UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

BRUNA LIKA TAMAKE

ANÁLISE DE DADOS PARA AUXILIAR NO DIAGNÓSTICO PRECOCE DE ACIDENTE VASCULAR CEREBRAL - AVC

BAURU
Dezembro/2020

BRUNA LIKA TAMAKE

ANÁLISE DE DADOS PARA AUXILIAR NO DIAGNÓSTICO PRECOCE DE ACIDENTE VASCULAR CEREBRAL - AVC

Trabalho de Conclusão de Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Dr. Clayton Reginaldo Pereira

BAURU
Dezembro/2020

Bruna Lika Tamake ANÁLISE DE DADOS PARA AUXILIAR NO DIAGNÓSTICO PRECOCE DE ACIDENTE VASCULAR CEREBRAL - AVC/ Bruna Lika Tamake. – Bauru, Dezembro/2020- 40 p. : il. (algumas color.) ; 30 cm.
Orientador: Prof. Dr. Clayton Reginaldo Pereira
Trabalho de Conclusão de Curso – Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Ciência da Computação, Dezembro/2020.

Bruna Lika Tamake

ANÁLISE DE DADOS PARA AUXILIAR NO DIAGNÓSTICO PRECOCE DE ACIDENTE VASCULAR CEREBRAL - AVC

Trabalho de Conclusão de Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Dr. Clayton Reginaldo Pereira
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Prof. Dra. Simone das Graças Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Dr. Leandro Aparecido Passos Júnior
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Bauru, 15 de dezembro de 2020.

Resumo

As doenças cardiovasculares são um grupo de doenças que podem atingir o coração e os vasos sanguíneos, estando entre as principais causas de morte no mundo. Alguns dos tipos que existem são: coronária, cardíaca reumática, cardiopatia congênita, cerebrovasculares, entre outros. Dentre os tipos de doenças cardiovasculares citados e existentes, as cerebrovasculares são as que mais atingem as pessoas e a mais neglicenciada no Brasil (LOTUFO et al., 2017), sendo uma das principais delas e objeto de estudo deste projeto: os Acidentes Vasculares Cerebrais (AVC). Este trabalho buscou por meio de técnicas de aprendizado de máquina como *K-Nearest Neighbors*, *Support Vector Machine* e *Random Forest*, prever a ocorrência de novos casos de AVC.

Palavras-chave: AVC, Acidentes Vasculares Cerebrais, aprendizado de máquina, predição.

Abstract

Cardiovascular diseases are a group of diseases that can affect the heart and blood vessels, being among the main causes of death in the world. Some of the types that exist are: coronary, rheumatic cardiac, congenital heart disease, cerebrovascular and others. Between the types of cardiovascular diseases mentioned above and the others that exists, the cerebrovascular are the ones that most affect people and the most neglected in Brazil (LOTUFO et al., 2017), the principal one and object of this project: the strokes (AVC). This project search through machine learning techniques such as K-Nearest Neighbors, Support Vector Machine and Random Forest, a way to predict the occurrence of new stroke cases.
Keywords: Stroke, AVC, Machine Learning, prediction.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under Curve*.
AVC	Acidente Vascular Cerebral.
AVCh	Acidente Vascular Cerebral Hemorrágico.
AVCi	Acidente Vascular Cerebral Isquêmico.
BMI	*Body Mass Indicator*.
CFM	Conselho Federal de Medicina.
FN	Falso negativo.
FP	Falso positivo.
GPL	*General Public License*.
IA	Inteligência Artificial.
KNN	K-Nearest Neighbors.
OMS	Organização Mundial de Saúde.
PEP	Prontuário Eletrônico de Paciente.
OPAS	Organização Pan Americana de Saúde.
PSF	*Python Software Foundation*.
RNA	Redes Neurais Artificiais.
RNM	Ressonância Nuclear Magnética.
ROC	*Receiver Operating Characteristic*.
SVM	Support Vector Machine.
TC	Tomografia Computadorizada.
TN	Verdadeiro negativo.
TP	Verdadeiro positivo.

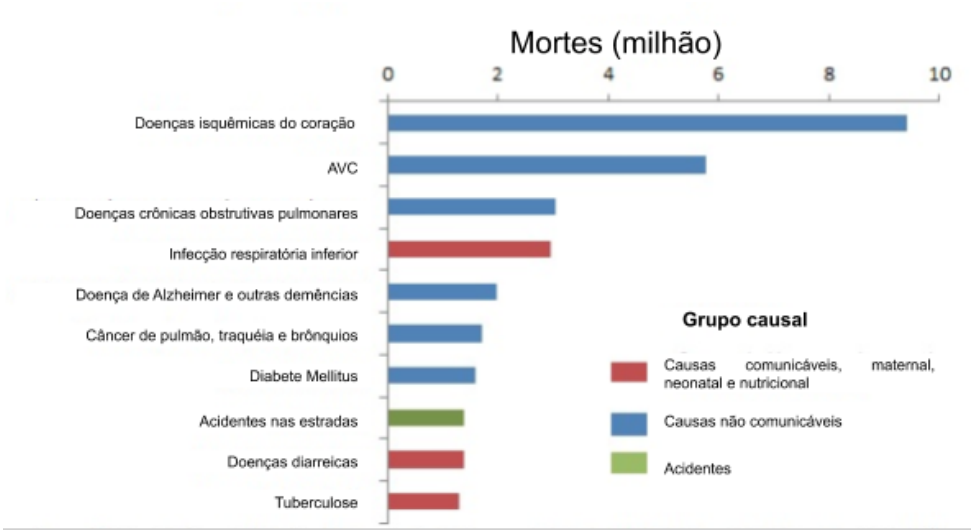
Sumário

1 Introdução

Líder entre as principais causas de morte do mundo, as doenças cardiovasculares apresentam como principal patologia subjacente a aterosclerose. A aterosclerose é caracterizada pelo acúmulo de lípides, células inflamatórias e elementos fibrosos das paredes das artérias e dependendo da artéria que atinge pode caracterizar um tipo diferente de doença cardiovascular. As principais artérias alvos são a aorta, coronária e cerebral, responsáveis pelas doenças cardiovasculares da aorta, coronárias e cerebrais, respectivamente (GOTTLIEB; BONARDI; MORIGUCHI, 2005).

Dentre os tipos de doenças cardiovasculares citados e existentes, as cerebrovasculares são as que mais atingem as pessoas e a mais neglicenciada no Brasil (LOTUFO et al., 2017). Em 2016 a Organização Mundial de Saúde (OMS) em parceria com a Organização Pan Americana de Saúde (OPAS), publicou um relatório sobre as dez principais causas de mortes no mundo.

Figura 1 – Gráfico das 10 principais causas de morte de 2016



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

Como é possível ver na Figura 1, os Acidentes Vasculares Cerebrais (AVC) encontram-se na segunda posição da classificação mundial. Segundo Bo Norrving[1], por ano, há o registro de cerca de 15 milhões de casos de AVC sendo que em cerca de 6 milhões ocorre o falecimento do paciente (NORRVING, 2014).

Iniciado na década de 70 e perdurado até os dias de hoje, as renovações na saúde vem ocorrendo de forma a incluir cada vez mais soluções tecnológicas no dia a dia dos profissionais

[1] Bo Norrving, professor de neurologia na *Lund University*, Suécia. Fundador do primeiro registro nacional de AVC o *Riksstroke The Swedish Stroke Register* <https://portal.research.lu.se/portal/en/persons/bo-norrving(202ac664-5da8-42c2-b110-96d26905695f).html>

da área. Em 2017, foi divulgado pelo Ministério da Saúde um plano nomeado e-Saúde, que tem como objetivo aumentar a qualidade e ampliar o acesso à saúde, de forma a qualificar as equipes, agilizar o atendimento e melhorar o fluxo de informações para apoio à decisão, incluindo tanto a decisão clínica, de vigilância em saúde, de regulação e promoção da saúde quanto a decisão de gestão(MINISTÉRIO DA SAÚDE, 2002).

Um resultado dessa busca pela renovação e regulamentado pelo Concelho Federal de Medicina (CFM) na resolução nº 1.638 em 2002, os Prontuários Eletrônicos do Paciente (PEP) tem trazido para a área de saúde uma forma mais rápida e barata de coletar, armazenar e processar os dados dos pacientes (MOURÃO; NEVES, 2007). Segundo o CFM, enquadra-se como PEP: documento único constituído de um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimento e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, que possibilita a comunicação entre os membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo (CONSELHO FEDERAL DE MEDICINA, 2002).

Concomitante as inovações tecnológicas no campo da saúde acompanhadas das tecnologias emergentes, a tecnologia em si tem avançado rapidamente em suas descobertas e aprimoramento de técnicas. Devido a grande quantidade de dados que são gerados diariamente, tornou-se necessário o desenvolvimento de uma nova área dentro da computação, nomeada Ciência de Dados.

Figura 2 – Diagrama de Venn sobre composição da Ciência de Dados

Fonte: <https://www.ironhack.com/br/data-analytics/data-science-x-data-analytics>

O Diagrama de Venn na Figura 2 foi criado pelo cientista de dados Drew Conway [2] e apresenta como seria a composição ideal da Ciência de Dados. A intersecção, onde se encontra

[2] <http://drewconway.com/>

a ciência de dados, representa a combinação das habilidades de modelagem e resumir conjuntos de dados (parte relacionada à matemática e estatística), habilidade de desenvolver *design* e algoritmos eficientes para armazenar, processar e visualizar os dados (parte relacionada à ciência da computação) e o entendimento de determinada área a fim de ser capaz de elaborar perguntas relevantes para serem respondidas (parte relacionada à expertise) (VANDERPLAS, 2016).

A utilização de conhecimentos de Ciências de Dados junto com o Aprendizado de Máquina na saúde já vem sendo explorada por pesquisadores realizando auxílio no diagnóstico de doença de Parkinson (PEREIRA et al., 2016), predição de casos de Hepatite A (SANTOS et al., 2005), diagnóstico de câncer de mama (SANTOS et al., 2020) entre outras patologias.

Este trabalho busca aplicar a estrutura da ciência de dados na área de conhecimento referentes as doenças cerebrovasculares, em especial o AVC. O objetivo é conseguir, por meio de conhecimentos de modelagem e métodos de aprendizado de máquina adquiridos no decorrer do desenvolvimento do trabalho de conclusão de curso, classificar os indivíduos portadores ou não de características que o levaria ser cometido pela doença bem como, predizer as possibilidades desse indivíduo sofrer um acidente vascular cerebral precoce.

2 Fundamentação Teórica

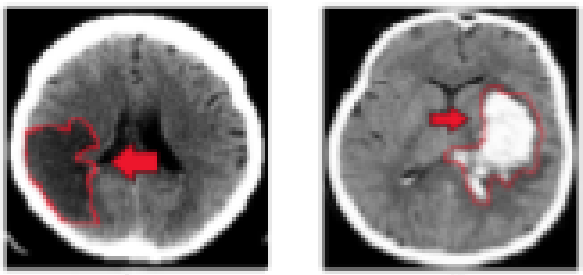
2.1 Acidente Vascular Cerebral

A patologia das doenças cerebrovasculares funcionam de forma que a doença atinge os vasos sanguíneos que irrigam o cérebro, inicialmente danificando o funcionamento do órgão e posteriormente afetando sua estrutura. A doença cerebrovascular mais conhecida e que servirá de base para o desenvolvimento deste projeto é o Acidente Vascular Cerebral.

Os AVCs são lesões cerebrais secundárias a um mecanismo vascular e não traumático, podendo ocorrer por conta da alteração do fluxo de sangue ao cérebro e que levam a morte das células nervosas da área atingida (ACADEMIA BRASILEIRA DE NEUROLOGIA, 2015). Dependendo da forma em que ocorrem podem ser do tipo isquêmico (AVCi) ou hemorrágico (AVCh).

Os AVCis ocorrem em 80% dos casos de AVC (SANARFLIX, 2019), sua causa está ligada a obstrução dos vasos sanguíneos que podem ocorrer proveniente de uma trombose[1] ou embolia[2]. Os AVCh por sua vez, ocorrem quando há o rompimento dos vasos sanguíneos no interior do cérebro, sendo ele o mais grave e que apresenta uma maior taxa de mortalidade. A diferença entre os tipos de AVC fica visualmente demonstrada pelas tomografias computadorizadas (TC) da Figura 3.

Figura 3 – Diferença entre AVCi(a) e AVCh(b) através de uma TC.



Fonte:Aguiar (2017)

Segundo a OMS, os AVCs estão entre as principais causas de morte, incapacitações e internações no mundo, sendo também a maior causa de incapacitação da população com mais

[1] A trombose ocorre quando há a formação de coágulos sanguíneos em um mais veias, levando a inchaços e dores pelo corpo (MINISTÉRIO DA SAÚDE, 2013).

[2] A embolia vem como consequência de uma trombose, onde um coágulo que tenha se formado no sistema venoso profundo, se desprende e atravessa as cavidades direitas do coração, podendo obstruir a artéria pulmonar ou um de seus ramos(VOLSCHAN et al., 2004).

de 50 anos (ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE, 2015). Se considerado apenas doenças cerebrovasculares e doenças cardíacas, no Brasil, esta vem a se tornar a primeira causa de morte (MARIA ELISABETH FERRAZ, 2019).

O diagnóstico da doença pode ser feito de duas formas: identificando um défice neurológico focal, repentino e não convulsivo do paciente com duração maior que 24 horas ou com a identificação de alterações nos exames de imagem como TC ou a Ressonância Nuclear Magnética (RNM), sendo este último o exame considerado padrão-ouro[3]. Apesar disso, o TC vem ganhando destaque visto que é um exame de maior agilidade de disponibilidade nos serviços (LIMA; PAGLIOLI; FILHO, 2012).

2.1.1 Fatores de Risco

Existem diversos fatores de risco que são considerados para a ocorrência de um AVC e que podem ser divididos em estilo de vida inadequado, patologias e fatores genéticos e fisiológicos.

2.1.1.1 Fatores de risco genéticos e fisiológicos

- Histórico familiar;
- Sexo masculino;
- Idade avançada.

2.1.1.2 Fatores de risco de estilo de vida inadequados

- Sedentarismo;
- Tabagismo;
- Uso excessivo de álcool;
- Uso de drogas ilícitas.

2.1.1.3 Fatores de risco considerados patologias

- Hipertensão arterial sistêmica (HAS): além de ser considerado um fator de risco, a HAS possui manifestações próprias. As Diretrizes Brasileiras de Hipertensão VI (DBH VI) classificam o HAS como uma condição clínica multifatorial caracterizada por níveis elevados e sustentados de pressão arterial (DHB VI, 2010).

[3] Um exame padrão-ouro é um exame que apresenta o diagnóstico e que serve de referência para outros exames realizados.

- Diabetes mellitus (destacando o tipo 2);
- Obesidade ou Sobrepeso;
- Colesterol alto (destaque para o LDL);
- Cardiopatias (em especial as arritmias cardíacas).

2.2 Aprendizado de Máquina

O aprendizado de máquina é uma área que trabalha na intersecção entre a estatística, inteligencia artificial (IA) e a ciência da computação na busca pela otimização da tarefa de tomar decisões a partir de dados ou experiências passadas e o padrão que estas apresentam por meio de um modelo parametrizado. Este modelo pode ser preditivo ou descritivo, possuindo as tarefas de predizer o futuro ou ganhar conhecimento através dos dados, respectivamente.

A forma com que os algoritmos de aprendizado de máquina trabalham pode ser dividida em três tipos:

- Aprendizado não supervisionado: nesse tipo de aprendizado a IA recebe dados de entrada não rotulados e a partir de similaridades e anomalias, busca encontrar classificações possíveis para estes dados.
- Aprendizado supervisionado: nesse tipo de aprendizado a IA recebe dados de entrada e saída previamente rotulados e a partir deles busca encontra a função que determina o comportamento das entradas para a saída.
- Aprendizado por reforço: nesse tipo de aprendizado não há tanta importância se os dados estão previamente rotulados ou não. O modelo de aprendizagem busca solucionar um problema ou cumprir uma tarefa a partir de suas ações e como o ambiente responde a elas, dando como resposta recompensas ou penalidades, caso estejam certas ou erradas, respectivamente. O objetivo é maximizar a recompensa final total a ser recebida em situações consideradas incertas e complexas.

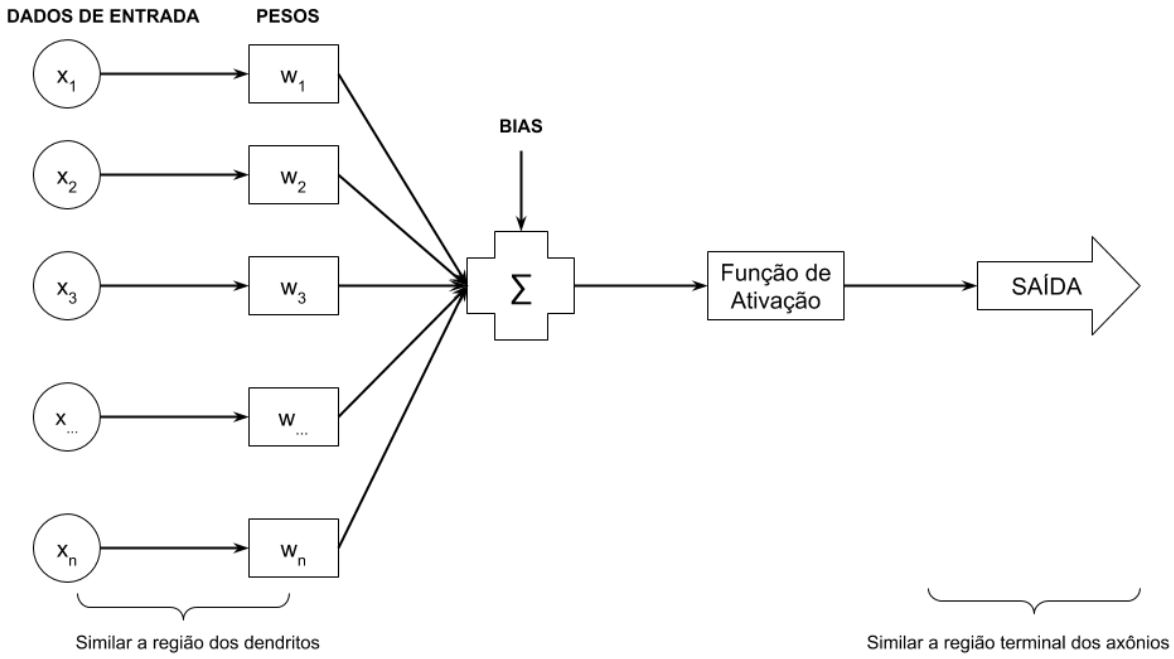
Os dados utilizados neste projeto, demonstrados na seção 3 deste documento, encontram-se previamente rotulados entre pessoas que sofreram um AVC, representado pelo atributo *stroke* equivalente a 1, e pessoas que não sofreram um AVC, representado pelo atributo *stroke* equivalente a zero. Como os dados estão rotulados e a tarefa principal deste projeto é uma previsão, os métodos de aprendizado de máquina a serem utilizados serão do tipo supervisionados.

Apesar de conseguirem realizar tomadas de decisões pontuais, os métodos de aprendizagem de máquina ainda necessitam de interferências humanas quando se trata de fornecer os

dados de forma perfeita para o método consumir, assim como quando é realizado uma predição ou aprendizado ruim. Na tentativa de tornar esse processo mais autônomo, semelhante ao funcionamento de um cérebro humano, são utilizadas as Redes Neurais Artificiais (RNA).

Uma RNA é um conjunto de neurônios artificiais, organizados de forma hierárquica e ligados entre si (KOVÁCS, 2002). Os neurônios utilizados pelas RNA possuem estrutura similar aos neurônios biológicos que os seres humanos possuem, como mostrado na Figura 4.

Figura 4 – Estrutura de uma Neurônio Artificial



Fonte:Elaborado pela autora

O processamento feito no neurônio ocorre a partir de uma combinação linear de entradas com pesos que passam por uma função de ativação e gera uma saída. Como um RNA é uma combinação entre arquitetura e algoritmos de aprendizado existem diversos modelos que podem ser elaborados dependendo da complexidade do problema a ser resolvido (KOVÁCS, 2002).

No tópico a seguir será melhor explorado os modelos utilizados para a tarefa de predição proposta neste trabalho.

2.2.1 Modelos de Aprendizado de Máquina Supervisionado

Dependendo do tipo de resultado que busca-se predizer, os algoritmos de aprendizado de máquina podem ser de dois tipos:

- Classificação: onde busca-se predizer a qual classe predeterminada um dado pertence. Podendo ainda ser dividido entre:

  - Binário: quando se tem apenas duas classes possíveis.
  - Multi classes: quando se tem mais de duas classes possíveis.

- Regressão: na regressão, busca-se predizer, ao invés de classes, um valor contínuo ou variável numérica dentro de um determinado intervalo.

Ao montar um modelo a partir de um conjunto de dados de treinamento e que seja igualmente preciso em um conjunto de dados novo, pode-se dizer que o modelo é capaz de ser generalizado. Um modelo complexo é capaz de realizar melhores predições em um conjunto de dados de treinamento, porém não é generalizável.

Existem diversos algoritmos de aprendizado de máquina supervisionados que podem ser explorados afim de realizar predições. Abaixo serão explicitados aqueles utilizados no projeto.

2.2.1.1 K Vizinhos mais Próximos - *K-Nearest Neighbors* (KNN)

Este algoritmo guarda as instâncias de todos os dados de treinamento e a partir disso analisa a classe de um número arbitrário **k** de pontos próximos nomeados "vizinhos próximos"a entrada realizada, a classe presente mais frequente é atribuída à entrada em questão. Para determinar a distância entre os pontos, o algoritmo usa por padrão utilizado pela biblioteca *Sklearn*, explicada na seção 3.2, é a distância Minkowski **(2.2)**, mas pode-se utilizar ainda a distância de Manhattan **(2.3)**, Euclidiana **(2.1)**, entre outras.

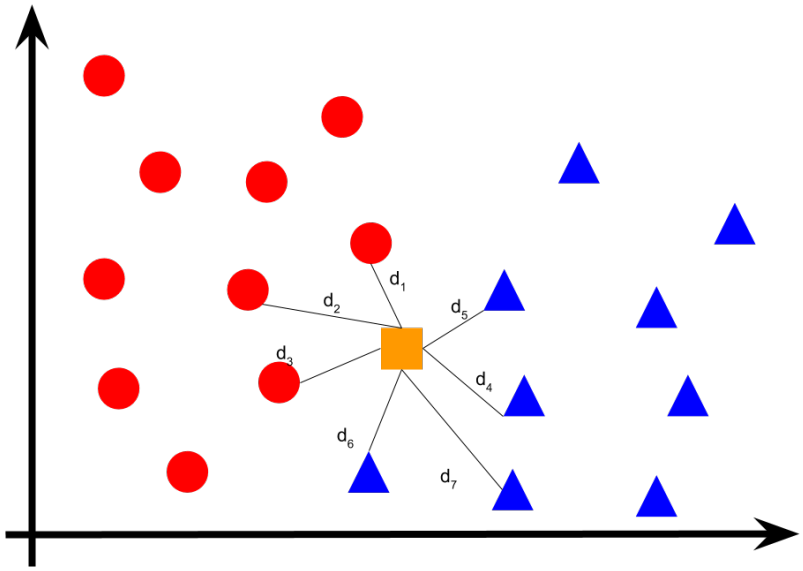
$$d_e(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)} \tag{2.1}$$

$$d_{mink}(x,y) = \sum_{i=1}^{n}(|x-y|^p)^(1/p) \tag{2.2}$$

$$d_{manh}(x,y) = \sum_{i=1}^{n}(|x_i - y_i|) \tag{2.3}$$

No exemplo criado abaixo e representado pelo gráfico de dispersão na Figura 5, existem duas classes distribuídas, os círculos e os triângulos, e uma entrada a qual deseja-se prever a classe a qual esta pertencerá, representada pelo quadrado, utilizando o método de knn para realizar a predição sendo o valor de vizinhos a serem analisados equivalente à três. Supondo que ao calcular as distâncias mais próximas, tem-se a seguinte sequência: $d_1 < d_5 < d_3 < d_4 < d_2 < d_6 < d_7$. Os pontos a serem considerados para a tomada de decisão serão as das distâncias d_1, d_5 e d_3, ou seja, as três menores distâncias, sendo as respectivas classes: círculo, triângulo e círculo, portanto, a classe de saída para a entrada quadrado analisando os três vizinhos mais próximos é equivalente a círculo.

Figura 5 – Gráfico de distribuição de classes.



Fonte:Elaborado pela autora

Para o problema em questão é importante notar que caso a quantidade de vizinhos fosse quatro, ocorreria um empate das classes presentes e se fosse a quantidade de vizinhos fosse equivalente a sete a classe prevista seria equivalente a classe triângulo. A principal vantagem da utilização do método é a sua fácil assimilação e implementação, além de que sua performance que tende a ser aceitável em conjunto de dados que não sofrem muitas alterações ou que não sejam significantemente grande (MÜLLER; GUIDO et al., 2016). A dificuldade, entretanto, reside no fato de ser difícil atribuir um valor ideal para a quantidade de vizinhos próximos a serem analisados, visto que se este for muito grande a classificação pode somente seguir a maioria presente no *dateset* e se for muito pequeno pode resultar em uma classificação imprecisa (RODRÍGUEZ et al., 2007).

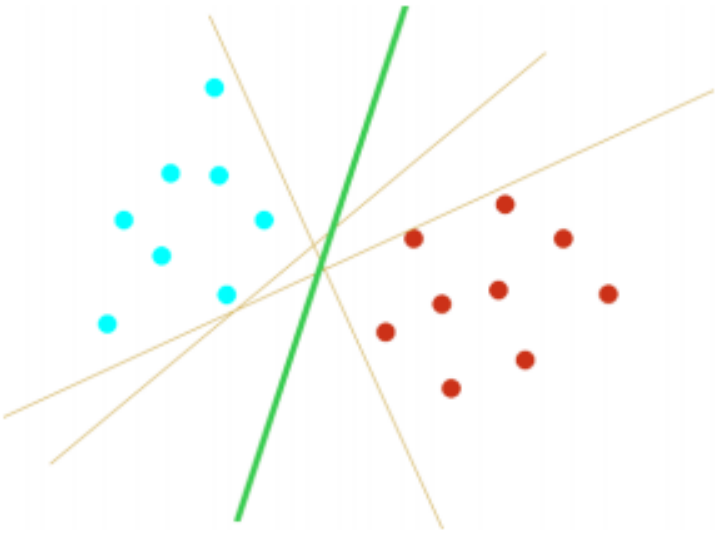
2.2.1.2 Máquinas de Vetores de Suporte - *Support Vector Machines* (SVM)

Para utilizar este algoritmo, os dados são colocados em um gráfico de dispersão a fim de se ter uma melhor noção da distribuição das classes. Existem, inicialmente, infinitos hiperplanos capazes de separar as classes mas o SVM seleciona o hiperplano ótimo para realizar essa separação de classes.

Para encontrar o hiperplano ótimo indicado em verde na Figura 6, busca-se a distância máxima, conhecida como margem, entre cada classe e o hiperplano em questão e aquele que apresenta menor erro na divisão das classes. Para realizar uma predição, é avaliado a classe associada a região que a entrada a ser predita se encontra.

No exemplo criado abaixo, assim como o exemplo dado do knn, existem duas classes

Figura 6 – Gráfico de hiperplanos em problema de classificação binário.



Fonte: Gunn et al. (1998)

dispersas no gráfico na Figura 7, círculo e triângulo. Utilizando o método de SVM, encontra-se o hiperplano que possui a maior distância dos pontos mais próximos de cada classe. No caso, o hiperplano ótimo é representado pelo traço cheio em verde e a distância máxima é representada por **m**, realizando desta forma a divisão de que a região superior do hiperplano pertence à classe círculo e a região inferior à classe triângulo.

A entrada que busca-se predizer a classe é representada pelo quadrado. No exemplo em questão, pela divisão realizada, pode-se dizer que a classe do quadrado é triângulo.

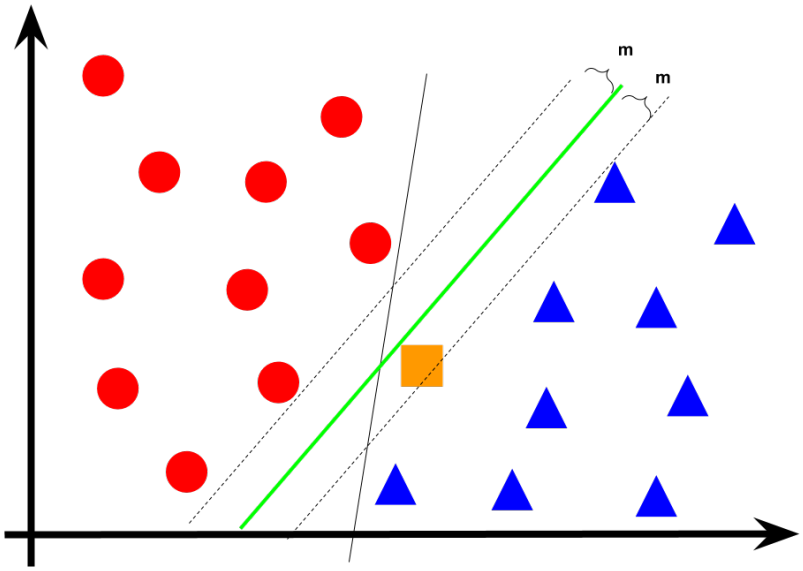
2.2.1.3 Floresta Aleatória

Floresta Aleatória é um método que cria uma combinação aleatória de árvores de decisão binárias para poder realizar a tarefa de classificação ou regressão. A busca pelas melhores características é feita em subconjuntos aleatórios de características sendo que cada árvore tenta prever um rótulo, o que resulta em uma maior diversidade de possibilidades (KHALILIA; CHAKRABORTY; POPESCU, 2011). A tarefa de predição é realizada avaliando as classes que cada árvore gerou para a entrada, a com maior frequência é a que será atribuída à entrada.

A estrutura base do algoritmo é apresentada na Figura 8, mas é difícil criar uma generalização para ele visto que a forma com que as árvores de decisão dentro da floresta aleatória são formadas depende do contexto e dos dados que possuem, o que pode ser predeterminado são as quantidades de árvores de decisão que irão compor o método.

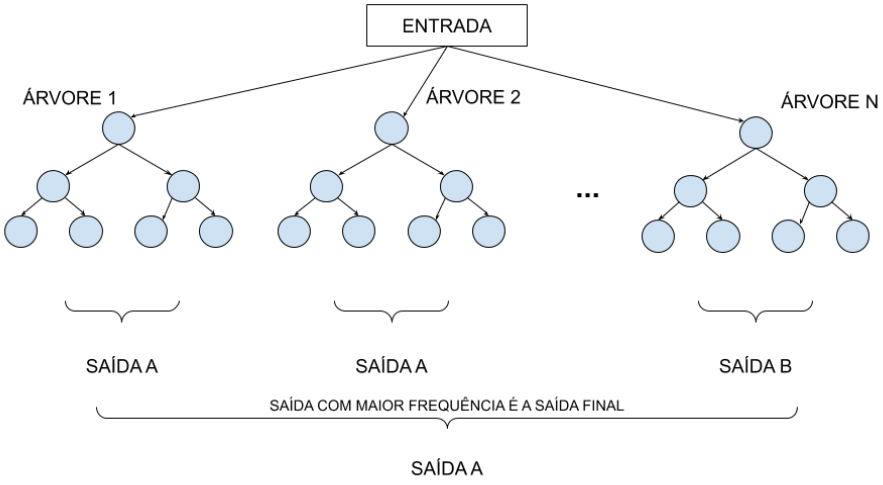
No exemplo criado abaixo, assim como nos exemplos dos métodos acima, existem duas classes, círculo e triângulo. Utilizando o método de florestas aleatórias com três árvores de decisões, uma estrutura possível seria a demostrada na Figura 9 . Pelas saídas fornecidas nas

Figura 7 – Gráfico de hiperplanos em problema de classificação círculo e triângulo.



Fonte: Elaborado pela autora

Figura 8 – Estrutura de uma Floresta Aleatória.



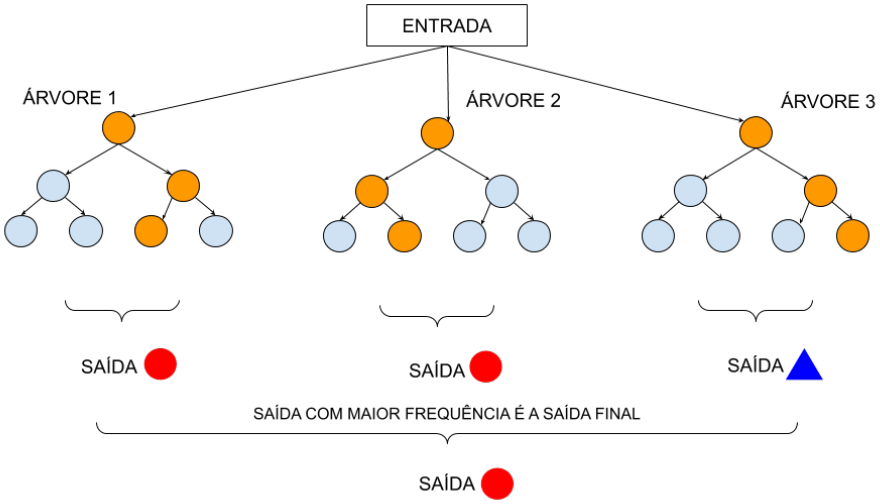
Fonte:Elaborado pela autora

árvores de decisão, a classe com maior frequência e atribuída a entrada é círculo.

2.3 Métricas de desempenho e qualidade

As métricas de desempenho e qualidade surgem com o propósito de garantir a qualidade e auxiliar a encontrar os métodos e dados mais adequados para o problema em questão.

Figura 9 – Estrutura da Floresta Aleatória exemplo para predição entre classe círculo ou triângulo.



Fonte:Elaborado pela autora

2.3.1 Matriz de Confusão

A matriz de confusão de uma hipótese **h** mostra uma medida efetiva do modelo de classificação pois mostra um número de classificação correta e as classificações preditivas para cada classe, sob um conjunto de exemplos T. Os resultados são totalizados em duas dimensões: classes preditivas e classes verdadeiras (MONARD; BARANAUSKAS, 2003).

Em problemas de classificação binária, a matriz de confusão a ser estruturada fica como a apresentada na Figura 10, na diagonal principal se tem os casos de acerto, onde a classe prevista é a mesma que a rotulada, chamados de verdadeiro positivo (TP) e verdadeiro negativo (TN); e na diagonal secundária encontram-se os casos de erro, onde a classe prevista não é a mesma que a rotulada, resultando os falso positivos (FP) e falso negativos (FN) (Diego Nogare, 2020).

Figura 10 – Matriz de confusão classificação binária.

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

Fonte:Diego Nogare (2020)

Existem diversas métricas criadas para avaliar o desempenho dos algoritmos de aprendi-

zado de máquina, os quatro utilizados neste trabalho são:

- Precisão: esta é uma medida de fidelidade, mostrando a taxa de que os elementos classificados como positivos são realmente positivos (MATOS et al., 2009). O cálculo desta métrica é realizada por meio de da seguinte divisão:

$$Prec = \frac{TP}{TP + FP} \tag{2.4}$$

- Revocação: esta por sua vez, é uma medida de completude, mostrando a taxa de acertos de verdadeiros, tanto positivo quanto negativo, indicando o total de informação relevante recuperada (MATOS et al., 2009). Esta métrica segue a seguinte função:

$$Revoc = \frac{TP}{TP + FN} \tag{2.5}$$

- Acurácia: esta é a métrica que avalia a taxa de acerto dos métodos, mas nem sempre pode ser o ideal para direcionar avaliações visto que depende muito do balanceamento dos dados (PRATI; BATISTA; MONARD, 2008). O cálculo para esta função:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

- *F1 Score*: esta métrica é a média harmônica ponderada entre a precisão e a revocação, sendo derivada por van Rijsbergen (1979) baseada na medida de eficiência(MATOS et al., 2009). Calculada pela função:

$$F_1Score = \frac{2.Prec.Revoc}{Prec + Revoc} \tag{2.7}$$

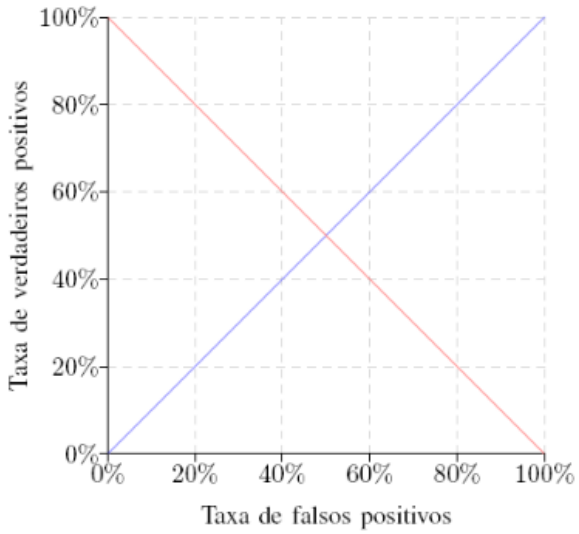
2.3.1.1 Curva ROC

Desenvolvido para avaliar o desempenho de classificadores, os gráficos da técnica de *Receiver Operating Characteristics* (ROC) permitem a visualização da taxa de acerto e erro dos classificadores (FAWCETT, 2006). Os gráficos da curva ROC são gerados no que se é conhecido como espaço ROC, mostrado na Figura 11, um espaço bidimensional que possui em seu eixo **y** a taxa de TP e em seu eixo **x** a taxa de FP.

Como o gráfico da curva ROC é apenas uma forma de mostrar a performance do classificador em questão, para que seja possível realizar comparações é preciso de um valor numérico à se comparar. Uma forma de se chegar a esse valor é avaliando a área abaixo da curva (AUC[4]). Como as grandezas do espaço ROC são em percentuais, criando um quadrado unitário, o valor da AUC pode variar entre 0,0 e 1,0. Sabendo que um classificador aleatório

[4] *Area Under Curve* (PRATI; BATISTA; MONARD, 2008)

Figura 11 – Espaço ROC.



Fonte: Prati, Batista e Monard (2008)

não é plotável no espaço ROC e sua área é equivalente a 0,5, as AUC encontradas são sempre maiores que 0,5 (PRATI; BATISTA; MONARD, 2008).

O critério para decidir qual o melhor classificador é aquele que tiver uma maior AUC, sendo considerado como o que possui uma melhor performance média, mas é possível que em algumas partes do gráfico um classificador seja melhor que outro mesmo que no contexto total o segundo seja considerado o melhor.

3 Metodologia

A metodologia envolvida para o desenvolvimento se inicia com uma revisão bibliográfica para entender a área e o problema em questão. Em seguida é realizado a busca por um *dataset* que possua o máximo de características relacionadas com o problema e o tratamento desses dados.

Após selecionar o *dataset*, é feito a análise e tratamento dos dados, verifica-se as características que os dados de cada atributo possui, a distribuição que apresentam, formas de lidar com dados faltantes e os *outliers*[1] que aparecem.

Por fim, busca-se a melhor forma e o melhor modelo para realizar a predição de classe e a análise dos resultados obtidos pelos modelos utilizados.

3.1 Base de Dados

Os dados utilizados nessa pesquisa foram retirados da plataforma *Github* da *University of British Columbia* encontrado no repositório *dsci100data*[2] destinado aos alunos do curso de Ciências de Dados oferecido pela universidade.

Cada registro da base de dados representa uma pessoa que foi entrevistada e possui as seguintes informações e tipos associados:

- Ordinal: chave de identificação (id);
- Binário:
  - Sexo da pessoa:
    1. Feminino;
    2. Masculino.
  - Possuidor de Hipertensão:
    1. Sim;
    2. Não.
  - Possuidor de doença do coração:
    1. Sim;
    2. Não.

[1] *Outliers* são dados que estão fora dos padrões em relação ao *dataset* em que se encontram. (MÜLLER; GUIDO et al., 2016)

[2] <https://github.com/UBC-DSCI/dsci100data/blob/master/raw_data/train_2v.csv>

  - Se a pessoa é casada:
    1. Sim;
    2. Não.
  - Sofreu um AVC:
    1. Sim;
    2. Não.

- Categórico:
  - Tipo de trabalho:
    1. Trabalha para empresa privada;
    2. Trabalha como autônomo;
    3. Trabalha para o governo;
    4. Nunca trabalhou;
    5. Criança.
  - *Status* de fumante:
    1. Fumante;
    2. Ex-fumante;
    3. Nunca fumou.

- Numérico: idade, média do nível de glicose, índice de massa corporal pelo padrão americano(BMI).

A Figura 12 ilustra melhor como os dados estão estruturados.

Figura 12 – Dados de vítimas de AVC

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

O *dataset* conta com 43.400 entradas no total, sendo 783 delas de pessoas que sofreram AVC, aproximadamente 2% do *dataset*, e 42.617 que não sofreram AVC, aproximadamente 98% do *dataset*. Na busca de tornar o *dataset* viável para uso em um método de aprendizado de máquina, foi criado um *subdataset* composto por uma quantidade **q** de pessoas que sofreram AVC e uma quantidade **2q** de pessoas que não sofreram AVC.

Para compor esse *subdataset*, foi separado o *dataset* original em dois, um apenas com entradas onde *stroke* fosse equivalente a 1, nomeado A, e outro com *stroke* equivalente a 0, nomeado B. Em seguida, foi selecionado e concatenado o *dataset* A completo com 783 entradas com as 1566 primeiras entradas do *dataset* B. Para que os dados não ficassem em blocos e enviesassem os resultados, o *dataset* final foi ordenado por meio da coluna de identificação **id**. O resultado foi um *subdataset* com 783 pessoas que sofreram AVC (34%) e 1566 pessoas que não sofreram AVC (66%), totalizando um *dataset* de 2349 entradas.

3.2 Ferramentas de Desenvolvimento

Para a realização desse projeto foram utilizadas as seguintes ferramentas de desenvolvimento.

3.2.1 Python

Criada em 1990 por Guido Van Rossum, *Pyhton*[3] é uma linguagem de programação de alto nível, orientada a objeto, com tipagem dinâmica e forte, interpretada e interativa. Um *software* de código aberto com licença compatível com a *General Public License* (GPL). As especificações da linguagem é mantida pela *Python Software Foundation* (PSF) (BORGES, 2014). A linguagem foi escolhida devido as bibliotecas (melhores descritas na próxima secção) que possui e que oferecem maior auxílio para as tarefas envolvidas pelo projeto.

3.2.1.1 Bibliotecas Pyhton

Abaixo serão discriminadas as bibliotecas utilizadas no desenvolvimento do projeto.

- *Scikit-learn*: atualmente na versão 0.23.2, esta é uma biblioteca em Python construída a partir das bibliotecas *Numpy, Scipy e Matplot*, voltada para Aprendizado de Máquina, sendo simples e eficiente para tarefas de predição visto que conta com diversas funções e métodos já implementados. Os métodos apresentados no site oficial[4] da biblioteca estão divididos em mais de 20 categorias, mas entre as principais estão: classificação, regressão, clusterização, redução de dimensionalidade, seleção de modelos, entre outros.

  Além disso, se tratando de métodos de aprendizado de máquina supervisionados como os utilizados neste projeto, a biblioteca conta com 12 tipos de modelos, incluindo um RNA, e suas variações.

- *Pandas*: A biblioteca *Pandas* tem seu nome derivado de *panel data* e é responsável por permitir a análise de dados utilizando *Python*; isso se deve à dois objetos principais:

[3] Site oficial: <https://www.python.org/>
[4] <https://scikit-learn.org/stable/>

DataFrames e *Series*. Os *DataFrames* são estruturas de dado tabular, orientadas à rótulos tanto para as linhas quanto para as colunas. As *Series* por sua vez são objetos do tipo *array* unidimensional que possuem rótulos (MCKINNEY, 2019).

- *Plotly*: A biblioteca Plotly é uma biblioteca para ciências de dados, podendo ser utilizada para o desenvolvimento de aplicativos, criar gráficos e figuras interativas e auxiliar na construção e execução de aplicações que utilizem IA.

- *Jupyter Notebook*: O *Jupyter* foi criado em 2014, sendo uma ferramenta de processamento interativo e independente de linguagem, atualmente contando com suporte para mais de 40 linguagens de programação. Além de oferecer um bom suporte para *Python*, o sistema de *notebooks Jupyter* permite a criação de *Markdowns* e conteúdos em HTML para maior organização (MCKINNEY, 2019) .

  É possível realizar todos os passos de desenvolvimento nos *notebooks Jupyter*, execução, depuração e testes de código.

4 Desenvolvimento

Nesta parte do documento, busca-se apresentar o processo e organização dos dados para realizar os experimentos propostos para esse projeto. Esta é uma fase essencial para viabilizar a realização das análises pois busca eliminar redundâncias, tratamento de valores categóricos e contínuos. O objetivo é possuir um *dataset* que possa proporcionar um conteúdo com informações no formato necessário para a análise dos algoritmos.

4.1 Análise dos Dados

Após separar o *subdataset*, foi elaborado duas tabelas: a primeira, ilustrada na Tabela 1 abaixo, apresenta aspectos como porcentagem de dados ausentes, quantidades de categorias existentes, moda e segunda moda, frequências das modas e a porcentagem da presença das modas dos atributos do tipo categóricos.

Tabela 1 – Tabela de informações dos atributos categóricos.

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.

A partir de Tabela 1, pode-se destacar os seguintes pontos quanto aos dados categóricos:

- Apesar dos dados de hipertensão e presença de doença do coração estarem desbalanceadas, pelas pesquisas bibliográficas realizadas, ambas são fatores de risco relacionados à doença, não sendo possível desconsiderá-las.

- Os dados de Status de Fumante possuem uma quantidade significativa de dados não informados de 611 pessoas (26,01%) não sendo possível apenas excluir todas as entradas que não possuem este atributo informado.

Logo na segunda Tabela 2, além de mostrar a porcentagem de dados ausentes, apresenta-se também informações como o máximo e o mínimo, média, mediana, desvio padrão e o primeiro e terceiro quantil[1], de atributos contínuos.

Tabela 2 – Tabela de informações dos atributos contínuos.

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.

Ao analisarmos a Tabela 2, pode-se notar que:

- Apesar de não se ter valores ausentes, no atributo de idade há a presença de *outliers*[2], como por exemplo o registro de idade equivalente a 0.8. Existem diversas formas de lidar com *outliers*, sendo possível excluir todas as entradas que as possuem (dependendo da porcentagem de presença), substituir todos os *outliers* por valores padrão como média, mediana ou moda, ou ainda realizar uma análise de cada caso presente, verificando se trata-se de erro humano, computacional ou um dado real.
- O atributo de BMI apresenta 8,34% das suas entradas faltantes, assim como a hipertensão e a presença de doença do coração o BMI também é um fator de risco diretamente ligado a doença, não sendo possível simplesmente eliminar as entradas que possuem tal atributo nulo.

4.2 Tratamento dos Dados

Após feita a análise dos dados e ter conhecimento dos principais pontos que inviabilizam o uso direto do *subdataset* nos métodos propostos de aprendizado de máquina, para contornar os problemas expostos na sessão acima foram propostas as seguintes decisões.

[1] Um quartil vem a ser um valor que divide um conjunto ordenado de dados em quatro partes iguais.
[2] Valores apresentados que são muito diferentes dos apresentados na série em questão, tratando-se de uma inconsistência.

Tabela 3 – Tabela de distribuição atributos categóricos por vítimas de AVC ou não.

AVC	QUANTIDADE	CATEGORIA	QUANTIDADE	AVC
0	917	Feminino	431	1
	649	Masculino	352	
	142	Hipertensão (1)	200	
	1424	Hipertensão (0)	583	
	76	D. Coração (1)	177	
	1490	D. Coração (0)	606	
	1017	Casado	703	
	549	Não Casado	80	
	937	T. Privado	441	
	229	Autônomo	251	
	211	Criança	2	
	182	T. Governamental	89	
	7	Nunca trabalhou	0	
	773	Urbano	399	
	793	Rural	384	
	240	Fumante	133	
	261	Ex- fumante	221	
	599	Nunca fumou	284	

Fonte: Elaborado pela autora.

A partir da Tabela 3, onde é possível ver como os dados categóricos estão distribuídos com relação ao atributo de AVC, sobre a quantidade de dados não informados de *status* de fumantes, decidiu-se que os campos nulos seriam considerado também uma categoria. Passa-se a ter, portanto, 4 categorias para este atributo: não fumantes, ex-fumantes, fumantes e não informados, distribuídos conforme apresentado na Tabela 4.

Ainda com relação aos atributos categóricos, quando se trata de métodos de aprendizado de máquina, estes apenas trabalham com números e algumas podem não ser boas para dados categóricos por não compreender categorias como as presentes em gênero (feminino e masculino) ou tipo de residência (urbano e rural). Além disso, para atributos em que a categoria é representada por algarismos, como em hipertensão e doenças do coração (1 e 0) os algoritmos interpretam os algarismos como números, interpretando que ter hipertensão é maior que não ter (1>0), por exemplo.

Para solucionar este problema, primeiro utiliza-se o método *LabelEncoder* para que

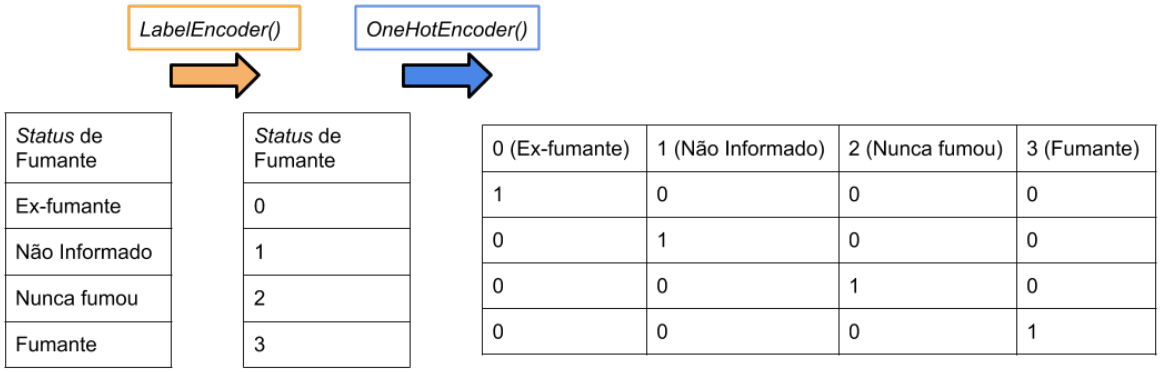
Tabela 4 – Tabela de distribuição da categoria *Status* de fumante por vítimas de AVC ou não.

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.

todas as categorias utilizadas sejam representadas por algarismos. Em seguida, aplica-se o *OneHotEncoder* responsável por transformar uma coluna em múltiplas colunas em que as entradas podem ser preenchidas com 1 ou 0, representando se pertence ou não a determinada categoria. Um exemplo das modificações que acontecem com os dados ao aplicar esses dois métodos pode ser visto na Figura 13.

Figura 13 – Aplicação de *LabelEncoder* e *OneHotEncoder* no atributo de *Status* de fumante.

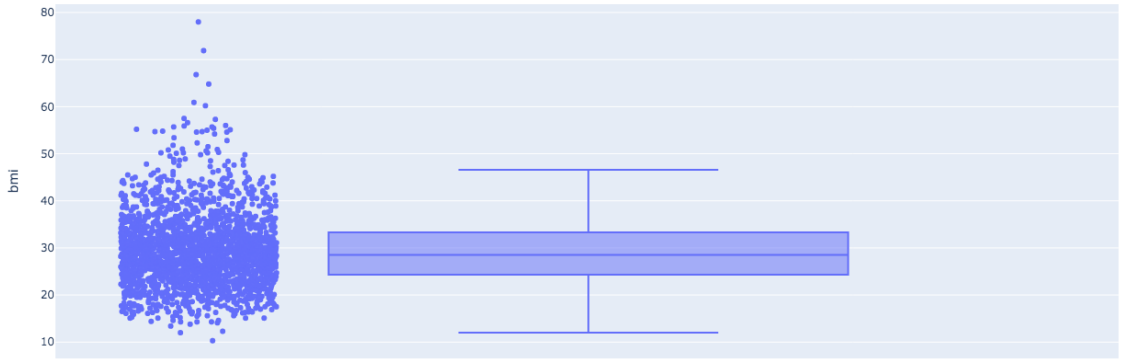


Fonte: Elaborado pela autora.

Quanto aos atributos contínuos, ao analisar os valores contidos no *boxplot* do atributo BMI na Figura 14, é possível notar que em sua grande maioria encontram-se dentro de uma faixa de valores fixa, apesar da grande variância entre si. Por conta disso, preencher os dados faltantes dessa categoria com valores como zero, máximo ou o mínimo poderiam enviesar significativamente os resultados dos métodos de aprendizado de máquina. Desta forma, os

dados foram preenchidos com o valor da mediana da categoria, equivalente a 28.5 ,ou seja, uma pessoa com sobrepeso (NIHISER et al., 2007).

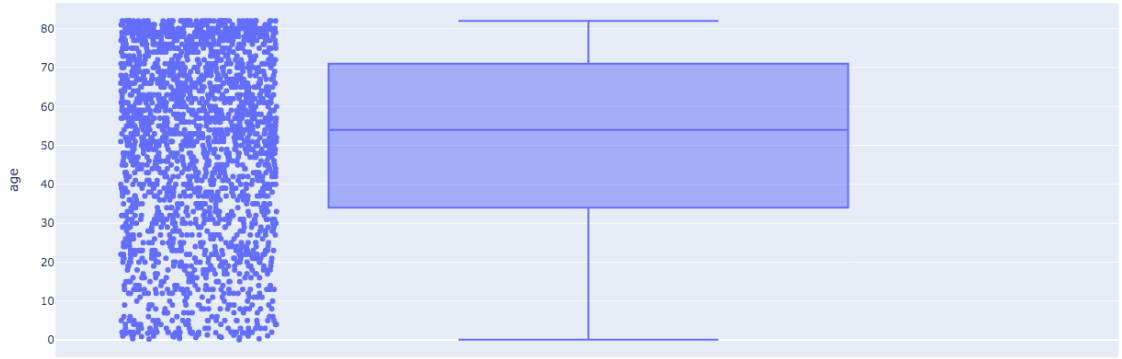
Figura 14 – Gráfico de Boxplot de BMI



Fonte: Elaborado pela autora.

Ao analisar as idades, é possível ver pela Figura 15 os *outliers* citados. Existem 32 valores que distam dos demais e as idades encontram-se no intervalo entre 0 e 2, como por exemplo 0,64 e 1,74. Pela forma com que a entrada aparece, provavelmente esses *outliers* foram criados por erro humano, por conta disso, as idades dessas entradas foram substituídas pelo valor anterior multiplicado por 100 e para casos em que o valor fosse maior que 100 foi retirado o primeiro algorismo, resultando em entradas como por exemplo 88 e 74 anos.

Figura 15 – Gráfico de Boxplot de Idade.



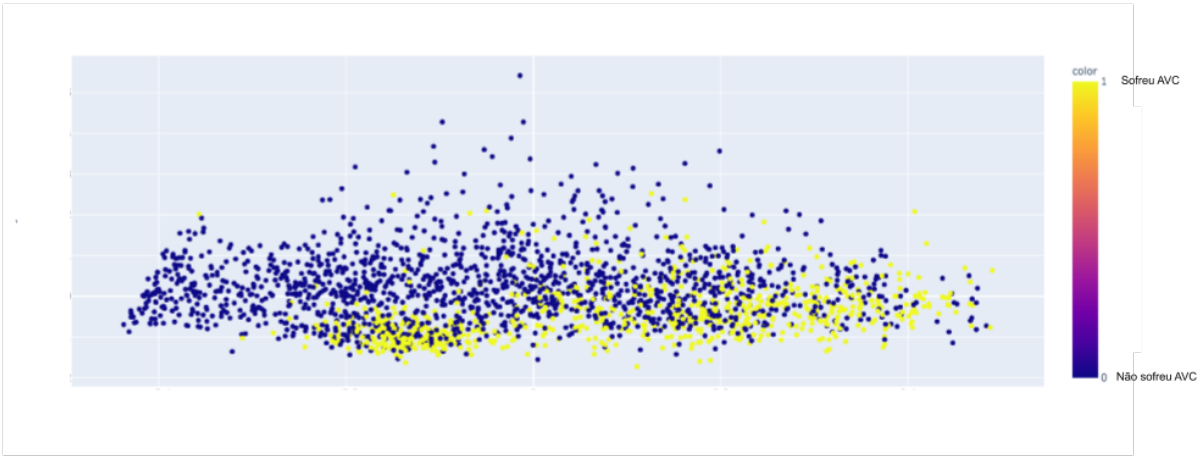
Fonte: Elaborado pela autora.

Por fim, visto que os atributos possuem grande diferenças de intervalo entre si, o *dataset*

foi normalizado com o objetivo de deixá-lo em uma escala comum mantendo as diferenças entre os intervalos de valores, utilizando o método *Normalizer()* do *sklearn.preprocessing*. [3]

Para entender a forma com que os dados ficaram distribuídos com relação a categoria a ser predita, o AVC, onde serão aplicados os algoritmos, foi gerado a Figura 16, um gráfico de dispersão do atributo em questão. As entradas referente a pessoas que sofreram AVC encontram-se de amarelo e os que não sofreram AVC de azul.

Figura 16 – Gráfico de Dispersão AVC.



Fonte: Elaborado pela autora.

4.3 Treinamento dos Modelos

Após o tratamento dos dados apresentado acima, os dados foram divididos 25% para teste e 75% para treinamento, equivalente a 1761 entradas para treinamento e 588 para teste.

Para cada um dos modelos foi realizada a instanciação do método a ser utilizado, preenchendo os parâmetros necessários como a quantidade de vizinhos no KNN, a profundidade da das árvores e a quantidade de estados aleatórios nas florestas aleatórias e a permissão para uso de métodos de probabilidade dentro da SVM.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.htmlsklearn.preprocessing.normalize

5 Resultados

Esta vem a ser a última fase deste trabalho, onde avalia-se os resultados obtidos da aplicação dos métodos de aprendizado de máquina no *dataset* com tratamentos a partir das métricas citadas na **sessão 2.3** deste documento.

As primeiras métricas para se avaliar são as relacionadas a matriz de confusão geradas pelos métodos, como apresentado na Figura 17.

Figura 17 – Conjunto de Matrizes de Confusão.

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

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

SVM		
	SEM AVC	AVC
SEM AVC	345	66
AVC	68	108

Fonte:Elaborado pela autora

A partir das matrizes de confusão apresentadas acima, tem-se a Tabela 5 , que apresenta as métricas de avaliação propostas para este trabalho.

Apresentando um percentual de 100% de acerto de TP e TN, o método de Floresta Aleatórias possui os valores de acurácia, precisão média, *recall* médio e *F1-Score* médio equivalentes a 100%. Mostrando-se o melhor método entre os três propostos. Em seguida tem-se o método de KNN, com um percentual de 67% e 83% de acertos de TP e TN, respectivamente, apresentando métricas de acurácia, precisão média, *recall* médio e *F1-Score* médio, respectivamente, 78%, 74%, 75% e 74,5%. O SVM por sua vez apresentou um percentual de acerto de TP e TN equivalente à 61% e 83%, respectivamente; quanto as métricas de acurácia, precisão média, *recall* médio e *F1-Score* médio apresentou 77%, 73%, 72,5% e 73%, respectivamente.

Como forma de complementação à análise das métricas da matriz de confusão, foi

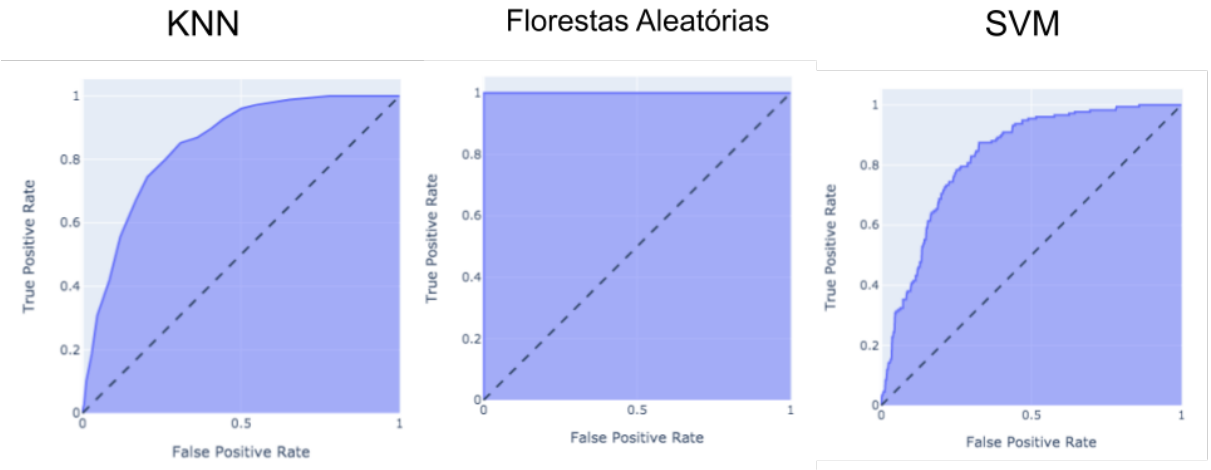
Tabela 5 – Métricas de Avaliação

	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.

proposta a análise da AUC gerada pelos métodos. Por meio da geração dos gráficos da curva *ROC* no espaço *ROC* apresentado na Figura 18, é possível confirmar o apresentado pelas métricas da Tabela 5.

Figura 18 – Gráfico de AUC ROC para KNN



Fonte:Elaborado pela autora

Pela forma de decisão utilizada para esta métrica, onde o método com a maior AUC é considerado o melhor para a tarefa de predição proposta. Como os valores de AUC gerados pelos métodos de KNN, SVM e Florestas Aleatórias são, respectivamente, 0,84 , 0,82 e 1,0. O método de aprendizado de máquina que pode ser considerado melhor para a tarefa de predição

de casos de AVC precoce, a partir do *dataset* apresentado, é o Florestas Aleatórias, com valor de AUC equivalente a 1,0, o valor máximo possível, representando que em todas as predições realizadas pelo método, este obteve sucesso.

6 Conclusão

Este trabalho de conclusão de curso buscou realizar um estudo de métodos de aprendizado de máquina com o objetivo de predizer as possibilidades de uma pessoa sofrer um AVC precoce, sendo que foi realizado uma pesquisa bibliográfica na área de previsão utilizando métodos de aprendizado de máquina na saúde para base teórica do trabalho.

Foi realizado um estudo sobre os Acidentes Vasculares Cerebrais com o objetivo de entender mais sobre a doença e sua fisiopatologia, os fatores de risco englobados e o comportamento dos dados para se ter formas de analisá-los. Em seguida, foram propostos três modelos de aprendizado de máquina para a tarefa de previsão.

Os resultados obtidos refletem que o melhor método a ser utilizado para esta tarefa é a de Florestas Aleatórias com uma acurácia de 100% e uma AUC equivalente a 1, mostrando que a partir do treino realizado, o método não apresentou erros nos testes. Contudo, é importante ressaltar que este trabalho foi realizado com base em um *dataset* público e sem documentação sobre a forma com que este foi elaborado, destinado à alunos de um curso universitário. Apesar de terem sido realizados tratamentos dos dados para tornar o *dataset* viável para utilização nos métodos de aprendizado de máquina, ainda faltam informações de fatores de risco que em vida real podem ter impacto significativo na ocorrência ou não da doença.

Por fim, conclui-se que o trabalho tem sua relevância no meio em que se encontra visto a alta taxa de ocorrência, a quantidade de pessoas que morrem ou ficam inválidas por conta dela e os elevados custos que trazem para o Estado, os pacientes e suas famílias, mas como esta diretamente relacionada a medicina e ao ser humano, é necessário que os resultados possuam uma maior confiabilidade e as decisões a serem tomadas estejam sempre acompanhadas de um especialista da área, não servindo como forma de auto-diagnosticação.

Para trabalhos futuros, fica interessante a realização das análises em uma base de dados que contenha mais atributos ligados aos fatores de risco da doença e que se tenha um histórico da forma com que os dados são coletados.

Referências

ACADEMIA BRASILEIRA DE NEUROLOGIA. *AVC ou Derrame Cerebral*. 2015. <http://www.cadastro.abneuro.org/site/publico_avc.asp>. Online; Acesso em: 10 de Março de 2020.

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

BORGES, L. E. *Python para desenvolvedores: aborda Python 3.3*. [S.l.]: Novatec Editora, 2014.

CONSELHO FEDERAL DE MEDICINA. *Resolução nº 1638/02*. 2002. <https://www.cremesp.org.br/?siteAcao=Revista&id=435>. Online; Acesso em: 26 de Julho de 2020.

DHB VI. *VI Diretrizes Brasileiras de Hipertensão*. 2010. <http://publicacoes.cardiol.br/consenso/2010/Diretriz_hipertensao_associados.pdf>. Online; Acesso em: 5 de Agosto 2020.

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.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

GOTTLIEB, M. G.; BONARDI, G.; MORIGUCHI, E. H. Fisiopatologia e aspectos inflamatórios da aterosclerose. *Scientia Medica*, v. 15, n. 3, p. 203–7, 2005.

GUNN, S. R. et al. Support vector machines for classification and regression. *ISIS technical report*, v. 14, n. 1, p. 5–16, 1998.

KHALILIA, M.; CHAKRABORTY, S.; POPESCU, M. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, Springer, v. 11, n. 1, p. 51, 2011.

KOVÁCS, Z. L. *Redes neurais artificiais*. [S.l.]: Editora Livraria da Fisica, 2002.

LIMA, M. R.; PAGLIOLI, R.; FILHO, J. R. H. Diagnóstico por imagem do acidente vascular encefálico. *Acta méd.(Porto Alegre)*, p. 9–9, 2012.

LOTUFO, P. A.; GOULART, A. C.; PASSOS, V. M. d. A.; SATAKE, F. M.; SOUZA, M. d. F. M. d.; FRANÇA, E. B.; RIBEIRO, A. L. P.; BENSENÕR, I. J. M. Doença cerebrovascular no brasil de 1990 a 2015: Global burden of disease 2015. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 20, p. 129–141, 2017.

MARIA ELISABETH FERRAZ. *Opinião: AVC é a segunda causa de mortalidade no Brasil*. 2019. <https://www.unifesp.br/reitoria/dci/releases/item/4108-avc-e-a-segunda-causa-de-mortalidade-no-brasil>. Online; Acesso em: 23 de Junho de 2020.

MATOS, P. F.; LOMBARDI, L. d. O.; CIFERRI, R. R.; PARDO, T. A.; CIFERRI, C. D.; VIEIRA, M. T. Relatório técnico "métricas de avaliaçao". *Universidade Federal de Sao Carlos*, 2009.

MCKINNEY, W. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. [S.l.]: Novatec Editora, 2019.

MINISTÉRIO DA SAÚDE. *Estratégia e-saúde para o Brasil*. 2002. <https://saudedigital.saude.gov.br/wp-content/uploads/2020/02/Estrategia-e-saude-para-o-Brasil_CIT_20170604.pdf>. Online; Acesso em: 03 de Agosto de 2020.

MINISTÉRIO DA SAÚDE. *Trombose: causas, sintomas, diagnóstico, tratamento e prevenção*. 2013. <https://saude.gov.br/saude-de-a-z/trombose-causas-sintomas-diagnostico-tratamento-e-prevencao>. Online; Acesso em: 05 de Agosto de 2020.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, Manole Ltda, v. 1, n. 1, p. 32, 2003.

MOURÃO, A. D.; NEVES, J. d. R. Impactos da implantação do prontuário eletrônico do paciente sobre o trabalho dos profissionais de saúde da prefeitura municipal de belo horizonte. *Anais do Simpósio de Excelência em Gestão e Tecnologia*, p. 22–24, 2007.

MÜLLER, A. C.; GUIDO, S. et al. *Introduction to machine learning with Python: a guide for data scientists*. [S.l.]: "O'Reilly Media, Inc.", 2016.

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.

NORRVING, B. *Oxford textbook of stroke and cerebrovascular disease*. [S.l.]: Oxford University Press, USA, 2014.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. *10 principais causas de morte no mundo*. 2015. <https://www.paho.org/bra/index.php?option=com_content&view=article&id=5638:10-principais-causas-de-morte-no-mundo&Itemid=0>. Online; Acesso em: 8 de Março de 2020.

PEREIRA, C. R.; PEREIRA, D. R.; SILVA, F. A.; MASIEIRO, J. P.; WEBER, S. A. T.; HOOK, C.; PAPA, J. P. A new computer vision-based approach to aid the diagnosis of parkinson's disease. *Computer Methods and Programs in Biomedicine*, Elsevier North-Holland, Inc., New York, NY, USA, v. 136, p. 79–88, 2016.

PRATI, R.; BATISTA, G.; MONARD, M. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008.

RODRÍGUEZ, J. E. R.; BLANCO, E. A. R.; CAMACHO, R. O. F. et al. Clasificación de datos usando el método k-nn. *revista Vínculos*, v. 4, n. 1, p. 4–18, 2007.

SANARFLIX. *Resumo de AVC: definição, fatores de risco, fisiopatologia, manifestações e mais*. 2019. <https://www.sanarmed.com/resumos/avc-definicao-fatores-fisiopatologia-manifestacoes>. Online; Acesso em: 10 de Março de 2020.

SANTOS, A. M. d.; SEIXAS, J. M. d.; PEREIRA, B. d. B.; MEDRONHO, R. d. A. Usando redes neurais artificiais e regressão logística na predição da hepatite a. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 8, p. 117–126, 2005.

SANTOS, C.; Afonso, L.; Pereira, C.; Papa, J. Breastnet: Breast cancer categorization using convolutional neural networks. In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.: s.n.], 2020. p. 463–468.

VANDERPLAS, J. *Python Data Science Handbook: Essential Tools for Working with Data*. [S.l.]: "O'Reilly Media, Inc.", 2016.

VOLSCHAN, A.; CARAMELLI, B.; GOTTSCHALL, C. A. M.; BLACHER, C.; CASAGRANDE, E. L.; MANENTE, E. Diretriz de embolia pulmonar. *Arq Bras Cardiol*, v. 83, n. Suppl 1, p. 1–8, 2004.