**UNIVERSIDADE VEIGA DE ALMEIDA – UVA**

**BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**Comparação de Desempenho entre algoritmos de Machine Learning**

**Alexandre de Souza Pereira**

**RIO DE JANEIRO**

**2021UNIVERSIDADE VEIGA DE ALMEIDA – UVA**

**Alexandre de Souza Pereira**

Monografia apresentada ao curso de Ciência da Computação da Universidade Veiga de Almeida, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador(a): Carlos Alberto Alves Lemos, DSc.

**Comparação de Desempenho entre algoritmos de Machine Learning**

**RIO DE JANEIRO**

**2021**



**UNIVERSIDADE VEIGA DE ALMEIDA - UVA**

**BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**Alexandre de Souza Pereira**

**Comparação de Desempenho entre algoritmos de Machine Learning**

Monografia apresentada como requisito parcial à conclusão do curso em Bacharel em Ciência da Computação.

APROVADA EM:

CONCEITO: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

BANCA EXAMINADORA:

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**PROF. DSc** Carlos Alberto Alves Lemos

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**PROF. DSc ou MSc NOME COMPLETO DO PROFESSOR DA BANCA**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**PROF. DSc ou MSc NOME COMPLETO DO PROFESSOR DA BANCA**

**Coordenação de Ciência da Computação**

Rio de Janeiro

*Dedico este trabalho aos meus pais. – Isso é um exemplo de dedicatória.*

**AGRADECIMENTOS**

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxx.

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx.

**“Epígrafe:** opcional, citação de uma frase ou pensamento, de até 5 linhas, relacionado ao tema do trabalho”

Nome do autor

**RESUMO**

O Resumo deve ser redigido em até uma página com fonte Times New Roman, corpo 12 e espaçamento entrelinhas simples. Deve conter, de forma explícita e concisa, a apresentação de todo o TCC desde a introdução, o objetivo, o método empregado, a pesquisa realizada, o desenvolvimento realizado, até os resultados atingidos e as conclusões. Não deve conter referências, citações, ilustrações, comentários, críticas ou julgamentos pessoais. Nessa mesma página devem ser listadas as palavras-chave que dizem respeito ao trabalho desenvolvido.

Palavras-Chave: Palavra1, Palavra2, Palavra3

**ABSTRACT**

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx.

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx.

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx.

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx.

Xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxx xxxxxx xxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx x xxxxxxxxxxxx xxxxxx xx xxxxxxx xx x xxxxxxxx xxx xxxxxxxxxx xx xx.

Keywords: word1, word2, word3

**Sumário**

[**1**](#_heading=h.3s49zyc) **INTRODUÇÃO** **12**

**1.1 OBJETIVO 13**

[**2**](#_heading=h.279ka65) **REVISÃO BIBLIOGRAFICA 14**

[**3**](#_heading=h.meukdy) **ALGORITMO DE *MACHINE LEARNING* 16**

**3.1 *ADABOOST*** 16

**3.2 ÁRVORE DE DECISÃO** 16

**3.3 FLORESTA ALEATÓRIA** 17

**3.4 REGRESSÃO LOGISTICA** 18

**3.5 K-NN** 18

**3.6 SVM** 19

[**4**](#_heading=h.1yyy98l) **ESTUDO DE CASOS** 20

[**4.1**](#_heading=h.4iylrwe) **TRATAMENTO DADOS 21**

[**4.2**](#_heading=h.1d96cc0) **SEPARAÇÃO DAS BASES 24**

[**4.3**](#_heading=h.3x8tuzt) **TREINAMENTO E TESTES 25**

5. CONCLUSÃO 30

5.1 DESENVOLVIMENTO FUTUROS 31

[**6.**](#_heading=h.1qoc8b1)BIBLIOGRAFIA 32

# INTRODUÇÃO

Na presente monografia, tem como objetivo principal fazer uma comparação entre as aplicações de modelos de aprendizado de máquina que melhor se encaixam com a base de dados referentes a qualidade de ar de distritos de Beijing. O aprendizado de máquina visa fazer o reconhecimento de padrões dentro da base de dados.  Porém antes de abordar o assunto dos modelos de aprendizado de máquina, pontos importantes a ressaltar.

O projeto de análise de dados potencialmente utilizado para diversos recursos, podendo agregar valor de mercado nos setores de uma companhia, não só em tecnologia quanto diversos setores como em recursos humanos, logística, vendas, compras, marketing e outros. Empresas conhecidas do nosso mercado costumam utilizar essa tecnologia de dados, para fazer diversas experiências para futuras aplicações ou serviços, como por exemplo Coca-Cola, Tim, Sodexo, Santander e entre outras.

A base de dados escolhida para esta monografia, consiste em dados referentes a qualidade do ar nos distritos de Beijing, referenciadas do ano de 2013 a 2017. Inicialmente em 2013 quando os indicadores poluentes atingiram valores extremos, a OMS (Organização Mundial Saúde), declarou que Beijing se encontrava alto periculosidade a vida humana, sendo assim chamado de alerta laranja. Posteriormente, no ano de 2015, Beijing alcançou a marca de maior emissor de dióxido de carbono (CO2) e assim assumindo o alerta vermelho segundo a OMS.

No entanto os indicadores de emissões de poluição estão ligados diretamente com o crescimento econômico do país. Relatórios feitos pelo governo chinês apontam que os impactos da COVID19 para a produção podem ser observados ao analisar os mesmo. "Para combater a poluição do ar por meio da reestruturação econômica se tornarão ainda mais importantes, já que o país tem como meta de zero emissões líquidas de carbono até 2060" Segundo Lauri Myllyvirta, Analista-chefe do CREA e coautor do relatório de poluentes utilizado pelo governo chinês. [GONÇALVES, 2020].

Processo de análise é constituído de grandes bases de dados de múltiplas variaveis, que envolve diversas fontes disponíveis, podendo ser planilhas, documentos, plataformas de gestão empresarial, dessa forma essa base pode ser chamada de *Big Data.*

Iniciando o projeto de análise de dados, é possível utilizar a linguagem de programação Python para o desenvolvimento de todo o projeto em questão. É necessário fazer uma prática chamada de imersão ou exploração dos dados, que consiste em pesquisar sobre o tema, analisar quais colunas estão descritas nas bases de dados e quais tipos de informação pode-se inferir a partir disto.

Há bibliotecas que facilitam a manipulação dos dados, uma vez que essas bases são muito grandes e é necessário fazer algumas manipulações para o tratamento dos dados. Visto que, o tratamento de dados é uma prática muito necessária, porque com base nas informações adquiridas na imersão dos dados é importante que consigamos identificar os *outliers* que podem expressar dados com uma alta disparidade ou pontos como ausência de dados.

## OBJETIVO.

Essa monografia, que tem como objetivo fazer a apresentação e comparação entre algoritmos relacionados ao aprendizado de máquina utilizando a linguagem Python. Para que das soluções exploradas seja possível fazer o reconhecimento de padrões das bases de dados. Sendo assim, podemos aplicar algoritmos de aprendizado de máquina, para que as tomadas de decisão do mundo real tenham um grau assertivo maior.

# Revisão Bibliográfica.

A revisão bibliográfica foi iniciada pela pesquisa de artigos acadêmicos no SciELO, selecionando artigo e trabalhos com palavras-chave*: machine learning*, *deep learning,* *random forest, artificial neural networks.*

Para iniciar o processo de seleção dos artigos que iriam compor a revisão bibliográfica do presente monografia, analisando por parte do resumo e tipo do estudo abordado no artigo acadêmico.

Na Tabela 01 é explicitado o resultado da pesquisa bibliográfica, categorizador por utilização de algoritmo no seu desenvolvimento dos respectivos artigos científicos.

Tabela 01 – Artigos Utilizados na Pesquisa: Artigo, Nome dos Autores e Palavras-Chave

|  |  |  |  |
| --- | --- | --- | --- |
| Nome | Autores | Palavras-Chave | Ano |
| Machine Learning na Medicina: Revisão e Aplicabilidade | Gabriela Miana de Mattos Paixão, Bruno Campos Santos, Rodrigo Martins de Araujo, Manoel Horta Ribeiro, Jermana Lopes de Moraes, Antonio L. Ribeiro | Aprendizado de Máquina, Medicina, Cardiologia | 2022 |
| Determining the geographical origin of lettuce with data mining applied to micronutrients and soil properties | Camila Maione, Eloá Moura Araujo, Sabrina Novaes dos Santos-Araujo, Alexys Giorgia Friol Boim, Rommel Melgaço Barbosa, Luís Reynaldo Ferracciú Alleoni | ICP–OES; traceability; tropical soils; heavy metals; feature selection | 2021 |
| Research on food safety sampling inspection system based on deep learning | Tzu-Chia, CHENShu-Yan YU | image processing; deep learning; network architectures; learning algorithms | 2021 |
| Ultimate Bending Strength Evaluation of MVFT Composite Girder by using Finite Element Method and Machine Learning Regressors | Zhihua Xiong, Jiawen Li, Houda Zhu, Xuyao Liu, Zhuoxi Liang | MVFT girder; ultimate bending strength; artificial neural networks; composite dowel; failure mode; LSSVM | 2022 |
| DETERMINATION OF QUALITY AND RIPENING STAGES OF ‘PACOVAN’ BANANAS USING VIS-NIR SPECTROSCOPY AND MACHINE LEARNING | Iara J. S. Ferreira, Sarah L. F. de O. Almeida, Acácio Figueiredo Neto, Daniel dos Santos Costa | quality attributes, non-destructive method, Musa spp | 2022 |
| Nome | Autores | Palavras-Chave | Ano |
| Forest yield prediction under different climate change scenarios using data intelligent models in Pakistan | A. Yousafzai, W. Manzoor, G. Raza, T. Mahmood, F. Rehman, R. Hadi, S. Shah, M. Amin, A. Akhtar, S. Bashir, U. Habiba, M. Hussain | climate change; forest yield; RF and KRR models; prediction; Gallies forest; Abbottabad | 2021 |
| Modelagem preditiva de propriedades mecânicas em concretos reforçados com fibra de aço utilizando redes neurais artificiais | Leonária Araújo Silva, Lucas Benício Rodrigues Araújo, Ana Karoliny Lemos Bezerra, Arthur Hermont Fonseca Murta, Lucas Feitosa de Albuquerque Lima Babadopulos, Marcelo Silva Medeiros Júnior | Concreto reforçado com fibra de aço (CRFA); Propriedades mecânicas; RNA; Análise de sensibilidade; Dosagem | 2022 |

Inicialmente, é necessário que seja escolhida a plataforma de desenvolvimento do projeto, assim optou-se pela escolha do Google *Collaboratory*, essa ferramenta tem grandes funcionalidades, sendo a conexão direta ao drive, sendo possível executar base de dados diretamente dos repositórios de referência. Esse ambiente de desenvolvimento é construído a partir de células, possibilitando um código arrumado em blocos, podem conter código ou texto e entre outras funcionalidades.

Inicializando o projeto de ciência de dados, faz-se necessário a imersão nos dados, neste momento a base de dados vem sendo observada de uma forma geral, como por exemplo quais as colunas a base possuem, quais tipos de variáveis, qual o provedor da base e como esses dados podem começar a agregar valor à organização. À medida que o projeto vai desenvolvendo inferências do mundo real são necessários para o enriquecimento e futuros resultados esperados por ele.

Sendo assim, é necessário fazer a utilização de algumas bibliotecas, como por exemplo: *Numpy, Pandas, Seaborn, Matplotlib, Plotly, Scikit Learn*. A parte de tratamento dos dados, é um dos passos mais demorados dos processos de desenvolvimento, pois existem muitos pontos a serem analisados, pontos como *outliers*, que são dados faltantes ou dados duvidosos. No entanto, quando se trata de aprendizado de máquina, é interessante fazer a manipulação dos dados, para que não seja necessário perder alguns dados.

À medida que os valores faltantes são encontrados, temos possibilidades de tratamento dos mesmo. Verificando qual o tipo de coluna que estamos tratando, se é coluna com variáveis numérico ou categóricas. Existem algumas formas de fazer o tratamento, sendo por meio de substituição de valores, exclusão de valores indefinidos ou até mesmo exclusão de toda a coluna do data frame.

A primeira forma de tratamento, em caso de variáveis numéricas pode-se substituir os valores faltantes pela média ou mediana dos valores existentes, para que assim não se perca volume de dados e comprometa os algoritmos de aprendizado de máquina.

A segundo forma consiste em excluir as linhas do data frame que possuam dados faltantes, essa medida pode ser utilizada quando os valores de determinada coluna são do tipo categórico, que acaba não possibilitando a substituição de valores.

A terceira forma consiste em fazer a exclusão de toda a coluna, sendo por motivos de alto índices de dados faltantes ou a coluna não apresenta nenhum nível de correlação com o resto do data frame.

Na linguagem de programação *Python*, é possível fazer operações matemáticas de forma simples e rápida, tendo em mente que nestes projetos de aprendizado de máquinas manipula-se grandes bases de dados. Para que seja possível fazer a manipulação, utiliza-se a biblioteca do Pandas para fazer a manipulação da massa de dados. Assim tornando a visualização inicial mais fácil.

O *data frame* é uma estrutura que guarda diversos tipos de variáveis, no estilo de matriz. Sendo assim podendo se orientar pelos índices das respectivas colunas distribuídas nela. Ao utilizar o *data frame*, pode ser necessário fazer uso de métodos estatísticos, como por exemplo a utilização de desvio padrão, mediana, média, quartil etc.

A biblioteca Numpy é geralmente utilizada quando se julga necessário fazer manipulações em matrizes ou vetores contendo diversos tipos de dados, assim aparecendo como um facilitador da linguagem. Sendo assim, existem muitas formas de utilizar as funções do Numpy, como por exemplo utilizar estatística básica, rápidas operações entre conjunto de dados, operações algébricas, simulações aleatórias, organização de dados, seleção entre outras formas.

A biblioteca Seaborn podendo ser utilizado para a montagem das informações dispostas nas bases de dados. É possível fazer a montagem de diversos tipos de gráficos como por exemplo: gráfico de barra, círculo, linhas, histograma.

A biblioteca Scikit Learn, também conhecida por sklearn, é utilizada para executar os algoritmos de *machine learning*, pré-processamento das informações e as métricas para avaliação. Neste projeto iremos abordar os a algoritmos de classificação, os algoritmos são o Randon Forest, *Logistic Regression, Decision Tree, K-Nearest Neighbors, AdaBoost e SVC*. Já as métricas escolhidas foram as *Accuracy, F1, Recall, Precision* e a matriz de confusão.

# ALGORITMOS DE MACHINE LEARNING.

Os modelos utilizados nesta monografia para o desenvolvimento, tendo como objetivo fazer a comparação de desempenho são algoritmos que buscam classificar informações semelhantes. Para obter os resultados dos treinamento dos modelos, será utilizado as métricas para julgar quanto de porcentagem o modelo foi assertivo, em cada uma dos treinos.

## AdaBoost: [???????]

## Árvore de Decisão - (*Decision Tree*)

É um algoritmo versátil, pois é possível fazer tarefas tanto de regressão como de classificação, sendo assim poderosos, para tratar de bases de dados complexas. Esse algoritmo se torna mais atraente, visto que para utilizá-lo não é necessária uma base de dados altamente tratada ou escalonada conforme a distribuição das informações presentes. À medida que o treinamento do algoritmo vai ocorrendo, existem ferramentas de visualização de árvores na biblioteca do *Scikit-Learn*, para assim entender como as classificações do modelo estão acontecendo.

Exemplo de plotagem visual de uma árvore de decisão. Caso a *pental length* menor que 2,45cm o algoritmo irá classificar como setosa, caso contrário o algoritmo irá fazer um número X de perguntas para que assim a classificação seja feita

Diagrama

Descrição gerada automaticamente

## Figura X: Árvore de Decisão da Iris. Fonte: Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow – 05/04/2022.

## Floresta Aleatória – (Random Forest):

## Regressão Logística:

A regressão logística pode ser utilizada tanto para algoritmos de regressão como também classificação. Normalmente esse modelo tem como objetivo, classificar uma ação como verdadeira, rotulada como 1, caso tenha a probabilidade estimada mais ou igual a 50%, caso contrário a ação seria falsa, rotulada como 0.

O funcionamento do modelo de regressão logística, tem como característica uma equação de reta que gera um gráfico sigmoide, ou seja, em formato de S. Como podemos observar o exemplo a seguir na figura X. Assim o treinamento e função de custo, pode ser expresso por uma função matemática chamada de log *loss.*

Diagrama

Descrição gerada automaticamente

Figura X: Função Logística. Fonte: Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow – 05/04/2022.

## K-Vizinho mais próximos:

Esse algoritmo também pode ser chamado de KNN ou k-NN, geralmente utilizado para tratar problemas de classificação, no entanto também pode ser utilizado para regressão. Neste tópico iremos abordar para modelos referente a classificação. A forma de classificar que esse algoritmo utiliza é por meio de votação de K vizinhos mais próximos, o número de vizinho geralmente é um número ímpar e a distância pode ter algumas opções de escolha, como as distância:

* Euclidiana
* Manhattan
* Minkowski
* Hamming

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Figura X: Classificação k-NN Fonte: <https://www.ibm.com/topics/knn> – 10/04/2022.

## MÁQUINA DE VETORES DE SUPORTE (SVM)

O algoritmo que pode ter algumas variações, sendo classificação linear, não linear, de regressão e identificador de outliers. No entanto, neste tópico são iram ser abordados os algoritmos de classificação linear e não linear. Este algoritmo de classificação é apropriado para base de dados complexos, sendo de pequena ou média dimensão.

Exemplificando o algoritmo de SVM como a figura X02, faz-se a classificação por meio da separação representada por pela linha continua, e a margem de segurança sendo a linha tracejada, a margem de segurança marca os pontos mais próximos das separações entre as classes.

Podemos observar que na figura mais a esquerda a margem de segurança não foi aplicada corretamente, no entanto, por mais que as classes possam vir a estar bem definidas, a margem não consegue aplicar a distância de segurança para definição de uma classe e outra. Já na aplicação mais à direita, a classe da base de dados se torna mais bem definida e a margem de segurança respeita, possibilitando a fixação dos pontos pertencentes a cada classe. A classificação abaixo é chamada de Classificação de margens largas.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Figura X02: Algoritmo de SVM. Fonte: Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow – 13/04/2022.

# Estudo de casos.

A base de dados escolhido, possui dados de qualidade do ar, de Beijing, composta por dados numéricos e categóricos, como por exemplo: ano, mês, dia, hora, concentração de PM2.5, PM10, SO2, NO2, CO, O3 por (ug/m^3), temperatura (°C), pressão, temperatura do ponto de orvalho (DEWP), chuva (mm), velocidade do vento (WSPM), direção do vento (wd), e o nome da estação que foi coletado.

Calendário

Descrição gerada automaticamente com confiança média

Figura X01: Apresentação de base de dados. Fonte: Autoral – 05/09/2021.

Dentre esses atributos citados anteriormente, há algumas variações dentre os valores como valores inteiros, reais, cadeia de caracteres. Os tipos das variáveis são separados por classificações como: numérica contínua, numérica discreta, categóricas nominais, categóricas ordinais.

Visualizando de forma mais prática, as variáveis numéricas têm como característica ser do tipo número inteiro para as discretas, como por exemplo dias, meses, anos. Já para o tipo contínuo, tem a característica de número real, como por exemplo temperatura, pressão atmosférica e velocidade do vento.

Seguindo, para as variáveis categóricas, iniciando pela nominal, são dados que não mensuráveis e sem ordem, como por exemplo cor dos olhos, gênero, nome, cor de pele. Já o tipo ordinal tem como característica fazer a classificação sob uma lógica, como por exemplo tamanho P, M, G.

Originalmente a base de dados estão organizadas em tabelas do Excel, separada por região de amostras coletadas, sendo assim o próximo passo o tratamento das bases de dados.

Como foi mostrado anteriormente, a apresentação da base de dados é feita utilizando a biblioteca Pandas, com a função *head*(x) que faz a apresentação dos cinco primeiros e últimos valores. Essa função não necessariamente precisa ter um valor definido, se tiver o retorno será do data frame com o número definido, caso tenha retornará como na figura X01.

## Tratamento e preparação dos dados.

Inicializando o procedimento de limpeza de dados, utilizando as bibliotecas de Python como Pandas e Numpy que possibilitam a visualização de dados, tratamento ou exclusão do dado. É possível fazer de algumas formas a verificações na base.

Nota-se que, o pré-processamento pode ser iniciado pela variáveis dos tipos numéricas, utilizando a função *describe()* do Pandas. Essa função descreve cada coluna da base em algumas métricas estudadas geralmente em estatísticas, como por exemplo: contagem, média, desvio-padrão, quartil, valor máximo e mínimo.

df.describe.transpose()

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Figura 02: plotando descrição da base. Fonte: Autoral – 12/09/2021.

Caso as variáveis possuam valores categóricos, utiliza-se a função *isnull*(), para retornar o data frame com todos os valores booleanos, se existe valor retorna *true*, caso contrário false. A fim de visualizar a quantidade de valores nulos do data frame, atrela-se a função *sum*() junto a *isnull*(), para que assim seja retornado por cada coluna as quantidades de valores nulos do data frame.

Como citado anteriormente, o processo de tratamento de dados é um trecho importante quando se trata de projeto de análise de dados. Com isso a base utilizada neste projeto estava segmentada em doze partes, sendo dados histórico de 2013 a 2017 das estações descritas a seguir: Aotizhongxin, Changping, Tiantan, Wanshouxigong, Dingling, Dongsi, Gucheng, Huairou, Guanyuan, Nongzhanguan, Shunyi, Wanliu.

Antes de agrupar todas essas bases em um grande conjunto, julgou-se necessário fazer o tratamento dos dados faltantes de todas as base em separado para que assim as particularidade de cada uma das regiões não tivesse alteração. Visto que ao agrupar todas as parte todas as partes as médias das tabelas ficariam diferentes por conta das particularidade de cada região do dado colhido tanto do número de instâncias do data frame.

Portanto o método de tratamento de dados escolhido foi o de substituição de valores inteiros e reais ausentes e exclusão de linhas. A substituição dos valores indefinidos foi feita por meio da média dos valores presentes na base de cada coluna respectivamente, já os valores que foram excluídos sendo do tipo categórico.

Seguindo o desenvolvimento, é necessário fazer com que a base de dados tenha somente informações que possam ser aproveitadas no processo de aprendizado de máquina, no qual serão os próximos passos. Com a necessidade de ajuste da base, utilizou-se a matriz de correlação, essa ferramenta é utilizada para verificar a relação entre as variáveis da base de dados, como podemos na figura 03.

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Figura 03: Matriz de correlação. Fonte: Autoral – 02/11/2021.

Ao utilizar a matriz de correlação, julgou-se necessário fazer a comparação entre as colunas para melhor adequar quando for aplicador os modelos de aprendizado de máquina. Com isso gera-se um valor das comparações fixado entre 1 e -1. No entanto, na matriz em questão os número estão no intervalo de 1 até -0,75.

Logo após a utilização da matriz de correlação, podemos observar que algumas colunas não se relacionavam entre si e podem ser excluídas do data frame. Portanto como por exemplo as colunas número da linha, ano, mês, dia e hora Não serão levadas em consideração neste trabalho

## Separação das Bases

Iniciando o processo de separação das bases de dados, sendo as base de treino e teste. Para que os modelos de aprendizado de máquina possam ser executados, as base de dados são submetidas a uma separação, como descrito anteriormente. No entanto, existe um problema que durante o treinamento, que é chamado de underfitting e overfitting. [TRADUÇÃO/SIGNIFICADO NO RODAPÉ]

Segundo os autores da Ian Goodfellow, Yoshua Bengio, Aaron Courville [Deep Learning (2017, MIT), Cap 5, Pag 111], “ 1. Fazer com que o erro de treino seja pequeno 2. Tornar pequeno o intervalo entre o treino e o erro de teste. Estes dois fatores correspondem aos dois desafios centrais na aprendizagem de máquinas: o underfitting e o overfitting. O underfitting ocorre quando o modelo não é capaz de obter um valor de erro suficientemente baixo no conjunto de formação. O overfitting ocorre quando o intervalo entre o erro de treino e o erro de teste é demasiado grande. Podemos controlar se um modelo é mais susceptível de se overfit ou underfit, alterando a sua capacidade”.

Os autores da situação apontam os fatores onde ocorrem os problemas, uma vez que esses problema são pontuados, é necessário tomar uma série de cuidados para que os modelos de aprendizado de máquina não fiquem tendenciosos. Conforme a separação de dados é feita, os grupos necessariamente precisam possuir características diferentes, de forma a base de teste oferecer um desafio ao modelo que está tentado estabelecer o resultado com base nos dados de treino.

Visto que os problemas citados anteriormente estão ligados diretamente com o balanceamento das quantidade de bases treino e teste, sendo assim fatores que podem estar diretamente ligados com o problema em questão. Geralmente, utiliza-se 70% e 30%, para base de treino e teste respectivamente.

Seguindo a linha de desenvolvimento e análise, algumas colunas foram retiradas do processo, por conta de não apresentarem valor quando o assunto é algoritmos de *machine learning,* sendo colunas como quantidade de chuva, direção do vento ( WD ) e velocidade do vento (WSPM)

## Treinamentos e Testes

Os treinamentos dos modelos que mencionamos anteriormente, foi apontado pelas métricas que os níveis de acerto do modelos alcançaram o valor menor que 50%, portanto a estratégia utilizada, foi fazer a diminuição do escopo das bases. Uma vez que iniciados os treinamentos utilizando doze cidades diferentes e após a visualização dos primeiros resultados não satisfatórios.

Optou-se por retomar os testes com uma menor variedade de cidades, a fim de aumentar os níveis assertivos dos modelos. Assim chegamos ao número de quatro cidades, tendo em vista que em só diminuir a quantidade, os níveis assertivos também subiram aproximadamente 25%. As bases de teste escolhidas para seguir rodando os modelos, tendo em vista que quanto menos dados faltantes melhor, as bases escolhidas estão situadas em Changping,Wanshouxigong ,Gucheng ,Nongzhanguan, respectivamente data frames 2, 7, 8, 10.

# Bibliografia.

* [1] Downey, A. B. Think stats exploratory data analysis in python. Version 2.1.0. – Ultimo acesso em: 20/05/2021.
* [2] McKinney, W. Python para análise de dados - tratamento de dados pandas, numpy e ipython. Novatec Editora; 1ª edição (maio 2019). - Acesso em: 20/05/2021.
* [3] Peng, R. D, Matsui, E. The Art of Data ScienceA - Guide for Anyone Who Works with Data.
* [4] A Programmer's Guide to Data Mining: The Ancient Art of the Numerati
* [5] VanderPlas, J. Python Data Science Handbook: Essential tools for working with data. O'Reilly Media; 1ª edição (21 novembro 2016)
* [6] Goodfellow, I; Bengio, Y; Courvelle, A. Deep Learning. 2017 MIT.
* [7] Grus, Joel. Data Science do Zero. O'Reilly Media; 1ª edição (21 novembro 2016)
* [8] <https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data> (Uci Dataset - Base de Dados) - Último acesso - 05/09/2021.
* [9] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5627385/> - Último acesso: 31/07/2021
* [10] <https://brasil.elpais.com/brasil/2015/12/07/internacional/1449490356_143778.html> - Último acesso - 01/08/2021
* [11] <https://www1.folha.uol.com.br/mercado/2020/12/retomada-puxada-pela-industria-explica-maior-poluicao-em-pequim-diz-relatorio.shtml> - Último acesso - 01/08/2021
* [12] <https://www.comciencia.br/uma-breve-trajetoria-da-questao-ambiental-recente-na-china/#_ftn8> - Último acesso - 01/08/2021
* [13] <http://climacom.mudancasclimaticas.net.br/discussoes-sobre-a-questao-ambiental-na-china-impactos-e-perspectivas/> - Último acesso - 01/08/2021
* [14] <https://pandas.pydata.org> - Último acesso – 05/09/2021.
* [16] https://numpy.org - Último acesso – 05/09/2021.
* [17] https://seaborn.pydata.org- Último acesso – 05/11/2021.
* [18] https://scikit-learn.org/stable/- Último acesso – 05/11/2021.
* [19] https://matplotlib.org- Último acesso – 05/11/2021.