Análise de Dados para auxiliar no diagnóstico precoce de Acidentes Vasculares Cerebrais
AVC

Bruna Lika Tamake RA: 171024427

Orientador: Prof. Dr. Clayton Reginaldo Pereira

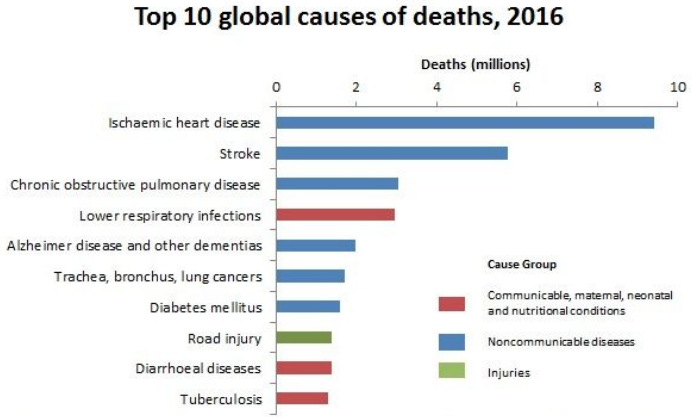
15 de Dezembro de 2020

Conteúdo

1. Problemática e Justificativa
2. Introdução teórica
3. Ferramentas
4. Base de Dados
5. Análise dos Dados
6. Tratamento dos Dados
7. Treinamento
8. Resultados
9. Conclusão

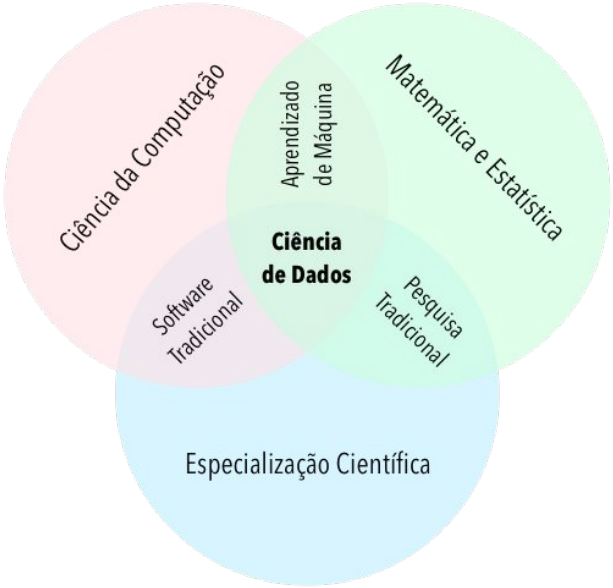
Problemática e Justificativa

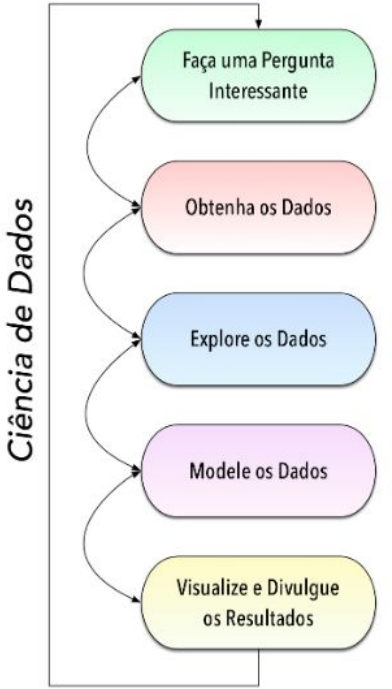
- Problemática:
  - Top 10 maiores causas de morte e invalidez no mundo.
  - Grande taxa de ocorrência mas baixa busca por informações e os cuidados necessários para quem tem chance de sofrer a doença.
- Justificativa:
  - AVC é tempo-dependente, quanto antes for tratada maiores são as chances de recuperação.



Fonte:<https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>

Introdução - Ciência de Dados



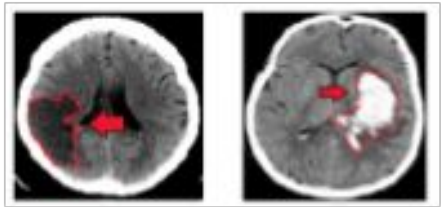


Fonte:<https://alfredbaudisch.medium.com/o-que-%C3%A9-ci%C3%AAncia-de-dados-data-science-7af5bdac101a>

Introdução - Acidente Vascular Cerebral

- Doenças que atingem os vasos sanguíneos que irrigam o cérebro:
  - rompimento → AVC hemorrágico
  - obstrução → AVC isquêmico

Diferença entre AVCi e AVCh em uma TC



Fonte: (AGUIAR,2017)

- Identificação:
  - Déficit neurológico focal, repentino e não convulsivo com duração maior que 24 horas.
  - Exames de imagem : Tomografia Computadorizada (TC) ou Ressonância Nuclear Magnética (RNM).
- Doença tempo-dependente.

Introdução - Acidente Vascular Cerebral

Fatores de Risco		
Genéticos e Fisiológicos	Estilo de Vida	Patológicos
Histórico Familiar	Sedentarismo	Hipertensão arterial sistêmica
Sexo Masculino	Tabagismo	Diabete Mellitus
Idade avançada	Uso excessivo de álcool	Obesidade/Sobrepeso
	Uso de drogas ilícitas	Colesterol alto
		Cardiopatias

Fonte: Elaborado pela autora.

Introdução - Métodos de Aprendizado de Máquina

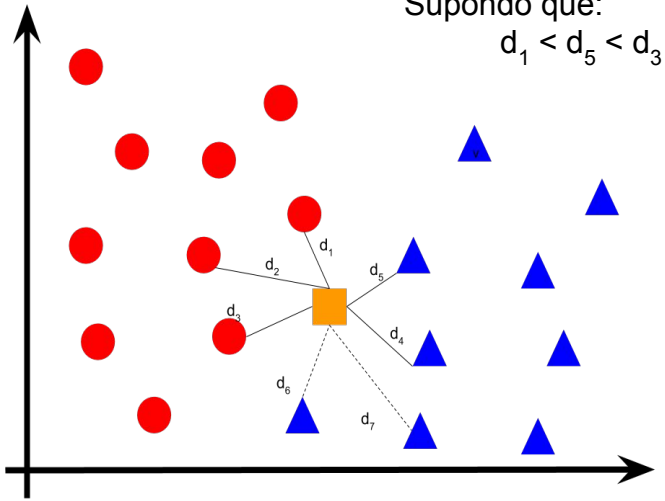
- Estatística + Inteligência Artificial + Computação
- Objetivo:
  - Prever a ocorrência de uma informação.
  - Ganhar conhecimento a partir dos dados.
- Tipos:
  - Aprendizado Supervisionado: dados de entrada já rotulados e busca-se a função que determina a o comportamento das entradas para as saídas.
  - Aprendizado Não-Supervisionado: dados de entrada não vem rotulados e busca-se formas possíveis de classificador esses dados.
  - Aprendizado por reforço: a IA busca uma forma de cumprir seu objetivo a partir de tomada de decisões e a forma com que o ambiente responde a essas decisões (recompensas ou penalidades).

Introdução - Aprendizado Supervisionado

- K Nearest Neighbor:

Supondo que:

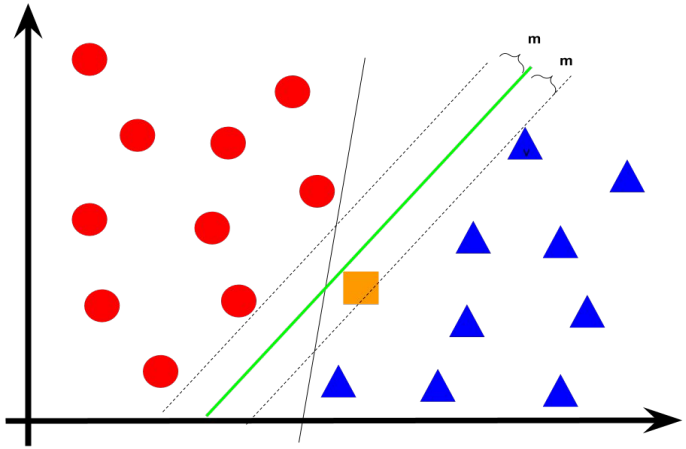
$$d_1 < d_5 < d_3 < d_4 < d_2 < d_6 < d_7$$



Fonte: Elaborado pela autora.

Introdução - Aprendizado Supervisionado

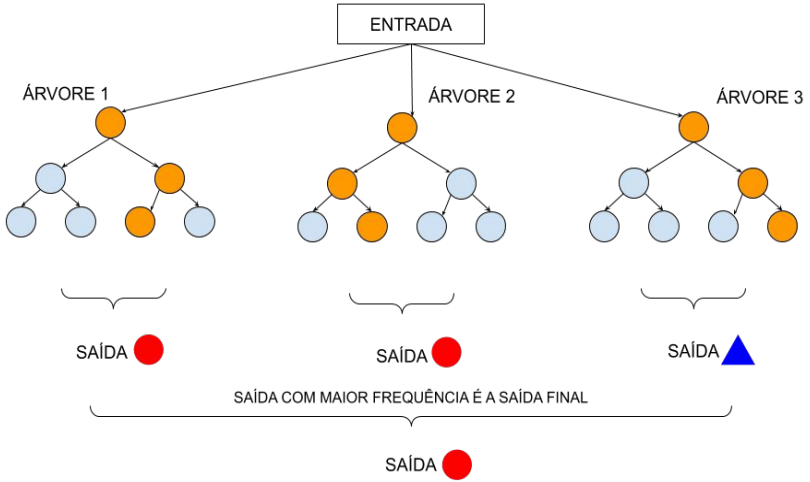
- Support Vector Machine (SVM)



Fonte: Elaborado pela autora.

Introdução - Aprendizado Supervisionado

- Florestas Aleatórias



Fonte: Elaborado pela autora.

Ferramentas

- Python: linguagem de programação de alto nível, orientada objeto com tipagem dinâmica e forte.
- Bibliotecas:
  - Pandas: trabalhar com Dataframes (tabelas de dados)
  - Scikit-Learn: métodos que facilitam a utilização dos métodos de aprendizado de máquina
  - Plotly: plotagem de gráficos e figuras
- Jupyter Notebook: ferramenta de processamento interativo e independente. Capaz de executar, depurar e testar códigos

Dataset

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose	bmi	smoking_status	stroke
12388	Female	62	0	0	Yes	Self-employed	Rural	92.78	18.6	never smoked	0
27584	Male	57	0	0	Yes	Private	Urban	90.37	30.7	formerly smoked	0
26727	Female	79	0	0	No	Private	Rural	88.92	22.9	never smoked	1
39002	Male	40	1	0	Yes	Private	Urban	84.57	32.7	smokes	0
54232	Female	60	1	1	Yes	Govt_job	Urban	86.54		smokes	0
58027	Female	71	0	0	No	Private	Urban	214.22	33.9	never smoked	0
35635	Male	63	0	0	Yes	Private	Urban	76.2	27.6	never smoked	0
9203	Female	52	1	0	Yes	Govt_job	Urban	92.72	53.4	smokes	0
34446	Female	81	0	0	Yes	Self-employed	Urban	74.92	26.2	smokes	0

Fonte: Elaborado pela autora.

Fonte: <https://github.com/>

Fonte: <https://www.ubc.ca/>

Dataset

- Chave de Identificação - Id
- Sexo:
  - Feminino;
  - Masculino.
- Hipertensão:
  - 1 → Sim;
  - 0 → Não.
- Doença do coração:
  - 1 → Sim;
  - 0 → Não.
- Casado:
  - Sim;
  - Não.
- Sofreu AVC:
  - 1 → Sim;
  - 0 → Não.
- Tipo de Trabalho:
  - Empresa privada;
  - Autônomo;
  - Governo;
  - Nunca trabalhou;
  - Criança.
- Status de Fumante:
  - Fumante;
  - Ex-fumante;
  - Nunca fumou.
- Idade.
- Média do nível de glicose.
- BMI → Body Mass Indicator.

Dataset

- Inicial:
  - 43.400 entradas:
    - 783 vítimas de AVC → ≅2%
    - 42.617 pessoas sem AVC → ≅98%

- Trabalho: **x** vítimas de AVC - **2x** pessoas sem AVC
  - 2.349 entradas:
    - 783 vítimas de AVC → ≅33%
    - 1.566 pessoas sem AVC → ≅67%

Análise dos Dados

Atributos Categóricos									
	Total	Faltante (%)	Categorias	Moda	Qnt. Moda	Freq. Moda	2ª Moda	Qnt. 2ª Moda	Freq. 2ª Moda
Gênero	2349	0	2	Feminino	1348	57,39	Masculino	1001	42,61
Hipertensão	2349	0	2	0	2007	85,44	1	342	14,56
Doença do Coração	2349	0	2	0	2096	89,23	1	253	10,73
Casamento	2349	0	2	Sim	1720	73,22	Não	629	26,78
Tipo de Trabalho	2349	0	5	Privado	1378	58,66	Autônomo	480	20,43
Tipo de Residência	2349	0	2	Rural	1177	50,10	Urban	1172	49,90
Status de Fumante	2349	26,01	3	Nunca fumou	883	37,59	Ex fumante	482	20,51
AVC	2349	0	2	0	1566	66,67	1	783	33,33

Fonte: Elaborado pela autora.

Análise dos Dados

Atributos Contínuos									
	Total	Faltante (%)	Mínimo	1º Quartil	Média	Mediana	3º Quartil	Máximo	Desvio Padrão
Idade	2349	0	0,08	34	50,85	54	71	82	22,95
Nível médio de Glicose	2349	0	55,01	78,43	113,58	94,39	126,35	271,74	51,24
BMI	2349	8,34	10,30	24,30	29,38	28,50	33,30	78	7,62

Fonte: Elaborado pela autora.

Tratamento dos Dados

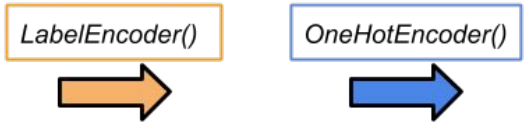
- Status de Fumante:
  - 26 % → 611 entradas
    - 466 pessoas sem AVC.
    - 145 vítimas de AVC.

AVC	Categoria	Quantidades
0 Não sofreu AVC	Nunca fumou	599
	Desconhecido	466
	Ex-fumante	261
	Fumante	240
1 Sofrem AVC	Nunca fumou	284
	Desconhecido	145
	Ex-fumante	221
	Fumante	133

Fonte: Elaborado pela autora.

Tratamento dos Dados

- Atributos Categóricos com Aprendizado de Máquina:
  - Fumante ? Ex-Fumante ?
  - 0 < 1 < 2



Status de Fumante
Ex-fumante
Não Informado
Nunca fumou
Fumante

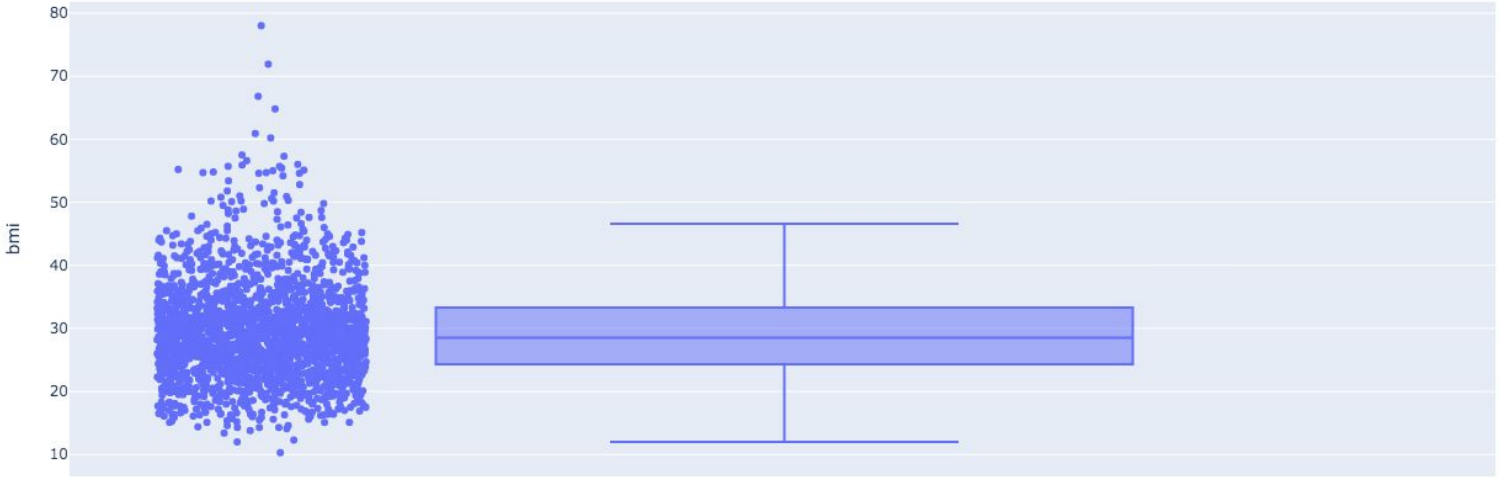
Status de Fumante
0
1
2
3

0 (Ex-fumante)	1 (Não Informado)	2 (Nunca fumou)	3 (Fumante)
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Fonte: Elaborado pela autora.

Tratamento dos Dados

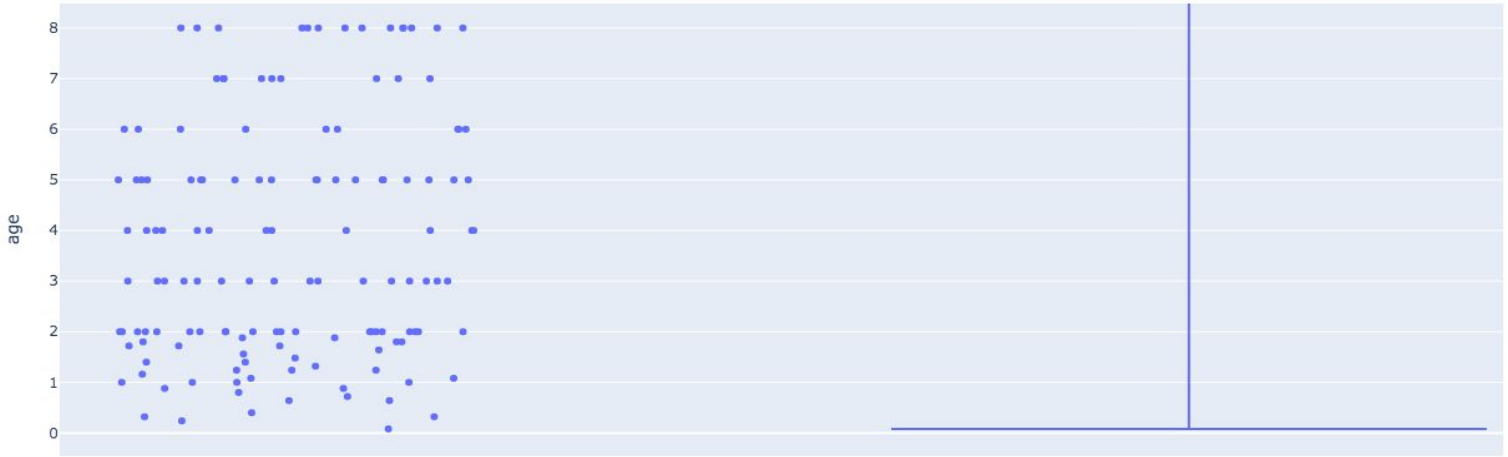
- BMI:
  - 8,34 % → 196 entradas
  - média : 28,5 → sobrepeso (NIHISER et al., 2007)



Fonte: Elaborado pela autora.

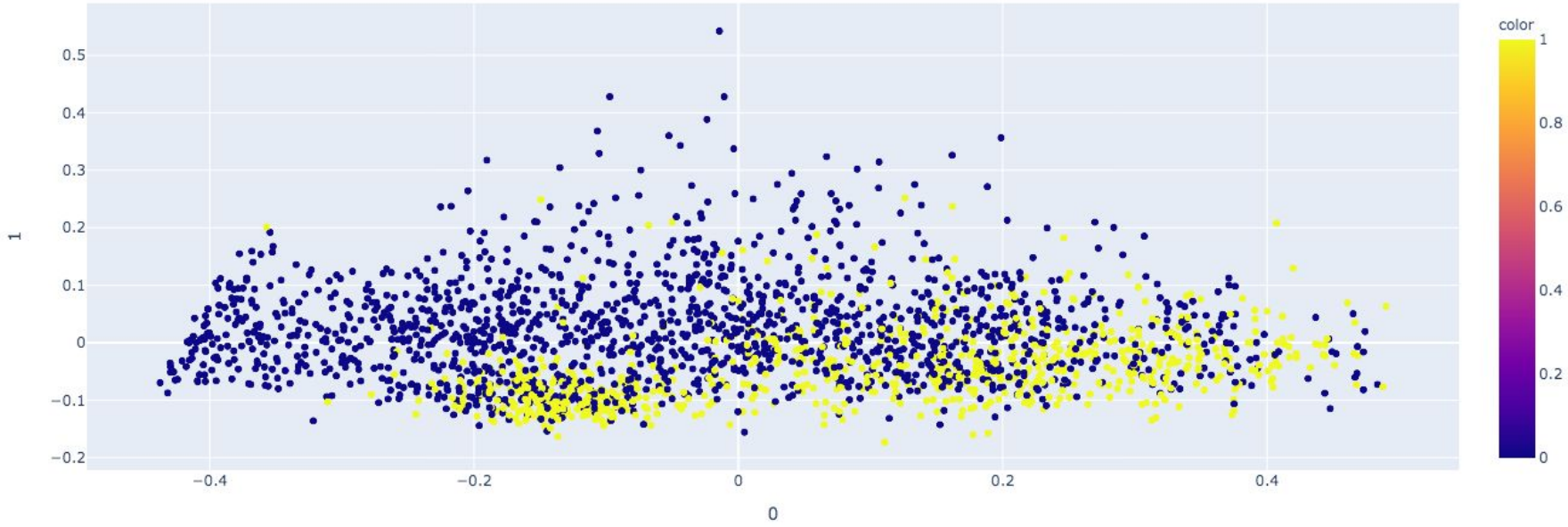
Tratamento dos Dados

- Idade
  - 32 entradas com idade não sendo valores inteiros
    - 0,08
    - 1,72



Fonte: Elaborado pela autora.

Distribuição dos Dados



Treinamento

- 25 % teste → 588 entradas
- 75% treinamento → 1.761 entradas
- Métodos utilizados:
  - K Nearest Neighbors : (K = 20)
  - Support Vector Machines
  - Floresta Aleatórias (default = 100 árvores)

Resultados

MATRIZ DE CONFUSÃO

	KNN		Random Forest		SVM	
	0	1	0	1	0	1
Acurácia	0,78		1,0		0,77	
Precisão	0,85	0,63	1,0	1,0	0,84	0,62
Recall	0,83	0,67	1,0	1,0	0,84	0,61
F1-Score	0,84	0,65	1,0	1,0	0,84	0,62
Suporte	411	176	411	176	411	176

Fonte: Elaborado pela autora.

Resultados

MATRIZ DE CONFUSÃO

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: (NOGARE, 2020)

K Nearest Neighbors		
	SEM AVC	AVC
SEM AVC	342	69
AVC	58	118

SVM		
	SEM AVC	AVC
SEM AVC	345	66
AVC	68	108

Florestas Aleatórias		
	SEM AVC	AVC
SEM AVC	411	0
AVC	0	176

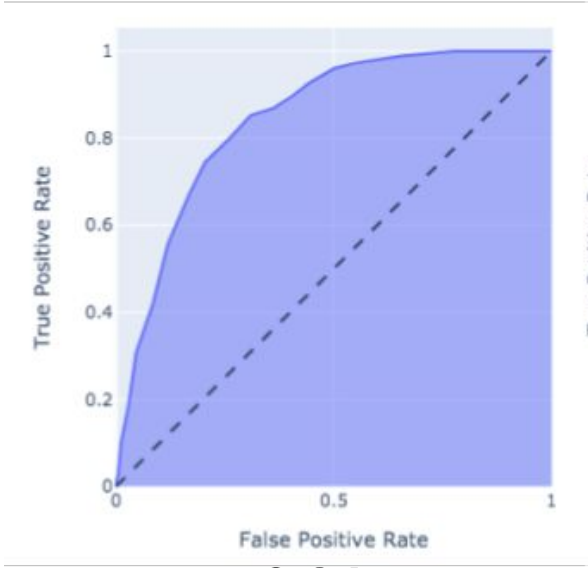
Fonte: Elaborado pela autora

$$\text{Prec} = \frac{VP}{VP + FP} \qquad \text{Revoc} = \frac{VP}{VP + FN} \qquad \text{ACURÁCIA} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{F1 Score} = \frac{2*\text{Prec}*\text{Rovoc}}{\text{Prec} + \text{Revoc}}$$

Resultados

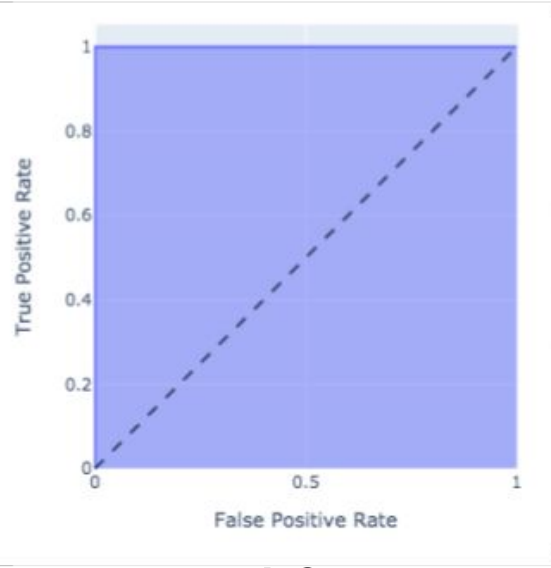
AUC ROC

KNN



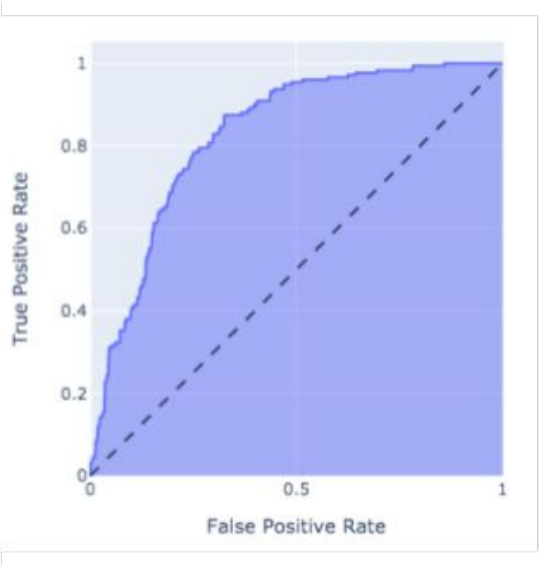
0,84

Florestas Aleatórias



1,0

SVM



0,82

Fonte: Elaborado pela autora

Conclusão

Florestas Aleatórias

	Random Forest	
	0	1
Acurácia	1,0	
Precisão	1,0	1,0
Recall	1,0	1,0
F1-Score	1,0	1,0
Suporte	411	176

Fonte: Elaborado pela autora

Trabalhos futuros

- Obter dataset com mais atributos ligados aos fatores de risco ligados a doença e se possível verídicos.

Referências Bibliográficas:

AGUIAR, C.Avaliação de acidente vascular cerebral em tomografia computadorizada utilizando algoritmo de otimização de formigas. Dissertação (Mestrado), 2017.

NIHISER, A. J.; LEE, S. M.; WECHSLER, H.; MCKENNA, M.; ODOM, E.; REINOLD, C.; THOMPSON, D.; GRUMMER-STRAWN, L. Body mass index measurement in schools. Journal of School Health , Wiley Online Library, v. 77, n. 10, p. 651–671, 2007.

Diego Nogare. Performance de Machine Learning – Matriz de Confusão. 2020. <http://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/>.Online; Acesso em: 14 de Outubro 2020

Obrigada!

Bruna Lika Tamake RA: 171024427
Orientador: Prof. Dr. Clayton Reginaldo Pereira