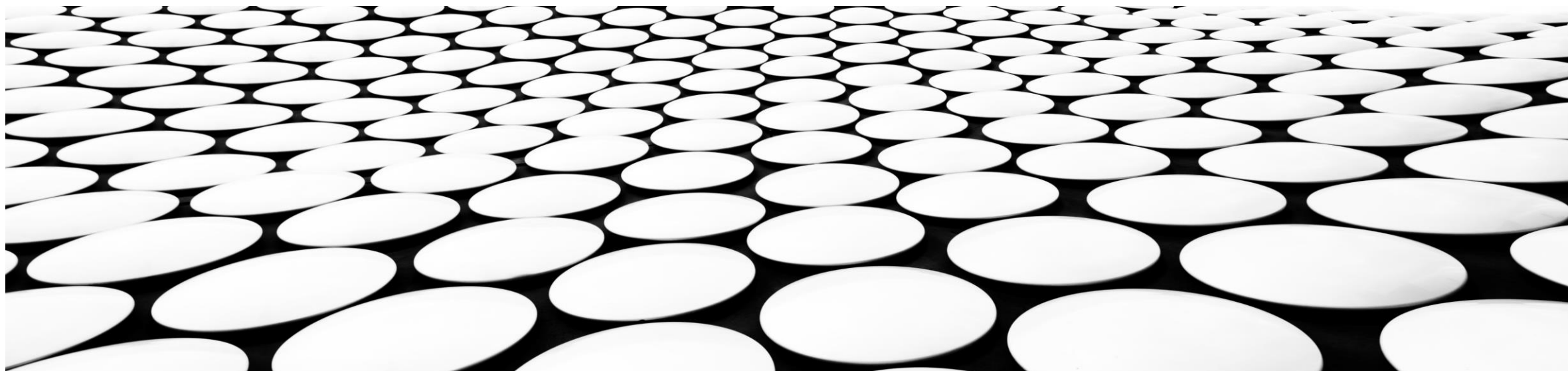


# UMA ABORDAGEM PREDITIVA PARA POTENCIAIS CASOS DE ACIDENTE VASCULAR CEREBRAL (AVC)

A PREDICTIVE APPROACH TO POTENTIAL CASES OF STROKE



Amanda C. Amorim (234942 - FCM)  
Marcia Jacobina A. Martins (225269 - IC)  
Regivaldo S. Ferreira (225153 - FEEC)

## RESUMO DO PROJETO

- O presente projeto intitulado **“Uma Abordagem Preditiva para Potenciais Casos de Acidente Vascular Cerebral (AVC)”** foi desenvolvido a partir da premissa de que o Acidente Vascular Cerebral, ou derrame cerebral é a segunda maior causa de morte e a primeira de incapacidade no Brasil.
- Causa seis milhões de óbitos por ano, deste total, 50% das pessoas ficam dependentes de outra para atividades do dia a dia e 70% não conseguem retornar mais ao trabalho.
- Suas causas são complexas e suas chances de acometimento podem ser aumentadas quando o indivíduo possui um ou mais fatores de risco relacionados à doença como: alta taxa de colesterol e triglicérides e doenças cardiovasculares, como hipertensão arterial e arritmias cardíacas, além de hábitos de vida prejudiciais à saúde como sedentarismo, tabagismo e etilismo.



# OBJETIVOS

- Identificar padrões nos dados dos pacientes para efetuar uma predição das pessoas que possuem risco de ter AVC durante a sua vida
- Buscar o melhor modelo estatístico para identificar a predisposição do paciente ao AVC, baseado na análise de determinadas características

# MOTIVAÇÃO





Efetuar o Diagnóstico precoce do ACIDENTE VASCULAR CEREBRAL (AVC) pode salvar muitas vidas e evitar que pessoas fiquem incapacitadas e dependentes

Detectar previamente o AVC é um desafio que tem mobilizado a ciência, já que as consequências da doença geram um enorme impacto econômico e social

## REFERENCIAIS TEÓRICAS

- O trabalho de (SITAR-TĂUT, et al., 2009) sugere a utilização de ferramentas de mineração de dados na área médica e busca prever de forma não invasiva doenças cardiovasculares, considerando fatores de risco
- Também é possível observar, no trabalho realizado por (Dongmei, et al., 2019), a busca por classificadores com o objetivo de identificar precocemente indivíduos com alto risco de diabetes
- O artigo de (Fisher, et al., 2016) demonstra a importância de medidas preventivas do AVC, a partir da análise de dados de pacientes que tiveram AVC e que com tratamento precoce poderiam ter sido evitados ou minimizados

# PRINCIPAIS FERRAMENTAS UTILIZADAS

	Python: Construção Principal do Projeto Desenvolvido
	Orange: Exploração dos cenários para análises estatísticas
	R: Construção teste do modelo Random Forest
	Knime: Apresentação detalhada dos testes em modo gráfico

# DIFICULDADES DA PESQUISA

- Variedade de base de dados
- Complexidade na base de dados encontrada
- Realização do tratamento ideal na base de dados
- Escolha dos critérios para seleção dos modelos

# MUDANÇA DE PERCURSO – BASE DE DADOS

## Stroke Prediction Dataset

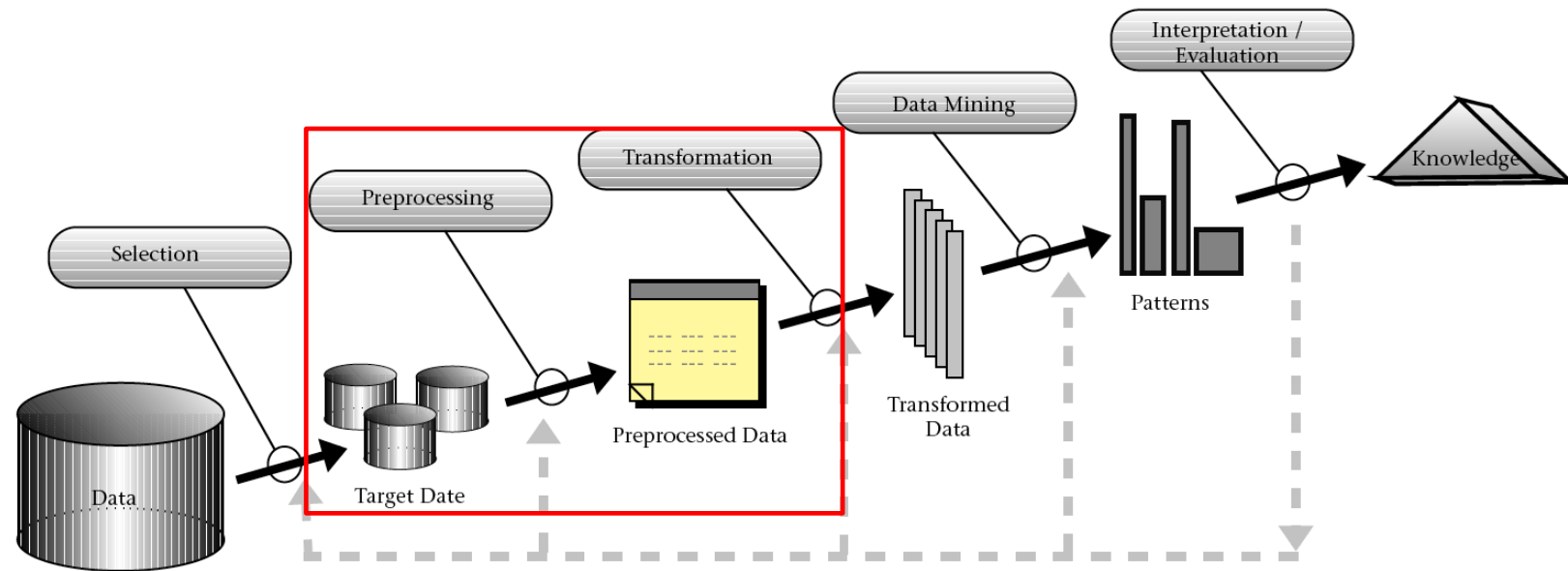
- Kaggle - Comunidade online de cientistas de dados e profissionais de aprendizado de máquina
- Base de dados de 5110 pacientes e 12 atributos usada para prever se um paciente tem probabilidade de desenvolver AVC com base nos parâmetros de entrada como sexo, idade, várias doenças e tabagismo. Cada linha dos dados fornece informações relevantes sobre o paciente.

## International Stroke Trial

- John Snow Labs - Empresa Privada Americana que trabalha com Inteligência Artificial voltada para a saúde
- Base do laboratório John Snow Labs dos EUA que contém dados de 10 mil pacientes de diversos países e 12 variáveis. Foi feito um teste de aspirina e AAS em pacientes com AVC e houve acompanhamento do paciente verificando se continuou viva ou se morreu após um período de 12 meses.



# ANALISE EXPLORATÓRIA DOS DADOS

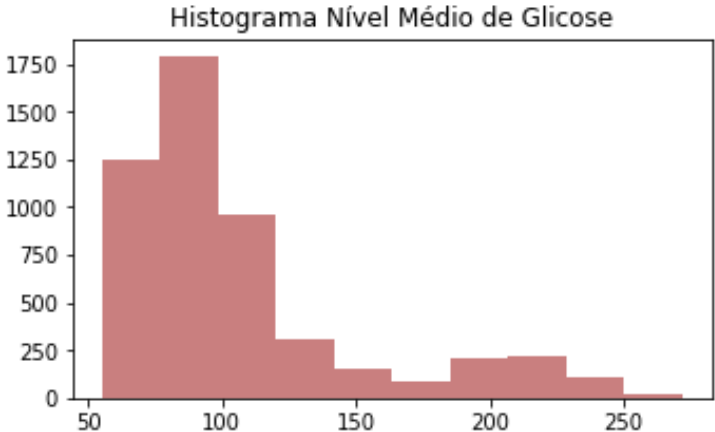
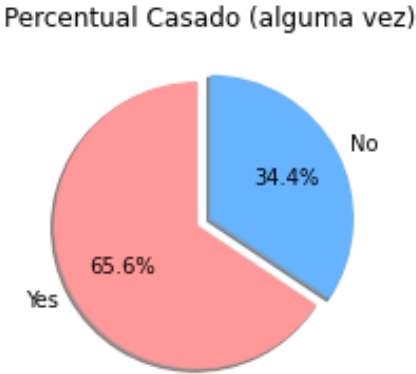
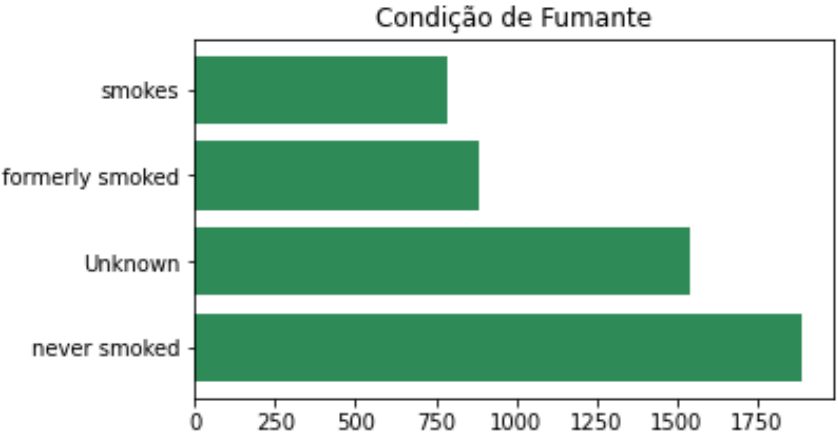
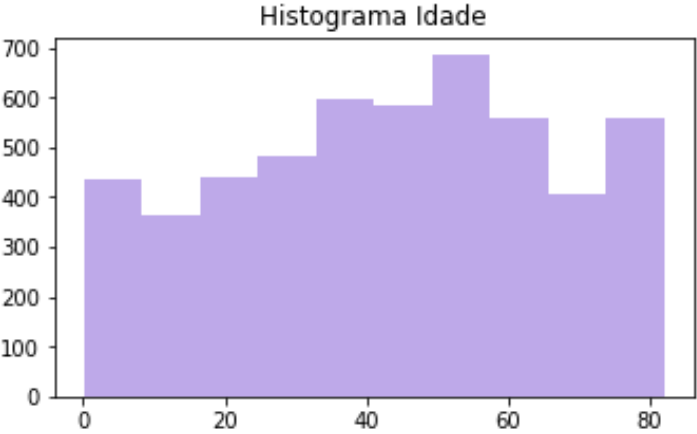
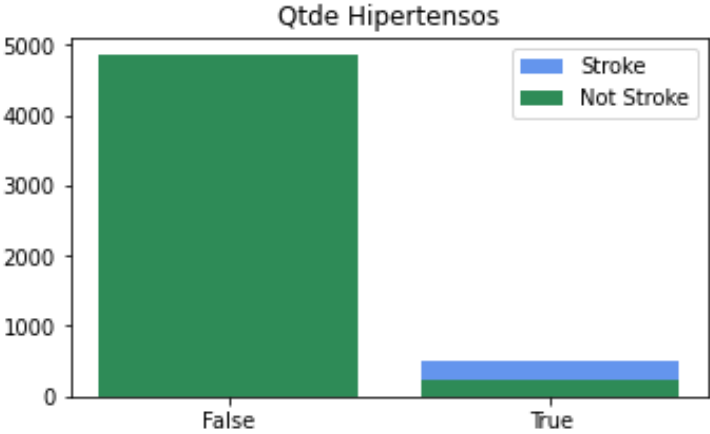
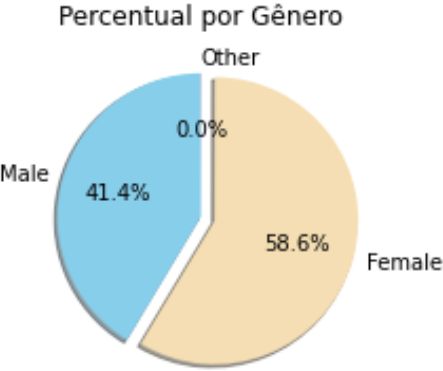


(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

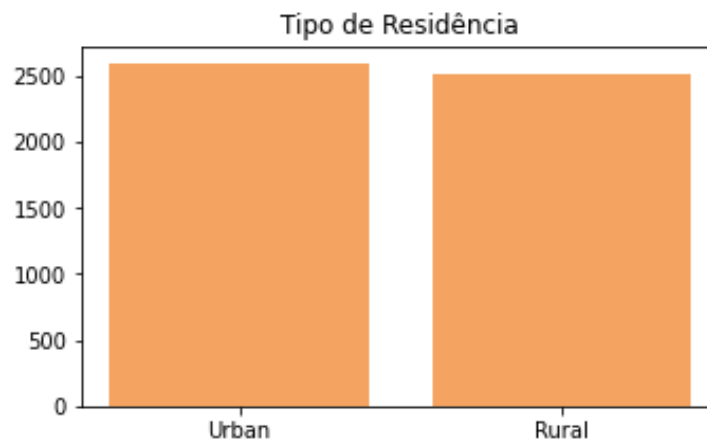
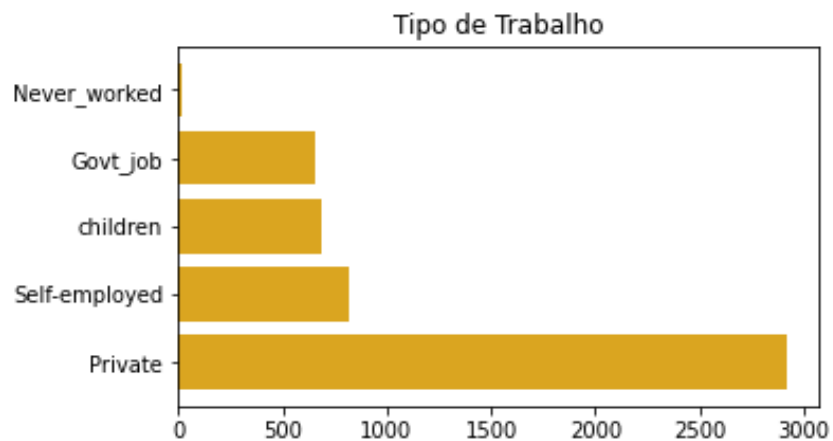
# BASE DE DADOS

- Base com 5110 pacientes e 12 atributos
- Usada para prever se um paciente tem probabilidade de desenvolver AVC com base nos parâmetros de entrada como sexo, idade, várias doenças e tabagismo
- A média da idade é em torno de 43 anos
- A maioria não é hipertenso e não tem doença do coração
- A média do nível de glicose fica em torno de 106
- O percentual do índice de massa corporal é de 28.8 e poucos pacientes tiveram AVC

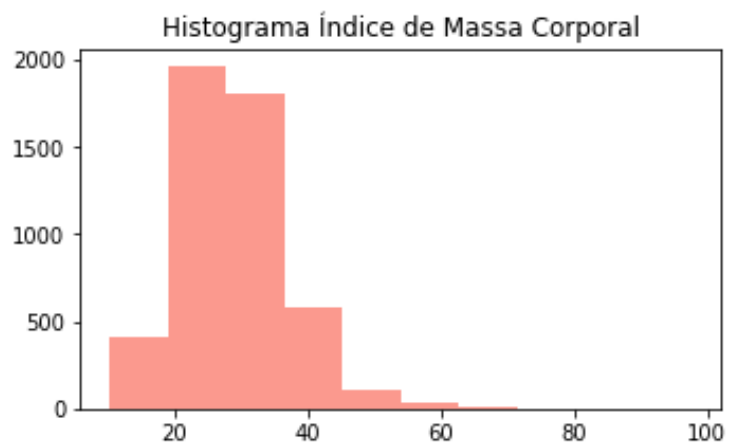
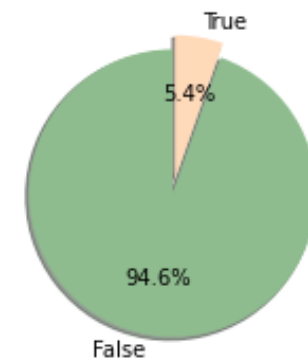
# ANÁLISE GRÁFICA



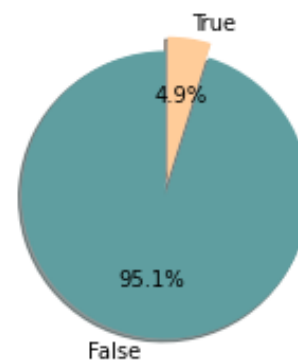
# ANÁLISE GRÁFICA



Percentual Doença do Coração

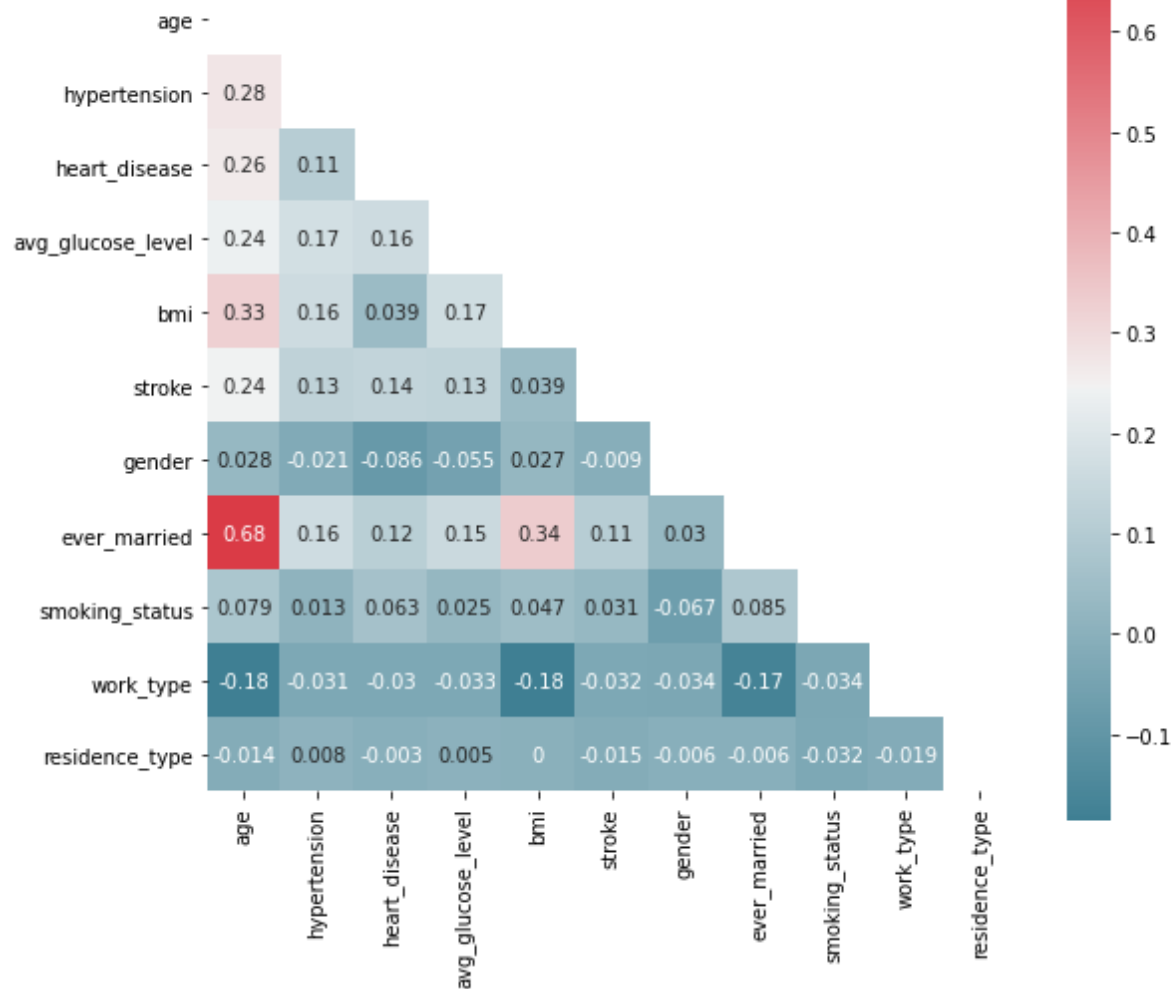


Percentual AVC

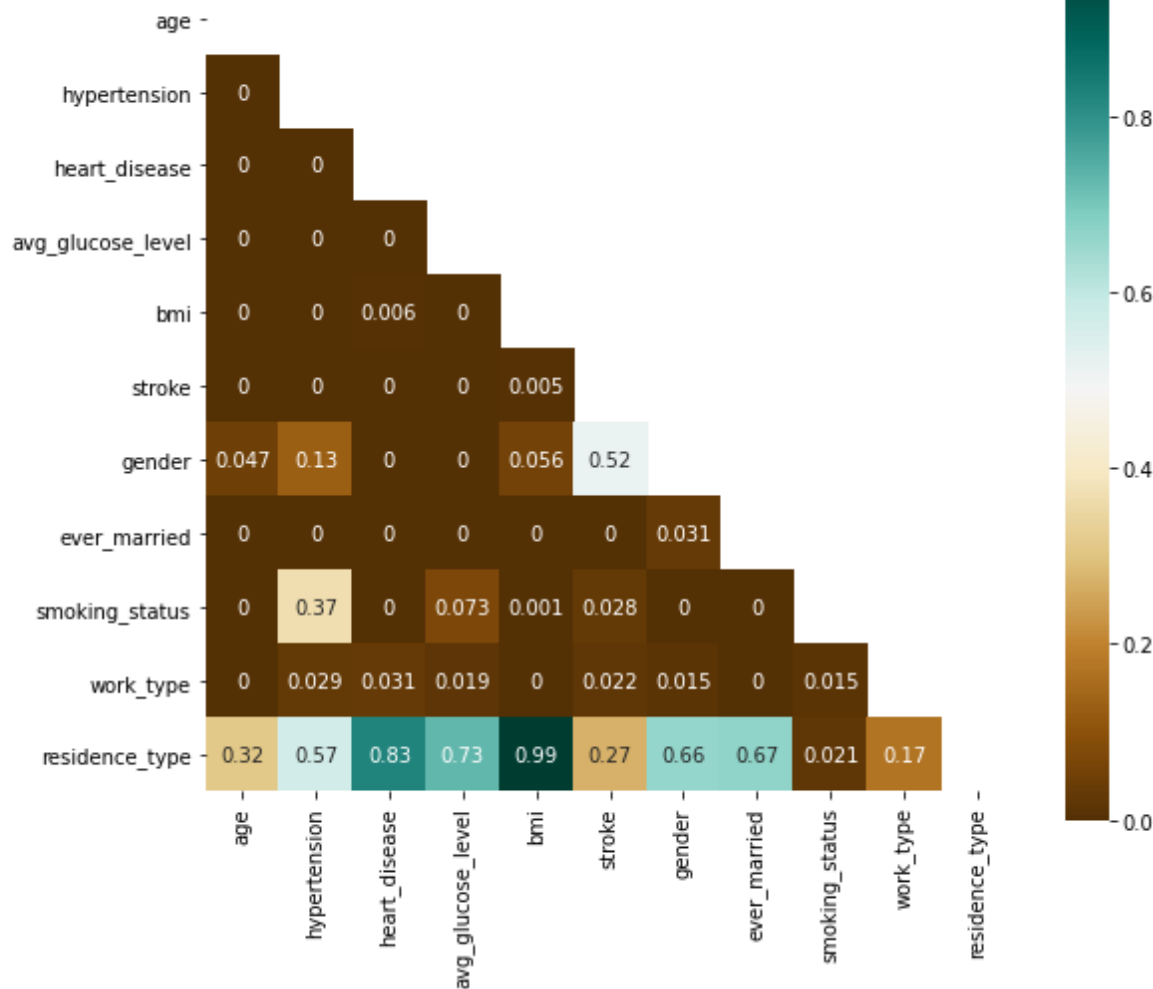


# ANÁLISE DA CORRELAÇÃO

Matriz Correlação PEARSON

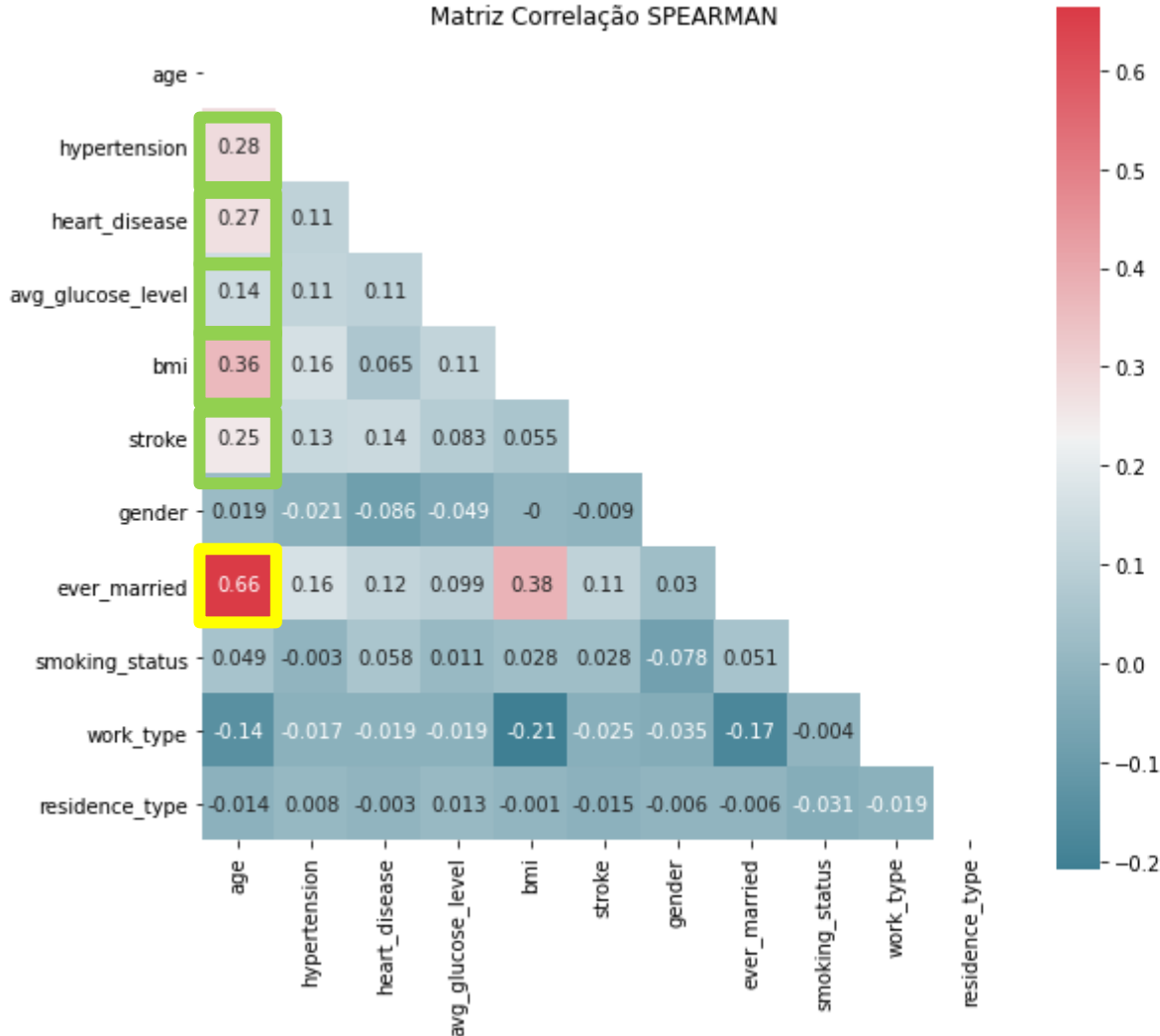


PValue Correlação PEARSON

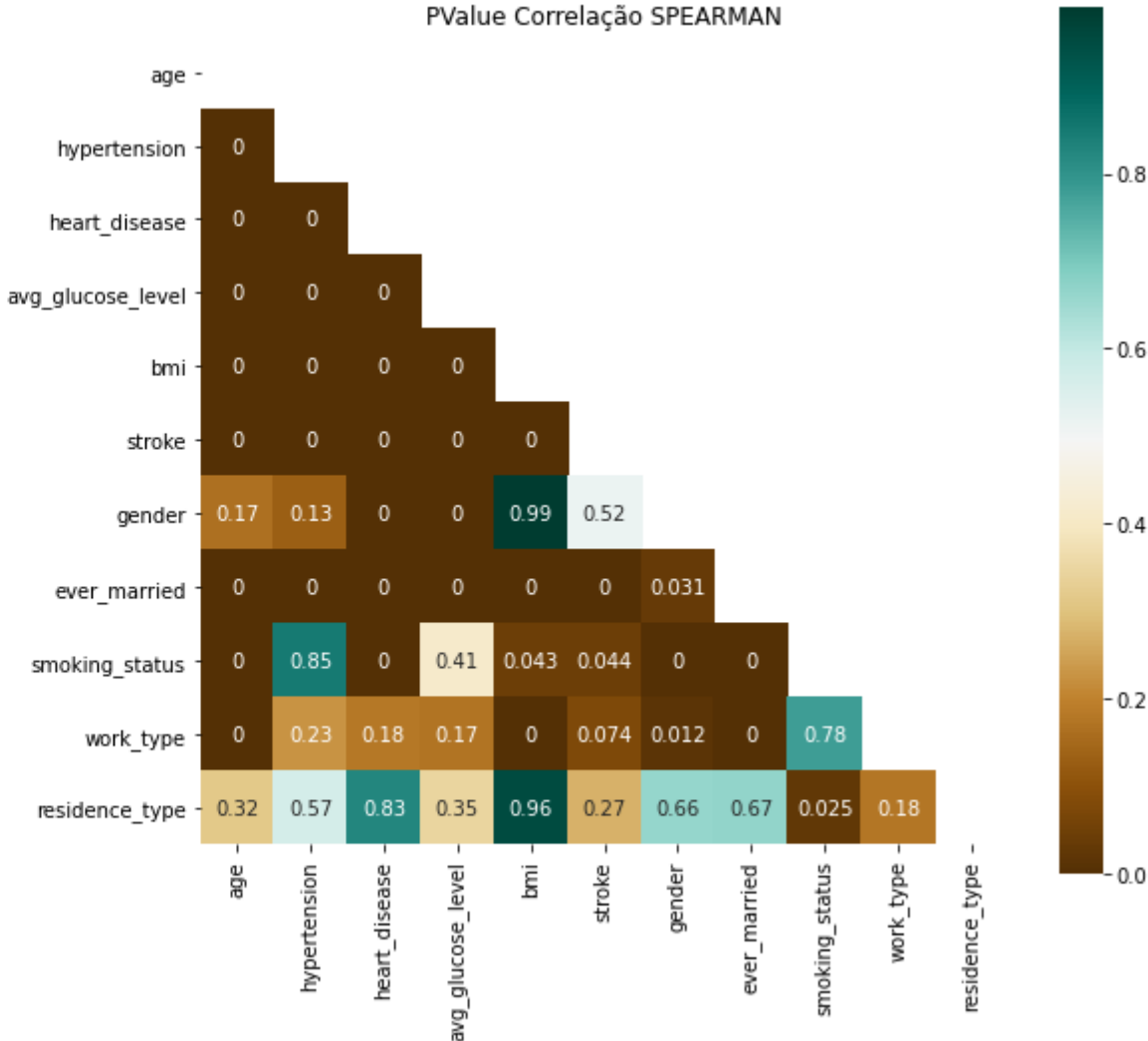


# ANÁLISE DA CORRELAÇÃO

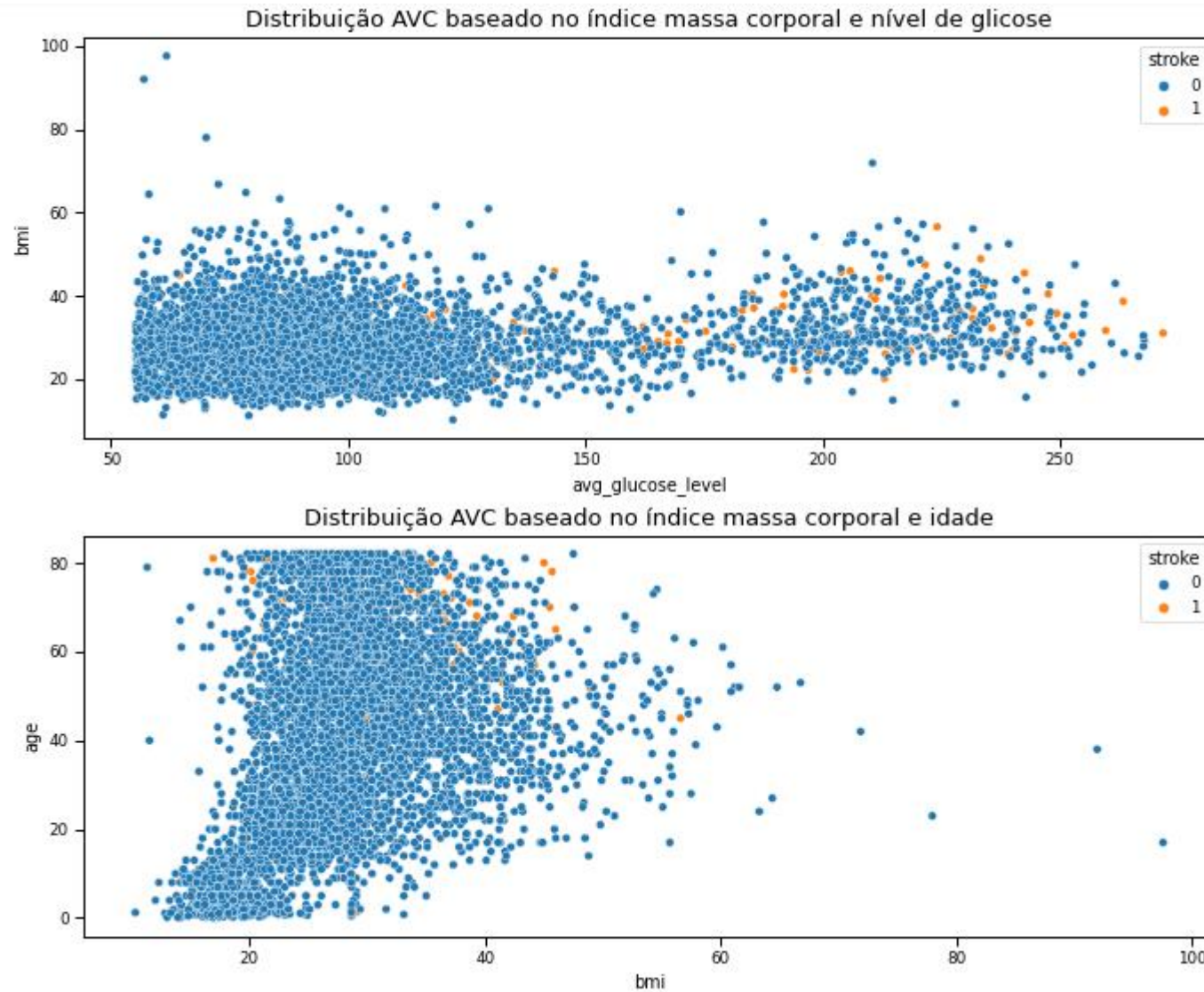
Matriz Correlação SPEARMAN



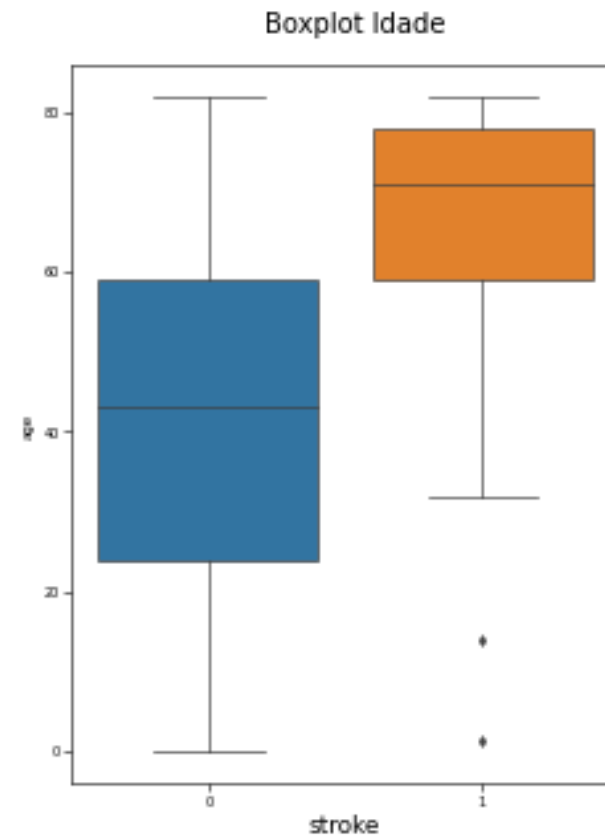
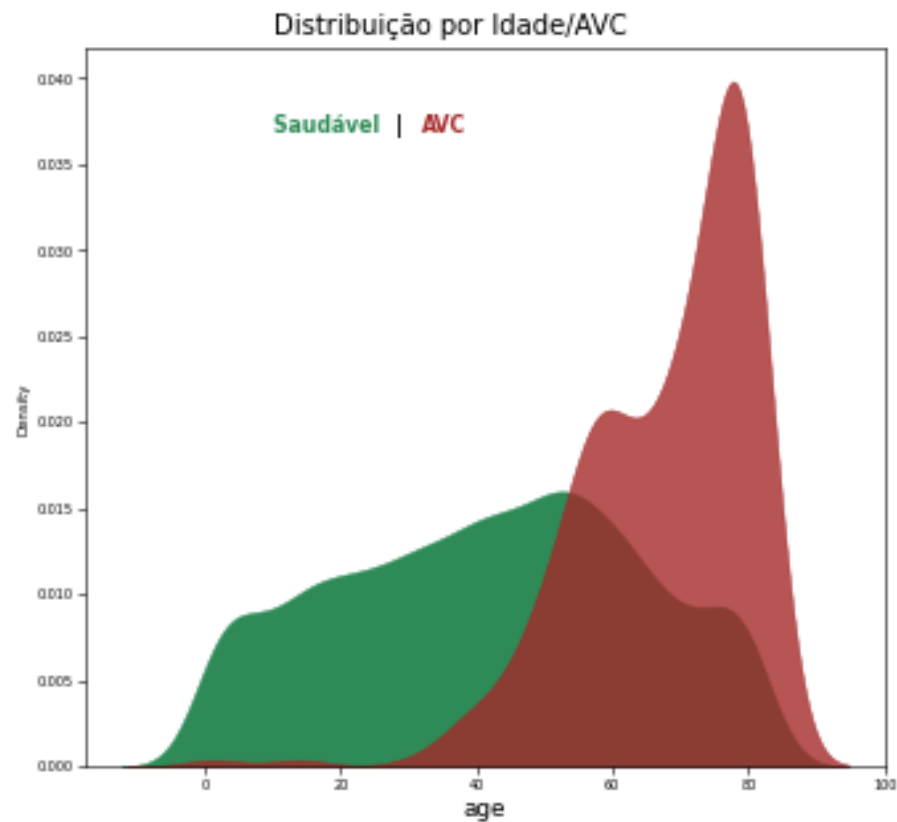
PValue Correlação SPEARMAN



# ANÁLISE DA CORRELAÇÃO



# ANÁLISE GRÁFICA PACIENTES COM PREDISPOSIÇÃO AO AVC

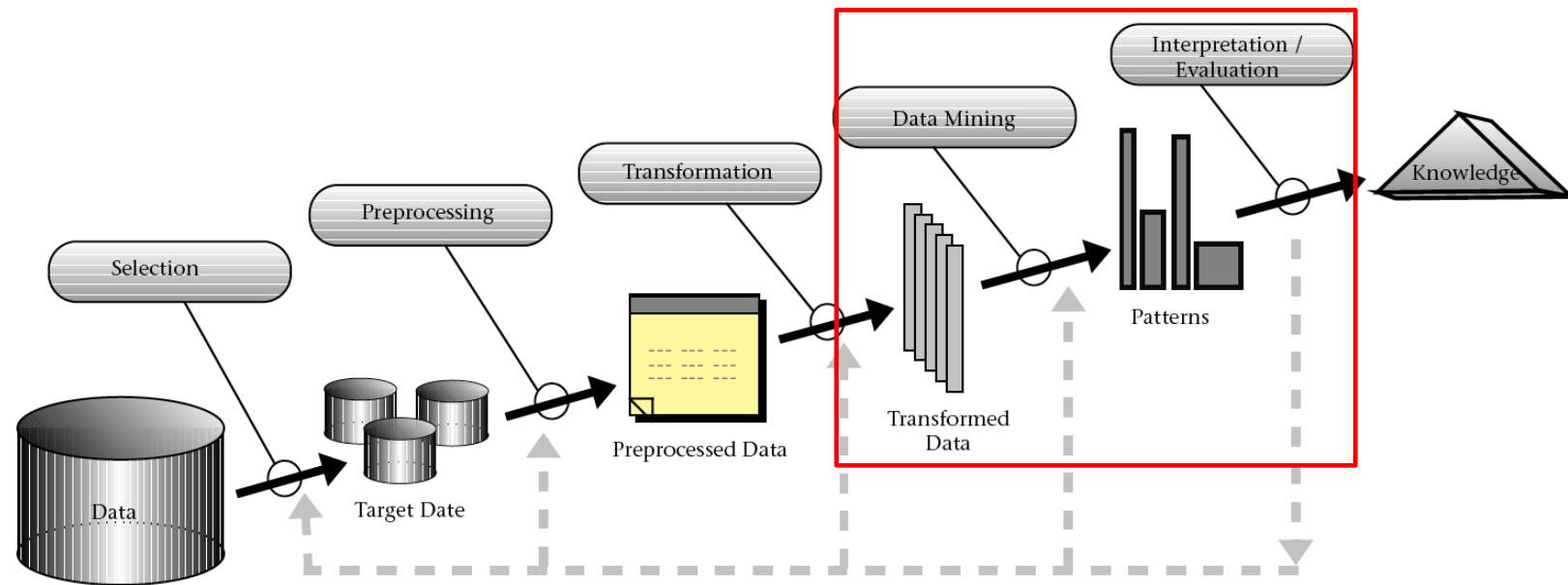




# CONCLUSÃO ANÁLISE EXPLORATÓRIA DE DADOS

- As features selecionadas como variáveis preditivas para os algoritmos foram: **idade** (age), **índice de massa corporal** (bmi), **nível médio de glicose no sangue** (avg\_glucose\_level) e **hipertensão** (hypertension)

# AVALIAÇÃO DOS MODELOS ESTATÍSTICOS



(Fayyad, Piatetsky-Shapiro, and Smyth 1996)

# ESTRATÉGIA

- Foram avaliados 5 modelos de aprendizado de máquina: **Regressão Logística, Support Vector Machine (SVM), Random Forest**, K-nearest neighbors e Naive Bayes
- As features adotadas: *age*, *bmi*, *hypertension*, *avg\_glucose\_level*
- Foram aplicadas 4 técnicas diferentes sobre os dados: **subdivisão em dados de teste e treino, validação cruzada com folds estratificados, amostragem de tamanhos iguais e sobreamostragem e busca dos melhores hiperparâmetros**
- Foi feita uma comparação entre os modelos baseada na matriz de confusão, nas métricas de avaliação, curva ROC e Precision-Recall

# TÉCNICA AVALIAÇÃO: SUBDIVISÃO EM DADOS DE TESTE E TREINO

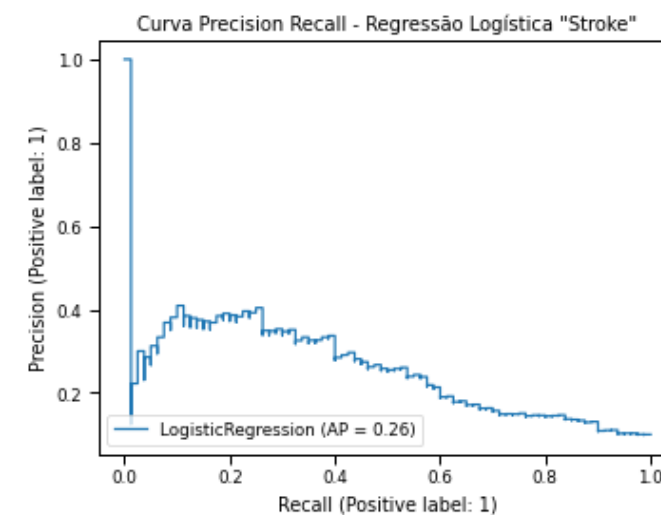
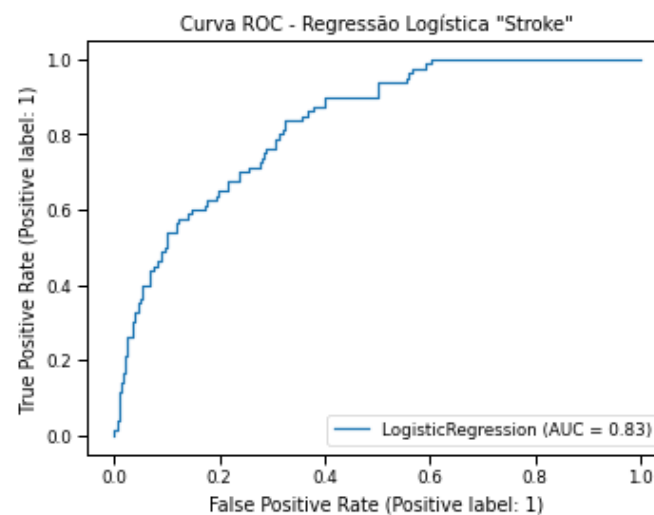
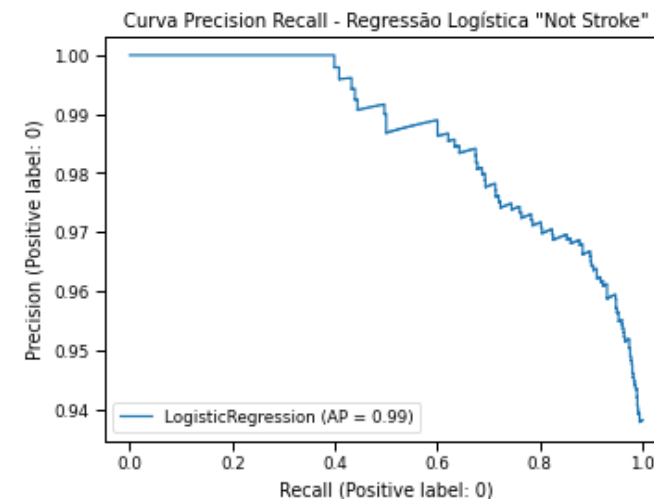
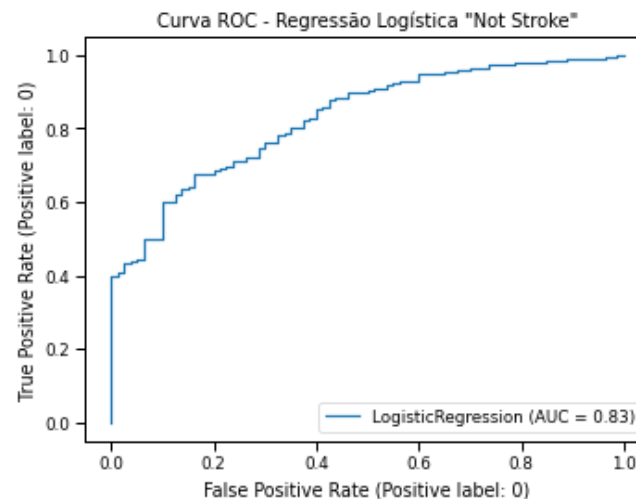
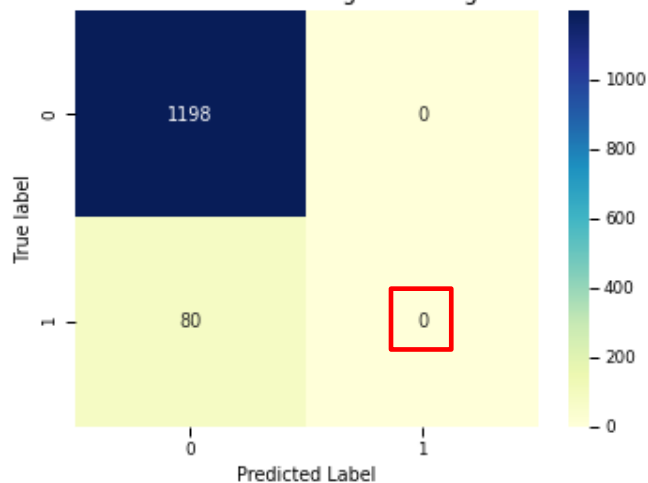
- O **train-test split** é uma técnica para avaliar a performance de um algoritmo de aprendizado de máquina que divide os dados em teste e treino
- Em função da correlação “fraca” existente entre algumas variáveis preditivas identificadas, foi efetuado um teste gradativo para validar se deveriam ser mantidas ou não
- Na matriz de confusão de todos os modelos os verdadeiros negativos são baixos ou zero (baixa especificidade), embora a sensibilidade seja 1
- Acurácia em torno de 93% sugere-se a possibilidade de **overfitting** ou base de dados desbalanceada
- A curva ROC de todos os modelos evidencia o resultado encontrado entre a especificidade e sensibilidade, o equilíbrio entre a sensibilidade e especificidade pode ser melhorado (AUC entre 0,68 e 0,84)

# RESULTADO COM REGRESSÃO LOGÍSTICA (TRAIN\_TEST\_SPLIT)

## Métricas do Modelo

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1198
1	0.00	0.00	0.00	80
accuracy			0.94	1278
macro avg	0.47	0.50	0.48	1278
weighted avg	0.88	0.94	0.91	1278
Accuracy Score:	0.9374021909233177			

Matriz de Confusão - Regressão Logística



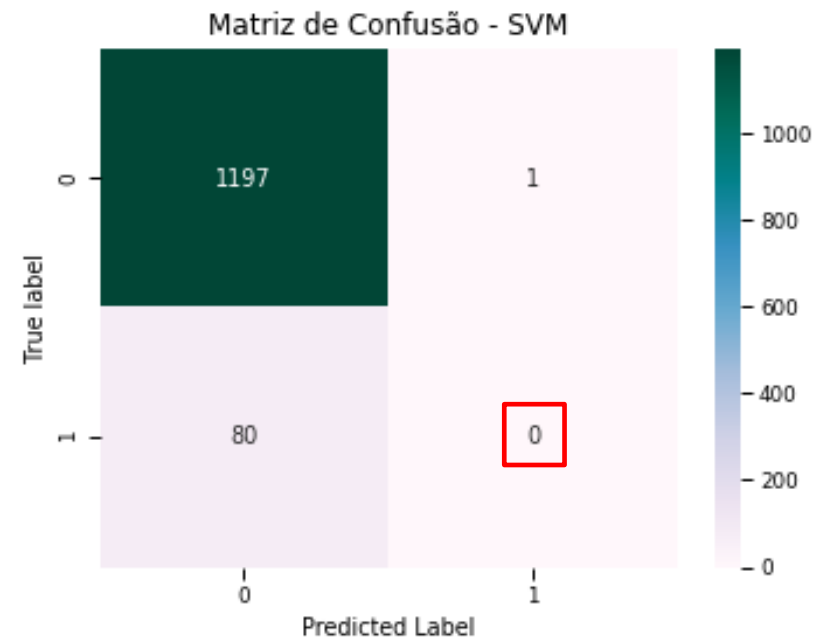
# TÉCNICA AVALIAÇÃO: VALIDAÇÃO CRUZADA E FOLDS ESTRATIFICADOS

- O **StratifiedKFold cross validation** é uma extensão da validação cruzada KFold e especificamente utilizada para problemas de classificação. Separa o conjunto de dados em dados de treino e de teste, subdividindo em folds estratificados. Para o caso em questão onde temos um grande desbalanceamento da base de dados essa técnica mostra-se bastante adequada
- Em todos os modelos, na matriz de confusão, os verdadeiros negativos são baixos ou zero (baixa especificidade), embora a sensibilidade tenha sido 1
- Acurácia em torno de 93% em todos os modelos e é sugestiva de **overfitting** ou base de dados desbalanceada

# RESULTADO COM SVM (CROSS VALIDATION)

## Métricas do Modelo

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1198
1	0.00	0.00	0.00	80
accuracy			0.94	1278
macro avg	0.47	0.50	0.48	1278
weighted avg	0.88	0.94	0.91	1278
Accuracy Score:	0.9366197183098591			



# TÉCNICA AVALIAÇÃO ® : AMOSTRAGEM DE TAMANHOS IGUAIS

- O **equal\_size\_sampling**, remove linhas do conjunto de dados de entrada de forma que os valores em uma coluna categórica sejam igualmente distribuídos
- Foi aplicada no modelo de Árvore de Decisão
- É possível testar a eficácia do modelo e da configuração escolhida
- Os resultados asseguram a utilização dos critérios considerando o modelo completo
- Além disso, endossa a possibilidade da existência de **overfitting** ou desbalanceamento na base completa



## RESULTADO COM ÁRVORE DE DECISÃO ® (EQUAL\_SIZE\_SAMPLING)

Rows Number : 124	0 (Predicted)	1 (Predicted)	
0 (Actual)	36	25	59.02%
1 (Actual)	12	51	80.95%
	75.00%	67.11%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
0	36	12	51	25	59.02%	75.00%	59.02%	80.95%	66.06%
1	51	25	36	12	80.95%	67.11%	80.95%	59.02%	73.38%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa ( $\kappa$ )	Correctly Classified	Incorrectly Classified
70.16%	29.84%	0.401	87	37

# TÉCNICA AVALIAÇÃO: SOBREAMOSTRAGEM E BUSCA DOS MELHORES HIPERPARÂMETROS

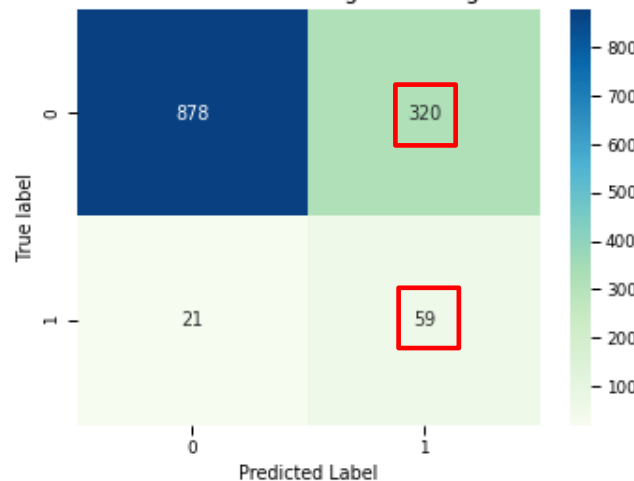
- A abordagem mais simples envolve a duplicação de exemplos na classe minoritária, embora esses exemplos não adicionem nenhuma informação nova ao modelo
- Este é um tipo de aumento de dados para a classe minoritária e é referido como **Synthetic Minority Oversampling Technique (SMOTE)**
- Buscou-se **os melhores hiperparâmetros** para os modelos avaliados através do GridSearch

# RESULTADO COM REGRESSÃO LOGÍSTICA (SMOTE)

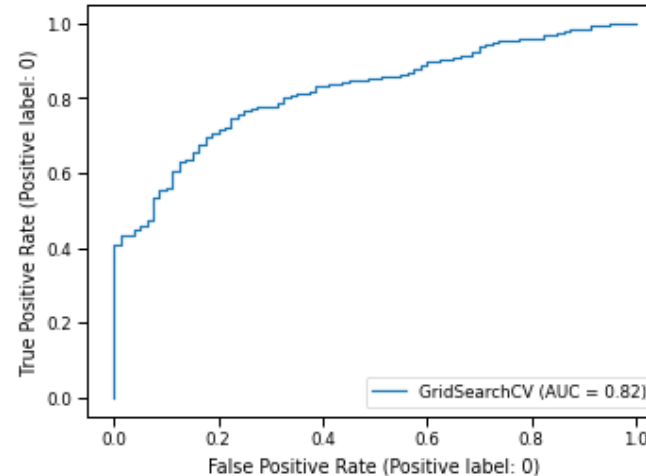
## Métricas do Modelo

	precision	recall	f1-score	support
0	0.98	0.73	0.84	1198
1	0.16	0.74	0.26	80
accuracy			0.73	1278
macro avg	0.57	0.74	0.55	1278
weighted avg	0.93	0.73	0.80	1278
Accuracy Score:	0.7331768388106417			

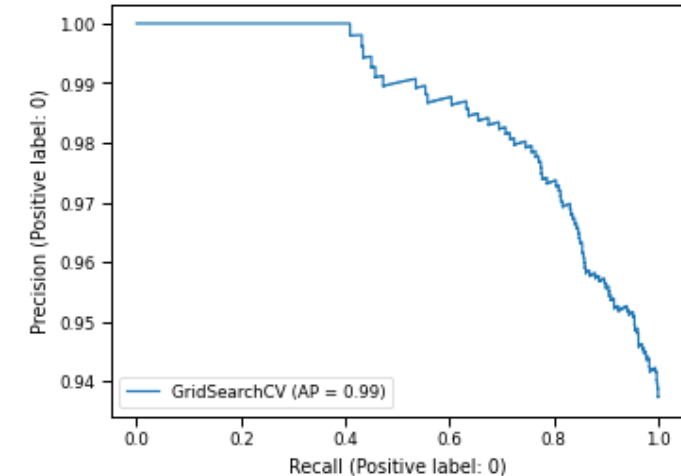
Matriz de Confusão - Regressao Logística



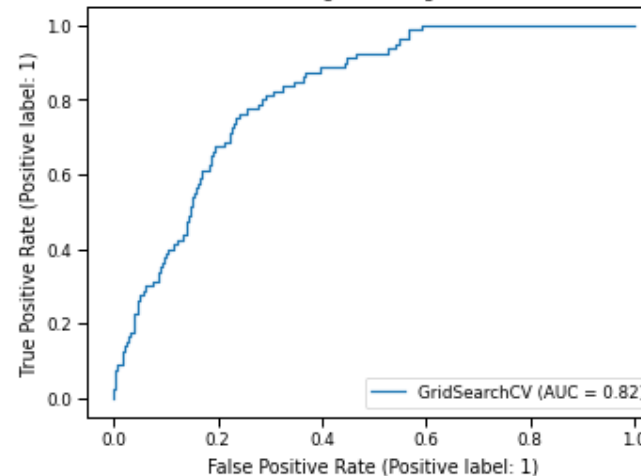
Curva ROC - Regressao Logística "Not Stroke"



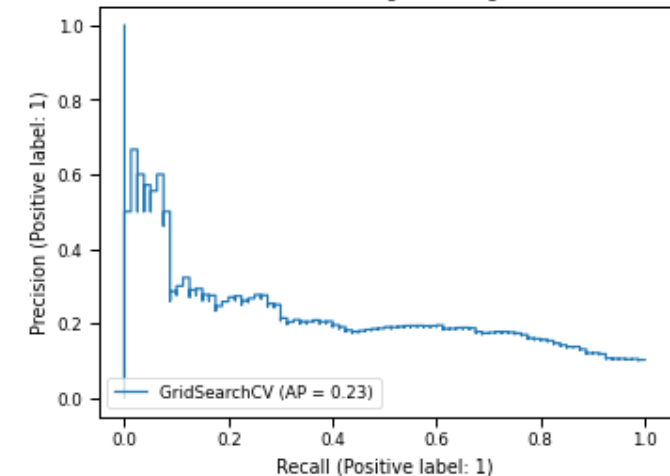
Curva Precision Recall - Regressao Logística "Not Stroke"



Curva ROC - Regressao Logística "Stroke"



Curva Precision Recall - Regressao Logística "Stroke"

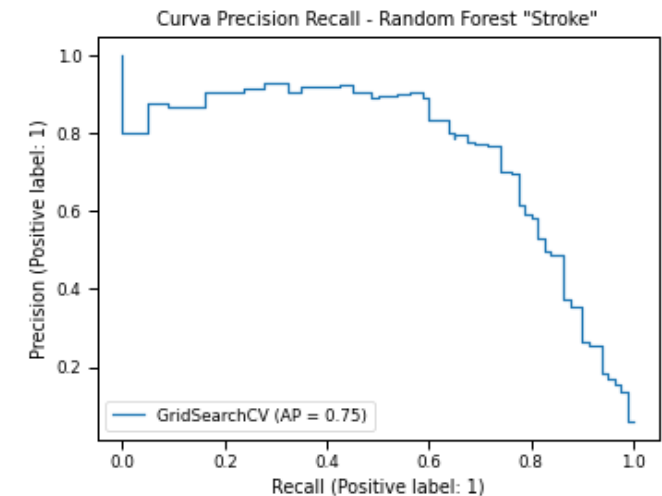
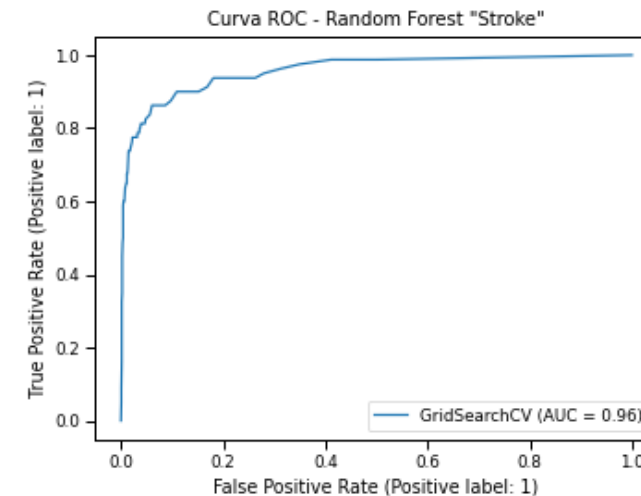
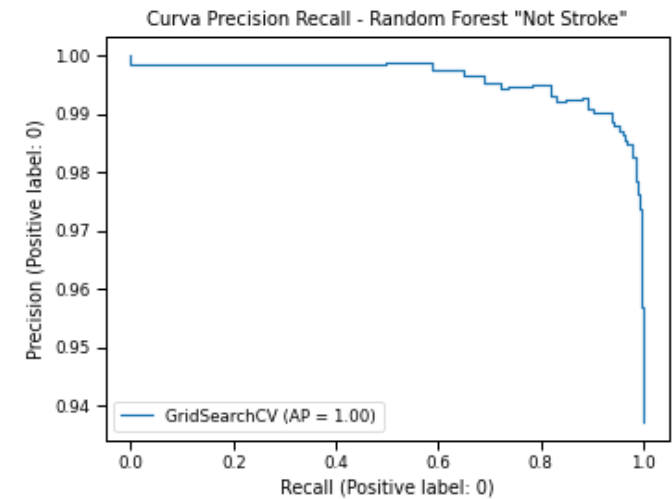
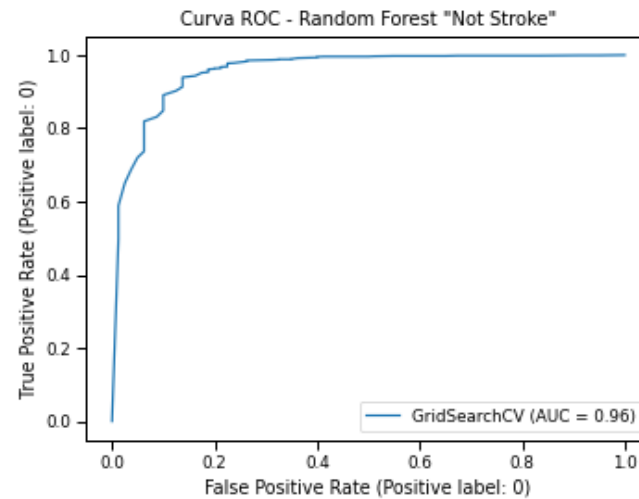
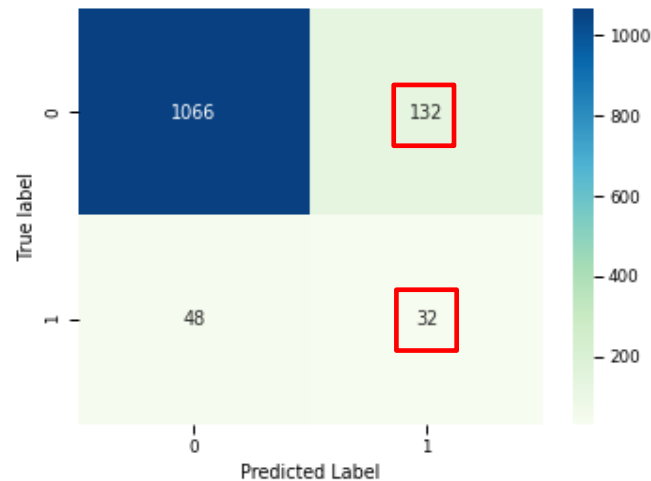


# RESULTADO COM RANDOM FOREST (SMOTE)

## Métricas do Modelo

	precision	recall	f1-score	support
0	0.96	0.89	0.92	1198
1	0.20	0.40	0.26	80
accuracy			0.86	1278
macro avg	0.58	0.64	0.59	1278
weighted avg	0.91	0.86	0.88	1278
Accuracy Score:	0.8591549295774648			

Matriz de Confusão - Random Forest



# RESULTADOS OBTIDOS

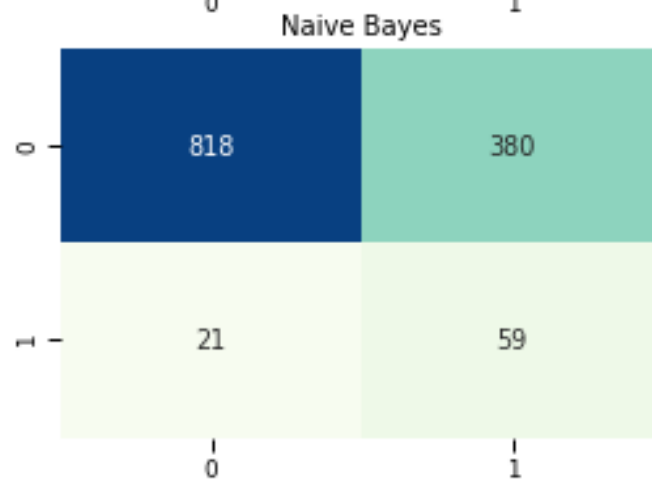
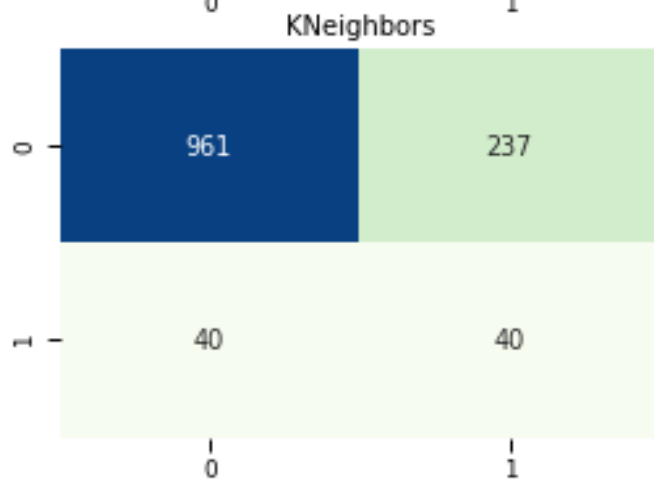
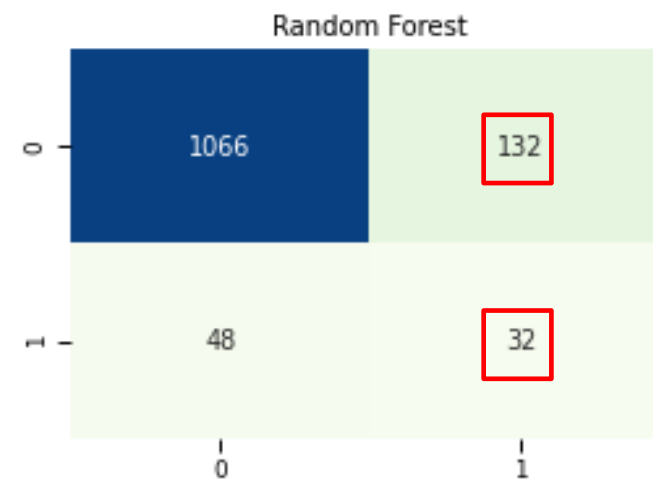
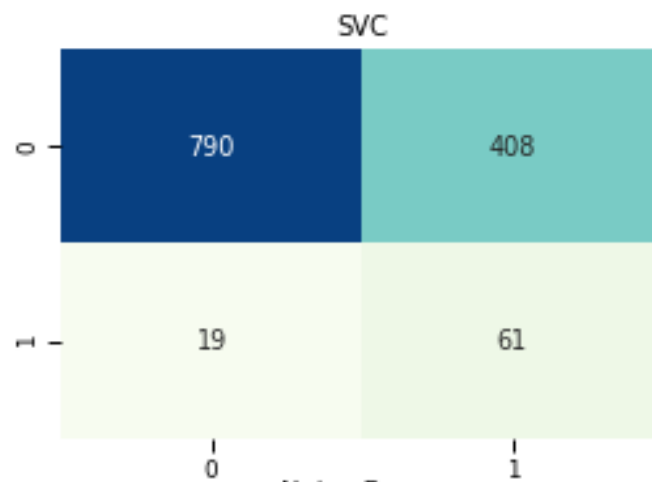
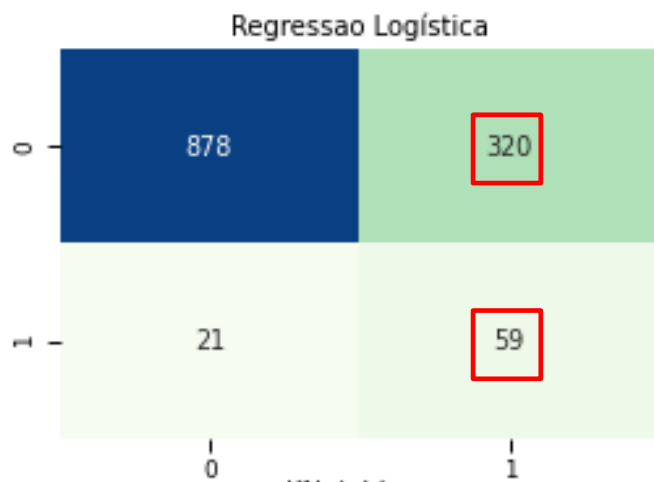
Métricas Modelos para Classe Not Stroke

	f1	precision	recall	roc_auc	ap	accuracy
LogisticRegression	0.84	0.98	0.73	0.82	0.99	0.73
SVC	0.79	0.98	0.66	0.81	0.99	0.67
RandomForest	0.92	0.96	0.89	0.96	1.0	0.86
KNeighbors	0.87	0.96	0.8	0.94	0.99	0.78
GaussianNB	0.8	0.97	0.68	0.81	0.98	0.69

Métricas Modelos para a Classe Stroke

	f1	precision	recall	roc_auc	ap
LogisticRegression	0.26	0.16	0.74	0.82	0.23
SVC	0.22	0.13	0.76	0.81	0.18
RandomForest	0.26	0.2	0.4	0.96	0.75
KNeighbors	0.22	0.14	0.5	0.94	0.66
GaussianNB	0.23	0.13	0.74	0.81	0.26

# RESULTADOS OBTIDOS



# DISCUSSÃO DOS RESULTADOS

- A acurácia dos modelos ficou em torno de 67% e 86%. O modelo de melhor acurácia foi o *RandomForest* (86%)
- A precisão para a classe dos não propensos ao AVC é bem elevada, entre 96% e 98%, mostrando que os acertos para esta previsão é muito bom em todos os algoritmos
- A melhor **relação entre sensibilidade e especificidade** foi a do algoritmo de **Regressão Logística (0,73 e 0,74)**. O **Random Forest (0,89 e 0,4)** teve uma sensibilidade alta, mas uma especificidade mais baixa
- Sobre a curva ROC, os melhores resultados são dos algoritmos Random Forest (0,96) e K-Neighbors (0,94)
- A matriz de confusão reforça o resultado das métricas, mostrando que os algoritmos Regressão Logística, SVC e Naive Bayes tiveram um Verdadeiro Negativo entre 59 e 61, ou seja, um maior acerto na classe minoritária
- O Random Forest teve o melhor Verdadeiro Positivo 1066 de 1198, seguido do KNeighbors e da Regressão Logística
- O desempenho geral do Random Forest foi o melhor, entretanto é necessário uma avaliação da métrica especificidade que tem relevância para responder à questão do projeto: prever as pessoas com predisposição ao AVC

# CONCLUSÃO

- Chegou-se a conclusão que é **mais importante identificar as pessoas propensas ao AVC**, já que elas poderão mudar seus hábitos alimentares e de vida para evitar que se concretize essa predição.
- Algumas técnicas foram aplicadas para avaliação dos algoritmos e, no final, a sobre amostragem junto com a busca dos melhores hiperparâmetros mostrou-se mais eficaz.
- As métricas com maior relevância para análise foram a sensibilidade e especificidade, em conjunto com a acurácia. A matriz de confusão e as curva ROC e Precision-Recall também foram avaliadas.
- No conjunto geral, **o algoritmo de Regressão Logística mostrou-se mais adequado** pelo equilíbrio apresentado entre as métricas de sensibilidade, especificidade e acurácia.



# LIÇÕES APRENDIDAS

- Para uma análise inicial do problema, ferramentas como o Orange e Knime possibilitam uma análise exploratória mais ágil, assim como a investigação de modelos de aprendizagem de máquina
- É de extrema importância a análise de correlação para identificar as variáveis preditoras
- Para a área de saúde, é fundamental o conhecimento sobre análise de dados para obter o melhor diagnóstico com eficácia e rapidez
- A descoberta do conhecimento é uma sequência de processos, otimizados quando realizados em conjunto

## REFERÊNCIAS

Acidente Vascular Cerebral Isquêmico. Hospital Albert Einstein. Disponível em: <<https://www.einstein.br/guia-doencas-sintomas/info/#4>> Acesso em: 11 abr 2021.

THRIFT, Amanda G. et al. Global stroke statistics. International Journal of Stroke, v. 12, n. 1, p. 13-32, 2017.

KIM, Joosup et al. Global stroke statistics 2019. International Journal of Stroke, v. 15, n. 8, p. 819-838, 2020.

XIANFANG, Wang et al. Predicting the types of ion channel-targeted conotoxins based on avc-svm model. BioMed research international, v. 2017, 2017.

PEI, Dongmei et al. Accurate and rapid screening model for potential diabetes mellitus. BMC medical informatics and decision making, v. 19, n. 1, p. 1-8, 2019.

Acidente Vascular Cerebral. Hospital Albert Einstein. Disponível em: <<https://www.einstein.br/doencas-sintomas/avc>> Acesso em: 11 abr 2021.

SITAR-TĂUT, A. et al. Using machine learning algorithms in cardiovascular disease risk evaluation. Age, v. 1, n. 4, p. 4, 2009.

ACIDENTE VASCULAR CEREBRAL. Sociedade Brasileira de Doenças Cerebrovasculares, 2020. Disponível em: [http://www.sbdcv.org.br/publica\\_avc.asp](http://www.sbdcv.org.br/publica_avc.asp). Acesso em: 10 abr. 2021.

## REFERÊNCIAS

Hankey GJ. Stroke. Lancet. 2017 Feb 11;389(10069):641-654. doi: 10.1016/S0140-6736(16)30962-X. Epub 2016 Sep 13. PMID: 27637676.

Fisher M, Moores L, Alsharif MN, Paganini-Hill A. Definition and Implications of the Preventable Stroke. JAMA Neurol. 2016 Feb;73(2):186-9. doi: 10.1001/jamaneurol.2015.3587. PMID: 26641201; PMCID: PMC4767801.

Thrift AG, Thayabaran Nathan T, Howard G, Howard VJ, Rothwell PM, Feigin VL, Norrving B, Donnan GA, Cadilhac DA. Global stroke statistics. Int J Stroke. 2017 Jan;12(1):13-32. doi: 10.1177/1747493016676285. Epub 2016 Oct 28. PMID: 27794138.

Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, Moran AE, Sacco RL, Anderson L, Truelsen T, O'Donnell M, Venketasubramanian N, Barker-Collo S, Lawes CM, Wang W, Shinohara Y, Witt E, Ezzati M, Naghavi M, Murray C; Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010) and the GBD Stroke Experts Group. Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet. 2014 Jan 18;383(9913):245-54. doi: 10.1016/s0140-6736(13)61953-4. Erratum in: Lancet. 2014 Jan 18;383(9913):218. PMID: 24449944; PMCID: PMC4181600.

GLOBAL STROKE ALLIANCE: UNIÃO E FORÇA CONTRA O AVC. Academia Brasileira de Neurologia, 2020. Disponível em: <https://www.abneuro.org.br/post/global-stroke-alliance-uni%C3%A3o-e-for%C3%A7a-contra-o-avc>. Acesso em: 10 abr. 2021.



# **OBIGADO**

São Paulo, 23 de Junho de 2021