UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Eliana Maria Silva de França

Estudo de técnicas preditivas para o auxílio a gestores na pandemia de COVID-19.

Recife

2022

Eliana Maria Silva de França

Estudo de técnicas preditivas para o auxílio a gestores na pandemia de COVID-19.

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Orientador: Prof. Dr. Rodrigo Gabriel Ferreira Soares

Recife

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

F815e França, Eliana Maria Silva de
Estudo de técnicas preditivas para o auxílio a gestores na pandemia de COVID-19 / Eliana Maria Silva de França. - 2022.
42 f. : il.

Orientador: Rodrigo Gabriel Ferreira Soares.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. Predição. 2. Aprendizado de Máquina. 3. COVID-19. 4. Gestão. I. Soares, Rodrigo Gabriel Ferreira, orient. II. Título

CDD 004

Eliana Maria Silva De França

Estudo de técnicas preditivas para o auxílio a gestores na pandemia de COVID-19.

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 27 de Maio de 2022.

BANCA EXAMINADORA

Rodrigo Gabriel Ferreira Soares (Orientador)
Departamento de Estatísitca e Informática
Universidade Federal Rural de Pernambuco

Cícero Garrozi
Departamento de Estatísitca e Informática
Universidade Federal Rural de Pernambuco

À Deus, meus pais, familiares, amigos e professores, por sempre estarem ao meu lado e acreditarem em mim.

Agradecimentos

Agradeço à Deus pela vida, por todas as pessoas que colocou no meu caminho, nesta jornada da vida, por todos os desafios e vitórias e por tudo que me deu, das coisas que sei e as que eu não sei.

Agradeço aos meus pais, Jozilda e Francisco, por me criarem e incentivarem a procurar o caminho da educação, e me mostrarem que dentro da escola pública é possível embasar uma boa formação e que apesar das dificuldades este é o melhor caminho a seguir. Agradeço por todo suporte e paciência, principalmente durante os anos da graduação.

Agradeço aos meus amigos e familiares, que sempre estiveram participando da minha vida acadêmica, mesmo que indiretamente, me dando dicas e conselhos para enfrentar os desafios da vida, minha tia Leninha e meu tio Valmir, minha melhor amiga Viviane, minhas amigas que conheci no primeiro estágio e levo pra vida, Aline e Hozana e os meus amigos da ETEPAM, Heloiza, Willian e Thiago.

Agradeço ao meu orientador, Prof. Dr. Rodrigo Soares, por ter aceitado me orientar neste trabalho de conclusão de curso, por todas as orientações, os ensinamentos, conselhos, que me deu desde o início da graduação e pela paciência e didática para retirar todas as minhas dúvidas.

Agradeço ao Prof. Dr. Cícero Garrozi, por aceitar participar da banca da minha defesa, por todos os ensinamentos durante o curso e por ter me ajudado na contratação do meu primeiro estágio neste curso, que foi o ponto inicial para a minha vida profissional.

Agradeço à todos que participam ou participaram do corpo docente, discente e técnico do Curso de Bacharelado em Sistemas de Informação da UFRPE, os quais são inúmeros e não consigo citar aqui.

Agradeço aos meus amigos que conheci durante a graduação, Lucas Sales, Jadiel Eudes, Everton Veloso, Evele Lemos, Wenderson Leonardo, Caroline Gomes, Jadeilson Rocha, Adailson Tavares, Anderson Rodrigues, Heitor Augusto e Erico André, pelo espírito de união, pelas injeções de ânimo, pelo companheirismo nas madrugadas e pelas amizades que a muito tempo transpassaram os muros da UFRPE.

Agradeço à oportunidade de iniciar minha carreira como desenvolvedora na Accenture Brasil, uma empresa que incentiva a formação e que fomenta políticas que colaboram para isto e com o desenvolvimento humano de suas pessoas. Agradeço à toda a equipe, e em especial, meus líderes diretos Fernanda Monegalha e Victor Sar-

denberg, pelos conselhos, paciência, conhecimento e compreensão, principalmente quando eu precisava de alguma ajuda por conta dos efeitos da graduação.

A todos que participaram desta jornada, eu agradeço!

“Suba o primeiro degrau com fé. Não é necessário que você veja toda a escada. Apenas dê o primeiro passo.”

(Martin Luther King)

Lista de Figuras

Lista de Tabelas

Sumário

Estudo de técnicas preditivas para o auxílio a gestores na pandemia de COVID-19.

[Eliana Maria Silva de França] [1], **[Rodrigo Gabriel Ferreira Soares]** [1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

eliana.franca@ufrpe.br, rodrigo.gfsoares@ufrpe.br

Resumo. *O objetivo principal deste trabalho é propor uma alternativa aos levantamentos estatísticos exploratórios, no suporte à tomada de decisão dos gestores, durante o enfrentamento à pandemia da COVID-19. Para tal, foi-se criada uma metodologia, utilizando aprendizado de máquina para fornecer uma nova ferramenta de predição de mortes causadas por COVID-19, a partir de dados abertos que contenham características sanitárias, demográficas e populacionais. De tal modo que, a partir deste estudo se possa desenvolver um modelo de inteligência artificial capaz de auxiliar no enfrentamento da pandemia de COVID-19. Dos 3 algoritmos de inteligência artificial utilizados (Decision Tree, Support Vector Machine e Multilayer Perceptron), o modelo baseado em Support Vector Machine foi o que apresentou o melhor desempenho, pois é o que possui o menor Erro Absoluto Médio, métrica utilizada para medir a qualidade de modelos de inteligência artificial baseados em regressão.*

Abstract. *The main objective of this work is to propose an alternative to exploratory statistical surveys, to support the decision-making of managers, during the confrontation of the COVID-19 pandemic. To this end, a methodology was created, using machine learning to provide a new tool for predicting deaths caused by COVID-19, from open data that contain sanitary, demographic and population characteristics. In such a way that, from this study, an artificial intelligence model can be developed capable of helping to face the COVID-19 pandemic. Of the 3 artificial intelligence algorithms used (Decision Tree, Support Vector Machine and Multilayer Perceptron), the model based on Support Vector Machine showed the best performance, because it has the lowest Mean Absolute Error, a metric used to measure the quality of regression-based artificial intelligence models.*

1. Introdução

Esta seção tem como objetivo introduzir a temática e motivação deste estudo, os quais serão descritos e explicados nas subseções a seguir.

Os tópicos a serem abordados são:

1. Contexto;
2. Problema e Motivação;
3. Objetivos;
4. Abordagem da Proposta.;
5. Impacto e Resultados;
6. Contribuições.

1.1. Contexto

A COVID-19 é uma doença infecciosa causada pelo vírus SARS-CoV-2. O termo COVID-19 vem do acrônimo da sigla em inglês que refere a (co)rona (vi)rus (d)isease, o que em tradução para o português significa "doença do coronavírus", quanto ao número 19 está ligado ao ano em que os primeiros casos foram publicamente divulgados, na cidade de Wuhan, província de Hubei, China, em dezembro de 2019.

O vírus se espalha, na maioria das vezes, de pessoa para pessoa por meio de gotículas respiratórias, produzidas quando uma pessoa infectada tem tosse, espirra ou fala. Há casos de pessoas infectadas que não apresentam sintomas, mas, geralmente, as pessoas infectadas apresentam sintomas respiratórios, que podem variar de leve a grave, estes últimos podem causar a falência da pessoa infectada.

As medidas para prevenção e mitigação da transmissão da COVID 19 são: distanciamento social, etiqueta respiratória e de higienização das mãos, uso de máscaras, limpeza e desinfeção de ambientes, isolamento de casos suspeitos e confirmados, quarentena dos contatos dos casos de COVID-19 e vacinação.

A fim de aumentar o distanciamento social, diversos setores tiveram suas atividades paralisadas ou mantidas de forma remota, este último foi e ainda é praticado por diversas empresas privadas e órgãos públicos.

Muito se é especulado sobre quando e como é seguro retornar às atividades presenciais. Diante deste cenário, algumas empresas privadas se adaptaram totalmente ao modelo de trabalho remoto, no entanto, algumas empresas e principalmente as repartições públicas sentem a necessidade de alguma modalidade de trabalho que envolve o convívio *in loco* dos seus colaboradores. Mas como saber quando e como é seguro retornar para os escritórios?

Existem diversas técnicas estatísticas e computacionais, que podem ajudar na tomada de decisão dos gestores, dentre estas técnicas, se destacam os métodos estatísticos e os métodos de aprendizado de máquina.

Os métodos estatísticos utilizam os dados para levantar comportamentos e tendências nos dados que refletem as condições do momento da coleta dos dados. Alguns levantamentos estatísticos frequentemente utilizados para medir a situação da COVID-19 são: média, média móvel, média ponderada e moda.

Algoritmos de aprendizado de máquina são eficientes para predizer resultados a partir de um modelo treinado, utilizando-se de uma base de dados como referência, quando estas bases possuem padrões de comportamentos, que ajudam a predizer comportamentos futuros.

1.2. Problema e Motivação

Uso de levantamentos estatísticos exploratórios no auxílio de tomada de decisão para retomada ou paralisação de atividades comerciais, empresariais e organizacionais durante a pandemia da COVID-19

Durante o ano de 2021, ocorreram constantes ciclos de abertura e fechamento de estabelecimentos, desde comércios e escritórios a repartições públicas. Muitas empresas foram impactadas pelas aberturas e fechamentos de seus negócios, pois não conseguiram se adaptar ao constante cenário de abertura e fechamento das empresas para atividades presenciais. Isto foi devido a flutuação do número de casos confirmados e de mortes causadas pela COVID-19.

Quando os casos confirmados e quantitativo de mortes decrescia, as organizações iniciavam a retomada de suas atividades presenciais, porém algumas semanas depois, por conta do aumento do número de casos e mortes, precisavam fechar os estabelecimentos, voltar a atividades remotas ou aumentar a rigidez nas medidas de manutenção de distanciamento social dentro das dependências físicas dos locais de trabalho. Quando os números diminuíam, as atividades eram retomadas e quando aumentavam, novamente, as atividades eram interrompidas ou afetadas de modo que impactava negativamente na administração e manutenção dos negócios.

A incerteza de como as atividades podem ser realizadas, geram custos, que em tempos de crise, podem ser catastróficos para as empresas e por consequências para as pessoas que dependem destes negócios.

Os impactos causados por esta incerteza envolvem horário de funcionamento dos estabelecimentos reduzido, compra de materiais sanitários e EPI's para os funcionários e usuários que utilizam os serviços, necessidade de fornecer apoios financeiros e estruturais para o trabalho remoto dos funcionários, aluguel de estabelecimentos que não estão sendo utilizados, necessidade de contratar serviços para prática do *delivery*, necessidade de investimento em serviços eletrônicos, tais quais certificados digitais e assinaturas eletrônicas para validar autenticidade de documentos, plataformas digitais para tratamento de documentações no geral, como em assinatura de contratos, abertura de editais, abertura e acompanhamento de processos, etc. Em cada um dos cenários apresentados há impacto direto nas receitas das empresas, que pelo ponto de vista micro e macroeconômico, impactam na estrutura econômica do país.

Muitas empresas, comércios, organizações e governos utilizaram os resultados de levantamentos estatísticos para auxiliar nas suas decisões sobre retomar ou paralisar as atividades presenciais. E isto foi visto, bastante, no ano de 2021, no qual ocorreu um ciclo de abertura e fechamento de estabelecimentos comerciais, um ciclo de retomada e paralisação de atividades presenciais em muitas empresas e organizações utilizando estas métricas para auxiliar na decisão.

Os levantamentos estatísticos exploratórios, como média, média ponderada, média móvel, moda e mediana, são importantes e ajudam a compreender situações, porém eles

apresentam um problema, eles refletem o passado, e no máximo o momento atual da coleta dos dados utilizados, no entanto, não são muito eficientes para realizar previsões do que pode ocorrer no futuro.

1.3. Objetivos

Este trabalho tem como objetivo demonstrar como foi obtido um modelo de máquina inteligente, com maior desempenho, para predizer o número de mortes por COVID-19 a partir das características sanitárias, populacionais e demográficas. De forma que seja possível propor uma ferramenta que auxilie na tomada de decisão de gestores, quanto a retomada ou paralisação de atividades comerciais, empresariais e organizacionais durante a pandemia da COVID-19.

1.3.1. Objetivos específicos

- Coletar bases de dados com características sanitárias, populacionais e demográficas, relacionadas a COVID-19.
- Estudar e selecionar técnicas de pré-processamento para regressão.
- Estudar e selecionar técnicas de extração de características para regressão.
- Executar experimentos para a abordagem proposta, utilizando métodos existentes na literatura.
- Analisar e avaliar os modelos através de medidas de avaliação de desempenho apropriadas para regressões.

1.4. Abordagem da Proposta

Quando se trata do futuro, é possível utilizar um outro método de análise e previsão, que são as técnicas de aprendizado de máquina. Estas técnicas podem, a partir de dados passados, prever com um certo nível de precisão o que poderá ocorrer no futuro.

Nos últimos vinte anos, a inteligência artificial vem sendo usada cada vez mais para auxiliar na tomada de decisão, em diversas áreas, incluindo a área da saúde. Existem diversos exemplos do uso da inteligência artificial na literatura científica que demonstram isto.

A revisão *"Introduction to Machine Learning in Digital Healthcare Epidemiology"* (ROTH et al., 2018), descreve como os algoritmos de aprendizados de máquina podem usar as mesmas técnicas utilizadas para o estudo da epidemiologia de saúde digital na epidemiologia da saúde. A revisão explica brevemente as diferenças de aplicação de estatística e aprendizado de máquina e como, no cenário atual, a inteligência artificial possui mais robustez computacional para lidar com *Big Data*, a grande variedade de tipos e a alta dimensionalidade dos dados.

A revisão *"Introduction to Machine Learning in Digital Healthcare Epidemiology"* (ROTH et al., 2018) ressalta que os algoritmos de aprendizado de máquina precisam de dados confiáveis para entregar resultados satisfatórios, incentiva que sejam realizados mais estudos e menciona que a barreira linguística é um desafio para maior amplificação do uso do aprendizado de máquina na epidemiologia da saúde.

Em *"Artificial intelligence can improve decision-making in infection management"* (RAWSON et al., 2019), é demonstrado como o aprendizado de máquina pode

contribuir no suporte e auxílio na tomada de decisão para o tratamento de infecção. É explicado como é possível reduzir vieses humanos e culturais, pois em momentos críticos, as decisões são tomadas em cima dos requisitos mínimos, em detrimento à solução ótima, como descreve Simon (SIMON, 1985), que explica que existem três limitações inevitáveis na decisão humana: (i) a informação disponível para tomar decisões é muitas vezes limitada e potencialmente não confiável; (ii) a mente humana tem uma capacidade limitada; e (iii) há apenas uma quantidade limitada de tempo para tomar uma decisão. O estudo defende que através da inteligência artificial a "racionalidade limitada", como define, é reduzida pela alta capacidade de processamento que um computador possui ao analisar dados.

No estudo *"Use of artificial intelligence in infectious diseases. In Artificial Intelligence in Precision Health"* (AGREBI; LARBI, 2020), é defendido como o uso de inteligência artificial impacta positivamente no estudo dos principais aspectos da infecção: diagnóstico, transmissão, resposta ao tratamento e resistência. Os autores demonstram como a inteligência artificial fornece instrumentos para prever melhor as epidemias, entender a especificidade dos patógenos e identificar alvos potenciais para o desenvolvimento de medicamentos. No estudo foi realizada uma revisão de uma série de aplicativos escolhidos seletivamente, indicando como a inteligência artificial está avançando e ajudando no enfrentamento à doenças infecciosas, especialmente em países de renda baixa.

Estes três estudos foram feitos em intervalos de poucos meses a um ano entre si. O que deixa claro como o avanço do estudo de aprendizado de máquina vem crescendo, de forma acelerada, na área da saúde, principalmente no combate a doenças infecciosas, como é o caso da COVID-19.

Inspirado nestes estudos e na motivação do cenário pandêmico atual, este estudo foi idealizado para utilizar as bases de dados abertos, para ajudar na previsão de mortes de COVID-19, e desta forma propor uma alternativa aos levantamentos estatísticos, no auxílio à tomada de decisão frente o enfrentamento a pandemia de COVID-19.

1.5. Impacto e Resultados

Este trabalho pretende fornecer os seguintes impactos para a sociedade:

- Influenciar positivamente nas tomadas de decisões de gestores de empresas públicas e privadas;
- Ser uma ferramenta para apoiar no embasamento para aplicação de políticas públicas, através de projeções nas características estudadas neste estudo;
- Possibilitar uma alternativa ao uso da estatística exploratória no apoio à tomada de decisão, no enfrentamento à COVID-19.

1.6. Contribuições

Este trabalho pretende auxiliar a comunidade científica através das seguintes contribuições:

- Fornecer uma análise de vários algoritmos de AI para a predição de mortes por COVID-19, a partir de características sanitárias, demográficas e populacionais.
- Demonstrar como o aprendizado de máquina pode ajudar a combater problemas que afligem a sociedade, através de bases de dados.
- Propagar o estudo de inteligência artificial no idioma brasileiro.

1.7. Organização do Trabalho

Este trabalho está organizado em 6 seções. Na seção 2 é feita uma abordagem teórica dos termos, algoritmos, técnicas utilizadas nos experimentos e medidas de avaliações utilizadas. Na seção 3 são apresentados os trabalhos correlatos, demonstrando as semelhanças e divergências com este estudo. Na seção 4 será explanado de maneira aprofundada a abordagem da proposta. Na seção 5 é explicado a metodologia seguida no trabalho, quais experimentos foram realizados, apresentado os métodos, ferramentas e tratamentos, assim como os resultados obtidos. Na seção 6 serão apresentadas as conclusões deste trabalho e as possibilidades de trabalhos futuros.

2. Referencial Teórico

Nesta seção serão explicados, brevemente, os termos técnicos, tecnologias e procedimentos utilizados neste estudo, assim como serão apresentados os algoritmos e técnicas que foram utilizadas para os experimentos e avaliação dos resultados obtidos.

As elucidações teóricas estão divididas nos seguintes grupos:

1. Conceitos e Terminologias Gerais de Aprendizado de Máquina;
2. Algoritmos Utilizados;
3. Análise de Componentes Principais;
4. Pré-processamento;
5. Hiper Parâmetros;
6. Medidas de Desempenho e Avaliação.

2.1. Conceitos e Terminologias Gerais de Aprendizado de Máquina

Nesta seção serão apresentados conceitos e terminologias comuns quando se trabalha com Aprendizado de Máquina.

2.1.1. Classificação

É um tipo de aprendizado de máquina supervisionado. Utilizado para categorizar instâncias de um determinado grupo de dados que possuem características semelhantes.

2.1.2. Regressão

É um tipo de aprendizado de máquina supervisionado, utilizado para prever o valor de uma característica contínua, variável dependente, a partir de características que influenciam o valor a ser predito, variáveis independentes, que também são contínuas. Enquanto que, nas classificações, o objetivo é rotular um indivíduo, nos algoritmos de regressão é determinar uma função que melhor descreve os dados, para prever o valor correto para um novo indivíduo.

2.1.3. Atributo

É uma característica de uma instância. Por exemplo, a idade de uma pessoa, a quantidade de pessoas fumantes, a densidade populacional, etc.

2.1.4. Instância

É o conjunto de características de um determinado objeto de estudo. Por exemplo, as características que classificam uma pessoa são: gênero biológico, idade, altura, etc.

2.1.5. Base de dados

São conjuntos de informações (instâncias), que possuem as mesmas características (atributos).

2.1.6. Dados abertos

São bases de dados disponibilizados por diversas organizações, de plataformas especializadas em dados a portais de transparência disponibilizados por instituições públicas.

2.1.7. Parâmetros do modelo

São parâmetros definidos pelo próprio modelo. Não podem ser definidos ou alterados pelo supervisor do modelo.

2.1.8. Inteligência Artificial

É a área da computação que estuda o aprendizado de máquinas artificiais. Ao contrário dos seres humanos, os computadores não são providos de criatividade e capacidade analítica, mas em contrapartida, possuem um alto poder de processamento de informações, o que humanos não possuem. Essa área da ciência da computação tem como objetivo unir as capacidades humanas com a velocidade de processamento dos computadores. O aprendizado de máquina pode ser dividido em três tipos: Aprendizado de máquina supervisionado, Aprendizado de máquina não supervisionado e Aprendizado de máquina semi-supervisionado. Neste estudo vamos abordar o Aprendizado de máquina supervisionado.

2.1.9. Aprendizado de máquina supervisionado

Um método de aprendizado de máquina supervisionado é um método que precisa que um ser humano ensine para a máquina como ele deve determinar uma dada instância, a partir de uma base de dados de treino. Podemos encontrar dois tipos de aprendizado de máquina supervisionado - classificação e regressão. A classificação ocorre quando as instâncias possuem um atributo que contém rótulos para cada instância de uma base de dados, por exemplo, determinar uma pessoa alta ou magra de acordo com suas características (atributos), já a regressão envolve cálculo de funções para determinar o valor e uma variável dependente a partir de variáveis independentes. Neste estudo vamos aplicar a regressão e faremos uma comparação no desempenho de três modelos de regressão

baseados em *Decision Tree*, Redes neurais (*Multilayer Perceptron*) e *Support Vector Machine*.

2.2. Algoritmos Utilizados neste estudo

Nesta seção serão apresentados os algoritmos utilizados neste estudo.

2.2.1. *Decision Tree*

Árvore de decisão é um algoritmo que fornece suporte à decisão. O algoritmo é semelhante a uma árvore natural invertida, iniciando pelo nó raiz e se ramificando até os nós folha. É um algoritmo baseado em lógica condicional, pois contém apenas instruções deste tipo. Tem como objetivo prever o valor de uma classe ou um valor numérico, aprendendo regras simples de decisão inferidas a partir de uma base de treinamento (BUITINCK et al., 2013). A figura 1 ilustra a estrutura e os componentes principais de uma árvore de decisão.

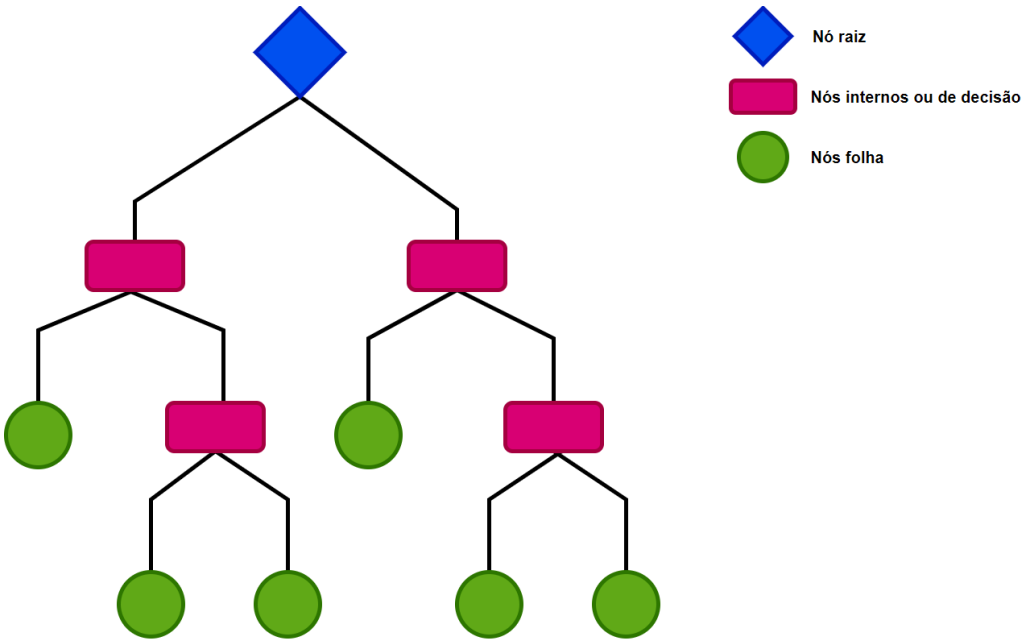

Figura 1. Representação dos componentes principais de uma árvore de decisão. Fonte da imagem: a autora.

As árvores de decisão, frequentemente, apresentam um impacto negativo, que é um super ajuste sobre os dados de treinamento, conhecido como *overfitting*. Este fenômeno faz com que o modelo seja ajustado para os dados de treinamento de tal forma que, o modelo se especializa aos dados de treinamento, apresentando ótimas métricas de avaliação, mas quando o modelo é executado em dados de testes, desconhecidos pelo modelo, ele apresenta uma péssima performance.

As floretas aleatórias são um *ensemble* das árvores de decisão, que reduz o *overfitting*, pois utiliza os mesmos hiper parâmetros. Uma floresta aleatória é um meta estimador que ajusta um número de árvores, em várias sub amostras do conjunto de dados, utilizando

a média para melhorar a precisão preditiva e controlar o ajuste excessivo. (BUITINCK et al., 2013).

2.2.2. *Support Vector Machine*

Máquina de Vetor de Suporte é um algoritmo baseado em álgebra linear. Classifica e realiza regressão baseada em dados, utilizando vetores de suporte como delimitadores de fronteira entre as classes, ou limites vetoriais. Descoberta pela primeira vez por Vladimir Vapnik e seus colegas em 1995 (VAPNIK, 1995), as *SVM*s são consideradas uma técnica não paramétrica porque dependem das funções do kernel.

O kernel é um dos hiper parâmetros das *SVM*s. É um conjunto de funções matemáticas que ajudam a encontrar um hiperplano no espaço dimensional superior. O kernel pode ser uma função linear, quando existe linearidade nos dados, ou não linear quando os dados não são lineares. Os kernels mais comumente utilizados são apresentados na tabela 1.

Tabela 1. Kernels da *SVM* e sua representação matemática. Fonte dos dados: (PEDREGOSA et al., 2011).

Nome do Kernel	Função Kernel
Linear (produto escalar)	$\langle x, x' \rangle$
Sigmoidal	$\tanh(\gamma\langle x, x' \rangle + r)$
Polinômio	$(\gamma\langle x, x' \rangle + r)^d$
Função de Base Radial	$\exp(-\gamma\|x - x'\|^2)$

A região entre os vetores de suporte é denominada como margem e é caracterizada pela presença de um hiperplano, que é a linha necessária para ajustar os dados. As figuras 2 e 3 demonstra um exemplo simples de hiperplano:

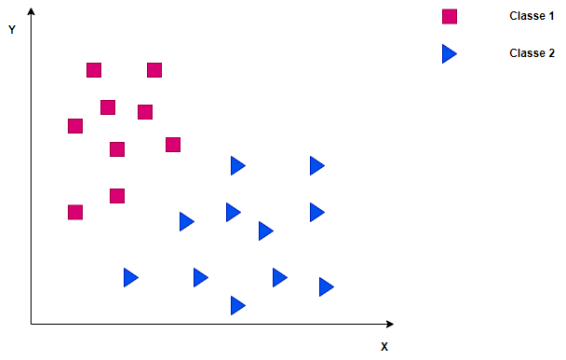

Figura 2. Uma classe de dois rótulos. Fonte da imagem: a autora.

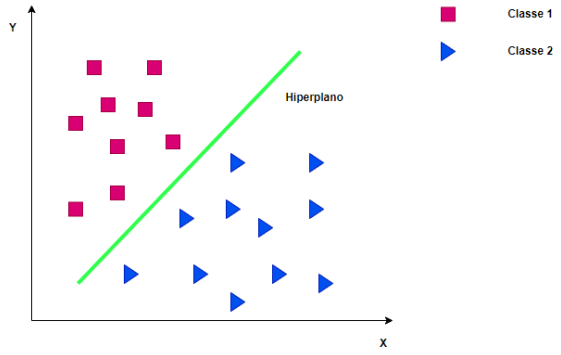

Figura 3. Separação simples de classes através de um hiperplano. Fonte da imagem: a autora.

As classes das figuras 2 e 3 estão idealmente separadas para fins de demonstração, no entanto, em situações reais os dados possuem mais entropia. Como pode ser visto nas figuras 4, 5 e 6:

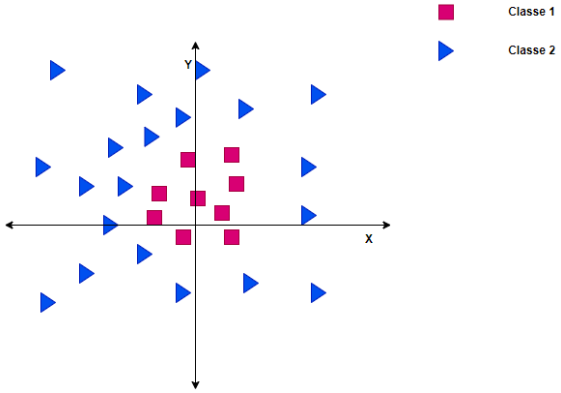

Figura 4. As classes não podem ser divididas facilmente por uma linha. Fonte da imagem: a autora.

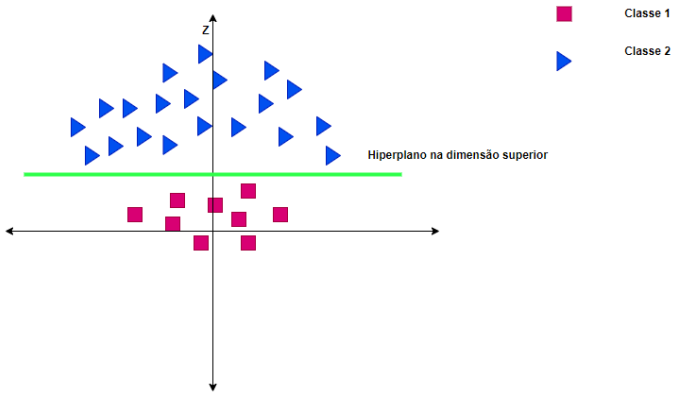

Figura 5. Como estratégia para poder dividir as classes, foi incluída uma terceira dimensão Z, um dimensão superior, e assim tracejar um hiperplano.

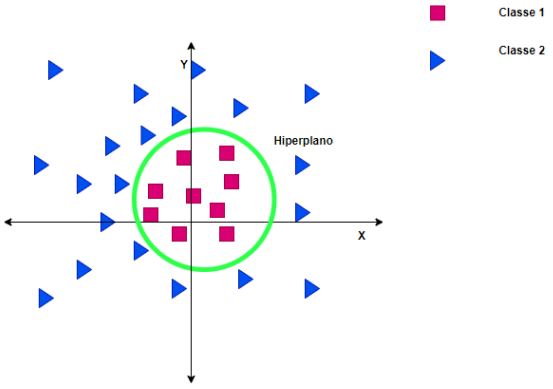

Figura 6. Ao transformar de volta ao plano original, o hiperplano consegue delimitar as classes. Fonte da imagem: a autora.

O exemplo visto nas figuras 4, 5 e 6, demonstra um simples exemplo de como as Máquinas de Vetor de Suporte trabalham para classificar os dados e de como o kernel eleva a uma dimensão superior para encontrar a margem entre os indivíduos.

No caso de regressões, o propósito não é dividir os dados em classes, mas prever valores discretos, encontrando o melhor hiperplano, que possui o número máximo de pontos (RAJ, 2020).

As máquinas de vetores de suporte regressivas, funcionam de modo reverso a classificações. A regressão *SVM* busca preencher a maior distância entre as instâncias de treinamento, limitando violações de margem, o que chamamos de largura de via. A largura de via é controlada pelo hiper parâmetro epsilon, ϵ.

Conforme é descrito em *Pattern Recognition and Machine Learning* (BISHOP, 2006), Máquinas de vetor de suporte preservam a propriedade de esparsidade, utilizando a maximização de margens para calcular a distância entre o hiperplano e as margens de modo que seja possível encontrar uma função $\mathrm{f}\,(\mathrm{x})$ que se desvie de $\mathrm{y_n}$ por um valor não maior que ϵ, para cada ponto de treinamento x e, ao mesmo tempo, seja o mais plano possível.

A figura 7 demonstra como é feita a maximização das margens.

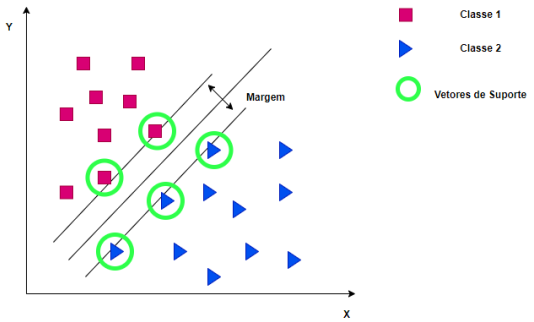

Figura 7. A maximização das margens ao redor do hiperplano. Os indivíduos circulados são os vetores de suporte, pois delimitam as margens. Fonte da imagem: a autora.

2.2.3. *Multilayer Perceptron*

É uma rede neural de múltiplas camadas. Uma rede neural é um algoritmo que tenta reproduzir de maneira artificial os neurônios humanos, suas interações e sinapses. Como um neurônio receptor humano recebe um sinal e este é transmitido pelo sistema nervoso humano, através de sinapses, um neurônio artificial, também conhecido como *perceptron*, transmite sinais através da rede neural utilizando funções de ativação que passam informações entre os neurônios artificiais.

Uma rede neural de múltiplas camadas, é um algoritmo de aprendizado de máquina supervisionado que aprende uma função $f(\cdot) : R^m \rightarrow R^o$. onde m é o número de dimensões para entrada e o é o número de dimensões para saída. Dado um conjunto de características $X = x_1, x_2, ..., x_m$ e um alvo y, ele pode aprender um aproximador de função não linear para classificação ou regressão, (PEDREGOSA et al., 2011).

Em *Pattern Recognition and Machine Learning* (BISHOP, 2006), é retratado como as redes neurais são um assunto discutido na área da matemática a muito tempo e que apesar de num primeiro momento as redes neurais tentarem replicar os conceitos biológicos, isto não implica uma limitação atualmente.

O livro (BISHOP, 2006) explica que o termo *"neural network"* teve sua origem na primeira metade do século XX (McCulloch and Pitts, 1943; Widrow and Hoff, 1960; Rosenblatt, 1962; Rumelhart *et al.*, 1986) e passou por diversas tentativas de criar sistemas extremamente semelhantes à sua contraparte biológica, no entanto, como isto não ocorreu conforme a expectativa, o estudo de redes neurais perdeu força, todavia, nos últimos anos, devido ao desenvolvimento de novas tecnologias, como as *GPUs*, o que gerou maior poder computacional, e com o aumento de fontes de dados, as redes neurais fornecem eficientes modelos estatísticos no reconhecimento de padrões, em particular a *Multilayer Perceptron*.

Quando falamos de redes neurais, podemos descrever como uma série de transformações funcionais. São construídas combinações lineares das variáveis de entrada, indicada como primeira camada da rede.

Primeiramente é construído uma combinação linear M das variáveis de entrada $x_1, ..., x_n$, que é representada na equação 1.

$$a_j = \sum_{i=1}^{n} w_{ji}^{(1)} x_i + w_{j^0}^{(0)} \tag{1}$$

Onde:

- $j = 1, ..., M$e o sobrescrito(1) indica que corresponde à primeira camada da rede neural;
- $w_{ji}^{(1)}$, se referem aos pesos;
- $w_{j^0}^{(0)}$, se refere a *biases*, que ajusta o limiar dos pesos w;
- as quantidades a_j são conhecidas como ativações. Cada ativação é transformada utilizando uma função de ativação não linear $h(.)$ para

$$z_j = h(a_j). \tag{2}$$

A Figura 8 demonstra a arquitetura de um neurônio artificial, indicando a primeira camada, responsável pelas entradas (*inputs*), a segunda camada, responsável pelos pesos (*weitghs*), as camadas ocultas foram abstraídas pelo somatório das transformações funcionais, em seguida há a função de ativação, e ao final a última camada, que exporta os *outputs*, a função objetivo do aprendizado. Esta figura apresenta apenas uma saída, no entanto, as redes neurais podem ter mais de uma saída.

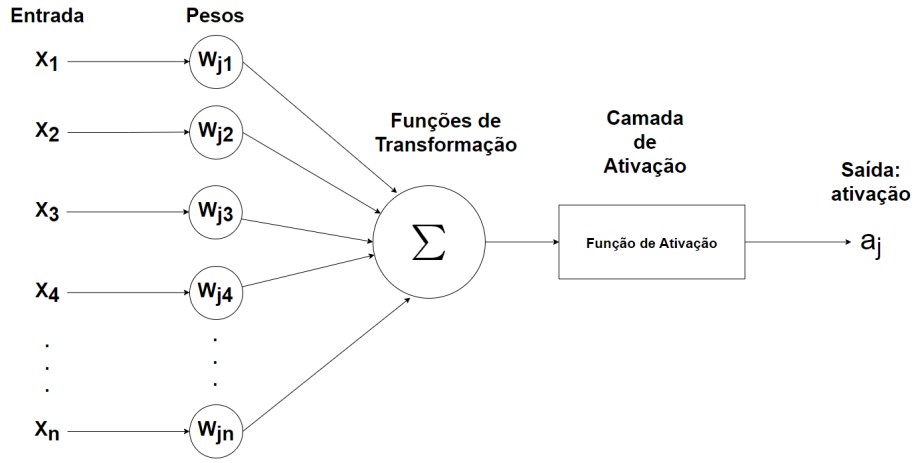

Figura 8. A esquematização de um neurônio Artificial. Fonte da imagem: a autora.

2.3. Análise de Componentes Principais

O problema conhecido como a Maldição da Dimensionalidade, é um dos maiores vilões do aprendizado de máquina porque quanto maior a quantidade de características que uma base de dados possui, maior é a complexidade computacional para o treinamento de um modelo de aprendizado de máquina.

A PCA ou *Principal Component Analyses*, em tradução livre Análise de Componentes Principais, é uma técnica amplamente utilizada para redução de dimensionalidade, compressão de dados, extrações de recurso e visualização de dados (Jolliffe, 2002). É também conhecida como a transformação de Karhunen-Loève.

Existem duas definições de PCA comumente usadas que dão origem ao mesmo algoritmo. A PCA pode ser definida como a projeção ortogonal dos dados em um espaço linear de menor dimensão, conhecido como subespaço principal, de modo que a variância dos dados projetados seja maximizada (Hotelling, 1933). De forma equivalente, pode ser definida como a projeção linear que minimiza o custo médio de projeção, definido como a distância quadrática média entre os pontos de dados e suas projeções (Pearson, 1901) (BISHOP, 2006).

2.4. Pré-processamento

Nesta subseção serão descritas as técnicas utilizadas para o tratamento dos dados.

2.4.1. Tratamentos para Dados Categóricos

Muitos algoritmos de aprendizado de máquina não podem operar diretamente nos dados categóricos, pois eles precisam que todas as variáveis de entrada e de saída sejam numéricas. Isto é principalmente uma restrição da implementação eficiente de algoritmos de aprendizado de máquina, em vez de limitações rígidas nos próprios algoritmos (BROWNLEE, 2020).

Algumas implementações de algoritmos de aprendizado de máquina exigem que todos os dados sejam numéricos. Por exemplo, scikit-learn (BUITINCK et al., 2013) tem esse requisito.

Os tratamentos mais encontrados utilizados para dados categóricos são:

1. Codificação Ordinal: Quando existe alguma ordenação natural entre os rótulos.
2. Codificação *One-Hot*: Quando não há nenhum grau de importância entre os dados.

2.4.2. Codificação *One-Hot*

Quando não existe relação ordinal para as variáveis, forçar um relacionamento ordinal por meio de uma codificação ordinal e permitir que o modelo assuma uma ordenação natural entre categorias pode resultar em desempenho ruim ou resultados inesperados. Por isto, quando as variáveis categóricas não possuírem grau de importância entre si, a codificação *One-Hot* deverá ser a escolha para o tratamento, caso este seja necessário, que é caso das regressões. Caso as variáveis categóricas possuam grau de importância entre si, a codificação ordinal é mais adequada.

A técnica de codificação *One-Hot* também é conhecida como binarização dos dados categóricos, pois ela cria uma coluna para cada categoria de uma característica específica, preenchendo a coluna com 0 ou 1, onde o 0 é utilizado para quando o registro da base de dados não pertence a categoria e 1 quando o registro pertence a categoria.

2.4.3. Imputação de Valores Faltosos

A ausência de valores é outro problema conhecido na área de inteligência artificial. Nos dias atuais há bilhões de dados espalhados em diversas bases de dados, no entanto, a qualidade dos dados deixam a desejar, principalmente por conta dos valores faltosos, e por isso se faz necessário lidar com eles antes de realizar o treinamento dos modelos de aprendizado de máquina.

Existem ao menos 3 motivos para a ausência dos dados:

1. Dados são perdidos, mas estão relacionados a alguns dos dados observados;
2. Dados são completamente perdidos e não estão relacionados a outras características das bases de dados;
3. Dados perdidos dependem de outras características das bases de dados, como por exemplo, pessoas com altos salários não gostam de revelar seus salários ou mulheres não gostam de revelar a sua idade.

Algumas técnicas são utilizadas para tratar a ausência de dados, no entanto, qual técnica é melhor, depende das características específicas das bases de dados e de suas variáveis.

Uma alternativa é eliminar os indivíduos que estão com valores faltosos, mas isto pode causar a perda de dados preciosos para o modelo a ser treinado. Por conta disso, a imputação dos dados faltosos é uma parte importantíssima do pré-processamento.

As imputações mais simples utilizam média, moda ou mediana para preencher os dados faltosos, porém reduz a variância do conjunto de dados. Há ainda outras abordagens, como séries temporais, regressão linear, imputação múltipla, imputação com *K-Nearest Neighbors*, entre outros métodos, que são implementados de acordo com os metadados das bases de dados (SWALIN, 2018).

2.5. Seleção de Hiper Parâmetros

Os Hiper Parâmetros não são definidos pelo próprio modelo, devem ser definidos ou alterados pelo supervisor do modelo. Normalmente estes parâmetros são ajustados para obter a melhor performance do modelo, ou seja, obter o melhor desempenho. Cada tipo de algoritmo de aprendizado possui seus próprios hiper parâmetros e estes são definidos de acordo com as decisões do supervisor, por tentativa e erro e pelas métricas de medida de desempenho. Também existem algoritmos que ajudam na escolha de hiper parâmetros, são eles o *Grid Search* e o *Randomized Search*.

2.5.1. *Grid Search*

Realiza uma busca ordenada de hiper parâmetros para um dado algoritmo de aprendizado de máquina.

2.5.2. *Randomized Search*

Realiza uma busca aleatória a partir de uma distribuição de probabilidade de valores possíveis para os hiper parâmetros de um dado algoritmo de aprendizado de máquina. Seu custo computacional é menor que o do *Grid Search* e possui melhores resultados (BERGSTRA; BENGIO, 2012).

2.6. Medidas de desempenho e avaliação

Nesta subseção serão apresentadas as medidas de avaliação utilizadas para verificar a qualidade e desempenho dos modelos treinados, em algoritmos baseados em regressão.

2.6.1. Erro Absoluto Médio - MAE

O MAE é calculado como a soma dos erros absolutos dividido pelo tamanho da amostra. É uma métrica de risco correspondente ao valor esperado da perda de erro absoluto ou $l1$-perda de norma. O MAE é definido na equação 3.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{amostra}}} \sum_{i=0}^{n_{\text{amostra}}-1} |y_i - \hat{y}_i| \,. \tag{3}$$

onde,

y : é o valor alvo real da base de treinamento;
$\hat{y}$: é o valor predito;
n_{amostra} : é o tamanho da amostra.

2.6.2. Erro Quadrado Médio - MSE

O MSE é calculado como a soma dos erros quadráticos dividido pelo tamanho da amostra. É uma métrica de risco correspondente ao valor esperado do erro quadrático. O MSE é definido na equação 4.

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{amostra}}} \sum_{i=0}^{n_{\text{amostra}}-1} (y_i - \hat{y}_i)^2. \tag{4}$$

onde,

y : é o valor alvo real da base de treinamento;
$\hat{y}$: é o valor predito;
n_{amostra} : é o tamanho da amostra.

2.6.3. Erro Percentual Absoluto Médio - MAPE

O MAPE, também conhecido como desvio percentual absoluto médio (MAPD), é uma métrica de avaliação para problemas de regressão. É uma métrica sensível a erros relativos. Por exemplo, não é alterado por uma escala global da variável de destino. O MAPE é definido na equação 5.

$$\text{MAPE}(y, \hat{y}) = \frac{1}{n_{\text{amostra}}} \sum_{i=0}^{n_{\text{amostra}}-1} \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)} \tag{5}$$

onde,

y : é o valor alvo real da base de treinamento;
$\hat{y}$: é o valor predito;
n_{amostra} : é o tamanho da amostra.
ϵ : é um número arbitrário pequeno,mas estritamente positivo para evitar resultados indefinidos quando y é zero.

3. Trabalhos Relacionados

Nesta seção serão descritos os trabalhos correlatos relevantes e como eles convergem e divergem deste estudo.

O artigo *"Analysis and prediction of covid-19 pandemic in bangladesh by using anfis and lstm network."* (CHOWDHURY; HASAN; HOQUE, 2021), descreve um estudo realizado com dados coletados em Bangladesh, usando as redes neurais ANFIS e LSTM, para analisar e prever casos novos de COVID-19, através do tempo.

O ANFIS, ou em livre tradução, Sistema Adaptativo de Inferência Neuro-Fuzzy, é uma rede neural modificada que utiliza a modificação do sistema de lógica fuzzy, a assimilação do fuzzy Takagi-Sugeno.

Esta arquitetura possui 5 camadas:

1. Camada de entrada: recebe os parâmetros e os constrói no modelo, assim como no modelo de sistema fuzzy;
2. Segunda camada: carrega os valores advindos da primeira camada, através das funções-membro (MF). Nos nós da segunda camada, a fixação do grau de atividade em regras fuzzy são concluídas.
3. Terceira camada: amplia o grau de atuação de quaisquer regulamentações.
4. Quarta camada: normaliza as funções, e os nós facilitam a produção das saídas.
5. A última camada de saída é responsável por exportar as saídas.

O LSTM, ou Memória de Longo Prazo, em livre tradução, é uma arquitetura que consiste em quatro portas:

1. Porta de entrada.
2. Porta de esquecimento.
3. Porta de controle.
4. Porta de saída.

Assim como nosso estudo, a abordagem do artigo *"Analysis and prediction of covid-19 pandemic in bangladesh by using anfis and lstm network."* (CHOWDHURY; HASAN; HOQUE, 2021) é desenvolver um modelo de previsão e facilitar o processo de tomada de decisão, analisando os dados para obter uma perspectiva futura. No entanto, este estudo visa identificar os casos de COVID-19 através do tempo, enquanto o nosso estudo prediz a quantidade de mortes, de acordo com variações das características sanitárias, demográficas e populacionais, pois é uma situação irreversível em termos de prevenção de vida. Ocorre também que o estudo foi realizado em 2020, e na época, havia dados insuficientes para aprendizado profundo. Os algoritmos utilizados para comparação neste estudo foram modificações de redes neurais, enquanto nosso estudo verifica as diferenças entre 3 tipos diferentes de algoritmos, redes neurais, arvores de decisão e máquina de vetor de suporte, na abordagem da regressão.

No artigo *"A study on office workplace modification during the covid-19 pandemic in the netherlands"* (HOU et al., 2021), é descrito um estudo exploratório sobre o ambiente de trabalho do escritório, nos Países Baixos, conhecidos como Holanda. O artigo visa investigar como as organizações respondem aos impactos causados pela pandemia de COVID-19 e como modificam a gestão do local de trabalho do escritório, de forma

estratégica e operacional, para atender às necessidades advindas do contexto pandêmico e ao desenvolvimento futuro no período pós-covid-19.

O artigo *"A study on office workplace modification during the covid-19 pandemic in the netherlands"* (HOU et al., 2021) desenvolve uma metodologia através de entrevistas a gerentes e analistas que precisaram modificar seus locais de trabalho, por conta da pandemia de COVID-19, com o propósito de fornecer uma lente prática para observar as mudanças futuras do ambiente de trabalho do escritório.

O estudo *"A study on office workplace modification during the covid-19 pandemic in the netherlands"* (HOU et al., 2021) é um trabalho analítico e exploratório, que visa identificar as percepções dos trabalhadores para identificar os desafios encontrados, no entanto, ele não fornece a proposta de uma ferramenta baseada em aprendizado de máquina que suporte as decisões no enfrentamento aos desafios encontrados.

4. Abordagem da Proposta

Estudo de técnicas preditivas para o auxílio a gestores na pandemia de COVID-19.

Nas seções anteriores foi visto o referencial teórico das ferramentas e técnicas que serão utilizados neste estudo. Também foram referenciados os trabalhos relacionados com a temática deste estudo, que demonstram como este estudo está inserido no meio do aprendizado de máquina e em temas relacionados à saúde, assim como podemos usar a inteligência artificial para ajudar as organizações a enfrentar os desafios encontrados na gestão.

Nesta seção vamos abordar o tema principal deste estudo.

Como foi demonstrado na introdução deste artigo, este estudo visa propor uma alternativa para auxílio na tomada de decisão de gestores durante a pandemia de COVID-19.

Este estudo pretende fornecer uma ferramenta que auxilie na tomada de decisão. Um sistema onde o usuário insira os valores para as *features* e que ele possa predizer a quantidade de mortes que resultaria. Por exemplo, um gestor pode projetar um cenário em que a prevalência de diabetes tenha aumentado e verificar se isso aumenta ou diminui o número de mortes. A partir deste resultado ele pode pensar em como criar medidas para que o número de mortes não aumente por conta da prevalência de diabetes.

O aumento de doses de vacinas administradas diminui ou aumenta a quantidade de mortes? Se o gestor de uma multinacional possui uma filial num continente A e neste continente tem um total Y de vacinas aplicadas, ele pode verificar que o número de mortes tende a aumentar ou diminuir e pode ajudar ele a tomar alguma decisão de voltar às atividades presenciais ou não.

O que geralmente ocorre, hoje em dia, é que as empresas estão olhando os dados oficiais do governo ou veiculações da mídia, que são baseados em estatística exploratória, o que de certa forma ajuda bastante na tomada de decisão, no entanto, estes dados tem um problema, eles são referentes ao passado e no máximo ao presente, mas não ao futuro. O futuro é incerto nessas análises.

Foi visto bastante no ano de 2021 um ciclo constante de abre e fecha do comércio, pois as tomadas de decisões estavam embasadas apenas nos levantamentos estatísticos

exploratórios.

Para fomentar este estudo foi realizada uma busca analítica e exploratória, chegando à base de dados *Data on COVID-19 (coronavirus) by Our World in Data*. (RITCHIE et al., 2020).

4.1. Conjunto de dados

Os dados utilizados neste estudo foram extraídos da base de dados disponibilizada pela *Our World in Data* (RITCHIE et al., 2020), uma organização mundial, que tem por objetivo propagar o estudo de dados para ajudar nos problemas que envolvem pobreza, doença, fome, mudança climática, guerra, riscos existenciais e desigualdade. Os dados são atualizados numa frequência que vai de diária a semanal e é mantida a partir de dados extraídos das fontes que podem ser verificados na tabela 2.

Tabela 2. Origem dos dados. Fonte dos dados: (RITCHIE et al., 2020)

.

Casos e mortes confirmados	Repositório de Dados COVID-19 do Centro de Ciência e Engenharia de Sistemas (CSSE) da Universidade Johns Hopkins (JHU) (DONG; DU; GARDNER, 2020).
Internações e internações em unidades de terapia intensiva (UTI)	São coletados de fontes oficiais e compilados pelo Our World in Data.
Vacinas contra o COVID-19	São coletados pela equipe Our World in Data a partir de relatórios oficiais (MATHIEU et al., 2021).
Outras variáveis	Coletados de várias fontes (Nações Unidas, Banco Mundial, Global Burden of Disease, Blavatnik School of Government, etc.) (HASELL et al., 2020).

As características selecionadas para predizer o número de mortes por COVID-19 estão descritas na tabela 3.

Tabela 3: Características selecionadas para predizer o número de mortes por COVID-19 Fonte dos dados:(RITCHIE et al., 2020).

Característica	**Tipo de dado**
Novos casos confirmados de COVID-19	Contínuo

Número de pacientes com COVID-19 em unidades de terapia intensiva (UTIs) em um determinado dia.	Contínuo
Número de pacientes com COVID-19 no hospital em um determinado dia	Contínuo
Número total de doses de vacinação COVID-19 administradas	Contínuo
Número total de pessoas que receberam pelo menos uma dose de vacina	Contínuo
Número total de pessoas que receberam todas as doses prescritas pelo protocolo de vacinação inicial	Contínuo
Número total de doses de reforço de vacinação COVID-19 administradas (doses administradas além do número prescrito pelo protocolo de vacinação)	Contínuo
Novas doses de vacinação COVID-19 administradas	Contínuo
Índice de Rigidez de Resposta do Governo	Contínuo
População	Contínuo
Número de pessoas dividido por área de terra, medido em quilômetros quadrados, ano mais recente disponível	Contínuo
Idade média da população, projeção da ONU para 2020	Contínuo
Parcela da população com 65 anos ou mais, ano mais recente disponível	Contínuo
Parcela da população com 70 anos ou mais em 2015	Contínuo
Produto interno bruto em paridade de poder de compra	Contínuo
Percentual da população que vive em extrema pobreza, ano mais recente disponível desde 2010	Contínuo
Taxa de mortalidade por doença cardiovascular em 2017 (número anual de mortes por 100.000 pessoas)	Contínuo
Prevalência de diabetes (% da população de 20 a 79 anos) em 2017	Contínuo
Proporção de mulheres que fumam, ano mais recente disponível	Contínuo
Proporção de homens que fumam, ano mais recente disponível	Contínuo

Percentagem da população com instalações básicas de lavagem das mãos nas instalações, ano mais recente disponível	Contínuo
Camas hospitalares por 1.000 pessoas, ano mais recente disponível desde 2010	Contínuo
Expectativa de vida ao nascer em 2019	Contínuo
Índice de desenvolvimento humano	Contínuo
Continente	Categórico

O Índice de Rigidez de Resposta do Governo é uma medida composta com base em 9 indicadores de resposta ao enfrentamento a COVID-19, são eles:

1. O fechamento de escolas;
2. Fechamentos de locais de trabalho;
3. Cancelamento de eventos públicos;
4. Restrições a reuniões públicas;
5. Fechamento de transporte público;
6. Requisitos de permanência em casa;
7. Campanhas de informação pública;
8. Restrições aos movimentos internos;
9. Controles de viagens internacionais.

Ele foi reescalonado para um valor de 0 a 100, sendo 100 a resposta mais rigorosa e 0 quando não há registro de nenhuma resposta. Este índice foi criado no projeto Oxford *COVID-19 Government Response Tracker (OxCGRT)* (HALE et al., 2021) .

5. Experimentos e Resultados

Na seção anterior, vimos a abordagem da proposta e a descrição dos dados utilizados. Nesta seção iremos apresentar a metodologia utilizada para este estudo e os resultados obtidos da metodologia aplicada.

5.1. Ambiente Experimental

Nesta subseção será apresentado o ferramental técnico utilizado neste estudo.

5.1.1. Linguagem de Programação e Bibliotecas Utilizadas

Todas as etapas que exigiram codificação, da análise dos dados à avaliação dos modelos treinados foram executadas utilizando a linguagem de programação Python, versão 3.7.13 (ROSSUM; DRAKE, 2009). As bibliotecas e suas aplicações no experimento estão descritas na tabela 4.

Tabela 4. Bibliotecas Python utilizadas no experimento. Fonte: a autora.

Biblioteca	Aplicação
Scikit-learn	Implementações de aprendizado de máquina
Pandas, SciPy e NumPy	Pré-processamento de dados e cálculos
Matplotlib e Seaborn	Plotagem de gráficos

5.1.2. Ambiente de Desenvolvimento

O ambiente de desenvolvimento utilizado para este estudo foi o Google Colab Pro +, devido à facilidade de utilização, direta no navegador, já possuir muitas bibliotecas pré-instaladas e na sua versão Pro + ter acesso à execução em segundo plano, e acesso a mais recursos computacionais, tais como mais memória RAM, maior disco, e poder utilizar GPUs e TPUs mais rápidas.

5.2. Pré-processamento

Como foi explicado na seção de fundamentação teórica, seção 2, as bases de dados possuem alguns problemas, como o da dimensionalidade, presença de dados faltosos e características que não podem ser utilizadas em alguns algoritmos, devido à limitação das implementações ou que não fazem sentido para o objetivo do estudo. Nesta subseção serão descritos os procedimentos realizados para lidar com o ruído dos dados antes de realizar o treinamento.

5.2.1. Análise da base de dados

O estudo utilizou a base de dados da *Our World in Data* (RITCHIE et al., 2020), extraída no dia 07 de maio de 2022, para prever o número de mortes de acordo com os valores das características sanitárias, populacionais e demográficas.

A base de dados original possui 67 características, que podem ser categóricas ou contínuas, 25 foram selecionadas, pois algumas características possuem redundância de dados e outras não são interessantes para o estudo, porque não estão relacionadas a características sanitárias, populacionais e demográficas.

Os registros que fazem referência a separações geopolíticas também foram descartados do treinamento. Estes dados não possuem valor preenchido na coluna continente. Tal informação é descrita pelos responsáveis pela manutenção da base de dados (RITCHIE et al., 2020).

Todas as características categóricas foram transformadas em dados numéricos para serem consideradas pela *API* do scikit-learn, utilizada neste estudo (PEDREGOSA et al., 2011) e (BUITINCK et al., 2013).

A característica localidade, que faz referência aos países, não se inclui nos cenários para descarte, mas, todavia, se fez necessária sua retirada, pois no tratamento

da codificação *one-hot*, a quantidade de características expandiu para 261, o que aumentou a complexidade e prejudicou a performance computacional na execução dos algoritmos, principalmente no caso da *SVM*, mesmo com uma subamostra da base de dados, deste modo, a característica continente ficou responsável pela delimitação geográfica dos dados.

5.2.2. Imputação de valores faltosos

Neste estudo, o *KNN Imputer*, com K = 2, foi o método de imputação escolhido, pois ele já foi utilizado para trabalhos relacionados à área da saúde (TROYANSKAYA et al., 2001).

5.2.3. Codificação *One-Hot*

A codificação *One-Hot* foi escolhida para o tratamento de variáveis categóricas, neste estudo, pois foi necessário transformar os dados categóricos em dados contínuos e para não ocorrer indução de grau de importância eles foram binarizados.

A característica continente foi binarizada e por conta disto, a base de dados possui 30 características, sendo cinco referentes aos continentes (*Asia, Africa, Europe, North America, Oceania* e *South America*). Quando o registro pertencer ao continente, ele terá o valor 1 e quando não pertencer, possuirá o valor 0.

5.2.4. Escalonamento de valores contínuos

Os dados contínuos foram escalonados para os valores ficarem no intervalo de 0 a 1, para padronização e evitar que a divergência de dimensionalidade das características influenciam negativamente nos treinamentos e por consequência impactem os resultados.

5.2.5. Sub amostragem dos dados

Devido a base de dados escolhida possuir 67 características e mesmo que, após a seleção manual dos dados, este número tenha caído para 30, a base de dados continuou com alta dimensionalidade, pois, por se tratar de uma base de dados mundial, e mesmo com a exclusão de dados referentes a continentes e separações geopolíticas, a base de dados continuou com 172872 registros, o que impacta no tempo de processamento de alguns algoritmos, como o *SVM* por exemplo.

Foi feita uma sub amostragem randômica de 40% de registros para este estudo, o que representa 69149 registros.

5.3. Treinamento dos modelos

Neste estudo foram utilizados 3 tipos de algoritmos de aprendizado de máquina (*decision tree*, *MLP* e *SVM*), utilizando regressão por se tratar de um estudo com foco na predição de um valor contínuo.

A fim de encontrar o modelo de inteligência artificial que apresente o melhor desempenho foi utilizado a busca aleatória, para encontrar os melhores hiper parâmetros, para cada algoritmo.

A busca aleatória foi executada através de um pipeline para executar o processo de treinamento, e possui as seguintes etapas:

1. Escalonamento da base de treinamento;
2. Análise de Componentes Principais;
3. Execução do algoritmo.

Ao final de cada treinamento foram extraídos os melhores hiper parâmetros dos modelos e as métricas dos algoritmos.

5.4. Seleção de Modelos

Nesta subseção, iremos mostrar as distribuições de probabilidades utilizadas para cada algoritmo treinado, assim como os melhores hiper parâmetros e as métricas dos desempenhos dos modelos selecionados pela busca aleatória.

5.4.1. *Decision Tree*

No primeiro momento foi utilizado o algoritmo *Decision Tree*, no entanto, usando apenas uma árvore de decisão, ocorreu muito *overfitting*, então para reduzir, foi usado o meta estimador Floresta Aleatória, que é uma extensão da árvore de decisão que usa várias árvores de decisão para moldar o modelo.

As distribuições de probabilidades são apresentadas na tabela 5, os hiper parâmetros selecionados na tabela 7 e as métricas do modelo selecionado pela busca aleatória na tabela 8.

Tabela 5. Distribuição de probabilidades para a Busca Aleatória para Floresta Aleatória. Fonte dos dados: (PEDREGOSA et al., 2011).

Hiper parâmetro	Distribuição de Probabilidades
Função para medir a qualidade da divisão de um nó	Erro quadrático médio Erro absoluto médio Redução no desvio de Poisson
Profundidade máxima	Inteiros de 1 a 1000
O número de recursos a serem considerados ao procurar a melhor divisão	Verificar a tabela 6
Número máximo de nós folhas	Inteiros de 1 a 1000
Fração mínima do número de instâncias de treinamento para em nó folha	Entre 0 e 2
Fração mínima do número de instâncias de treinamento para divisão de nó.	Entre 0 e $0,5$

Tabela 6. Possíveis valores para o número de recursos a serem considerados ao procurar a melhor divisão. Fonte dos dados: (PEDREGOSA et al., 2011)

Função	Descrição
Inteiro	Considera os recursos em cada divisão
Ponto Flutuante	Considera os recursos em cada divisão
Automático	Utiliza o quadrático
Quadratico	Utiliza a raiz quadrada da quantidade de características
Logartimo na base 2	Utiliza a função logarítmica de base 2 na quantidade de características
Nenhum	Utiliza a quantidade de características

Tabela 7. Hiper Parâmetros Selecionados Pela Busca Aleatória Para Floresta Aleatória. Fonte: a autora.

Hiper parâmetro	Valor
Função para medir a qualidade da divisão de um nó	Erro quadrado médio
Profundidade máxima	752
O número de recursos a serem considerados ao procurar a melhor divisão	Nenhum
Número máximo de nós folhas	981
Fração mínima do número de instâncias de treinamento para em nó folha	$0,259862$
Fração mínima do número de instâncias de treinamento para divisão de nó	$0,562235$

Tabela 8. Métricas do modelo treinado da Floresta Aleatória. Fonte: a autora.

Métrica	Valor
Erro Absoluto Médio - MAE	$1,203738e^{-02}$
Erro Quadrado Médio - MSE	$1,450833e^{-03}$
Erro percentual absoluto médio - MAPE	$5,366393e^{+12}$

5.4.2. *Support Vector Machine*

O treinamento com *Support Vector Machine* foi o mais desafiador, pois foi o algoritmo que mais requereu tempo de processamento. *SVM's* são muito boas para lidar com dados contínuos, mas não são boas para treinamentos com uma grande base de dados. Mesmo com a base de dados sub amostrada, o processamento demandou mais tempo que os demais treinamentos e não foi possível verificar o kernel polinomial na busca aleatória.

Como a *SVM* apresentou problemas de tempo de execução, a busca aleatória foi executada por kernel. Portanto foram realizadas 4 execuções, uma para cada kernel, no entanto o kernel polinomial ainda apresentou o problema e não pode ser concluído.

A distribuição de probabilidades está descrita na tabela 9. Os hiper parâmetros selecionados pela busca aleatória para os kernels Função de Base Radial, sigmoidal e Linear, respectivamente nas tabelas, 10, 12 e 14 e as métricas dos modelos treinados, respectivamente nas tabelas 11, 13 e 15.

Tabela 9. Distribuição de probabilidades para a Busca Aleatória para *Support Vector Machine*. Fonte dos dados: (PEDREGOSA et al., 2011).

Hiper parâmetro	Distribuição de Probabilidades
C	Uniforme entre $1e^{-04}$ e $1e^{+04}$
Épsilon	Uniforme entre $1e^{-04}$ e $1e^{+04}$
Kernel	Polinomial, Função de Base Radial ou Sigmoidal
Grau do polinômio (kernel polinomial)	Inteiro entre 2 e 3
Gama (inverso da variância $sigma^2$, Kernel Gaussiano e Sigmoidal)	Uniforme entre 0.5 e 1

Tabela 10. Hiper Parâmetros Selecionados Pela Busca Aleatória Para *Support Vector Machine* - Kernel Função de Base Radial. Fonte: a autora.

Hiper parâmetro	Valor
C	$0,004865$
Épsilon	$0,003821$
Kernel	Função de Base Radial
Gama (inverso da variância $sigma^2$, Kernel Gaussiano e Sigmoidal)	$0,56602$

Tabela 11. Métricas do modelo treinado da *Support Vector Machine* - Kernel Função de Base Radial. Fonte: a autora.

Métrica	Valor
Erro Absoluto Médio - MAE	$6,866216e^{-03}$
Erro Quadrado Médio - MSE	$1,112506e^{-03}$
Erro percentual absoluto médio - MAPE	$7,747901e^{+12}$

Tabela 12. Hiper Parâmetros Selecionados Pela Busca Aleatória Para *Support Vector Machine* - Kernel Sigmoidal. Fonte: a autora.

Hiper parâmetro	Valor
C	$0,920401$
Épsilon	$7546,173904$
Kernel	Sigmoidal
Gama (inverso da variância $sigma^2$, Kernel Gaussiano e Sigmoidal)	$0,758226$

Tabela 13. Métricas do modelo treinado da *Support Vector Machine* - Kernel Sigmoidal. Fonte: a autora.

Métrica	Valor
Erro Absoluto Médio - MAE	$4,750538e^{-01}$
Erro Quadrado Médio - MSE	$2,265819e^{-01}$
Erro percentual absoluto médio - MAPE	$1,054522e^{+15}$

Tabela 14. Hiper Parâmetros Selecionados Pela Busca Aleatória Para *Support Vector Machine* - Kernel Linear. Fonte: a autora.

Hiper parâmetro	Valor
C	$0,003646$
Épsilon	$0,007291$

Tabela 15. Métricas do modelo treinado da *Support Vector Machine* - Kernel Linear. Fonte: a autora.

Métrica	Valor
Erro Absoluto Médio - MAE	$8,374376e^{-03}$
Erro Quadrado Médio - MSE	$8,584649e^{-04}$
Erro percentual absoluto médio - MAPE	$9,380100e^{+12}$

5.4.3. *Multilayer Perceptron*

Das três abordagens a *Multilayer Perceptron* foi o algoritmo que apresentou menos problemas durante a execução dos experimentos. Apenas o pré-processamento foi necessário para o seu treinamento.

As distribuição de probabilidades estão descritas na tabela 16, os hiper parâmetros selecionados, na tabela 17 e as métricas do modelo treinado da *Multilayer Perceptron*, na tabela 18.

Tabela 16. Distribuição de probabilidades para a Busca Aleatória para *Multilayer Perceptron*. Fonte dos dados: (PEDREGOSA et al., 2011).

Hiper parâmetro	Distribuição de Probabilidades
Número de neurônios escondidos	Inteiro entre 1 a 100
Função de ativação	Identidade, Logística e Tangencial
Taxa de aprendizado	Exponencial de$1e^{-6}$ a $0,1$
Número de épocas	Inteiro entre 1 a 1000
Taxa de momento	Uniforme entre $0,5$ e 1

Tabela 17. Hiper Parâmetros Selecionados Pela Busca Aleatória Para a *Multilayer Perceptron*. Fonte: a autora.

Hiper parâmetro	Valor
Número de neurônios escondidos	75
Função de ativação	Logística
Taxa de aprendizado	$0,014816$
Número de épocas	226
Taxa de momento	$0,241785$

Tabela 18. Métricas do modelo treinado da *Multilayer Perceptron*. Fonte: a autora.

Métrica	Valor
Erro Absoluto Médio - MAE	$2,047666e^{-02}$
Erro Quadrado Médio - MSE	$6,842645e^{-04}$
Erro percentual absoluto médio - MAPE	$4,036332e^{+13}$

5.5. Análise dos Resultados

Como os modelos treinados, foram obtidos a partir de algoritmos que utilizam regressão, a principal métrica escolhida para determinar a qualidade dos modelos foi o erro absoluto médio (MAE), e a segundaria, o erro quadrado médio (MSE), pois ambas são métricas que verificam os resultados preditos frente aos valores reais durante os testes.

O erro percentual absoluto médio (MAPE) não foi utilizado para avaliar o desempenho dos modelos, pois o ϵ tende a tornar o valor do MAPE muito alto quando os valores reais dos dados estão próximos a 0, e como a base de treinamento foi escalonada em valores de 0 a 1, os resultados da métrica não foram considerados neste estudo.

Dos 3 algoritmos de inteligência artificial utilizados, os modelos baseados em *Support Vector Machine*, foram os que apresentaram o melhor desempenho, pois o modelo com kernel Função de Base Radial possui o menor erro absoluto médio e também os modelos, com os kernels Função de Base Radial e linear, possuem 2 dos 3 menores erros quadrados médios.

O Segundo melhor modelo foi o baseado em *Multilayer Perceptron*, com erro absoluto médio e erro quadrado médio também baixos. O Terceiro melhor foi a Floresta Aleatória. O modelo com o pior desempenho entre os estudados foi a *Support Vector Machine* com kernel sigmoidal.

A diferença entre os modelos baseados nos 3 algoritmos utilizados neste estudo são apresentados nas figuras 9 e 10, onde são verificados os valores dos erros absolutos médios e os erros quadrados médios, respectivamente.

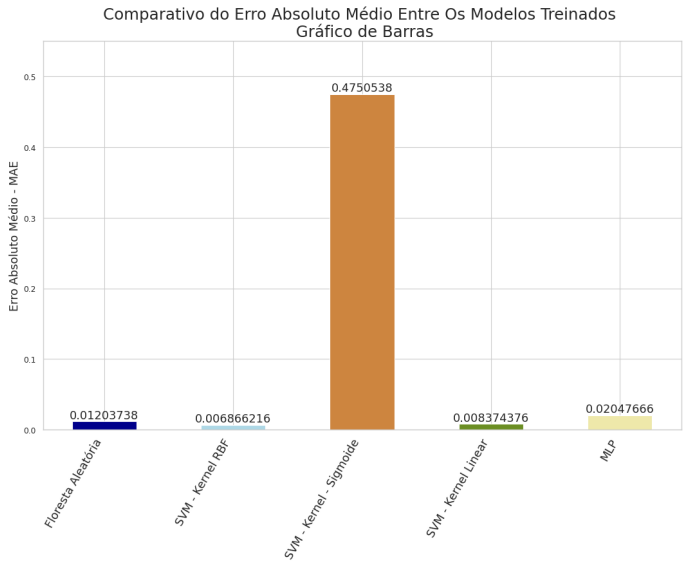

Figura 9. Comparativo do Erro Absoluto Médio Entre Os Modelos Treinados. Fonte da imagem: a autora.

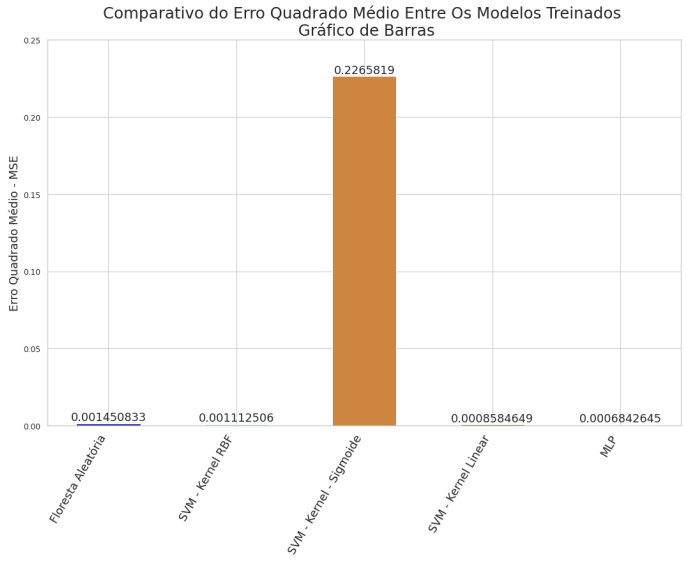

Figura 10. Comparativo do Erro Quadrado Médio Entre Os Modelos Treinados. Fonte da imagem: a autora.

Como é possível verificar nos gráficos das figuras 9 e 10, os erros absolutos médios

e os erros quadrados médios são valores baixos, bem próximos de 0, e mesmo que a base de dados esteja escalonada entre 0 e 1, os menores erros estão à 3 e 4 casas a direita da vírgula, em notação científica: e^{-03} e e^{-04}.

O menor valor do MAE foi de $6,866216e^{-03}$ para a *SVM* utilizando o kernel função de base radial, seguido de $8,374376e^{-03}$ para a *SVM* utilizando o kernel linear, $1,203738e^{-02}$ para a floresta aleatória e $2,047666e^{-02}$ para o *MLP*.

No caso do MSE, o modelo da rede neural, *MLP* foi o que teve melhor desempenho, com $6,842645e^{-04}$ de erro quadrado médio, seguido de $8,584649e^{-04}$ para a *SVM* com Kernel linear, $1,112506e^{-03}$ para a *SVM* com kernel função de base radial e $1,450833e^{-03}$ para a floresta aleatória.

O modelo com o pior desempenho é o baseado em *SVM* com kernel sigmoidal. O que pode se levantar a hipótese de problemas de processamento na busca aleatória, como foi o caso do kernel polinomial, que não chegou a concluir o processamento por falta de recursos computacionais do ambiente experimental utilizado.

No caso da *SVM* com Kernel sigmoidal será necessário maior empenho no pré-processamento e de maior poder computacional para verificar a hipótese levantada.

6. Conclusões

No início deste trabalho, o objetivo principal era propor uma alternativa aos levantamentos estatísticos exploratórios, no suporte à tomada de decisão dos gestores, durante o enfrentamento à pandemia da COVID-19. Para tal, foi realizado um estudo, utilizando aprendizado de máquina para fornecer uma nova ferramenta de predição de mortes causadas por COVID-19, a partir de dados abertos que contenham características sanitárias, demográficas e populacionais, de tal modo que, a partir deste estudo se possa desenvolver um modelo de inteligência artificial capaz de auxiliar no enfrentamento da pandemia de COVID-19.

Os objetivos foram alcançados, pois conseguiu-se concluir o experimento, e os modelos desenvolvidos apresentaram métricas satisfatórias, para um primeiro momento. E, a partir destes modelos poderemos avançar para fornecer uma ferramenta que auxilie na tomada de decisão dos gestores, assim contribuindo para a sociedade e a ciência.

Esta conclusão nos leva a planejar os próximos experimentos deste estudo, pois mesmo que os modelos baseados em *Support Vector Machine*, com exceção do modelo com kernel sigmoidal, sendo os que apresentaram maior qualidade, foi o algoritmo que precisou de mais atenção devido à limitações. Portanto é importante preparar um ambiente computacional mais poderoso, para que possam ser realizados mais experimentos, E também verificar quais outras técnicas de pré-processamento de dados podem ser executadas para o aumento da performance do treinamento.

As próximas etapas deste estudo, além do citado anteriormente, seriam abordar uma nova metodologia para a seleção de hiper parâmetros, utilizando algoritmos genéticos e comparar com a busca aleatória, por exemplo, e indo um pouco mais além, criar de fato um sistema protótipo de uma aplicação *web* ou *mobile*, utilizando técnicas de *UX Design* e inserir inteligência artificial *online* para que o sistema possa se adaptar mais facilmente às mudanças constantes, que é uma das maiores características do contexto pandêmico atual.

Referências

AGREBI, S.; LARBI, A. Use of artificial intelligence in infectious diseases. In: *Artificial Intelligence in Precision Health*. [S.l.]: Elsevier, 2020. p. 415–438. ISBN 978-0-12-817133-2. Citado na página 16.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, v. 13, n. 10, p. 281–305, 2012. Disponível em: ⟨http://jmlr.org/papers/v13/bergstra12a.html⟩. Citado na página 26.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado 3 vezes nas páginas 22, 23 e 24.

BROWNLEE, J. *Ordinal and One-Hot Encodings for Categorical Data*. 2020. ⟨https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/⟩. (Accessed on 05/23/2022). Citado na página 25.

BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2013. p. 108–122. Citado 4 vezes nas páginas 19, 20, 25 e 33.

CHOWDHURY, A. A.; HASAN, K. T.; HOQUE, K. K. S. Analysis and prediction of covid-19 pandemic in bangladesh by using anfis and lstm network. *Cognitive Computation*, v. 13, n. 3, p. 761–770, May 2021. ISSN 1866-9964. Disponível em: ⟨https://doi.org/10.1007/s12559-021-09859-0⟩. Citado na página 28.

DONG, E.; DU, H.; GARDNER, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*, v. 20, n. 5, p. 533–534, fev. 2020. Citado na página 30.

HALE, T. et al. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, v. 5, n. 4, p. 529–538, Apr 2021. ISSN 2397-3374. Disponível em: ⟨https://doi.org/10.1038/s41562-021-01079-8⟩. Citado na página 32.

HASELL, J. et al. A cross-country database of covid-19 testing. *Scientific Data*, v. 7, n. 1, p. 345, Oct 2020. ISSN 2052-4463. Disponível em: ⟨https://doi.org/10.1038/s41597-020-00688-8⟩. Citado na página 30.

HOU, H. C. et al. A study on office workplace modification during the covid-19 pandemic in the netherlands. *Journal of Corporate Real Estate*, Emerald Publishing Limited, v. 23, n. 3, p. 186–202, Jan 2021. ISSN 1463-001X. Disponível em: ⟨https://doi.org/10.1108/JCRE-10-2020-0051⟩. Citado 2 vezes nas páginas 28 e 29.

MATHIEU, E. et al. A global database of covid-19 vaccinations. *Nature Human Behaviour*, v. 5, n. 7, p. 947–953, Jul 2021. ISSN 2397-3374. Disponível em: ⟨https://doi.org/10.1038/s41562-021-01122-8⟩. Citado na página 30.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 8 vezes nas páginas , 20, 23, 33, 35, 36, 37 e 39.

RAJ, A. *Unlocking the True Power of Support Vector Regression | by Ashwin Raj | Towards Data Science*. 2020. ⟨https://towardsdatascience.com/unlocking-the-true-powe

r-of-support-vector-regression-847fd123a4a0/⟩. (Accessed on 05/22/2022). Citado na página 22.

RAWSON, T. M. et al. Artificial intelligence can improve decision-making in infection management. *Nature human behaviour*, England, v. 3, n. 6, p. 543–545, jun. 2019. ISSN 2397-3374. Citado na página 15.

RITCHIE, H. et al. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. Https://ourworldindata.org/coronavirus. Citado 3 vezes nas páginas , 30 e 33.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 32.

ROTH, J. A. et al. Introduction to Machine Learning in Digital Healthcare Epidemiology. *Infection Control & Hospital Epidemiology*, v. 39, n. 12, p. 1457–1462, dez. 2018. ISSN 0899-823X, 1559-6834. Citado na página 15.

SIMON, H. A. H. A. Simon: Models of bounded rationality. *Volume 1: Economic Analysis and Public Policy. Volume 2: Behavioural Economics and Business Organization*, MIT Press., (Cambridge, MA, "1:"and "2:", n. 6, 1985. Citado na página 16.

SWALIN, A. *How to Handle Missing Data. "The idea of imputation is both... | by Alvira Swalin | Towards Data Science*. 2018. ⟨https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4⟩. (Accessed on 05/23/2022). Citado na página 26.

TROYANSKAYA, O. et al. Missing value estimation methods for DNA microarrays. *BIOINFORMATICS*, v. 17, n. 6, 2001. Citado na página 34.

VAPNIK, V. *The Nature of Statistical Learning Theory*. [S.l.]: Springer, New York, 1995. Citado na página 20.