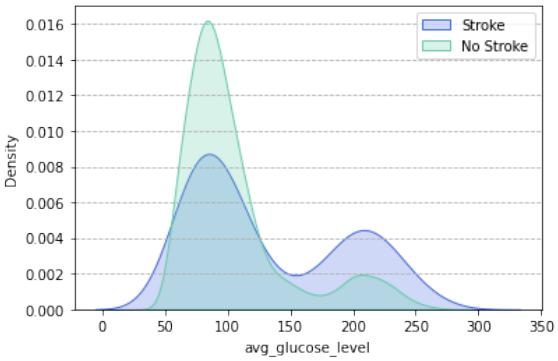
Acurácia e a curva ROC foram as métricas de avaliação aplicadas nesse projeto, também foi criada uma matriz de confusão para uma melhor visualização dos acertos e erros, ela está representada da maneira [[TP,FN][FP,TN]]. A acurácia diz quantos dados foram classificados de forma correta, independente da classe (stroke: 0 ou 1), já a curva ROC é mostrada na forma de um gráfico, ela é construida medindo a taxa de FP(falso positivo) e a taxa de TP(verdadeiro positivo), quanto mais proxima a curva estiver do canto superior esquerdo, melhor será a predição do modelo utilizado.

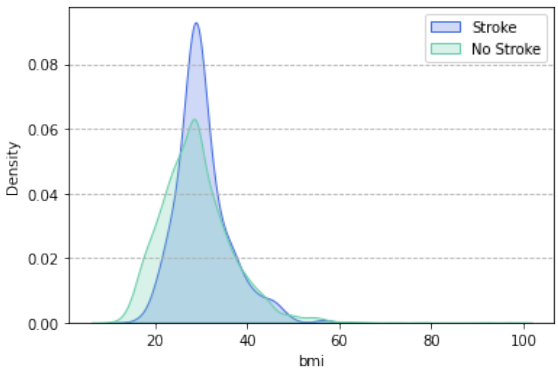
2.5. Métodos de avaliação

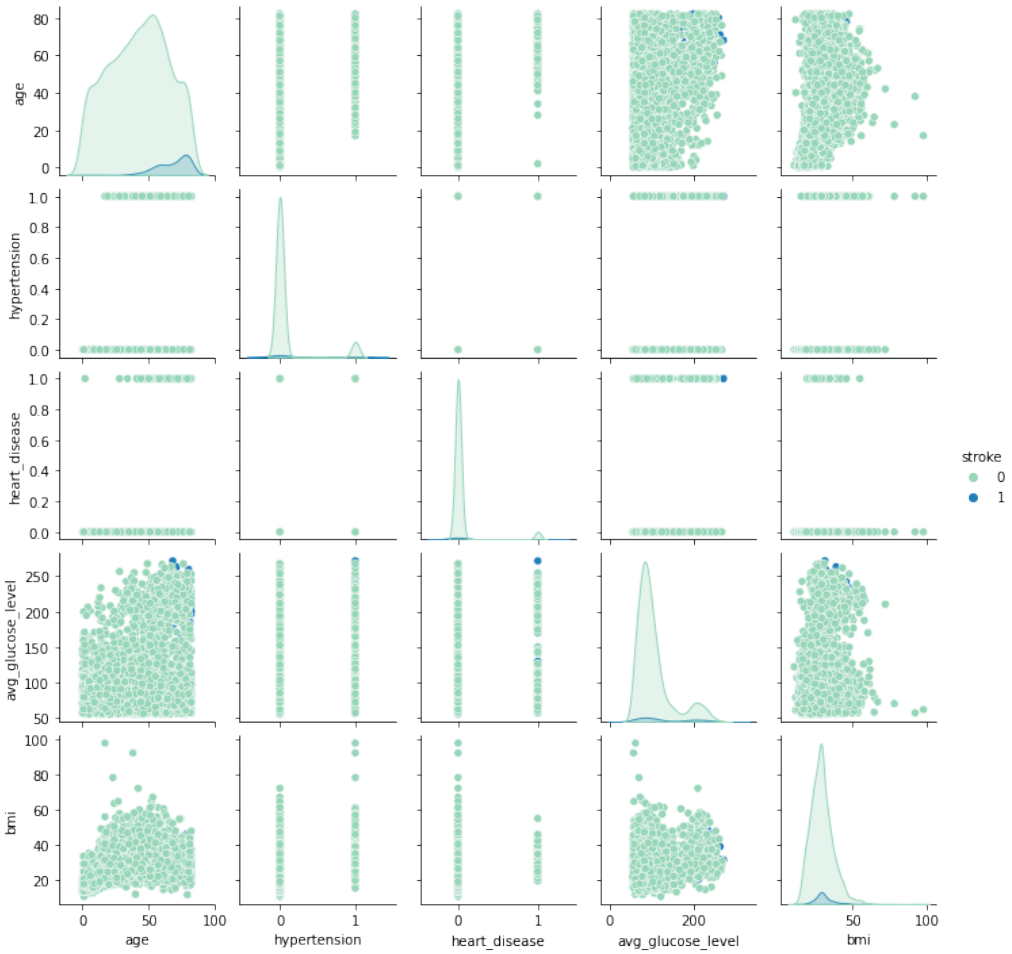
Para o método de avaliação foi utilizado o Hold-out, esse método divide o conjunto de dados de forma aleatória em uma base de treino e outra de teste, o tamanho da base de treino é sempre maior. No projeto foi utilizado 75% para treino e 25% para teste.

3. Resultados

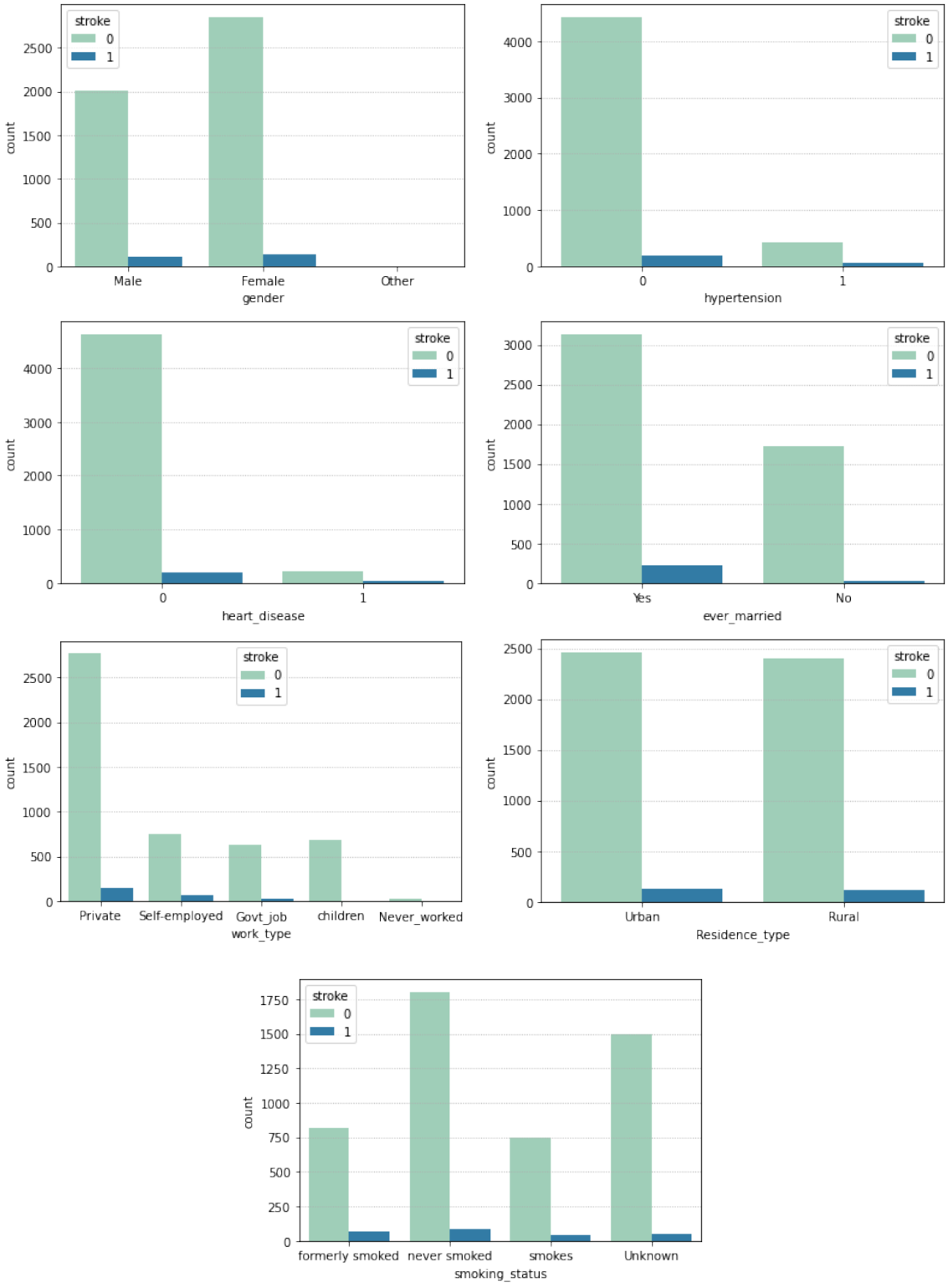




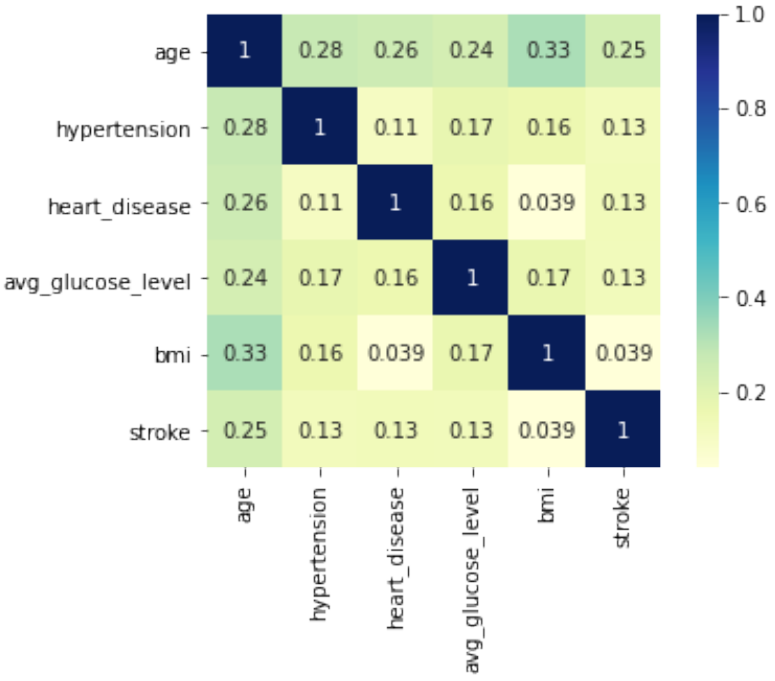




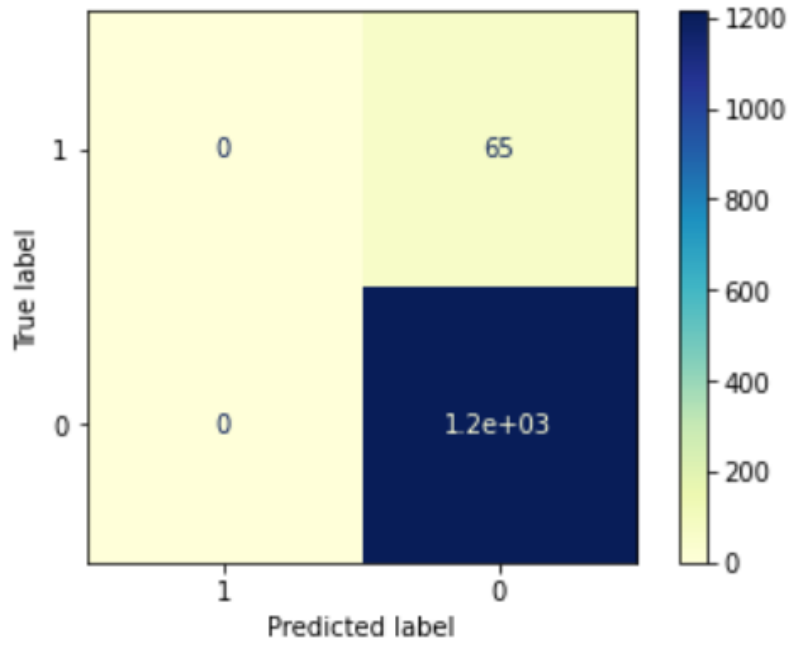
- A maioria dos casos de AVC está presente em pessoas acima dos 40 anos.
- Os dados parecem estar desbalanceados (apenas alguns pontos onde stroke = 1).
- Existem poucos outliers nos gráficos BMI/age e BMI/avg_glucose_level, por esse motivo não há necessidade de remove-los.
- O gráfico BMI/avg_glucose_level mostra que pessoas com menos de 150 avg_glucose_levels são menos propensas a AVC do que pessoas com mais de 150.

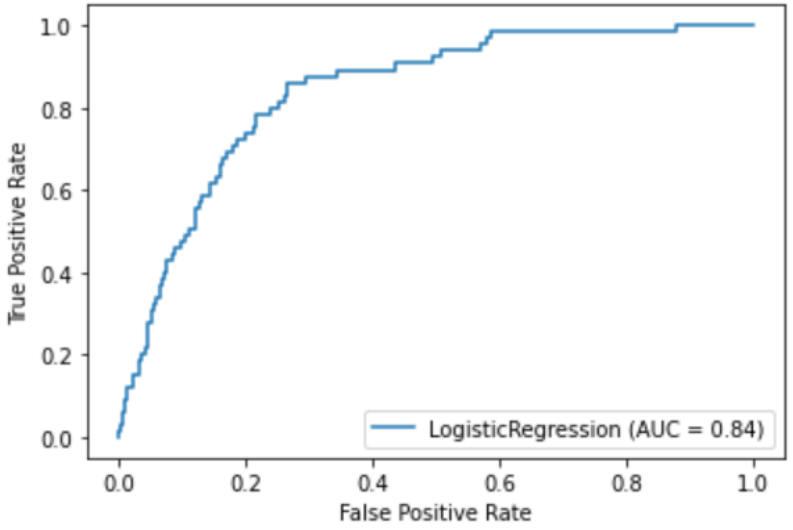


- O número de homens e mulheres que tiveram AVC é bem próximo.
- Há pessoas com algum tipo de doença cardíaca que também tiveram AVC.
- Pessoas casadas demonstram ser mais suscetíveis a ter AVC do que as solteiras.
- Pessoas que não são hipertensas também possuem risco de ter AVC.
- Funcionários privados aparentam ter mais casos de AVC do que pessoas com outro tipo de emprego. As crianças praticamente não tem casos de AVC.
- Há um diferença muito baixa entre as pessoas que vivem em áreas urbanas e rurais que tiveram AVC, essa coluna pode ser descartada.
- Pessoas que fumam ou já fumaram demonstram ser mais vulneráveis a ter AVC, pois mesmo com uma amostra total menor do que a amostra das pessoas que nunca fumaram, os casos da doença são bem próximos.



- A maior correlação está presente entre as variáveis age e bmi, porém é muito baixa, o que descarta as chances de multicolinearidade ou dados redundantes.

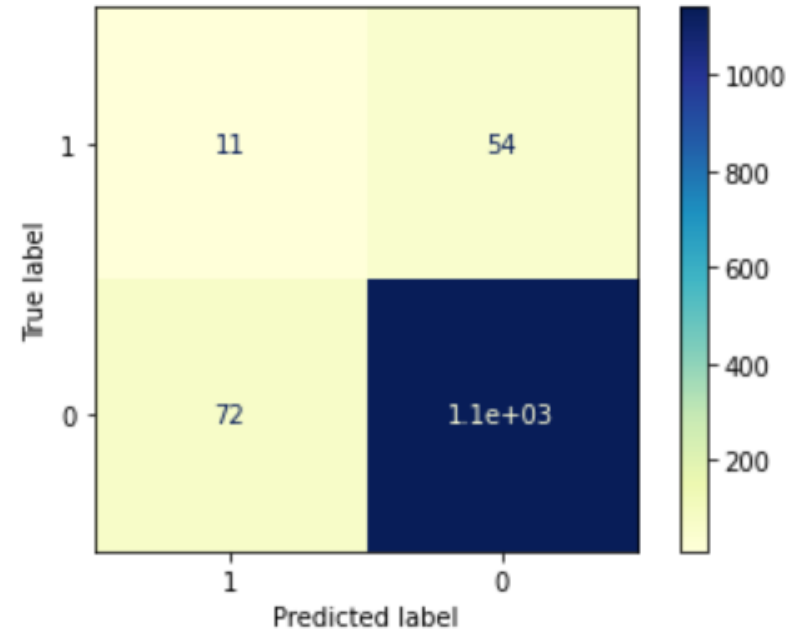


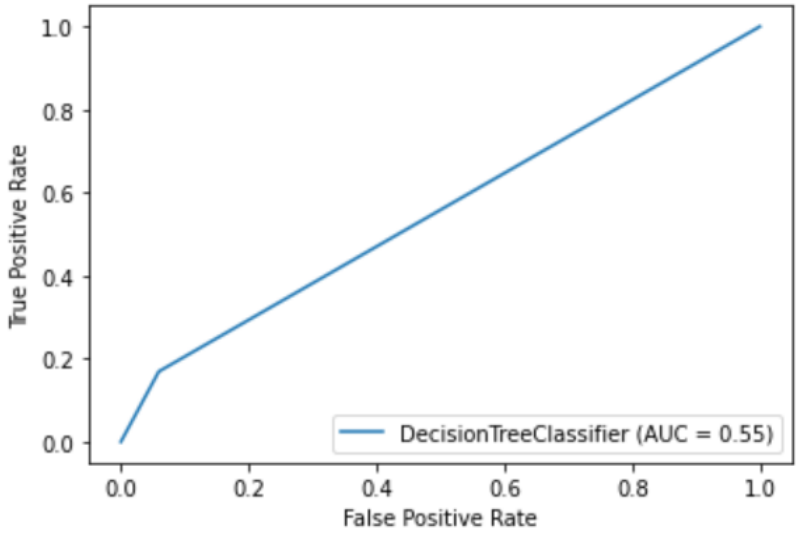


- A acurácia para o modelo preditivo da regressão logistíca foi de aproximadamente 95%.

- Na matriz de confusão tivemos: TP = 0, FN = 65, FP = 0 e TN = 1213.

- Já na curva ROC, tivemos uma área sob a curva de 84%.





- A acurácia para o modelo preditivo da árvore de decisão foi de aproximadamente 95%.
- Na matriz de confusão tivemos: TP = 11, FN = 54, FP = 72 e TN = 1141.
- Já na curva ROC, tivemos uma área sob a curva de 55%.

4. Conclusão

O objetivo desse estudo foi criar um modelo preditivo com uma boa eficiência na classificação de casos de AVC. Após análise dos resultados obtidos, foi constado que os valores de acurácia foram muito elevados, isso se deve ao fato do grande desbalanceamento do conjunto de dados, onde grande parte dos valores de ”stroke” eram iguais a 0. Por isso o resultado obtido com a curva ROC e o valor de AUC passam mais confiança na avaliação do modelo, já que esse tipo de métrica funciona melhor com dados desbalanceados que a acurácia. Portanto o algoritmo de regressão logistíca se saiu melhor neste tipo de problema de classificação, com AUC = 84%.

5. Referências

MEDIUM. **Métricas de Avaliação em Machine Learning: Classificação**. Disponível em: https://medium.com/kunumi/métricas-de-avaliação-em-machine-learning-classificação-49340dcdb198. Acesso em: 14 mai. 2021.

TOWARDSDATASCIENCE. **Understanding the Confusion Matrix from Scikit learn**. Disponível em: https://towardsdatascience.com/understanding-the-confusion-matrix-from-scikit-learn-c51d88929c79. Acesso em: 14 mai. 2021.

EDICIPLINAS. **Regressão Logística**. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf. Acesso em: 11 mai. 2021.

LEG. **Métodos de Reamostragem**. Disponível em :http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html. Acesso em: 14 mai. 2021.

CIÊNCIA DE DADOS. **Documentos**. Disponível em: https://sites.google.com/ic.ufal.br/ciencia-de-dados/documentos?authuser=1. Acesso em: 11 mai. 2021.

Link do dataset utilizado e GitHub do projeto:

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
https://github.com/hecpinheiro/Projeto-Ciencia-de-Dados