# MESTRADO

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

# TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO ELABORADO PARA A OBTENÇÃO DE GRAU DE MESTRE

IMPACTO DEMOGRÁFICO E FINANCEIRO DA PANDEMIA *COVID-19* EM PORTUGAL – PREVISÃO DO NÚMERO DE MORTES E DO PSI

ALEXANDRE POEIRAS ARRIAGA

**ORIENTAÇÃO:**

PROFESSOR DOUTOR CARLOS J. COSTA

OUTUBRO - 2022

# Agradecimentos

O primeiro dos agradecimentos, nunca poderia deixar de ser para os meus pais, Tomás e Mariana, sem vocês nada disto seria possível, obrigado por me possibilitarem este momento, por toda a educação que me deram e por todo o amor e carinho que me deram durante toda a vida. Vou sempre tentar seguir todos os valores que me ensinaram e transmiti-los durante toda a vida.

À minha namorada, companheira dos bons e maus momentos, à pessoa que tem aturado todo o meu mau feitio ao longo destes dois anos, e que nunca deixou de estar ao meu lado, apesar de todas as frustrações. Vou-te levar no coração toda a vida.

Aos meus avós, Manuel e Isabel, e especialmente a minha avó Catarina, que apesar de já não estar presente, vai estar sempre no meu coração. Obrigado por todo o carinho que me deram, sou o que sou muito graças a vocês.

Por último um especial agradecimento ao Prof. Carlos J. Costa, obrigado por todos os conselhos, e pela sabedoria que me transmitiu durante este processo. Obrigado por me permitir a realização de um artigo científico, algo verdadeiramente prazeroso de conseguir.

*"Other things may change us, but*
*we start and end with the family."*-
Anthony Brandt

# Resumo

A pandemia de *COVID-19* é uma das maiores crises de saúde do século XXI, afetou completamente o quotidiano da sociedade e impactou toda a população mundial, económica e socialmente. O uso de algoritmos de *machine learning* para o estudo de dados relativamente a esta pandemia tem sido bastante frequente nos mais variados artigos publicados nos últimos tempos. Nesta dissertação foi analisado o impacto de diversas variáveis (número de casos, temperatura, pessoas totalmente vacinadas, número de vacinações diárias e vários indicadores da mobilidade) no número de mortes causadas pela *COVID-19* ou *SARS-CoV-2* em Portugal e no índice da bolsa Portuguesa, o PSI, de forma a encontrar os modelos preditivos mais adequados. Foram utilizados vários algoritmos, como o *OLS*, *Ridge*, *MLP*, *Gradient Boosting* e *Random Forest* através do *software* de programação *Python*. A análise foi dividida em dois modelos, o primeiro referente à previsão do número de mortes causadas pela *COVID-19* e o segundo à previsão do PSI. No primeiro modelo foram usadas as variáveis originais, enquanto no segundo modelo foi feita uma Análise de Componentes Principais, que posteriormente foram usados para a regressão do modelo. O método utilizado para o processamento dos dados foi o CRISP-DM. Os dados foram obtidos através de uma base de dados pública. Por último, referir, que o Gradient Boosting foi o que obteve melhores resultados para ambos os modelos, de acordo com as métricas de precisão utilizadas. É de salientar também a maior eficácia dos algoritmos de *Ensemble* e de redes neuronais em comparação com os algoritmos lineares na previsão dos dados utilizados.

**Keywords:** *COVID-19*; óbitos; *PSI*; *machine learning*; Portugal

# Abstract

The *COVID-19* pandemic is one of the biggest health crises of the 21st century, it has completely affected society's daily life, and has impacted populations worldwide, both economically and socially. The use of machine learning algorithms to study data from the *COVID-19* pandemic has been quite frequent in the most varied articles published in recent times. In this dissertation it was analyzed the impact of several variables (number of cases, temperature, people fully vaccinated, number of daily vaccinations and several mobility variables) on the number of deaths caused by *COVID-19* or *SARS-CoV-2* in Portugal and on the number of the Portuguese stock index, PSI, to find the most appropriate predictive model. Several algorithms were used, such as OLS, Ridge, MLP, Gradient Boosting and Random Forest through Python programming software. The analysis was divided into two models, the first referring to the prediction of the number of deaths caused by *COVID-19* and the second to the PSI prediction. In the first model, the original variables were used, while in the second model, a Principal Component Analysis was made, that were later used for the regression of the model. The method used for data processing was CRISP-DM. Data were obtained from an open access database. Finally, it should be noted that Gradient Boosting was the algorithm that obtained the best results according to the precision metrics that were used. It is worth highlighting the greater efficiency of the Ensemble and neural networks algorithms compared to the linear algorithms in the prediction of the data used.

**Keywords:** COVID-19; deaths; PSI; machine learning; Portugal

# Índice

# Índice de Figuras

# Índice de Tabelas

# Índice de Anexos

# Lista de Siglas e Abreviaturas

*ARIMA – Auto Regressive Integrated Moving Average*

*COVID-19 - Coronavirus Disease 2019*

*CRISP-DM - Cross-Industry Standard Process for Data Mining*

*GB – Gradient Boosting*

*LASSO - Least Absolute Shrinkage and Selection Operator*

*MAPE – Mean Absolute Percent Error*

*$M_d$AE – Median Absolute Error*

*ML – Machine Learning*

*MLP - Multi-Layer Perceptron*

*MSE – Mean Square Error*

*OLS - Ordinary Least Squares*

*PSI – Portuguese Stock Index*

*RF – Random Forest*

*RMSE – Root Mean Square Error*

*SARIMA – Seasonal Auto Regressive Integrated Moving Average*

*SARS-CoV-2 - Severe Acute Respiratory Syndrome Coronavirus*

*TFM – Trabalho Final de Mestrado*

# 1. Introdução

## 1.1. Contexto Demográfico e Financeiro

Um surto de uma doença causada por um vírus é considerado uma pandemia quando afeta uma ampla área geográfica e tem um alto nível de infeção que pode levar a muitas mortes. (Almalki et al., 2022) Ao longo da história da humanidade ocorreram várias pandemias, algumas com maior taxa de mortalidade do que outras, como a gripe espanhola (1918), a gripe asiática (1957), a gripe de Hong Kong (1968) e a gripe suína (2009). (Rustagi et al., 2022) A pandemia mais impactante deste século é a pandemia de *COVID-19*. *COVID-19* é uma doença respiratória causada pelo vírus *SARS-CoV-2*, (Almalki et al., 2022) que afeta todas as faixas etárias, mas tem consequências mais graves em indivíduos mais velhos e/ou pessoas com condições médicas pré-existentes. (Sarirete, 2021) Os primeiros casos registados datam de 31 de dezembro de 2019 na cidade de Wuhan, China. (Sohrabi et al., 2020) Esta doença espalhou-se rapidamente por todo o mundo, em Portugal, o primeiro caso foi registado a 2 de março de 2020. (Milhinhos e Costa, 2020) Qualquer pessoa que teste positivo para esta doença pode ser sintomática ou assintomática. Os sintomas da *COVID-19* podem ser febre, cansaço, tosse e em casos mais graves falta de ar e problemas pulmonares. (Rustagi et al., 2022)

A pandemia teve um grande impacto nos mercados financeiros em todo o mundo, como por exemplo em março de 2020, o mercado de ações dos Estados Unidos da América teve de ativar o mecanismo de amortização e rebalanceamento das ordens de compra e de venda de ações, denominado *circuit breaker mechanism* (Chen et al.) quatro vezes em dez dias, algo que em toda a história apenas tinha sido ativado uma vez em 1997 (Zhang et al., 2020). Na Europa e na Ásia os mercados de ações também tiveram uma queda bastante acentuada, como por exemplo a descida de 10% do índice principal do Reino Unido, em março de 2020 ou a descida de 20% do principal índice do Japão em dezembro de 2019 (Zhang et al., 2020). Em Portugal o índice *PSI*, que está representado pelas empresas portuguesas com uma capitalização de mercado de cem milhões de euros em *free float*[1], teve uma descida de aproximadamente 15% desde o início da pandemia

[1] ***free float*** *– A capitalização é calculada multiplicando o preço das ações pelo número de ações disponíveis no mercado.* ***Fonte:*** Free-Float Methodology Definition (investopedia.com)

(considera-se a data registada da primeira infeção, 2 de março de 2020) até ao fim desse mesmo mês, segundo dados de («PSI 5 487,44 | Euronext Live quotes preços»).

A pandemia da *COVID-19* levou a uma mudança drástica no quotidiano da população mundial, devido às medidas de confinamento implementadas pelos governos. As rotinas da população foram completamente alteradas, levando a mudanças de hábitos já implementados para outros completamente diferentes, como por exemplo existiu um aumento substancial das atividades ao ar livre, e uma diminuição das atividades em espaços fechados, devido à prevenção da disseminação do vírus (Li et al., 2021). Uma das mudanças mais importantes em Portugal foi a implementação do trabalho remoto, algo que para a maioria das empresas portuguesas era algo impensável antes do início da pandemia, tendo em conta que em 2019, apenas 6,5% dos trabalhadores portugueses utilizavam esse método de trabalho. (Andrade e Petiz Lousã, 2021) Isso levou a uma diminuição das idas ao escritório e por consequência da utilização dos transportes públicos também.

## 1.2. Objetivos e Metodologia

O objetivo deste TFM é encontrar dois modelos adequados para estimar o número de mortes diárias causadas pelo vírus *SARS-CoV-2* e os valores de fecho diários do *PSI* e posteriormente encontrar o algoritmo de *machine learning* com maior poder preditivo, de acordo com as métricas escolhidas para esse efeito, de forma a estudar o impacto demográfico e financeiro da pandemia *COVID-19* em Portugal. Para esse propósito optou-se por seguir uma metodologia bastante utilizada em todo o mundo, denominada *CRISP-DM*. (Costa e Aparicio, 2020, Costa e Aparicio, 2021). Através desta metodologia, foram selecionadas várias variáveis relacionadas com o objetivo do TFM, após a análise das mesmas foram criadas várias variáveis a partir das originais (através de *lags* ou variáveis polinomiais) e foram estimados os modelos para a previsão do número de óbitos e dos valores diários de fecho do *PSI*. Posteriormente os dados foram previstos através de vários algoritmos de *machine learning* e foram interpretados os resultados obtidos de forma a inferir acerca da importância de cada uma das variáveis preditoras. Foi também medida a precisão de cada um dos algoritmos, de forma a perceber qual o algoritmo com maior poder preditivo. Por fim, foram comparados os resultados

obtidos neste TFM, com os de outros autores e foram tiradas as conclusões pertinentes para o estudo.

# 2. Revisão da Literatura

## 2.1. Impacto Demográfico e Financeiro da pandemia *COVID-19*

O estudo do impacto demográfico e financeiro que a pandemia da *COVID-19* tem sido bastante frequente nos últimos anos. (Xie e Li, 2020) mostrou que a densidade populacional está positivamente relacionada com o número de infeções e mortes por *COVID-19* nos EUA. (Khan et al.) através do estudo das características demográficas de vários países, concluiu que o número de óbitos causados por esta doença está relacionado com a distribuição etária, rácio de pobreza, percentagem de mulheres fumadoras, nível de obesidade e temperatura média anual de cada país. Arriaga e Costa (2023) estuda o impacto da vacinação, do número de casos e da temperatura no número de óbitos relacionados à *COVID-19*, tendo aferido que a temperatura foi a variável com maior impacto, na estimação dos óbitos, estando correlacionada negativamente com os mesmos. Já foram feitos diversos estudos também para medir o impacto que a vacina da *COVID-19* veio trazer em termos de consequências demográficas, como por exemplo no estudo (Haas et al., 2022) acerca do impacto da vacina *Pfizer-BioNTech BNT162b2 mRNA* em Israel, concluindo que se não tivesse existido todo o processo de vacinação descrito, teriam existido três vezes mais hospitalizações e mortes relacionadas com a doença. (Watson et al., 2022) estudou o impacto do primeiro ano de vacinação em termos globais, concluindo que esta prevenção permitiu a redução de aproximadamente 19 milhões de mortes em relação ao que seria esperado se não existisse o processo de vacinação.

Em relação ao impacto financeiro, diversos autores já abordaram o tema, como por exemplo, (Estrada et al., 2021) que estimou, através de um simulador, que os efeitos da crise que esta pandemia causaria iriam ser similares aos da Crise de 1929. (Zhang et al., 2020) optou por uma abordagem relacionada com os *stocks index*, indicando a alta volatilidade existente nos mesmos após a perdas económicas associadas à pandemia. (Zhang et al., 2020) também conclui que os riscos do mercado financeiro têm em conta a gravidade com que a pandemia atingiu cada país. (Arriaga e Costa, 2023) realça a

importância que a imposição de políticas governamentais bem delineadas e sem atrasos de forma a diminuir o impacto negativo nos *stock index*.

Os dados da mortalidade por *COVID-19* podem ser previstos por vários métodos de previsão, como algoritmos de *machine learning* ou *statistical forecast* (Almalki et al.,2022) Diversos estudos utilizaram os modelos *ARIMA* e *SARIMA*, considerando o comportamento sazonal presente na série da mortalidade. (Perone,. 2022 e Chaurasia e Pal, 2022) Dentro dos algoritmos de *machine learning* várias abordagens foram utilizadas pelos autores nos diversos artigos já feitos, através do *Random Forest* e das redes neuronais (Gupta et al., 2021), ou da Regressão Linear e Polinomial (Rustagi et al., 2022). Também o *Gradient Boosting* já foi utilizado por (Saba et al., 2021) para modelar e prever os dados da mortalidade resultante da *COVID-19*, provando ser um algoritmo bastante eficiente.

Foram escolhidas diversas variáveis preditivas para estimar o impacto demográfico da pandemia *COVID-19* em Portugal. Os dados da vacinação foram escolhidos para prever a mortalidade, dado o impacto que a vacinação teve desde o seu princípio no número de mortes e de infeções causadas por *COVID-19* (Haas et al., 2022), o número diário de novos casos e essa mesma variável ao quadrado foram escolhidos de forma a que fosse possível estudar o impacto que esta variável teve antes e depois do início do processo de vacinação no número de óbitos, tendo em conta que após o processo de vacinação, apesar do número de casos ter tido um aumento bastante elevado, o número de mortes não seguiu o comportamento que tinha tido antes do processo de vacinação ter começado, como se pode observar nas Figuras 8 e 9. Pode-se então aferir que o número de infeções, não tem uma relação linear com o número de mortes, daí ter-se acrescentado um termo polinomial no modelo (Ostertagová, 2012). Por último foi utilizada a temperatura média diária devido ao padrão sazonal presente nos dados como é possível observar na Figura 12 e no artigo (Li et al., 2021)

Relativamente à previsão dos valores dos *Stock Index*, *como* o *PSI*, vários autores seguiram abordagens de *machine learning*, por exemplo através do uso de redes neuronais (Moghaddam et al., 2016), ou de algoritmos de *ensemble* como o *Gradient Boosting* (Xue et al., 2020) ou o *Random Forest* (Polamuri, S. R., Srinivasi, K., & Mohan, A. K. (2019).), mostrando todos eles uma grande eficácia a nível de previsão dos dados.

Outras abordagens foram usadas, através de algoritmos de *statistical forecast*, como os *ARIMA* (Banerjee, 2014, Samadani & Costa, 2021). Vários artigos usam modelos híbridos entre os ARIMA e vários modelos de *machine learning*, como por exemplo (Wang e Guo, 2020) devido à eficácia de ambos em diferentes situações.

Após uma pesquisa entre os numerosos artigos já publicados acerca do impacto que a *COVID-19* teve economicamente e financeiramente, uma grande parte destes aborda o tema relacionando indicadores de mobilidade com indicadores financeiros (Kartal et al., 2021) e económicos (Sampi e Jooste, 2020), daí a escolha ter recaído por este tipo de variáveis. Em relação ao número de casos, já ficou provado em diversos artigos como (Pavlyshenko, 2020), que existe uma relação de causa-efeito entre ambas as variáveis. Por fim a vacinação foi escolhida de forma a funcionar como variável atenuadora do efeito que o número de casos tem no *PSI*, funcionando como um ponto de viragem nesse efeito.

## 2.2. A importância da Aprendizagem Automática (*Machine Learning*)

De acordo com o artigo Aparicio, et al. (2022)., "A premissa de machine learning é que um programa de computador pode aprender e adaptar-se a novos dados sem a necessidade de intervenção humana". Em machine learning não existe um algoritmo que possa prever com o menor erro todos os tipos de dados, (Mahesh, 2018). ou seja, para cada tipo de dados existem algoritmos mais adequados que outros para prever dados futuros. A escolha do melhor algoritmo também depende do problema que estamos a enfrentar e do número de variáveis usadas no modelo (Mahesh, 2018).

Existem vários tipos de aprendizagem utilizada pelos algoritmos de *machine learning*, os supervisionados, não supervisionados, semi-supervisionados e por *reinforcement*. Os do tipo supervisionado realizam um mapeamento das variáveis dependentes e independentes, para prever dados futuros desconhecidos da variável dependente. (Cord, & Cunningham, 2008, Aparicio et al, 2019). Os semi-supervisionados utilizam dados não classificados (não precisam de intervenção humana) conjuntamente com dados classificados (precisam de intervenção humana) para prever dados futuros. Esse tipo de aprendizagem pode ser mais eficiente, pois precisa de muito menos

intervenção humana na construção dos modelos. Por último, os algoritmos de aprendizagem por *reinforcement* produzem uma série de ações considerando o ambiente onde estão inseridos para maximizar *"as recompensas futuras que recebe (ou minimizar as punições) ao longo da sua vida"*, (Zhu, 2005). Por último, mas não menos importantes, em algoritmos de aprendizagem não supervisionada, os dados de entrada são inseridos, mas não obtêm resultados alvo supervisionados, nem recompensas do seu ambiente (Aparicio, et al. 2022). Um exemplo deste tipo de algoritmos é o *K-means*.

Este TFM focou-se na utilização de algoritmos com aprendizagem supervisionada, o *OLS*, *LASSO*, *Ridge*, *Gradient Boosting*, *MLP* e *Random Forest*, e também não supervisionada, *Principal Component Analysis* (Figura 1).

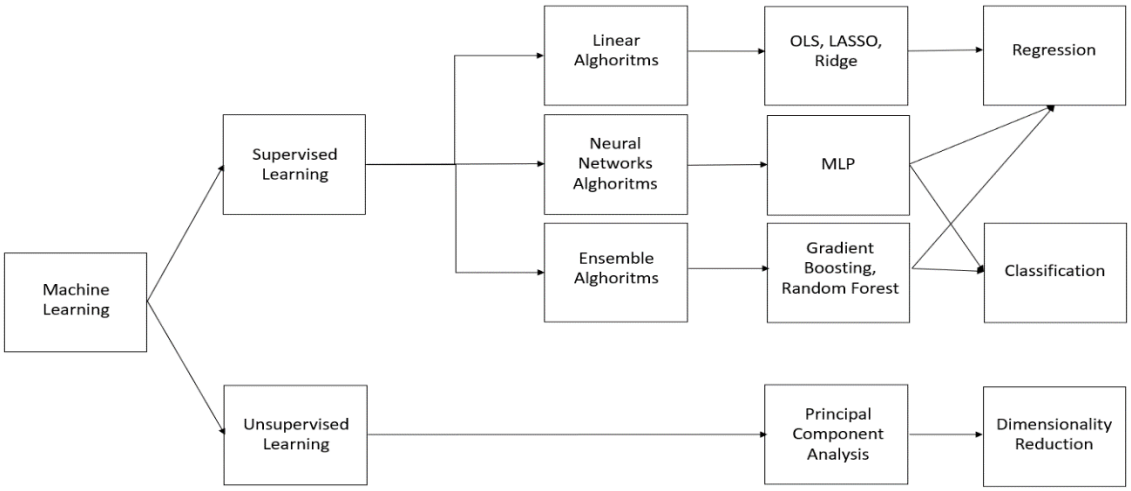



***Figura 1**: Algoritmos de machine learning do TFM*
***Fonte:** Elaboração Própria*

### 2.2.1. *Supervised Learning* (*Linear Alghoritms)*

Existem três algoritmos de regressão linear, do tipo *Supervised*, utilizados neste TFM, como demonstra a Figura 1, *OLS*, *LASSO* e *Ridge*.

***OLS* ou Regressão Linear** é um dos algoritmos de *machine learning* com mais fácil compreensão. A regressão linear pode ser simples (quando apenas uma variável independente é usada no modelo) ou múltipla (quando duas ou mais variáveis preditivas são usadas para prever a variável dependente) (Saleh, 2022) O modelo estrutural da Regressão Linear pode ser representado por:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon \quad (1)$$

onde $Y$ representa a variável dependente, e $X_m$ as variáveis independentes. Os parâmetros $\beta_m$ são os coeficientes estimados pelo modelo de regressão e o parâmetro ε é o erro associado à estimação do modelo.

A **Regressão de *Ridge*** é um algoritmo usado quando se enfrenta problemas de multicolinearidade[2] entre as variáveis preditivas do modelo (Saleh, 2022). Este tipo de regressão linear é bastante similar com o *OLS*, com a diferença que na estimação dos $\beta$, é adicionado um termo $k$ aos elementos diagonais da matriz de correlação. (McDonald, 2009) Para este tipo de regressão os dados, tanto das variáveis preditivas como da variável independente têm de ser estandardizados, ou seja, realizar uma subtração da média e dividir pelo desvio padrão das observações originais. (McDonald, 2009)

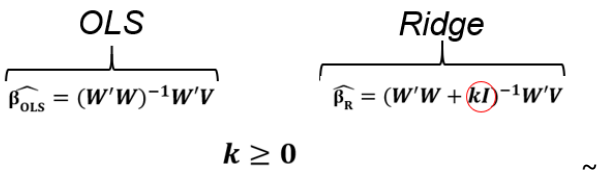



***Figura 2:*** *Comparação da estimação dos coeficientes* $\tilde{\widehat{\boldsymbol{\beta}}}$ *(OLS v.s Ridge)*
***Fonte:*** *Elaboração Própria*

Através da fórmula da estimação do parâmetro $\widehat{\boldsymbol{\beta}_{\boldsymbol{R}}}$ de *Ridge* da Figura 2, pode-se observar que se o valor de $k$ for igual a zero, então o estimador vai ser igual ao estimador do *OLS*.

Por último temos o algoritmo denominado ***Least Absolute Shrinkage and Selection Operator* ou *LASSO***. Este algoritmo permite melhorar a precisão do modelo por meio da seleção e regularização de variáveis. Esse processo é chamado de *variable shrinkage*, no qual o objetivo é reduzir o número de variáveis preditivas presentes no modelo.0 O estimador $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}}$ de *LASSO* pode ser estimado através do seguinte problema de minimização:

[2] ***multicolinearidade:*** *Condição que sucede quando uma ou mais variáveis independentes do modelo podem prever outra variável independente de uma forma eficiente e linear.* ***Fonte:*** *(McDonald, 2009)*

$$1º:\ \left\|V - W^{3}B\right\|^{2} + k\|B\|_{1} \rightarrow min,$$
$$2º: Para\ qualquer\ k > 0, existe\ t(k) > 0\ tal\ como$$
$$\|V - WB\|^{2} \rightarrow mi\ n\ sujeito\ a\ \|B\|_{1} \leq t(k), \quad (2)$$
$$onde\ \|B\|_{1} = \sum_{j=1}^{i-1} |\beta_{j}|.$$

Ao contrário da regularização feita na regressão de *Ridge*, a de *LASSO* gera coeficientes exatamente iguais a zero, quando o parâmetro $k$ é demasiado grande, melhorando o modelo em termos de interpretação. (Melkumova e Shatskikh, 2017)

### 2.2.2. *Supervised Learning* (*Neural Network Alghoritms*)

Continuando ainda nos algoritmos com *Supervised Learning*, vai ser abordado o algoritmo denominado por *MLP*, um algoritmo de redes neuronais como indica a Figura 1.

***Multilayer Perception (MLP)*** é um algoritmo de *machine learning* que usa redes neuronais artificiais. Como refere o artigo (Borghi et al., 2021), "A experiência da rede é armazenada pelos pesos sinápticos entre os neurônios e a sua performance é avaliada, por exemplo, pela capacidade de generalizar comportamentos, reconhecer padrões, corrigir erros ou executar previsões". Este algoritmo associa vários neurónios, formando redes neurais que permitem realizar diversas funções para melhorar a previsão. (Borghi et al., 2021) O *MLP* é composto por três etapas principais, a informação entra na rede através da input layer, e é libertada pela output layer, passando pela etapa intermédia denominada de hidden layer. (Park e Lek, 2016). O número de neurónios que entram para a rede dependem do número de variáveis independentes do modelo, enquanto que os que são libertados a partir da output layer dependem do número de variáveis dependentes. (Park e Lek, 2016).

[3] *W e V representam as matrizes estandardizadas, das matrizes originais X e y, respetivamente. A matriz X é representativa das variáveis independentes, enquanto a matriz Y corresponde à variável dependente.*

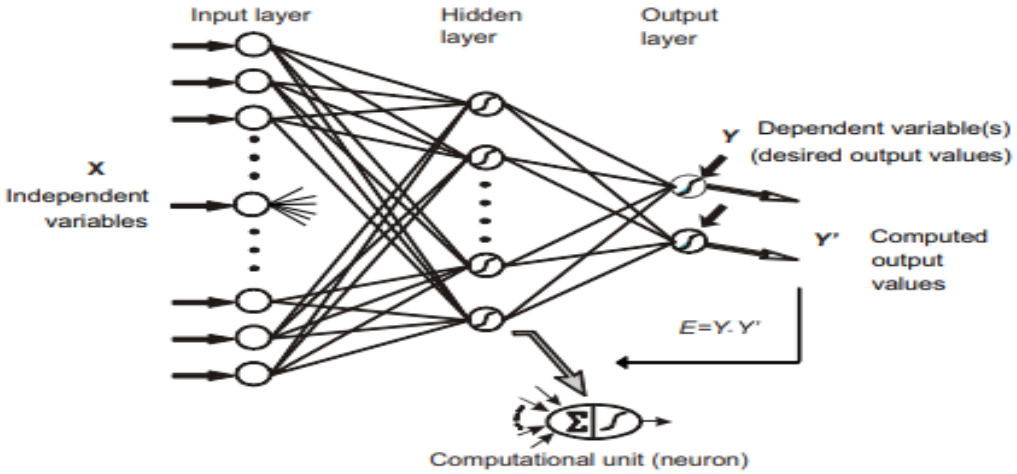



***Figura 3:*** *Arquitetura do algoritmo MLP*
***Fonte:*** *(Park e Lek, 2016)*

### 2.2.3. *Supervised Learning* (*Ensemble Alghoritms)*

Restam abordar apenas dois algoritmos utilizados neste TFM, são eles o *Gradient Boosting* e o *Random Forest*. Estes algoritmos são do tipo *Ensemble*, e pertencem também ao grupo dos algoritmos com *Supervised Learning*, como mostra a Figura 1.

O ***Gradient Boosting*** **(GB)** pode ser usado para fins de classificação e regressão. Este algoritmo é um algoritmo de *ensemble*, que começou a ser utilizado na otimização de uma função de custo e tem sido utilizado em diversas áreas, como na deteção de roubo de energia. 0) Este método tem sido muito utilizado em variados estudos sobre a pandemia de *COVID-19*. (Shrivastav, & Jha, 2021) O GB é um algoritmo, que através de várias iterações combina uma série de modelos com uma taxa de aprendizagem, com o objetivo de minimizar erros de previsão. Em cada um dos modelos resultantes das iterações, descarta os preditores mais fracos e escolhe os mais eficientes (Gumaei, et. Al, 2021). O modelo aditivo do GB pode ser representado da seguinte forma:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \tag{3}$$

onde $F_{m-1}$ representa o modelo anterior, e $h_m$ é a taxa de aprendizagem usada para diminuir os erros da previsão (Gumaei, et. Al, 2021). $\rho_m$ é um multiplicador que pode ser representado da seguinte forma:

$$\rho_m = arg\,\min_{\rho} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \rho h_m(x_i)) \quad (4)$$

onde $y_i$ é a classificação da classe de destino. (Gumaei et al., 2021)

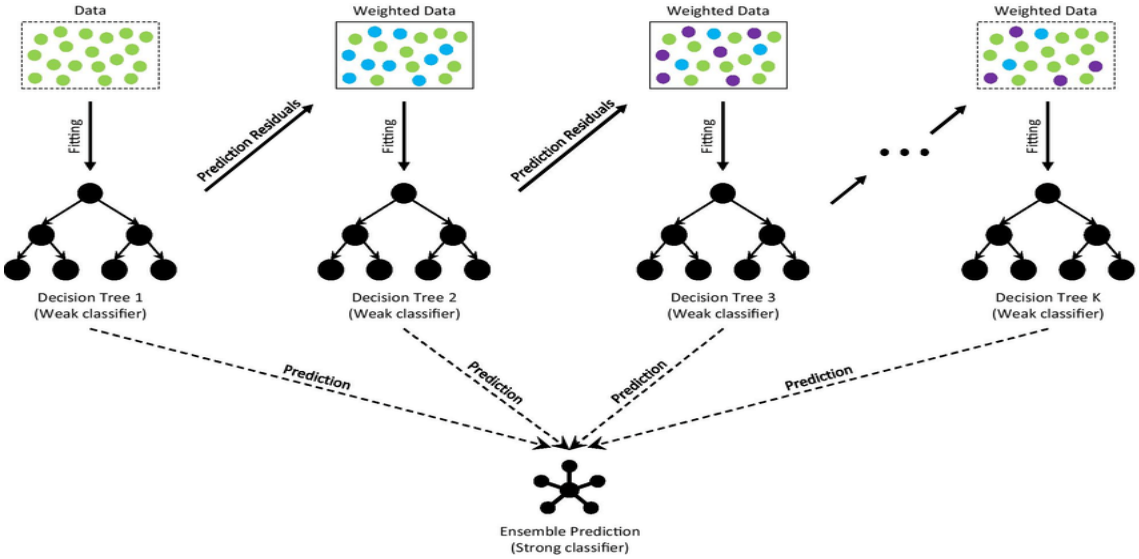



***Figura 4:*** *Arquitetura do algoritmo Gradient Boosting*
***Fonte:*** *(Deng et al., 2021)*

***Random Forest* (*RF*)** é outro algoritmo de *ensemble*, como o Gradient Boosting, que usa árvores de decisão em segundo plano. As árvores de decisão são criadas com uma base de amostra aleatória dos dados de treino. (Gupta et al., 2021) A diferença entre o RF e o GB é que o RF não usa uma taxa de aprendizagem, usa apenas a média de todas as árvores geradas. (Yeşilkanat, 2020)

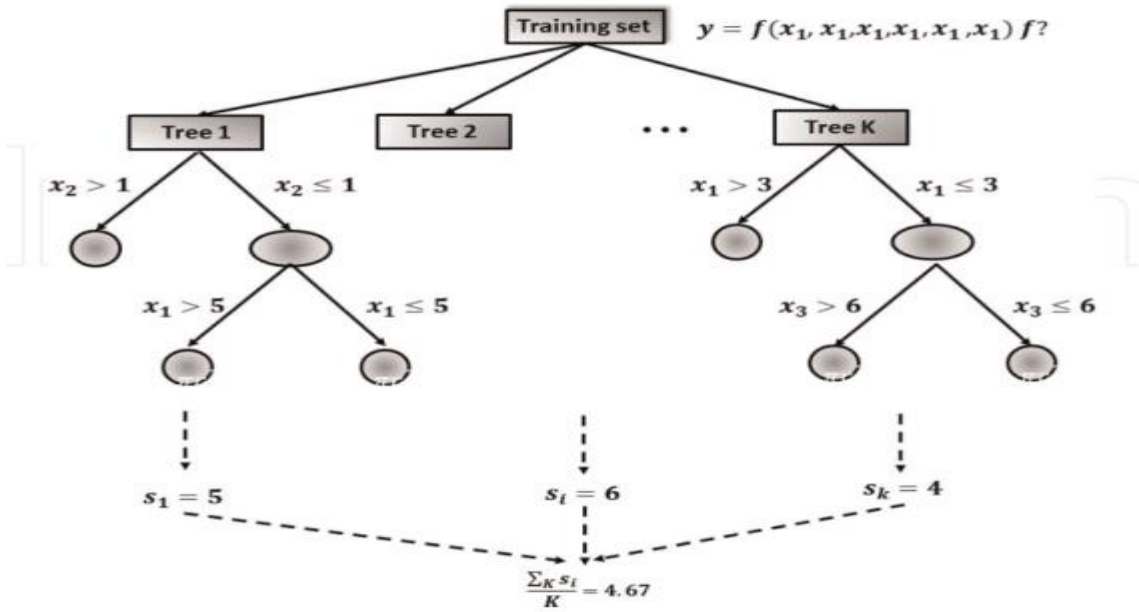



***Figura 5:*** *Arquitetura do algoritmo Random Forest*
***Fonte:*** *(Ornella et al., 2020)*

### 2.2.4. *Unsupervised Learning*

Foram também utilizados de *Unsupervised Learning* neste TFM, como é o caso da *Principal Component Analysis (PCA)*.

A ***PCA*** é um tipo de *machine learning* não supervisionada que tem como objetivo a redução da dimensionalidade dos dados de forma a ser possível uma melhor interpretação dos mesmos. Esta redução gera um ou mais fatores latentes (fatores não observáveis derivados das variáveis originais), denominados componentes principais através do seguinte problema de otimização (Martin-Barreiro et al., 2021):

$$\begin{aligned} & max(Z = \boldsymbol{w}^T \boldsymbol{V} \boldsymbol{w}), \\ & subject\ to\ \boldsymbol{w}^T \boldsymbol{w} = 1, \\ & \boldsymbol{V} = \frac{1}{n} \boldsymbol{X}^T \boldsymbol{X}, \end{aligned} \tag{5}$$

em que $\boldsymbol{V}$ representa a matriz de covariâncias da matriz dos dados $\mathbf{X}$, centrada através das colunas, e a condição $\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$ é indicativa que o vetor $\mathbf{w}$ tem norma unitária. A solução deste problema de maximização será o maior valor próprio (*eigenvalue*) λ da matriz $\mathbf{V}$, o que significa que o primeiro componente principal é o vetor próprio (*eigenvector*) $\boldsymbol{w}_1$ de norma unitária associado ao maior valor próprio $\lambda_1$, o segundo componente principal será o vetor próprio $\boldsymbol{w}_2$ associado ao segundo maior valor próprio $\lambda_2$ e assim por diante.

Os *loadings* dos componentes principais correspondem à correlação entre os fatores latentes e as variáveis observadas, e podem ser obtidos através da seguinte transformação (Risvik, 2017):

$$\boldsymbol{L} = \boldsymbol{\lambda}\sqrt{\boldsymbol{V}} \tag{6}$$

A *PCA* é uma técnica bastante utilizada para combater a multicolinearidade entre as variáveis observáveis, através da redução da dimensão dos dados, criando fatores latentes que podem ser relacionados às variáveis, criando um agrupamento das variáveis correlacionados entre si. (Graham, 2003)

Para a escolha do número de componentes principais foi utilizado o método de **Análise Paralela de Horn**. A Análise Paralela de Horn é uma derivação do método de Kaiser, que consiste na seleção do número de componentes baseada nos *eigenvalues* (valores próprios) de cada um dos componentes. (Brown, 2009) O método de Kaiser diz-nos que devemos reter os componentes com *eigenvalues* superiores a 1. (Brown, 2009) A Análise Paralela de Horn é um método um pouco mais robusto, pois através de simulações de Monte Carlo (Harrison, 2010) gera dados artificiais com uma distribuição normal e calcula os seus *eigenvalues*. (Çokluk e Koçak, 2016) Após esse processo é feita uma comparação enter os *eigenvalues* dos dados reais e dos dados simulados, e são retidos os componentes em que os *eigenvalues* da amostra real são superiores aos da amostra simulada. (Çokluk e Koçak, 2016)

### 2.2.5. *Cross Validation - K Fold*

O método de validação ***K-Fold*** é um método de *Cross Validation* que produz uma divisão do *dataset* em vários subconjuntos $\boldsymbol{k}$ de aproximadamente igual dimensão através de um processo de amostragem sem substituição. Esses subconjuntos são novamente divididos, de forma a seguir duas etapas distintas, na primeira etapa é feito um treino do modelo nos $\boldsymbol{k} - 1$ subconjuntos, denominados de *training set*, para posteriormente ser feita a validação do modelo, através da segunda etapa, no subconjunto restante, denominado de *validation set*. O processo é repetido até que todos os subconjuntos tenham sido utilizados para a validação do modelo. Por último é realizada uma média aritmética da performance de cada uma das iterações. (Berrar, 2019)

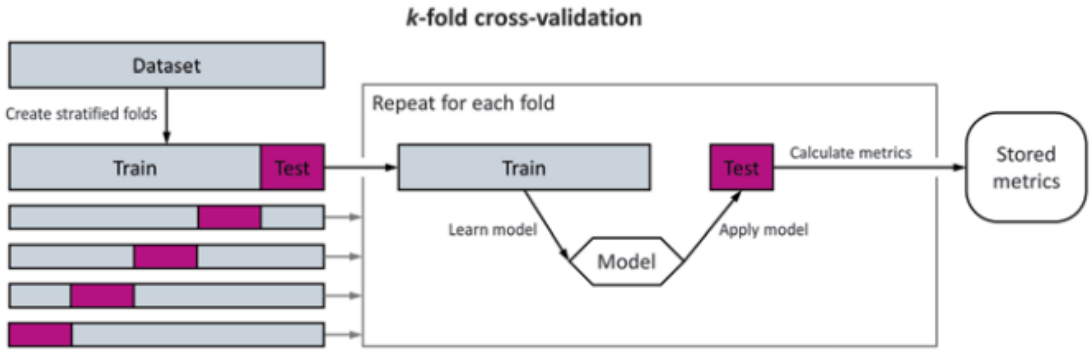



***Figura 6:*** *Processo K-Fold Cross Validation*
***Fonte:*** *(Kubben et al., 2019)*

## 2.3. Métricas de Precisão

Para que seja possível avaliar a *performance* de cada um dos algoritmos, é necessária a utilização de diversas métricas de precisão, de forma a realizar uma comparação entre os dados previstos e os dados originais.

Neste TFM foram utilizadas as métricas utilizadas foram as seguintes: *Mean Absolute Error (MAE), Mean Squared Error (MSE), Median Absolute Error ($M_d$AE), Explained Variance Score* e *$R^2$ Score.*

O ***Mean Absolute Error* ou Erro Absoluto Médio**, é uma medida inserida no grupo das medidas de erro absoluto, e corresponde à média dos resíduos absolutos todas as observações. (Shcherbakov et al., 2013) Pode ser representado pela seguinte fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|, \tag{7}$$

em que $n$ representa o número de observações e $\varepsilon_i$ os resíduos da observação $i$ .

O ***Mean Squared Error* ou Erro Quadrático Médio**, é bastante idêntico ao *MAE,* com uma ligeira diferença, os resíduos nesta medida são elevados à potência quadrática. (Shcherbakov et al., 2013) Esta medida também pertence ao grupo das medidas de erro absoluto e pode ser representada da seguinte forma:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i^2|, \tag{8}$$

O ***Median Absolute Error* ou Erro Mediano Absoluto** é a última das medidas de erro absoluto utilizadas neste TFM. Nesta medida de precisão é calculada a mediana dos resíduos de todas as observações, ao invés do que acontece nas duas medidas anteriores, que é calculada uma média aritmética. (Shcherbakov et al., 2013) A sua representação pode ser feita através de:

$$M_dAE = median\sum_{i=1}^{n}|\varepsilon_i|, \tag{9}$$

De seguida apresentam-se medidas de precisão mais robustas, e bastantes similares, como é o caso do $R^2$ *Score*, o *Adjusted* $R^2$ *Score* e a *Explained Variance Score*. O ***$R^2$* ou Coeficiente de Determinação** é uma das medidas de precisão mais utilizadas em *estatística*. (Redell, 2019) Esta medida já foi caracterizada de diferentes formas consoante a aplicação dada à mesma, principalmente se existe ou não o termo de interceção no modelo. Como exemplo de uma dessas caracterizações é outra das medidas de precisão usada no TFM, a *Explained Variance Score*.

O $R^2$ representa a proporção da variância explicada na variável dependente que é previsível a partir das variáveis independentes (Chicco et al., 2021), podendo estar representado na seguinte forma:

$$R^2 = \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}\varepsilon^2} \tag{10}$$

onde $X_i$ representa as observações da variável dependente e $Y_i$ as observações da variável independente. Esta métrica é escalada entre $]-\infty; 1]$. (Chicco et al., 2021)

O ***Adjusted $R^2$*** *é uma derivação do* $R^2$ referido no ponto anterior, em que o principal objetivo é substituir os estimadores enviesados, $\sum_{i=1}^{n}(X_i - Y_i)^2$ e $\sum_{i=1}^{n}\varepsilon^2$ por estimadores não enviesados, $\sum_{i=1}^{n}(X_i - Y_i)^2/N - p - 1$ e $\sum_{i=1}^{n}\varepsilon^2/N - 1$ respetivamente. (Karch, 2020) Devido aos estimadores serem não enviesados, esta medida permite comparar modelos com um número de variáveis diferentes de uma forma eficiente, tendo em conta que com o $R^2$ esta comparação não pode ser feita, devido ao facto de que quando aumentamos o número de variáveis do modelo o $R^2$ irá sempre aumentar gradualmente (Akossou e Palm). O *Adjusted* $R^2$ pode então ser representado através da seguinte fórmula:

$$R^2_{Adj} = 1 - \frac{N-1}{N-p-1}\,(1 - R^2), \tag{11}$$

onde $N$ representa o número de observações da variável dependente e $p$ representa o número de variáveis independentes no modelo.

A última métrica de precisão utilizada foi a ***Explained Variance* Score**. Segundo Pedegrosa, et. Al, 2012, a única diferença entre a *Explained Variance Score* e o $R^2$ *Score*

acontece quando a primeira medida *"não tem em conta o deslocamento sistemático da previsão".* A fórmula desta métrica de precisão é representada da seguinte forma:

$$EVS = 1 - \frac{Var\,(Y-\hat{Y})}{Var(Y)} \tag{12}$$

## 2.4. Testes Estatísticos

### 2.4.1. *Variance Inflation Factor (VIF)*

O *Variance Inflation Factor* é uma medida que tem como objetivo medir a multicolinearidade existente entre as variáveis. Os *VIF's* de cada variável ganharam esta denominação devido ao facto de explicarem quanto do aumento da variação dos coeficientes do modelo é devido a variáveis colineares independentes, correlacionadas entre si. (Craney e Surles, 2002) Os *VIF's* podem ser representados da seguinte forma:

$$VIF_i = \frac{1}{1 - r_i^2}, \quad i = 1, \dots, p-1, \tag{13}$$

em que $p$ representa o número de variáveis independetes e $r_i^2$ representa o valor do $R^2$ obtido para a regressão da $i$-éssima variável preditiva sobre as outras $p-2$ variáveis. Habitualmente, consideram-se valores elevados, $VIF \geq 5$ ou $VIF \geq 10$. (Craney e Surles, 2002)

### 2.4.2. Teste *Durbin-Watson*

Para que seja possível saber se o modelo capturou a informação dos dados de uma forma adequada, é necessário estudar os resíduos resultantes da previsão. Os resíduos do modelo não devem estar correlacionados, o que significaria que existe informação em falta que não foi considerada, e a média destes deve ser zero, pois se isto não for verificado significa que a previsão está enviesada (Hyndman e Athanasopoulos, 2018). Para que essas condições sejam verificadas, a série dos resíduos tem de ser um *white noise*, ou seja os erros têm de ser provenientes de fatores externos ao modelo e não da forma estrutural do mesmo. (Chen, 2016)

Para verificar essas mesmas condições existe um teste bastante conhecido, denominado de Teste *Durbin-Watson*. A estatística do teste pode ser representada por:

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}, \quad (14)$$

em que $\varepsilon_i$ representa o valor dos resíduos na $i$-éssima observação e $\varepsilon_{i-1}$ o valor dos resíduos na observação $i$-éssima$-1$. Os valores da estatística *DW* situam-se entre 0 e 4, sendo que se assumir o valor 2, significa que não existe autocorrelação entre os resíduos (Chen, 2016).

# 3. Metodologia

Neste TFM foi seguida uma metodologia *standard* bastante utilizada em projetos de *machine learning*, denominada de *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*), com o objetivo de inferir acerca do impacto demográfico e financeiro que a pandemia *COVID-19* teve em Portugal, através de algoritmos de *machine learning*, construídos com várias variáveis relacionadas com o tema. Outros dos objetivos é também analisar com o algoritmo de *machine learning* com maior taxa de precisão através das métricas escolhidas. Este tipo de metodologia contém seis etapas principais: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation e Deployment*. (Schröer et al., 2021)

A primeira fase denominada por ***Bussiness Understanding***, consiste na compreensão do objetivo principal do estudo. Primeiramente deve-se identificar o tipo de *data mining* e explicá-lo, que neste estudo é realizar uma previsão do número de mortes por *COVID-19* e do *PSI*, através de regressões lineares e não lineares, com a ajuda de algoritmos de *machine learning*. De seguida é necessário aferir sobre quais as medidas de desempenho a usar para classificar o sucesso ou insucesso de cada um dos algoritmos nessa previsão.

Após a conclusão desta primeira fase, deve-se prosseguir para a segunda fase denominada de ***Data Understanding***, que consiste na exploração dos dados. Nessa exploração deve-se fazer uma análise exaustiva, testando a qualidade dos dados e realizando uma análise estatística e descritiva dos dados. Para este TFM todo este processo será detalhado nos capítulos seguintes, para cada um dos modelos construídos.

Após a análise e compreensão dos dados, deve-se prosseguir para a preparação dos mesmos, avançando para a próxima etapa, denominada de ***Data Preparation***. Nesta etapa deve-se analisar a fiabilidade dos dados, e aplicar diferentes métodos para melhorar a qualidade dos mesmos, algo que também será detalhado nos capítulos seguintes para cada um dos modelos.

Quando os dados já tiverem sido todos escolhidos e ultrapassadas todas as etapas de pré processamento, é realizada a modelação dos dados, através da quarta etapa denominada de ***Modeling***. Esta etapa consiste na escolha dos algoritmos utilizados para

atingir o objetivo proposto, e também na escolha de como os dados serão utilizados no modelo (p.e. através de *lags*[4] das variáveis, variáveis polinomiais ou variáveis de interação). Essa escolha pode depois ser avaliada através de diversos critérios, como por exemplo o teste *t* para verificar se as variáveis são todas significativas individualmente (Kim, 2015), ou através do *Variance Inflation Factor (VIF)*, que tem como objetivo medir a multicolinearidade existente entre as variáveis preditivas do modelo. (Murray et al., 2012) Após a avaliação e aprovação do modelo final, é necessário realizar uma *hyperparameter optimization*[5] dos algoritmos, para posteriormente realizar o treino e o teste dos modelos.

Após todo o processo de *Modeling* é necessário avaliar os resultados obtidos, através da quinta fase denominada ***Evaluation***. Com esse propósito foram escolhidas as métricas de precisão: *MAE, MSE, $M_d$AE, $R^2$ Score e EVS*. Todas estas métricas já foram abordadas anteriormente neste TFM e não serão mais aprofundadas neste capítulo.

Por último resta a fase do ***Deployment*** que consiste em implementar todo o processo realizado até aqui e os resultados obtidos em algo concreto, como a construção de um *software*, e a manutenção do mesmo. (Schröer et al., 2021) As aplicações dos resultados obtidos neste TFM serão abordadas nos capítulos posteriores.

## 3.1. Modelo 1 – Previsão dos óbitos associados à *COVID-19*

Para prever os dados da mortalidade associada à doença *COVID-19*, foram usados o número de infeções diárias, o total de pessoas totalmente vacinadas (com pelo menos duas doses da vacina) e também o número de vacinas administradas diariamente em Portugal, presentes na base de dados *Our World in Data* («COVID-19 Data Explorer»). Os dados da temperatura foram obtidos através da base de dados *National Centers for Environmental Information* e referem-se à temperatura média registada na estação metereológica LISBOA GEOFISICA. (Menne et al., 2012) As figuras seguintes mostram os gráficos de todas as variáveis de 2 de março de 2020 a 28 de fevereiro de 2022, tal como a decomposição sazonal da variável da mortalidade:

[4] ***lag:*** *deslocar os valores um ou mais passos à frente.* ***Fonte:*** Time Series as Features | Kaggle
[5] ***hyperparameter optimization:*** *ajuste dos hiperparâmetros dos algoritmos de forma a maximizar o desempenho do mesmo.* ***Fonte:*** ***(Elgeldawi et al., 2021)***

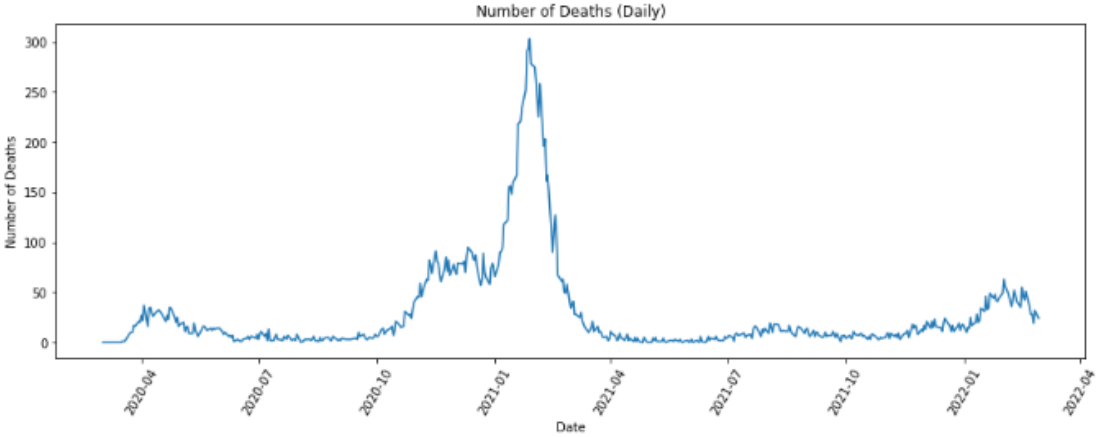



***Figura 7:*** *Número diário de mortes relacionadas com o vírus SARS-CoV-2 em Portugal*
***Fonte:*** *Elaboração Própria*

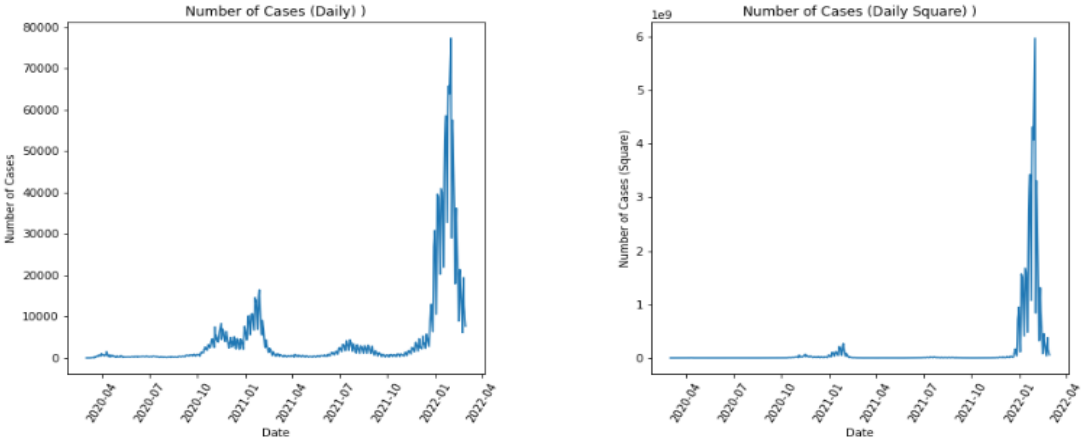



***Figura 8:*** *Número de infeções e número de infeções ao quadrado em Portugal*
***Fonte:*** *Elaboração Própria*

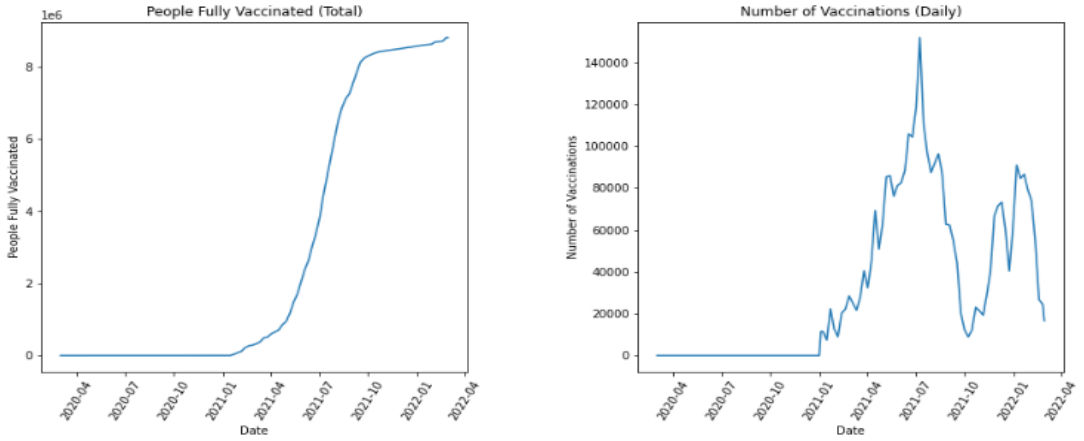



***Figura 9:*** *Total de pessoas totalmente vacinadas e número diário de vacinações em Portugal*
***Fonte:*** *Elaboração Própria*

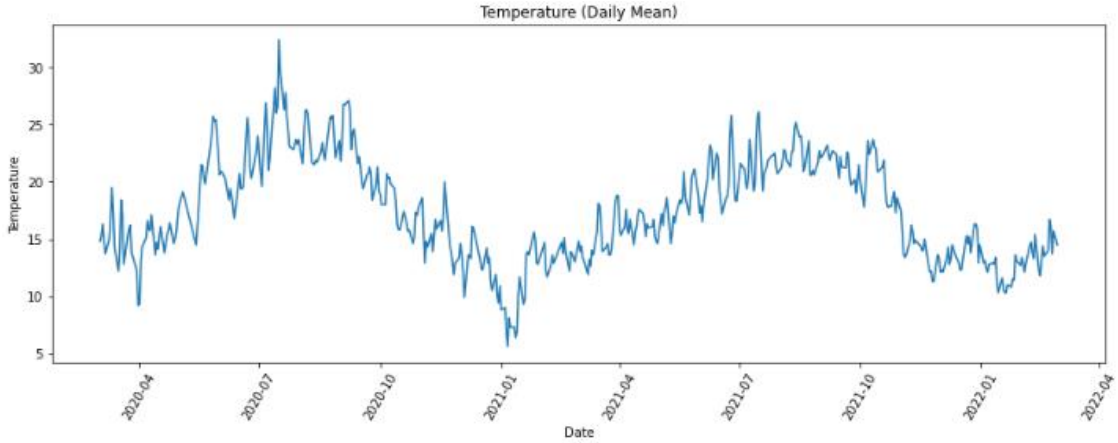



***Figura 10:*** *Temperatura média diária em Portugal*
***Fonte:*** *Elaboração Própria*

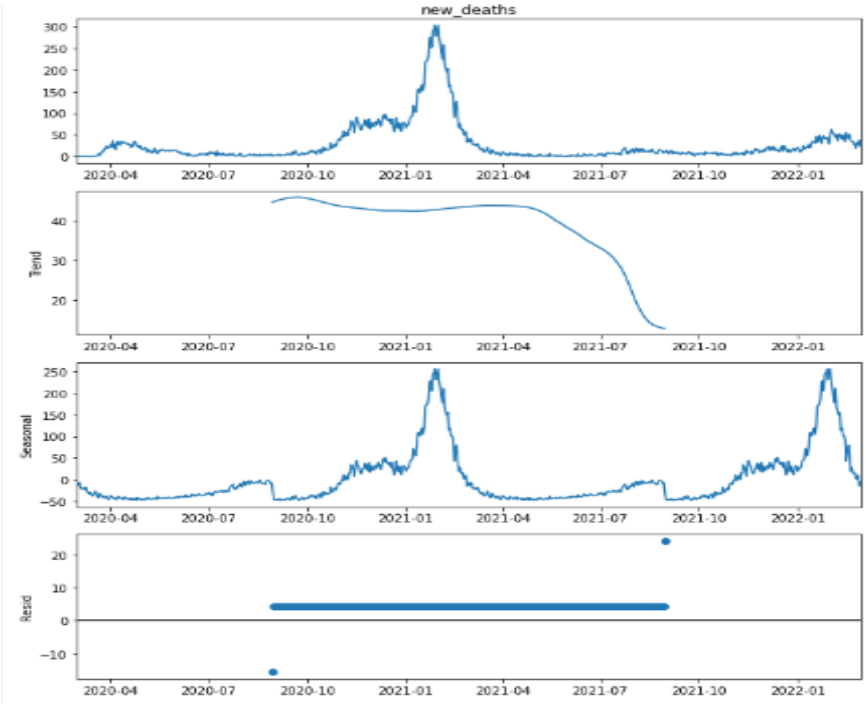



***Figura 11:*** *Decomposição Sazonal da séria das Temperaturas média diárias em Portugal*
***Fonte:*** *Elaboração Própria*

Como existiam dados em falta na base de dados, decidiu-se proceder à sua substituição através de dois métodos diferentes, o primeiro foi substituir os dados iniciais em falta por zero, e a segunda foi através do método *interpolate* do python, para preencher os restantes dados em falta. (Albon,2017) Foram também removidos todos os dados registados aos fins de semana (Sábado e Domingo) tendo em conta a inconsistência dos dados nesses mesmos períodos. Por fim, para que todas as variáveis tivessem a mesma escala e para que fosse possível medir o impacto de cada uma das variáveis no modelo, foi realizada uma estandardização dos dados através da função *StandardScaler* do módulo *scikit-learn* do software *python*. (Avila e Hauck, 2017)

Devido a estudos já efetuados, sabe-se que o impacto da vacinação (Dyer, 2021), do número de infeções (Jin, 2021) e da temperatura (Tapia-Muñoz et al., 2022), foram geradas nove novas variáveis relacionadas com as variáveis originais do número de novos casos diários, número de vacinações diárias e a média das temperaturas diárias, com *lag's* de 7, 14 e 21 dias. Em relação ao total de pessoas totalmente vacinadas não foram feitos os *lag's* devido ao facto de se tratar de um total e não de números diários. Após a criação das novas variáveis, foi seguido um critério de escolha das variáveis para o modelo através da correlação que cada uma delas tinha com a variável dependente, ou seja, o *lag* da variável com maior correlação foi o escolhido.

O próximo passo foi dividir os dados em amostras de treino e teste, de forma aleatória, de forma a combater o *overfitting*[6], sendo que amostra de treino corresponde a 80% da amostra total e a amostra de teste a 20% da mesma.

Posteriormente foi testado o *VIF* entre as variáveis independentes, de forma a compreender se existia correlação entre as mesmas. Por último observou-se o *p-value* do teste *t* para comprovar a existência de variáveis não significativas.

Na fase seguinte foi realizada uma *hyperparameter optimization* dos algoritmos com o objetivo de estimar os parâmetros ótimos para cada um. Essa *hyperparameter optimization* foi realizada para os algoritmos: *Ridge*, *LASSO*, *Gradient Boosting*, *MLP* e *Random Forest*, inserindo dados aleatórios para os parâmetros dos algoritmos e realizando um grande número de iterações, através de vários métodos de *cross-validation*, (Avila e Hauck, 2017) tentando atingir a convergência. Após terem sido estimados os dados da previsão para a fase de teste, realizou-se um teste de *Durbin Watson* aos resíduos, para testar a existência de autocorrelação. Também foi calculada a média de todas as observações dos resíduos, de forma a inferir se o valor estaria próximo de 0. (Hyndman e Athanasopoulos, 2018) Por último, foi avaliada a eficiência da previsão, através da comparação de algumas medidas de validação como o *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, *Median Absolute Error* ($M_dAE$), *Explained Variance Score* e o $R^2$ *Score*.

[6] ***overfitting:*** *ocorre quando os algoritmos têm uma boa performance na fase de treino e uma má performance na fase de teste.* ***Fonte:*** ***(Ying, 2019)***

## 3.2. Modelo 2 – Previsão dos valores diários de fecho do *PSI*

O modelo 2, consiste em estimar os valores do *Portuguese Stock Index (PSI)* a partir de dados da mobilidade, vacinação e número de infeções causadas pela *COVID-19*. Os dados do *PSI* foram obtidos através de um módulo do *python* denominado *yfinance* (Aroussi, 2022), que é basicamente um *Application programming interface* (*API*)[7] que permite aceder à base de dados do à base de dados do *Yahoo Finance* («Yahoo Finance - Stock Market Live, Quotes, Business & Finance News»). Os dados da mobilidade foram obtidos através de utilizadores com conta *Google*, que ativaram a definição "Histórico de localizações" da sua conta, e representam a variação percentual das deslocações a mercearias e farmácias, parques, estações de transportes públicos, retalho e lazer, residencial e locais de trabalho, em comparação aos valores da mediana para o dia da semana correspondente registados entre os dias 3 de janeiro e 6 de fevereiro de 2020. («COVID-19 Community Mobility Report») Ainda relacionado com a mobilidade são usados os dados do *stringency index* presentes na base de dados acerca da *COVID-19*, *Our World in Data* («COVID-19 Data Explorer»), que representa nove medidas de confinamento adotadas pelos governos, escaladas de 0 a 100, sendo que 0 representa o valor mais baixo de confinamento e 100 o valor mais alto, o valor final diário é calculado através da média dos valores dessas 9 medidas («COVID-19 Data Explorer»). Em relação aos dados da vacinação e do número de infeções, foram utilizados os mesmos do modelo anterior.

Como foi explicado no Modelo 1, o número de casos teve uma influência completamente diferente em termos de proporção após o começo do processo de vacinação. Neste modelo ao invés de se criar uma variável polinomial desse número de infeções, foram criadas duas variáveis de interação entre a variável e uma variável *dummy*[8] criada para o propósito. Essas variáveis *dummy's* denominadas de *before_vaccination*, que tomava o valor 1 se a data da observação foi antes do processo de vacinação ter começado e 0 se foi depois, e *after_vaccination*, em que o processo era revertido. Após a criação destas variáveis multiplicou-se o número de novos casos por

[7] ***API:*** *serve para "expor serviços ou dados fornecidos por uma aplicação de software através de um conjunto de recursos pré-definidos, tais como métodos, objectos ou URIs." **Fonte:** (Meng et al., 2018)*

[8] ***Variável Dummy:*** *variável que não é medida, convencionalmente, numa escala numérica. **Fonte:** (Suits, 1957)*

ambas gerando as variáveis *new_cases_before_vaccination* e *new_cases_after_vaccination*.

Nas figuras seguintes estão representados os gráficos de todas as variáveis de 2 de março de 2020 a 28 de fevereiro de 2022:

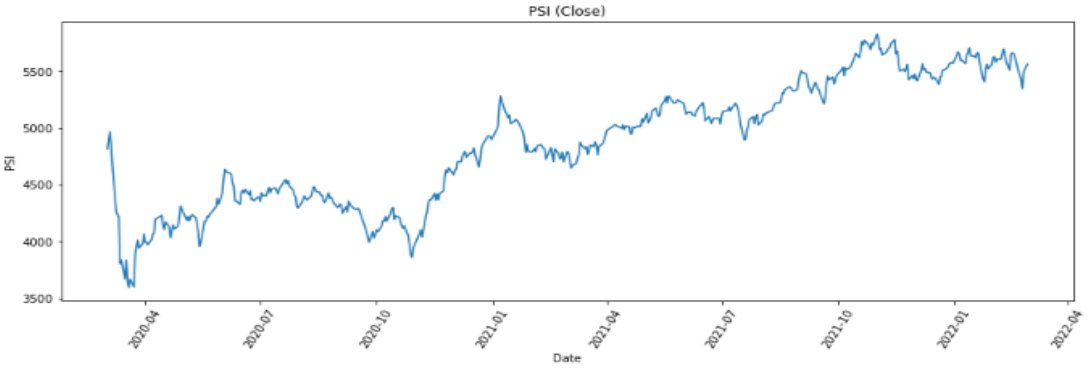



***Figura 12:*** *Dados diários do fecho do PSI*
***Fonte:*** *Elaboração Própria*

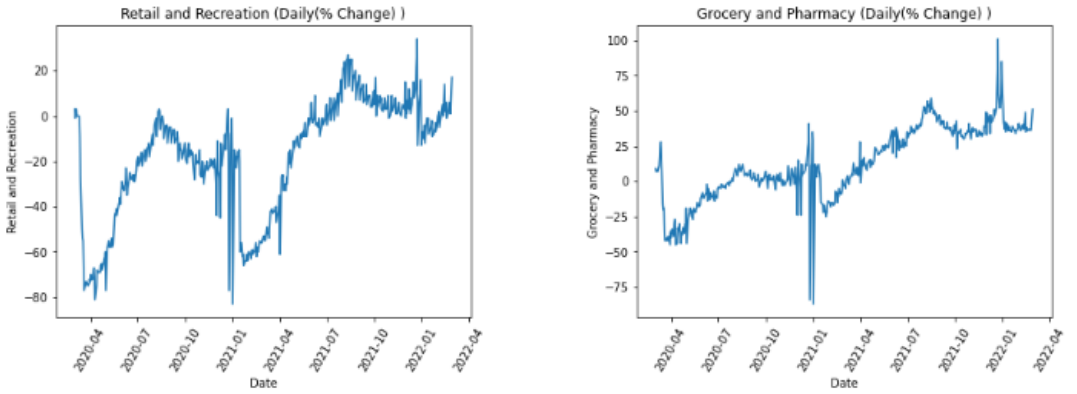



***Figura 13:*** *Dados diários da variação da mobilidade no retalho e lazer, e supermercados e farmácias*
***Fonte:*** *Elaboração Própria*

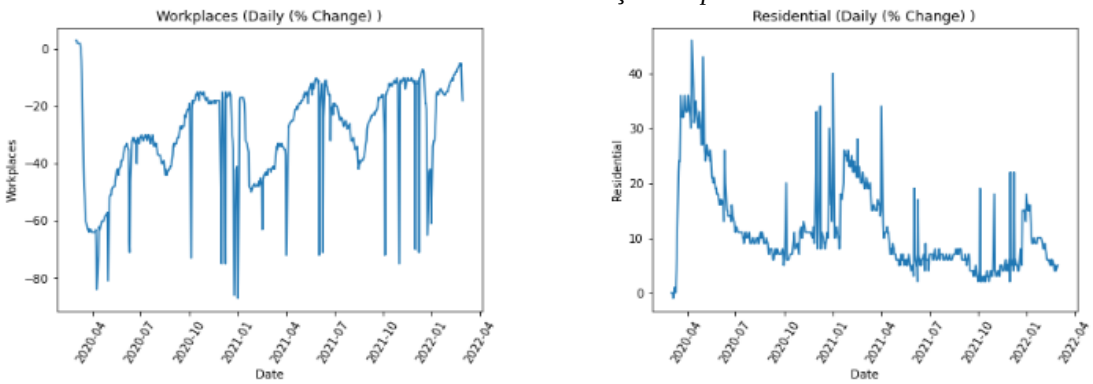



***Figura 14:*** *Dados diários da variação da mobilidade nos locais de trabalho e visitas residenciais*
***Fonte:*** *Elaboração Própria*

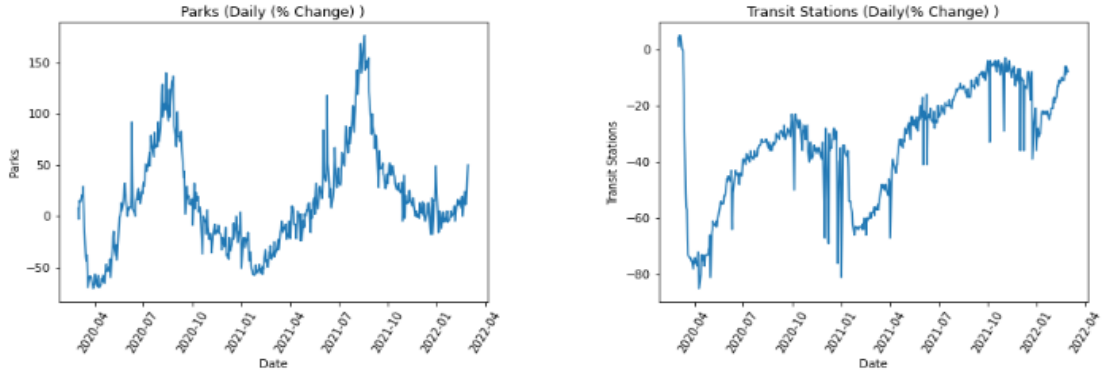



***Figura 15:*** *Dados diários da variação da mobilidade nos parques e estações de transportes públicos*
***Fonte:*** *Elaboração Própria*

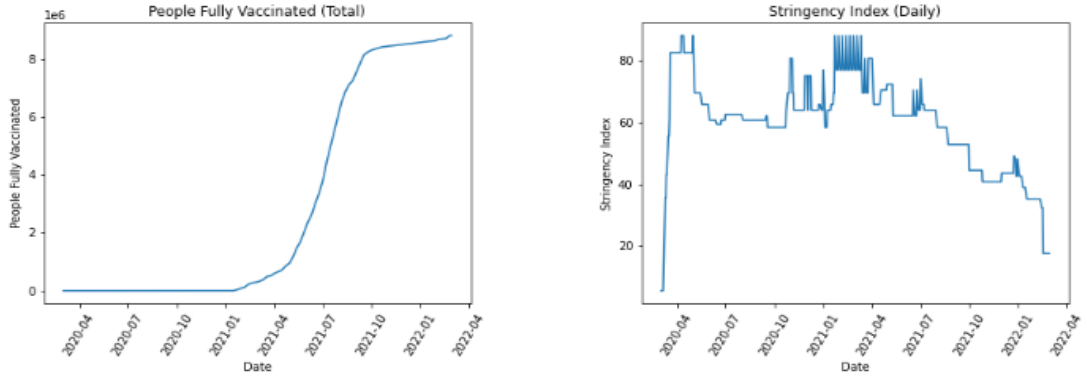



***Figura 16:*** *Total de indivíduos totalmente vacinados e valores diários do stringency index*
***Fonte:*** *Elaboração Própria*

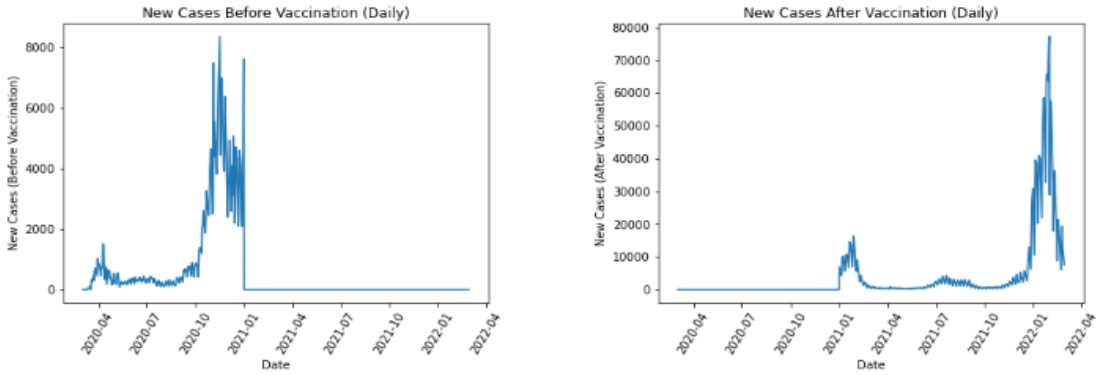



***Figura 17:*** *Número de novos casos antes e depois do processo de vacinação*
***Fonte:*** *Elaboração Própria*

Tal como nas variáveis do Modelo 1, procedeu-se à substituição dos *missing values*, substituindo os dados iniciais em falta por zero, e através do método *interpolate* do *python*, para preencher os restantes dados em falta. (Albon,2017) Foram também removidos todos os dados registados aos fins de semana (Sábado e Domingo) tendo em conta a inconsistência dos dados nesses mesmos períodos, principalmente porque não existem dados do *PSI* nesses mesmos dias. Para que todas as variáveis tivessem a mesma escala e para que fosse possível medir o impacto de cada uma das variáveis no modelo,

foi realizada uma estandardização dos dados através da função *StandardScaler* do módulo *scikit-learn* do software *python*, mais uma vez. (Avila e Hauck, 2017)

Após os processos descritos anteriormente verificou-se o *VIF* das variáveis em toda a amostra, verificando-se a existência de elevados valores do *VIF* para algumas variáveis, como pode ser percetível através da visualização do Anexo 12, muito devido à alta correlação entre as variáveis da mobilidade, como se pode observar no Anexo 13. Posto isto, e para que fosse possível combater a multicolinearidade existente entre as variáveis preditivas, elaborou-se uma Análise de Componentes Principais, de forma a conseguir agrupar as variáveis em componentes, e posteriormente proceder à regressão com esses mesmos componentes. O critério de escolha dos componentes do número de componentes foi realizado através da Análise Paralela de Horn.

De forma a realizar a validação do modelo com os componentes escolhidos, foi feita uma divisão dos dados em amostras de treino e teste, novamente de forma aleatória, com 80% dos dados a serem representados através da amostra de treino e 20% através da amostra de teste. Posteriormente foi feita uma regressão linear do modelo, para que fosse possível obter o *p-value* do teste *t* para comprovar a existência de variáveis não significativas.

Após a remoção das variáveis não significativas, foi realizada uma *hyperparameter optimization* como aconteceu no Modelo 1, através dos mesmos métodos de *cross-validation*. Após a previsão ter sido efetuada através de todos os algoritmos, testaram-se novamente os resíduos através do mesmo processo do Modelo 1 e foram avaliados os dados obtidos nessa previsão através das métricas estabelecidas.

# 4. Resultados

## 4.1. Modelo 1 – Previsão dos óbitos associados à *COVID-19*

De forma a associar-se o nome atribuído às variáveis no *software python* à sua descrição foi concebida a seguinte tabela.

***Tabela 1:*** *Descrição das variáveis do software python – Modelo 1*

| Variáveis *python* | Descrição |
|---|---|
| ***new_deaths*** | Número diário de mortes ***(Variável Dependente)*** |
| ***people_fully_vaccinated*** | Total de pessoas totalmente vacinadas (pelo menos duas doses da vacina) |
| ***new_vaccinations*** | Número de vacinações diárias |
| ***new_vaccinations_lag7d*** | Número de vacinações diárias com *lag* de 7 dias |
| ***new_vaccinations_lag14d*** | Número de vacinações diárias com *lag* de 14 dias |
| ***new_vaccinations_lag21d*** | Número de vacinações diárias com *lag* de 21 dias |
| ***new_cases*** | Número diário de infeções |
| ***new_cases_lag7d*** | Número diário de infeções com *lag* de 7 dias |
| ***new_cases_lag14d*** | Número diário de infeções com *lag* de 14 dias |
| ***new_cases_lag21d*** | Número diário de infeções com *lag* de 21 dias |
| ***new_cases_square*** | Número diário de infeções ao quadrado |
| ***new_cases_square_lag7d*** | Número diário de infeções ao quadrado com *lag* de 7 dias |
| ***new_cases_square_lag14d*** | Número diário de infeções ao quadrado com *lag* de 14 dias |
| ***new_cases_square_lag21d*** | Número diário de infeções ao quadrado com *lag* de 21 dias |
| ***temperature*** | Temperatura media diária |
| ***temperature_lag7d*** | Temperatura media diária com *lag* de 7 dias |
| ***temperature_lag14d*** | Temperatura media diária com *lag* de 14 dias |
| ***temperature_lag21d*** | Temperatura media diária com *lag* de 21 dias |

Como mencionado no capítulo anterior, as variáveis foram selecionadas através da correlação que cada variável independente tinha com o número de mortes associadas à *COVID-19*. Através dos Anexos 1,2,3 e 4, que representam a matriz de correlações divididas pelo tipo de *lag* efetuado, pode-se observar essas mesmas correlações. O modelo foi então construído com as variáveis: *new_cases, people_fully_vaccinated, new_cases_lag7d, new_cases_square_lag7d, new_vaccinations_lag21d* e *temperature_lag21d.* Após a construção do modelo foram calculados os *VIF's* para cada uma das variáveis (Anexo 5). Observou-se que ambas as variáveis do número de infeções tinham

valores superiores a 10, que representa um elevado valor do *Variance Inflation Factor*, mas devido a uma das variáveis ser a variável polinomial de grau dois da outra, essa evidência foi ignorada. (Allison, 2012)

De seguida foi realizada a estimação de um modelo de regressão linear (*OLS*) conforme mostra a Figura 6.

**OLS Model**

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | new_deaths | R-squared: | 0.493 |
| Model: | OLS | Adj. R-squared: | 0.487 |
| Method: | Least Squares | F-statistic: | 75.91 |
| Date: | Thu, 15 Sep 2022 | Prob (F-statistic): | 1.99e-55 |
| Time: | 23:22:08 | Log-Likelihood: | -448.27 |
| No. Observations: | 396 | AIC: | 908.5 |
| Df Residuals: | 390 | BIC: | 932.4 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0123 | 0.038 | 0.322 | 0.748 | -0.063 | 0.087 |
| people_fully_vaccinated | -0.2985 | 0.055 | -5.451 | 0.000 | -0.406 | -0.191 |
| new_cases_square_lag7d | -0.9474 | 0.117 | -8.130 | 0.000 | -1.177 | -0.718 |
| new_cases_lag7d | 1.1393 | 0.130 | 8.794 | 0.000 | 0.885 | 1.394 |
| new_vaccinations_lag21d | -0.1138 | 0.050 | -2.277 | 0.023 | -0.212 | -0.016 |
| temperature_lag21d | -0.4070 | 0.046 | -8.830 | 0.000 | -0.498 | -0.316 |

| | | | |
|---|---|---|---|
| Omnibus: | 170.221 | Durbin-Watson: | 1.912 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 801.159 |
| Skew: | 1.836 | Prob(JB): | 1.07e-174 |
| Kurtosis: | 8.923 | Cond. No. | 7.66 |

***Figura 18:*** *OLS – Modelo 1*
***Fonte:*** *Elaboração Própria*

A partir da Figura 19 observa-se que o número diário de casos de *COVID-19* com um *lag* de 7 dias tem um impacto positivo (sendo o maior coeficiente do modelo), enquanto o número de casos ao quadrado com um *lag* de 7 dias já tem um coeficiente negativo e próximo dos valores do anterior, o que vai de encontro ao que seria esperado, tendo em conta a relação não linear existente entre esta variável preditiva e a variável dependente. Todas as variáveis de vacinação têm um impacto negativo no número de mortes, resultado que vai de encontro ao que seria esperado, tendo em conta a diminuição do número de mortes após o início do processo de vacinação, como já tinha sido provado em alguns artigos e como se pode visualizar graficamente nas Figuras 8 e 10. Sendo que o total de indivíduos totalmente vacinados tem uma maior importância em relação ao número diário de vacinações com um *lag* de 21 dias. Por fim, pode-se inferir que as temperaturas com um *lag* de 21 dias têm um impacto negativo no número de óbitos, e são

a terceira variável com maior coeficiente do modelo, provando o comportamento sazonal da variável dependente, já referenciado anteriormente.

Para a estimação dos demais algoritmos, utilizou-se uma *hiperparameter optimization* através dos algoritmos de *cross-validation*, *RidgeCV (Ridge)*, *LassoCV (LASSO)*, *RandomizedSearchCV (Gradient Boosting, MLP e Random Forest)*, do módulo *scikit-learn* (Avila e Hauck, 2017). Pode-se observar os resultados obtidos nesse processo nos Anexos 6,7,8,9 e 10, e também os gráficos da convergência dos valores do parâmetro *k* de *Ridge* e *LASSO* nos Anexos 11 e 12. É de realçar que os parâmetros ótimos de *Ridge* e *LASSO* são bastante próximos de 0, logo ambas as regressões se aproximam da regressão do método *OLS*.

Passando agora para a identificação do modelo com melhor poder preditivo, podemos observar na tabela abaixo as informações referentes a cada algoritmo.

***Tabela 2:*** *Medidas de precisão dos algoritmos – Modelo 1*

| **Algoritmo** | ***MAE*** | ***MSE*** | ***$M_dAE$*** | ***EVS*** | ***R2*** |
|---|---|---|---|---|---|
| ***OLS*** | 0.460 | 0.333 | 0.399 | 0.366 | 0.359 |
| ***Ridge*** | 0.460 | 0.333 | 0.399 | 0.366 | 0.359 |
| ***LASSO*** | 0.464 | 0.338 | 0.415 | 0.359 | 0.350 |
| ***Gradient Boosting*** | 0.116 | 0.033 | 0.064 | 0.937 | 0.936 |
| ***MLP*** | 0.125 | 0.043 | 0.063 | 0.917 | 0.916 |
| ***Random Forest*** | 0.113 | 0.034 | 0.073 | 0.936 | 0.935 |

Observando a Tabela 2, podemos inferir que o Gradient Boosting foi o melhor algoritmo preditivo, obtendo os melhores *scores* em todas as medidas, exceto no *MAE* em que foi superado pelo *Random Forest*, e no $M_dAE$ em que foi superado pelo *MLP*. O *Random Forest e o MLP* também obtiveram bons resultados, sendo RF superior a MLP em todas as medidas de pontuação, exceto no $M_dAE$. Isso indica que esses três algoritmos podem ser candidatos para fazer uma boa previsão futura dos dados diários de mortalidade por *COVID-19*. Por outro lado, os algoritmos lineares geraram resultados aquém do esperado, o que leva a crer que a relação entre as variáveis preditivas e a variável independente não é linear.

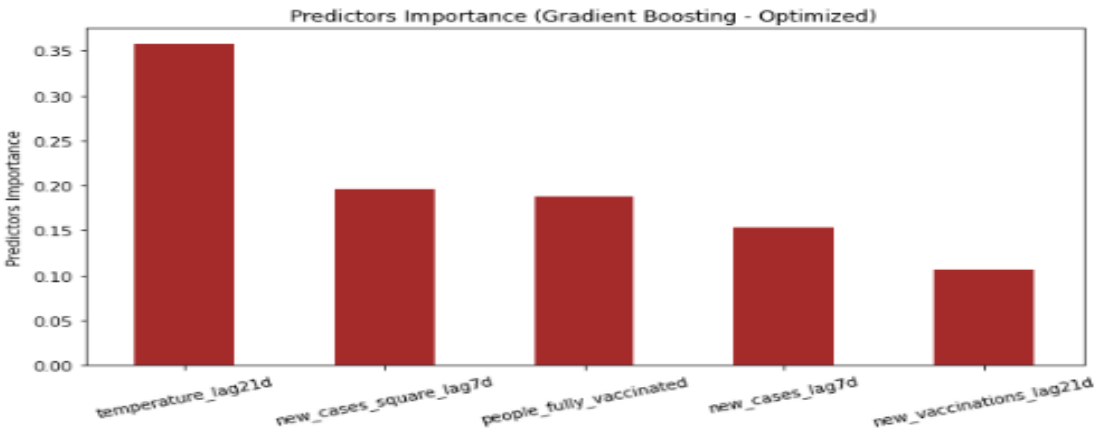



***Figura 19:*** *Gradient Boosting – Importância dos Preditores – Modelo 1*
***Fonte:*** *Elaboração Própria*

Na Figura 20 podemos observar a importância de cada uma das variáveis preditivas, através do algoritmo *Gradient Boosting*. A temperatura com um *lag* de 21 dias é a variável com maior importância para a previsão de óbitos por *COVID-19*, ao contrário do que aconteceu no OLS, em que o número de casos diários com *lag* de 7dias foi a variável com maior coeficiente. As variáveis do número de infeções com um *lag* de 7 dias e o total de indivíduos vacinados têm uma importância relativamente similar. Tendo em conta que este foi o algoritmo com maior poder preditivo e os coeficientes dados pelo *OLS*, podemos dizer que a temperatura média e as pessoas vacinadas desempenharam um papel preponderante na redução de mortes relacionas com o vírus *SARS-CoV2*.

Por fim, podemos observar na Tabela 3 os resultados do teste de *Durbin-Watson* e a média dos resíduos, para testar sua qualidade.

***Tabela 3:*** *Resultados do teste Durbin-Watson e média dos resíduos – Modelo 1*

| **Modelo** | **Teste *Durbin-Watson*** | **Média dos Resíduos** |
|---|---|---|
| ***OLS*** | 1.852 | -0.061 |
| ***Ridge*** | 1.852 | -0.061 |
| ***LASSO*** | 1.843 | -0.065 |
| ***Gradient Boosting*** | 1.886 | 0.022 |
| ***MLP*** | 1.804 | 0.014 |
| ***Random Forest*** | 1.933 | 0.020 |

Os valores da Tabela 2 mostram que os resíduos não estão correlacionados (estatística de teste entre 2±0.5), e sua média é próxima de 0 em todos os algoritmos. (McKinney et al., 2011) Podemos dizer que todos os modelos capturam adequadamente as informações presentes nos dados. (Hyndman e Athanasopoulos, 2018)

Por último, está representado na figura abaixo o gráfico dos dados previstos pelo melhor algoritmo (*Gradient Boosting*) e os dados reais do número de mortes relacionadas com a *COVID-19* nos dados da fase de teste, onde se pode observar um grande ajustamento dos dados previstos em relação aos reais.

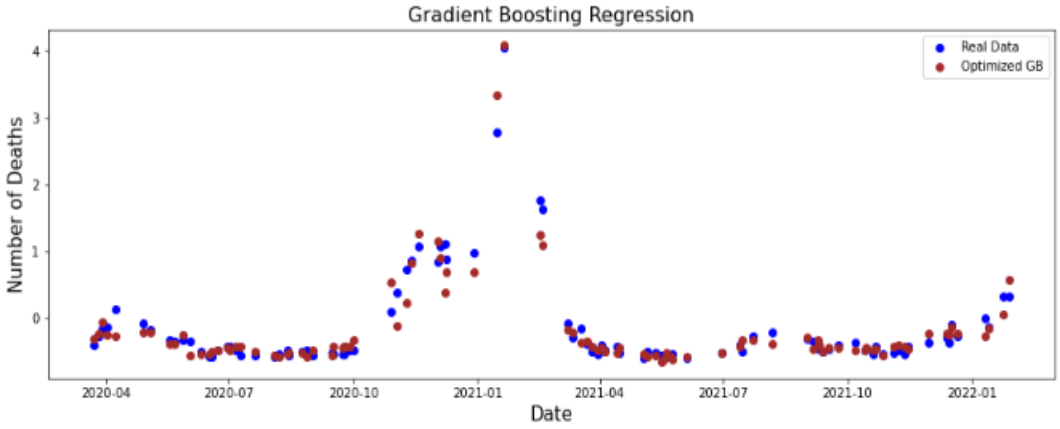



***Figura 20:*** *Dados previstos v.s dados reais do número de óbitos por COVID-19 em Portugal na fase de teste*
***Fonte:*** *Elaboração Própria*

## 4.2. Modelo 2 – Previsão dos valores diários de fecho do *PSI*

Para que seja possível associar as variáveis criadas no *python* com a descrição das mesmas, foi elaborada a seguinte tabela:

***Tabela 4:*** *Descrição das variáveis do software python – Modelo 2*

| Variáveis *python* | Descrição |
|---|---|
| ***psi*** | Valor de fecho diário do *stock index PSI* ***(Variável Dependente)*** |
| ***retail_and_recreation*** | Variação percentual da mobilidade no retalho e lazer |
| ***grocery_and_pharmacy*** | Variação percentual da mobilidade nos supermercados e farmácias |
| ***parks*** | Variação percentual da mobilidade nos parques |
| ***transit_stations*** | Variação percentual da mobilidade nas estações de transportes públicos |
| ***workplaces*** | Variação percentual da mobilidade nos locais de trabalho |
| ***residential*** | Variação percentual da mobilidade nas visitas a residências |
| ***people_fully_vaccinated*** | Total de pessoas totalmente vacinadas (pelo menos duas doses da vacina) |
| ***stringency_index*** | Índice de confinamento governamental |
| ***new_cases_before_vaccination*** | Número diário de infeções antes do processo de vacinação começar |
| ***new_cases_after_vaccination*** | Número diário de infeções após o processo de vacinação começar |

Começando pela Análise de Componentes Principais efetuada às variáveis, foram obtidos os seguintes resultados através da Análise Paralela de Horn:

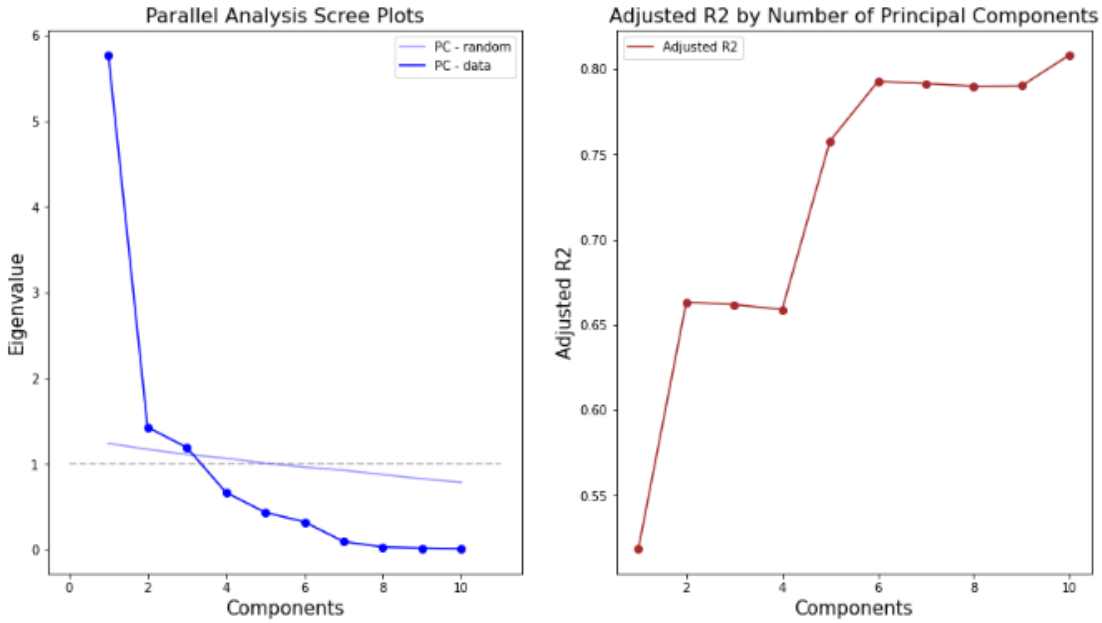



***Figura 21:*** *Análise Paralela de Horn e Adjusted $R^2$ por número de componentes*
***Fonte:*** *Elaboração Própria*

A Análise Paralela de Horn sugere a escolha de três componentes principais, como é possível visualizar na Figura 22. Através da Figura 22 também podemos observar que o *Adjusted $R^2$* é bastante similar para dois, três e quatro componentes, tendo uma subida acentuada quando são escolhidos cinco componentes principais. A escolha final nesta fase do processo foram os três componentes sugeridos pelo critério utilizado.

Após a escolha dos componentes foi realizada uma regressão linear com os mesmos, obtendo o seguinte *output*:

```
OLS Model
                            OLS Regression Results
==============================================================================
Dep. Variable:                    psi   R-squared:                       0.685
Model:                            OLS   Adj. R-squared:                  0.683
Method:                 Least Squares   F-statistic:                     289.2
Date:                Tue, 27 Sep 2022   Prob (F-statistic):           1.10e-99
Time:                        17:24:05   Log-Likelihood:                -338.75
No. Observations:                 403   AIC:                             685.5
Df Residuals:                     399   BIC:                             701.5
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0040      0.028      0.144      0.886      -0.051       0.059
PC1           -0.2994      0.011    -26.283      0.000      -0.322      -0.277
PC2            0.3153      0.023     13.563      0.000       0.270       0.361
PC3           -0.0034      0.026     -0.131      0.896      -0.054       0.047
==============================================================================
Omnibus:                       14.337   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               15.149
Skew:                          -0.460   Prob(JB):                     0.000513
Kurtosis:                       2.764   Cond. No.                         2.47
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Adj R2 Score before removing not significant variables
R2 Score: 0.6210354286540132
```

***Figura 22:*** *Modelo OLS com três componentes principais e respetivo Adjusted $R^2$ – Modelo 2*
***Fonte:*** *Elaboração Própria*

Pode-se observar através da Figura 23 que o *PC3* tem um *p-value* bastante próximo de 1, o que significa que o componente não é significativo para o modelo, segundo o teste *t* (Anexo 22). Removendo este componente do modelo foram obtidos os seguintes resultados, de forma a verificar se o *Adjusted* $R^2$ sofreu um aumento:

**OLS Model**

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | psi | R-squared: | 0.685 |
| Model: | OLS | Adj. R-squared: | 0.683 |
| Method: | Least Squares | F-statistic: | 434.8 |
| Date: | Tue, 27 Sep 2022 | Prob (F-statistic): | 4.70e-101 |
| Time: | 17:24:05 | Log-Likelihood: | -338.75 |
| No. Observations: | 403 | AIC: | 683.5 |
| Df Residuals: | 400 | BIC: | 695.5 |
| Df Model: | 2 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0040 | 0.028 | 0.143 | 0.886 | -0.051 | 0.059 |
| PC1 | -0.2994 | 0.011 | -26.315 | 0.000 | -0.322 | -0.277 |
| PC2 | 0.3152 | 0.023 | 13.580 | 0.000 | 0.270 | 0.361 |

| | | | |
|---|---|---|---|
| Omnibus: | 14.562 | Durbin-Watson: | 2.055 |
| Prob(Omnibus): | 0.001 | Jarque-Bera (JB): | 15.439 |
| Skew: | -0.466 | Prob(JB): | 0.000444 |
| Kurtosis: | 2.772 | Cond. No. | 2.47 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

**Adj R2 Score after removing not significant variables**
R2 Score: 0.6246691176340715

***Figura 23:*** *Modelo OLS com dois componentes principais e respetivo Adjusted* $R^2$ *– Modelo 2*
***Fonte:*** *Elaboração Própria*

Através da Figura 24 pode-se observar que de facto o *Adjusted* $R^2$ teve um ligeiro aumento, logo a variável foi removida definitivamente do modelo. Após esta regressão decidiu-se analisar as correlações entre as variáveis e os dois componentes restantes, para que fosse possível atribuir um significado a cada um dos componentes. Para esse propósito foi gerado o seguinte gráfico através do *software python*:

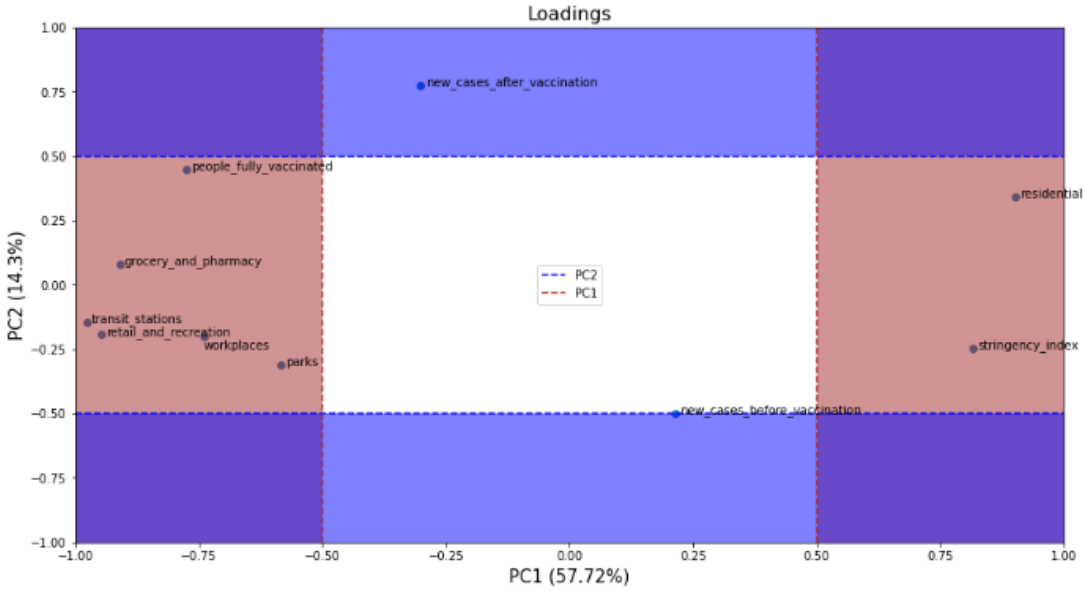



***Figura 24:*** *Correlação das variáveis com o número de componentes*
***Fonte:*** *Elaboração Própria*

Com a ajuda da Figura 25 pode-se observar que o *PC1* está bastante correlacionado negativamente com os dados da vacinação e da mobilidade exceto o *stringency index* e as visitas a residências. Este componente pode ser denominado então como "Inverso das deslocações a locais públicos e do número de vacinados". Em relação ao *PC2*, é visível que está bastante correlacionado com o número de infeções registadas após o início do processo de vacinação e ligeiramente correlacionado (um pouco acima de 0,5), com o número de casos registados antes desse mesmo processo ter começado, pode-se considerar então que o *PC2* explica o "Número de infeções registadas".

Avançando agora para a análise dos coeficientes dos componentes na regressão, pode-se observar através da Figura 24 que o *PC1* e o *PC2* têm um peso bastante similar para a estimação do *PSI*, sendo que o primeiro tem um coeficiente negativo, enquanto o segundo tem um coeficiente positivo. Pode-se então inferir que a diminuição das deslocações a locais públicos e o aumento do número de casos antes da vacinação, levou a uma diminuição do *PSI*, enquanto que o aumento da vacinação permite um crescimento do *PSI*, tal como o aumento do número de casos após a vacinação, muito relacionado com o aumento da vacinação.

Posteriormente a todo este processo descrito anteriormente, foi realizada a *hyperparameter optimization* dos restantes algoritmos através do processo descrito no Modelo 1. Os valores dessa otimização podem ser visualizados nos Anexos 15, 16, 17, 18 e 19 e os gráficos de convergência de *Ridge* e *LASSO* nos Anexos 20 e 21. De realçar

novamente que, tal como aconteceu no Modelo 1, os valores de *k* para a otimização de *Ridge* e *LASSO* situam-se muito próximo de 0, assemelhando-se novamente ao algoritmo *OLS*.

Com os algoritmos otimizados foi realizada a regressão do modelo através de todos, obtendo os seguintes resultados nas métricas de precisão utilizadas:

***Tabela 5:*** *Medidas de precisão dos algoritmos – Modelo 2*

| **Algoritmo** | ***MAE*** | ***MSE*** | ***$M_dAE$*** | ***EVS*** | ***R2*** |
|---|---|---|---|---|---|
| ***OLS*** | 0.495 | 0.370 | 0.486 | 0.633 | 0.632 |
| ***Ridge*** | 0.495 | 0.370 | 0.486 | 0.633 | 0.632 |
| ***LASSO*** | 0.495 | 0.370 | 0.485 | 0.632 | 0.632 |
| ***Gradient Boosting*** | 0.282 | 0.199 | 0.167 | 0.803 | 0.803 |
| ***MLP*** | 0.271 | 0.203 | 0.135 | 0.799 | 0.798 |
| ***Random Forest*** | 0.316 | 0.224 | 0.150 | 0.779 | 0.777 |

Com a visualização da Tabela 2, é percetível que o Gradient Boosting foi novamente o algoritmo com maior eficiência na previsão, obtendo os melhores *scores* em todas as medidas, exceto no *MAE* em que foi superado pelo *MLP* e no $M_dAE$ em que foi superado pelo *MLP e pelo Random Forest*. O *Random Forest e o MLP* provaram ser também bastante eficientes para a previsão do *PSI*, sendo que o *MLP* superou o *Random Forest*, ao contrário do que tinha acontecido no primeiro modelo. Isso indica que esses três algoritmos podem ser candidatos para fazer uma boa previsão futura dos dados diários de fecho do *PSI*. Os algoritmos lineares geraram resultados bastante aceitáveis também, o que leva a crer que existe uma relação linear mais significante entre os componentes principais determinados e o *PSI* em comparação com a relação existente entre as variáveis preditivas e a variável dependente do Modelo 1. De notar também os resultados quase iguais dos três primeiros algoritmos, devido ao facto relatado no parágrafo anterior, dos valores muito próximo de 0 do parâmetro optimizador.

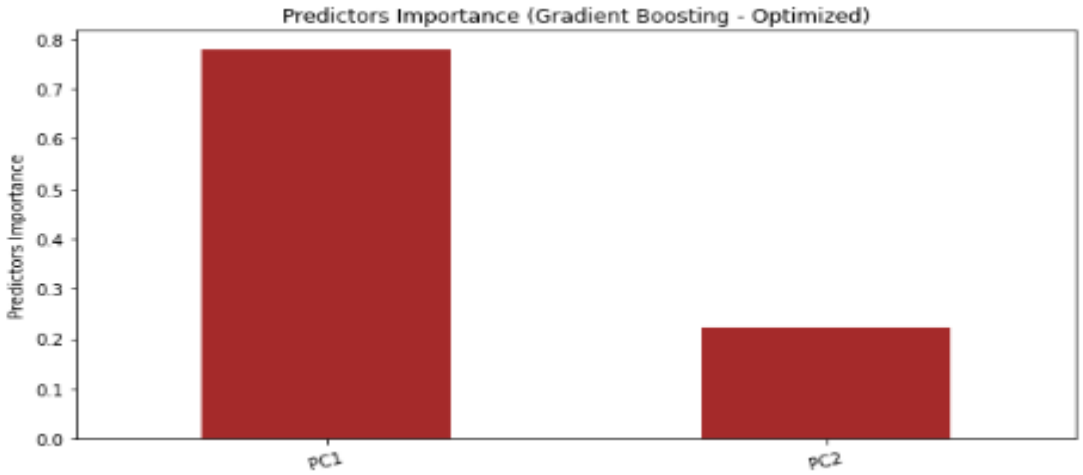



***Figura 25:*** *Gradient Boosting – Importância dos Preditores – Modelo 2*
***Fonte:*** *Elaboração Própria*

Na Figura 26 pode-se observar a importância que o GB atribuiu a cada um dos preditores. De notar a enorme diferença entre o *PC1* e o *PC2*, ao contrário do que acontecia no *OLS*, onde ambos os componentes tinham a mesma importância, apesar de sinais diferentes. O GB dá então muito mais importância à variação da mobilidade e à vacinação, do que ao número de infeções diária, em termos de estimar os valores do *PSI* eficientemente.

Para testar se os algoritmos captaram a informação fundamental dos dados, foi realizado novamente um Teste *Durbin-Watson* aos resíduos de cada algoritmo, obtendo os seguintes resultados:

***Tabela 6:*** *Resultados do teste Durbin-Watson e média dos resíduos – Modelo 2*

| **Modelo** | **Teste *Durbin-Watson*** | **Média dos Resíduos** |
|---|---|---|
| ***OLS*** | 1.636 | -0.002 |
| ***Ridge*** | 1.635 | -0.002 |
| ***LASSO*** | 1.636 | -0.002 |
| ***Gradient Boosting*** | 2.055 | 0.016 |
| ***MLP*** | 2.046 | 0.005 |
| ***Random Forest*** | 1.979 | 0.026 |

Os valores obtidos mostram que os resíduos não estão correlacionados (estatística de teste entre 2±0.5), e a sua média é próxima de 0 em todos os algoritmos, conclui-se então que todos os algoritmos capturam adequadamente as informações presentes nos dados.

Por último representa-se na Figura 27 o gráfico dos dados previstos pelo melhor algoritmo (*Gradient Boosting*) e os dados reais do *PSI* nos dados da fase de teste, onde se pode observar ajustamento bastante bom dos dados previstos em relação aos reais.

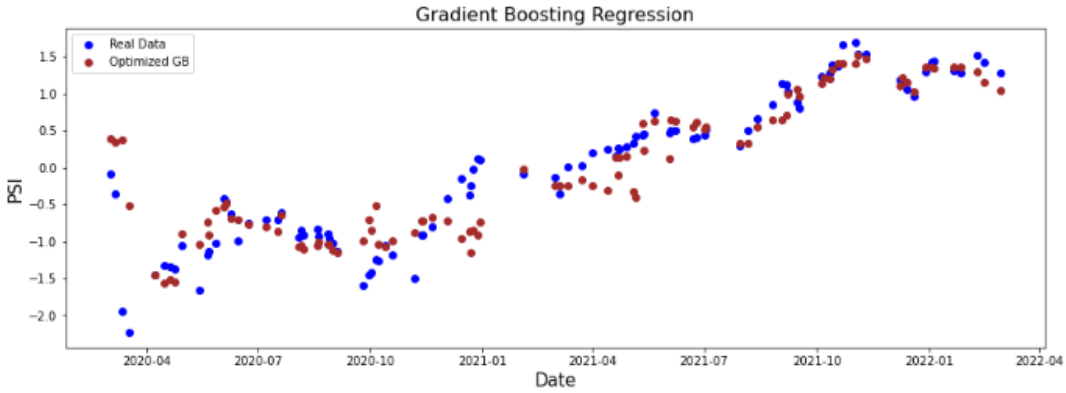



***Figura 26:*** *Dados previstos v.s dados reais dos valores de fecho do PSI na fase de teste*
***Fonte:*** *Elaboração Própria*

## 4.3. Discussão dos Resultados

Em relação ao primeiro modelo, o *GB* foi o algoritmo com maior poder preditivo, algo que não aconteceu no artigo (Gupta et al., 2021) em que o *Random Forest* destacou-se como o algoritmo mais robusto para prever os óbitos causados pela pandemia. No artigo (Rustagi et al., 2022) foi utilizado um algoritmo, que neste TFM não foi escolhido para a análise, denominado de *Support Vector Machine (SVM)*, acabando por ser aquele com melhores resultados, a sua utilização poderá ser considerada em trabalhos futuros, estudando o impacto da vacinação. Ainda em relação ao artigo (Rustagi et al., 2022), ficou verificado que a vacinação tem um papel fundamental no decréscimo do número de óbitos, algo que também ficou salientado neste TFM, tendo em conta que é o terceiro preditor mais importante (aproximadamente 20%), no *GB*. No artigo (Saba et al., 2021) o *GB* não obteve a melhor *performance* para qualquer dos países utilizados, ao contrário do que foi verificado no presente trabalho em que superou os outros algoritmos em quase todas as medidas de precisão utilizadas, para a previsão do número de mortes em Portugal. Ainda no artigo (Saba et al., 2021) verifica-se que o *RF* foi um dos algoritmos que obteve melhor performance, algo que também foi verificado neste TFM para Portugal, obtendo valores muito similares ao *GB*, sendo assim considerado o segundo modelo com maior poder preditivo. No artigo (Li et al., 2021) é referido que a propagação do vírus SARS-CoV-2 é menor na existência de temperaturas mais altas, algo que ficou salientado nos

resultados obtidos neste TFM, sendo que a média da temperatura diária registada em Portugal, tem um coeficiente negativo no *OLS* e a variável com mais importância no *GB*.

Relativamente ao segundo modelo, no artigo (Moghaddam et al., 2016) é seguido uma abordagem para a estimação do *stock index NASDAQ*, através de redes neuronais, obtendo resultados bastantes robustos, algo que também aconteceu neste TFM, sendo o segundo algoritmo com maior poder preditivo. No artigo (Xue et al., 2020) foi seguida uma abordagem através de *SVM* e *GB*, obtendo melhores resultados para o *SVM*, algoritmo que não foi usado neste TFM, podendo a sua utilização ser considerada para trabalhos futuros relacionados com os *Stock Index*, como foi também referido para o primeiro modelo. No artigo (Pavlyshenko, 2020) fica patente o impacto que a pandemia *COVID-19* teve nos *Stock Index*, algo bastante patente também na estimação deste segundo modelo, obtendo elevados valores de precisão dos modelos estimados, sendo que o *PC1*, que representa o inverso das deslocações a locais públicos a vacinação, tem uma relação negativa com *o PSI* no *OLS*, e é o preditor com maior importância no *GB* (aproximadamente 80%), algoritmo que obteve melhores resultados. Significando que com o decréscimo das deslocações a locais públicos, relacionado com as políticas de confinamento impostas pelo governo Português, o *PSI* tem uma resposta negativa. Enquanto que em relação ao processo de vacinação, o *PSI* tem uma resposta positiva ao aumento do número de indivíduos totalmente vacinados em Portugal.

# 5. Conclusão e Trabalhos Futuros

Os objetivos deste TFM foram inferir sobre o impacto demográfico e financeiro da pandemia *COVID-19* em Portugal, e analisar qual o algoritmo com maior poder preditivo para esse propósito. Foram utilizadas várias variáveis preditivas, de forma a atingir os objetivos, relacionadas com o processo de vacinação, número de infeções, mobilidade, temperatura e políticas de confinamento. As variáveis dependentes selecionadas, relacionadas com tais impactos foram o número de óbitos relacionados com a *COVID-19* e os números de fecho diário do *PSI*.

Para a estimação do primeiro modelo, foi criada uma variável, que representa o número de infeções com grau polinomial 2. Foram usados *lag's* de 7, 14 e 21 dias do número diário de vacinações e de ambas as variáveis do número diário de infeções. As variáveis que apresentaram maior correlação com o número diário de óbitos, dentro do mesmo grupo, foram escolhidas e foram utilizados diversos algoritmos de *machine learning* para a previsão. O algoritmo com maior poder preditivo foi o *Gradient Boosting*, e a variável preditiva com maior importância para a estimação foi a temperatura média diária em Portugal, com um *lag* de 21 dias.

Em relação ao segundo modelo, tendo em conta a correlação elevada registada entre as variáveis preditivas escolhidas, foi realizada uma *PCA*. Através dessa *PCA* surgiram dois componentes principais, em que o primeiro representava o inverso das deslocações a locais públicos e do número de vacinados e o segundo o número de infeções registadas. O algoritmo com maior poder preditivo foi o *Gradient Boosting* novamente, atribuindo maior importante ao *PC1*, ou seja, ao inverso das deslocações a locais públicos e do número de vacinados.

Ambos os modelos obtiveram resultados robustos para os algoritmos não lineares. No primeiro modelo os algoritmos lineares ficaram aquém do esperado, fica assim patente a existência de uma relação não linear entre as variáveis preditivas e a variável dependente. No segundo modelo a precisão dos modelos lineares e não lineares aproximaram-se bastante, sendo que todos obtiveram bons resultados em termos de precisão.

Os resíduos de todos os algoritmos, para ambos os modelos, não estavam correlacionados e tinham média próxima de zero, pode-se então afirmar que todos os algoritmos capturaram adequadamente as informações presentes nos dados.

Ficou bastante patente o impacto demográfico e financeiro que a pandemia teve em Portugal, sendo que o processo de vacinação foi o fator impulsionador da recuperação desses fatores. Para trabalhos futuros, fica a ideia de realizar uma análise de sentimentos de uma determinada rede social, para estudar a percentagem de vacinação de cada país, tendo em conta essa análise, algo que foi pensado inicialmente para este TFM, mas não foi possível realizar.

# Referências Bibliográficas

Akossou, A. & Palm, R. (2013). Impact of Data Structure on the Estimators R-Square And Adjusted R-Square in Linear Regression. . *Int. J. Math. Comput,* 20(3), 84-93.

Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. O'Reilly.

Allison, P. (2012). When Can You Safely Ignore Multicollinearity? *Statistical Horizons*. Obtido a 10 de outubro de 2022. https://statisticalhorizons.com/multicollinearity/

Almalki, A., Gokaraju, B., Acquaah, Y., & Turlapaty, A. (2022). Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues. *Healthcare*, *10*(2), 324. https://doi.org/10.3390/healthcare10020324

Andrade, C., & Petiz Lousã, E. (2021). Telework and Work–Family Conflict during COVID-19 Lockdown in Portugal: The Influence of Job-Related Factors. *Administrative Sciences*, *11*(3), 103. https://doi.org/10.3390/admsci11030103

Aparicio, J. T., Romao, M., & Costa, C. J. (2022). Predicting Bitcoin prices: The effect of interest rate, search on the internet, and energy prices. *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–5. https://doi.org/10.23919/CISTI54924.2022.9820085

Aparicio, S, Aparicio, J. & Costa, C. (2019) "Data Science and AI: Trends Analysis," 2019 *14th Iberian Conference on Information Systems and Technologies (CISTI),* pp. 1-6, https://doi.org/10.23919/CISTI.2019.8760820.

Aroussi, R. (2022). *Download market data from Yahoo! Finance's API* [Python]. https://github.com/ranaroussi/yfinance (Original work published 2017)

Arriaga, A., & Costa, C. J. (2023). Modelling and predicting daily COVID-19 (SARS-CoV-2) mortality in Portugal. *Proceedings of International Conference on Information Technology and Applications*. Springer Singapore

Avila, J., & Hauck, T. (2017). Scikit-learn cookbook: Over 80 recipes for machine learning in Python with scikit-learn. Packt Publishing.

Banerjee, D. (2014). Forecasting of Indian stock market using time-series ARIMA model. *2014 2nd International Conference on Business and Information Management (ICBIM)*, 131–135. https://doi.org/10.1109/ICBIM.2014.6970973

Berrar, D. (2019). Cross-Validation. Em *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Borghi, P. H., Zakordonets, O., & Teixeira, J. P. (2021). A COVID-19 time series forecasting model based on MLP ANN. *Procedia Computer Science*, *181*, 940–947. https://doi.org/10.1016/j.procs.2021.01.250

Bousquet, O., Luxburg, U. von, & Rätsch, G. (2004). *Advanced lectures on machine learning: ML Summer Schools* 2003, Canberra, Australia, February 2-14, 2003 [and] Tübingen, Germany, August 4-16, 2003: revised lectures. Springer.

Brown, J. D. (2009). Questions and answers about language testing statistics: 5.

Chaurasia, V., & Pal, S. (2022). Application of machine learning time series analysis for prediction COVID-19 pandemic. *Research on Biomedical Engineering*, *38*(1), 35–47. https://doi.org/10.1007/s42600-020-00105-4

Chen, H., Petukhov, A., & Wang, J. (2018). *The Dark Side of Circuit Breakers*. 57.

Chen, Y. (2016). Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression. *PLOS ONE*, *11*(1), e0146865. https://doi.org/10.1371/journal.pone.0146865

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Çokluk, Ö., & Koçak, D. (2016). Using Horn's Parallel Analysis Method in Exploratory Factor Analysis for Determining the Number of Factors. *Educational Sciences: Theory & Practice*. https://doi.org/10.12738/estp.2016.2.0328

Cord, M., & Cunningham, P. (2008). *Machine learning techniques for multimedia: Case studies on organization and retrieval*. Springer.

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A Methodology to Boost Data Science. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C.J., Aparicio, J.T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In: , *et al. Advances in Parallel & Distributed Processing, and Applications. Transactions on Computational Science and Computational Intelligence*. Springer, Cham. https://doi.org/10.1007/978-3-030-69984-0_7

*COVID-19 Data Explorer*. (sem data). Our World in Data. Obtido 9 de outubro de 2022, de https://ourworldindata.org/explorers/coronavirus-data-explorer

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, *14*(3), 391–403. https://doi.org/10.1081/QEN-120001878

Deng, H., Zhou, Y., Wang, L., & Zhang, C. (2021). Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC Medical Informatics and Decision Making*, *21*(1), 338. https://doi.org/10.1186/s12911-021-01701-9

Dyer, O. (2021). Covid-19: Moderna and Pfizer vaccines prevent infections as well as symptoms, CDC study finds. *BMJ*, n888. https://doi.org/10.1136/bmj.n888

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. *Informatics*, *8*(4), 79. https://doi.org/10.3390/informatics8040079

Estrada, M. A. R., Koutronas, E., & Lee, M. (2021). Stagpression: The Economic and Financial Impact of the COVID-19 Pandemic. *Contemporary Economics*, *15*(1), 19–33. https://doi.org/10.5709/ce.1897-9254.433

Ferreira, P. J. S. (2010). *Principios de econometría*. Rei dos livros.

Google LLC "Google COVID-19 Community Mobility Reports". Obtido a: 27 de setembro de 2022.
https://www.google.com/covid19/mobility/

Graham, M. H. (2003). CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION. *Ecology*, *84*(11), 2809–2815. https://doi.org/10.1890/02-3114

Gumaei, A., Al-Rakhami, M., Mahmoud Al Rahhal, M., Raddah H Albogamy, F., Al Maghayreh, E., & AlSalman, H. (2020). Prediction of COVID-19 Confirmed Cases Using Gradient Boosting Regression Method. *Computers, Materials & Continua*, *66*(1), 315–329. https://doi.org/10.32604/cmc.2020.012045

Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics*, *4*(2), 116–123. https://doi.org/10.26599/BDMA.2020.9020016

Haas, E. J., McLaughlin, J. M., Khan, F., Angulo, F. J., Anis, E., Lipsitch, M., Singer, S. R., Mircus, G., Brooks, N., Smaja, M., Pan, K., Southern, J., Swerdlow, D. L., Jodar, L., Levy, Y., & Alroy-Preis, S. (2022). Infections, hospitalisations, and deaths averted via a nationwide vaccination campaign using the Pfizer–BioNTech BNT162b2 mRNA COVID-19 vaccine in Israel: A retrospective surveillance study. *The Lancet Infectious Diseases*, *22*(3), 357–366. https://doi.org/10.1016/S1473-3099(21)00566-1

Harrison, R. L. (2010). Introduction to Monte Carlo Simulation. AIP Conference Proceedings 1204, 17, 6. https://doi.org/10.1063/1.3295638

Hyndman R. & Athanasopoulos J. (2028) *Forecasting: Principles and Practice (2nd ed)*. OTexts

Jin, R. (2021). The Lag between Daily Reported Covid-19 Cases and Deaths and Its Relationship to Age. *Journal of Public Health Research*, *10*(3), jphr.2021.2049. https://doi.org/10.4081/jphr.2021.2049

Karch, J. (2020). Improving on Adjusted R-Squared. *Collabra: Psychology*, *6*(1), 45. https://doi.org/10.1525/collabra.343

Kartal, M. T., Kiliç Depren, S., & Depren, Ö. (2021). How Main Stock Exchange Indices React to Covid-19 Pandemic: Daily Evidence from East Asian Countries. *Global Economic Review*, *50*(1), 54–71. https://doi.org/10.1080/1226508X.2020.1869055

Khan, W., Hussain, A., Khan, S. A., Al-Jumailey, M., Nawaz, R., & Liatsis, P. (2021). Analysing the impact of global demographic characteristics over the COVID-19 spread using class rule mining and pattern matching. *Royal Society Open Science*, *8*(1), 201823. https://doi.org/10.1098/rsos.201823

Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, *68*(6), 540. https://doi.org/10.4097/kjae.2015.68.6.540

Kubben, P., Dumontier, M., & Dekker, A. (Eds.). (2019). *Fundamentals of Clinical Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-99713-1

Li, M., Zhang, Z., Cao, W., Liu, Y., Du, B., Chen, C., Liu, Q., Uddin, Md. N., Jiang, S., Chen, C., Zhang, Y., & Wang, X. (2021). Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. *Science of The Total Environment*, *764*, 142810. https://doi.org/10.1016/j.scitotenv.2020.142810

Li, Y., Li, M., Rice, M., Zhang, H., Sha, D., Li, M., Su, Y., & Yang, C. (2021). The Impact of Policy Measures on Human Mobility, COVID-19 Cases, and Mortality in the US: A Spatiotemporal Perspective. *International Journal of Environmental Research and Public Health*, *18*(3), 996. https://doi.org/10.3390/ijerph18030996

Mahesh, B. (2018). Machine Learning Algorithms—A Review. 9(1), 7.

Martin-Barreiro, C., Ramirez-Figueroa, J. A., Cabezas, X., Leiva, V., & Galindo-Villardón, M. P. (2021). Disjoint and Functional Principal Component Analysis for Infected Cases and Deaths Due to COVID-19 in South American Countries with Sensor-Related Data. *Sensors*, *21*(12), 4094. https://doi.org/10.3390/s21124094

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93–100. https://doi.org/10.1002/wics.14

McKinney, W., Perktold, J., & Seabold, S. (2011). *Time Series Analysis in Python with statsmodels*. 107–113. https://doi.org/10.25080/Majora-ebaa42b7-012

Melkumova, L. E., & Shatskikh, S. Ya. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746–755. https://doi.org/10.1016/j.proeng.2017.09.615

Meng, M., Steinhardt, S., & Schubert, A. (2018). Application Programming Interface Documentation: What Do Software Developers Want? *Journal of Technical Writing and Communication*, *48*(3), 295–330. https://doi.org/10.1177/0047281617721853

Menne, Matthew J., Imke Durre, Bryant Korzeniewski, Shelley McNeill, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E.Gleason, and Tamara G. Houston (2012): Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. NOAA National Climatic Data Center. Obtido 9 de outubro de 2022. https://doi.org/10.7289/V5D21VHZ

Milhinhos, A., & Costa, P. M. (2020). On the Progression of COVID-19 in Portugal: A Comparative Analysis of Active Cases Using Non-linear Regression. *Frontiers in Public Health*, *8*, 495. https://doi.org/10.3389/fpubh.2020.00495

Moghaddam, A. H., Moghaddam, M. H., & Esfandyari, M. (2016). Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, *21*(41), 89–93. https://doi.org/10.1016/j.jefas.2016.07.002

Murray, L., Nguyen, H., Lee, Y.-F., Remmenga, M. D., & Smith, D. W. (2012). Variance inflation factors in Regression Models with Dummy Variables. *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1034

Ornella, L., Kruseman, G., & Crossa, J. (2020). Satellite Data and Supervised Learning to Prevent Impact of Drought on Crop Production: Meteorological Drought. in G. Ondrasek (Ed.), *Drought—Detection and Solutions*. IntechOpen. https://doi.org/10.5772/intechopen.85471

Ostertagová, E. (2012). Modelling using Polynomial Regression. *Procedia Engineering*, *48*, 500–506. https://doi.org/10.1016/j.proeng.2012.09.545

Park, Y.-S., & Lek, S. (2016). Artificial Neural Networks. Em *Developments in Environmental Modelling* (Vol. 28, pp. 123–140). Elsevier. https://doi.org/10.1016/B978-0-444-63623-2.00007-4

Pavlyshenko, B. M. (2020). *Regression Approach for Modeling COVID-19 Spread and its Impact On Stock Market* (arXiv:2004.01489). arXiv. https://doi.org/10.48550/arXiv.2004.01489

Pedregosa et al., *(2011)* Scikit-learn: Machine Learning in Python*, Journal of machine Learning research, 12, 2825-2830.*

Perone, G. (2022). Using the SARIMA Model to Forecast the Fourth Global Wave of Cumulative Deaths from COVID-19: Evidence from 12 Hard-Hit Big Countries. *Econometrics*, *10*(2), 18. https://doi.org/10.3390/econometrics10020018

*PSI 5 487,44 | Euronext Live quotes preços*. (sem data). Obtido 24 de setembro de 2022, de PSI 5 487,44 | Euronext Live quotes preços

Redell, N. (2019). *Shapley Decomposition of R-Squared in Machine Learning Models* (arXiv:1908.09718). arXiv. http://arxiv.org/abs/1908.09718

Polamuri, S. R., Srinivasi, K., & Mohan, A. K. (2019). Stock Market Prices Prediction using Random Forest and Extra Tree Regression. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 1224–1228). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijrte.c4314.098319

Risvik, H. (2007). Principal Component Analysis (PCA) & NIPALS algorithm. 6.

Rustagi, V., Bajaj, M., Tanvi, Singh, P., Aggarwal, R., AlAjmi, M. F., Hussain, A., Hassan, Md. I., Singh, A., & Singh, I. K. (2022). Analyzing the Effect of Vaccination Over COVID Cases and Deaths in Asian Countries Using Machine Learning Models. *Frontiers in Cellular and Infection Microbiology*, *11*, 806265. https://doi.org/10.3389/fcimb.2021.806265

Saba, T., Abunadi, I., Shahzad, M. N., & Khan, A. R. (2021). Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types. *Microscopy Research and Technique*, *84*(7), 1462–1474. https://doi.org/10.1002/jemt.23702

Saleh, H. (2022). *Machine Learning-Regression*. Thesis for: 4th year seminar https://doi.org/10.13140/RG.2.2.35768.67842

Samadani S. & Costa, C. (2021) "Forecasting real estate prices in Portugal : A data science approach," 2021 *16th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1-6, http://doi.org/10.23919/CISTI52073.2021.9476447.

Sampi, J., & Jooste, C. (2020). *Nowcasting Economic Activity in Times of COVID-19: An Approximation from the Google Community Mobility Report*. World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-9247

Sarirete, A. (2021). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Procedia Computer Science*, *194*, 280–287. https://doi.org/10.1016/j.procs.2021.10.083

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. E. (2013). A survey of forecast error measures. *World applied sciences journal*, *24*(24), 171-176.

Shrivastav, L. K., & Jha, S. K. (2021). A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. *Applied Intelligence*, *51*(5), 2727–2739. https://doi.org/10.1007/s10489-020-01997-6

Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., & Agha, R. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International Journal of Surgery*, *76*, 71–76. https://doi.org/10.1016/j.ijsu.2020.02.034

Suits, D. B. (1957). Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association*, *52*(280), 548–551. https://doi.org/10.1080/01621459.1957.10501412

Tapia-Muñoz, T., González-Santa Cruz, A., Clarke, H., Morris, W., Palmeiro-Silva, Y., & Allel, K. (2022). COVID-19 attributed mortality and ambient temperature: A global

ecological study using a two-stage regression model. *Pathogens and Global Health*, *116*(5), 319–329. https://doi.org/10.1080/20477724.2021.2007336

Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications*, *17*(3), 205–221. https://doi.org/10.23919/JCC.2020.03.017

Watson, O. J., Barnsley, G., Toor, J., Hogan, A. B., Winskill, P., & Ghani, A. C. (2022). Global impact of the first year of COVID-19 vaccination: A mathematical modelling study. *The Lancet Infectious Diseases*, *22*(9), 1293–1302. https://doi.org/10.1016/S1473-3099(22)00320-6

Xie, Z., & Li, D. (2020). *Health and Demographic Impact on COVID-19 Infection and Mortality in US Counties* (p. 2020.05.06.20093195). medRxiv. https://doi.org/10.1101/2020.05.06.20093195

Xue, P., Lei, Y., & Li, Y. (2020). Research and prediction of Shanghai-Shenzhen 20 Index Based on the Support Vector Machine Model and Gradient Boosting Regression Tree. *2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, 58–62. https://doi.org/10.1109/ICICAS51530.2020.00019

*Yahoo Finance—Stock Market Live, Quotes, Business & Finance News*. (sem data). Obtido a 27 de setembro de 2022, de https://finance.yahoo.com/

Yeşilkanat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons & Fractals*, *140*, 110210. https://doi.org/10.1016/j.chaos.2020.110210

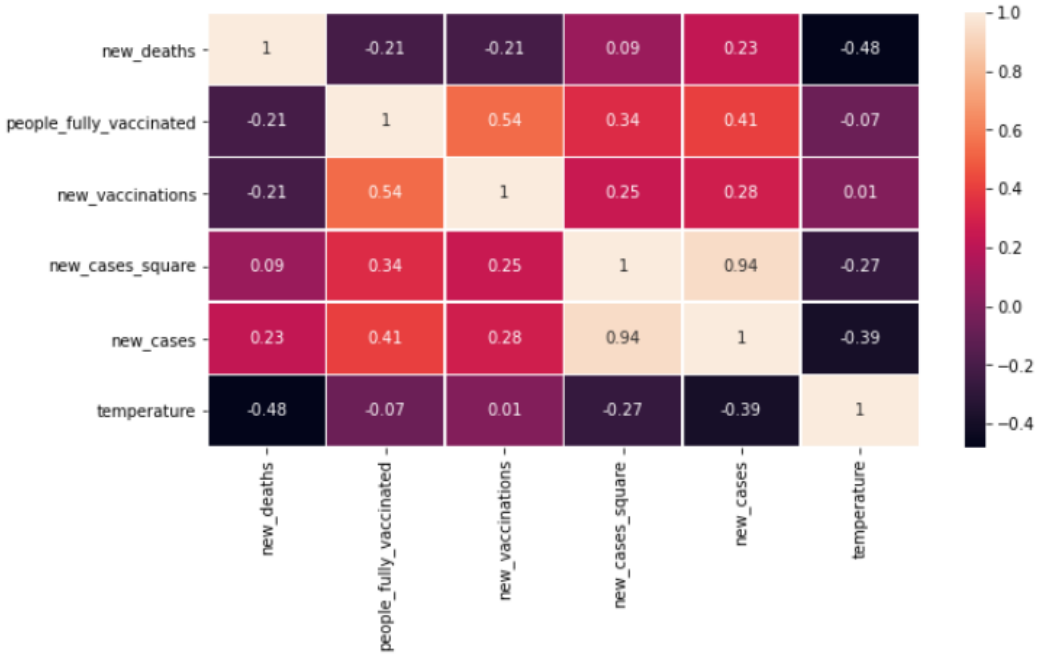
Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*, 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Zhang, D., Hu, M., & Ji, Q. (2020). Financial markets under the global pandemic of COVID-19. *Finance Research Letters*, *36*, 101528. https://doi.org/10.1016/j.frl.2020.101528

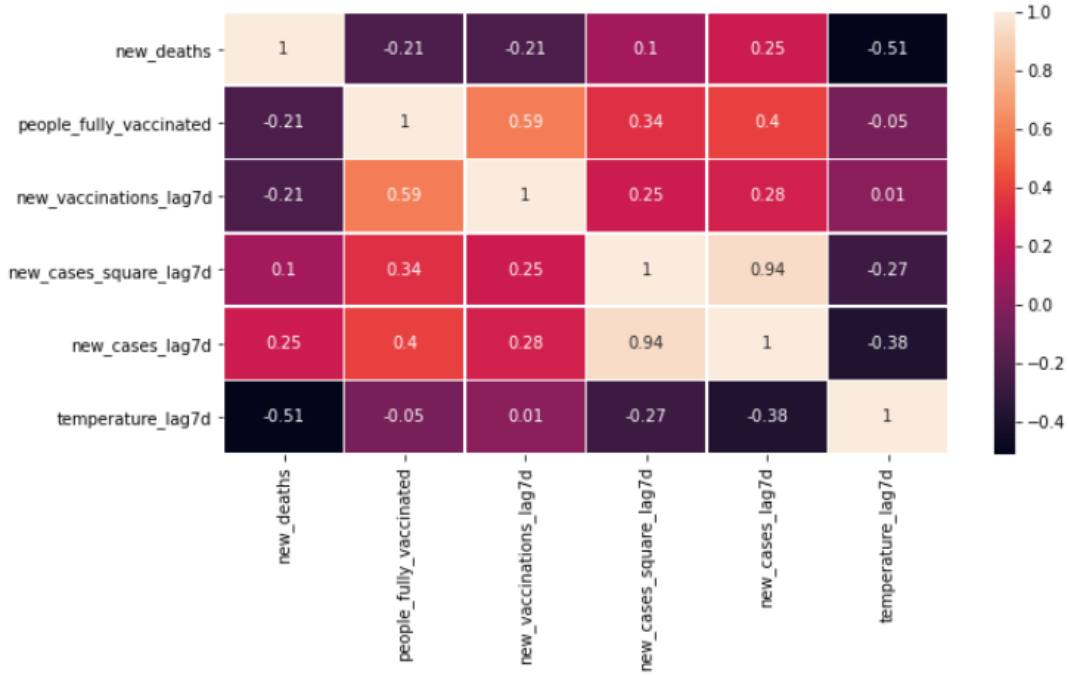
Zhang, D., Hu, M., & Ji, Q. (2020). Financial markets under the global pandemic of COVID-19. *Finance Research Letters*, *36*, 101528. https://doi.org/10.1016/j.frl.2020.101528

Zhu, X. (Jerry). (2005). *Semi-Supervised Learning Literature Survey* [Technical Report]. University of Wisconsin-Madison Department of Computer Sciences. https://minds.wisconsin.edu/handle/1793/60444

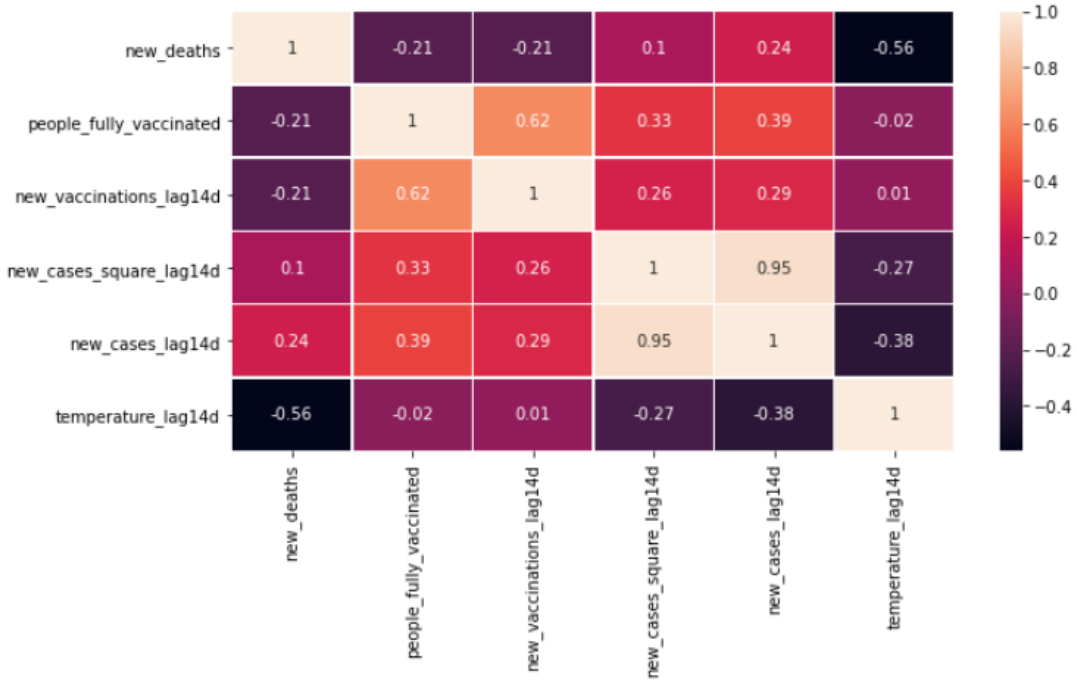
# Anexos



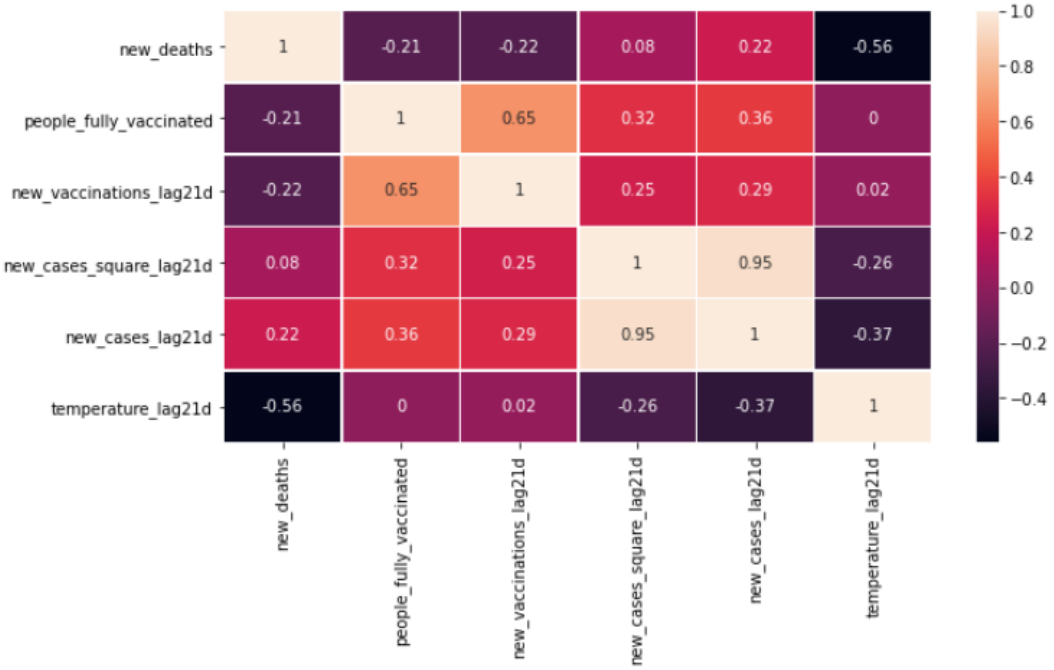

***Anexo 1:*** Matriz de correlações do Modelo 1 – Sem *lag*




***Anexo 2:*** Matriz de correlações do Modelo 1 – *Lag* 7 dias




***Anexo 3:*** Matriz de correlações do Modelo 1 – *Lag* 14 dias




***Anexo 4:*** Matriz de correlações do Modelo 1 – *Lag* 21 dias

**VIF (Variance Inflation Factors):**

| | feature | VIF |
|---|---|---|
| 0 | const | 1.003816 |
| 1 | people_fully_vaccinated | 2.097835 |
| 2 | new_cases_square_lag7d | 10.759980 |
| 3 | new_cases_lag7d | 12.982356 |
| 4 | new_vaccinations_lag21d | 1.804289 |
| 5 | temperature_lag21d | 1.458455 |

***Anexo 5:*** *VIF's Modelo 1*

**Optimal Alpha (Ridge):** 0.01

***Anexo 6:*** *Modelo 1 – Hyperparameter Optimization: Ridge*

**Optimal Alpha (Lasso):** 0.0053109756270552045

***Anexo 7:*** *Modelo 1 – Hyperparameter Optimization: LASSO*

| Predictors | Best Values |
|---|---|
| learning_rate | 0.076550 |
| max_depth | 8.000000 |
| max_features | 2.000000 |
| min_samples_leaf | 6.000000 |
| min_samples_split | 7.000000 |
| n_estimators | 772.000000 |
| subsample | 0.874072 |

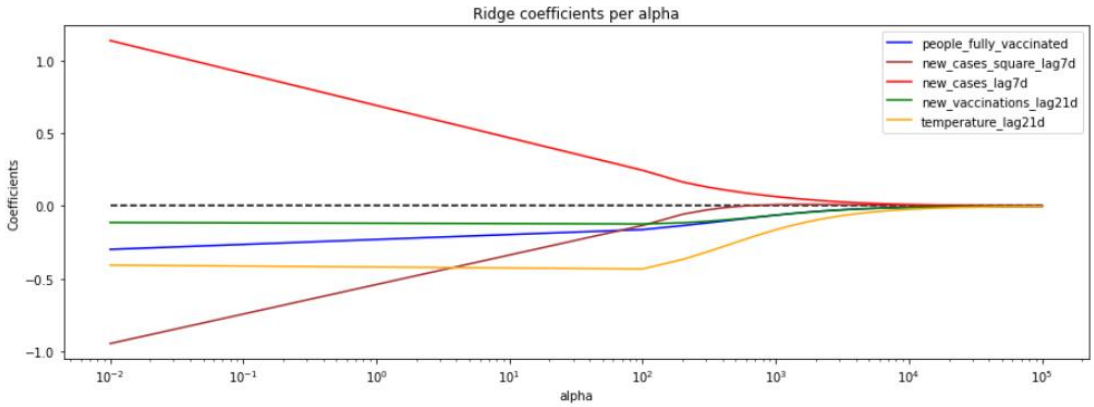
***Anexo 8:*** *Modelo 1 – Hyperparameter Optimization: Gradient Boosting*

| Predictors | Best Values |
|---|---|
| activation | logistic |
| alpha | 0.069459 |
| hidden_layer_sizes | 17 |
| solver | lbfgs |

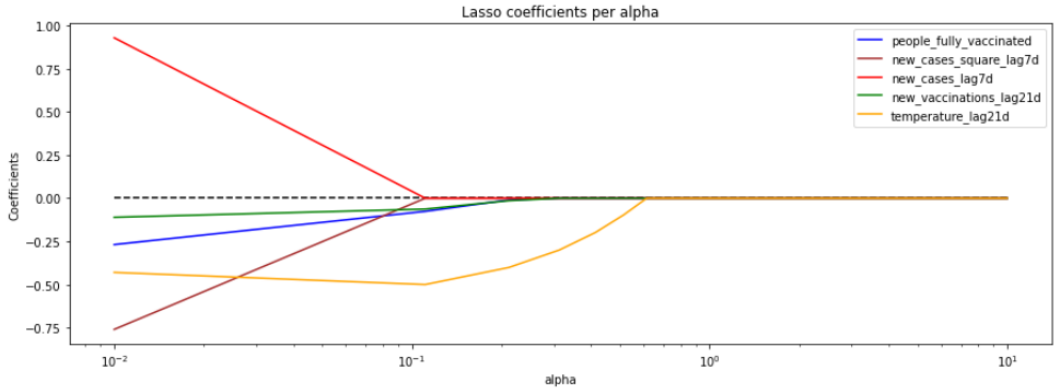
***Anexo 9:*** *Modelo 1 – Hyperparameter Optimization: MLP*

| Predictors | Best Values |
|---|---|
| max_depth | 6 |
| max_features | 2 |
| min_samples_leaf | 5 |
| min_samples_split | 7 |
| n_estimators | 776 |

***Anexo 10:*** *Modelo 1 – Hyperparameter Optimization: Random Forest*




***Anexo 11:*** *Convergência dos coeficientes de Ridge com o aumento do parâmetro* ***k*** *– Modelo 1*

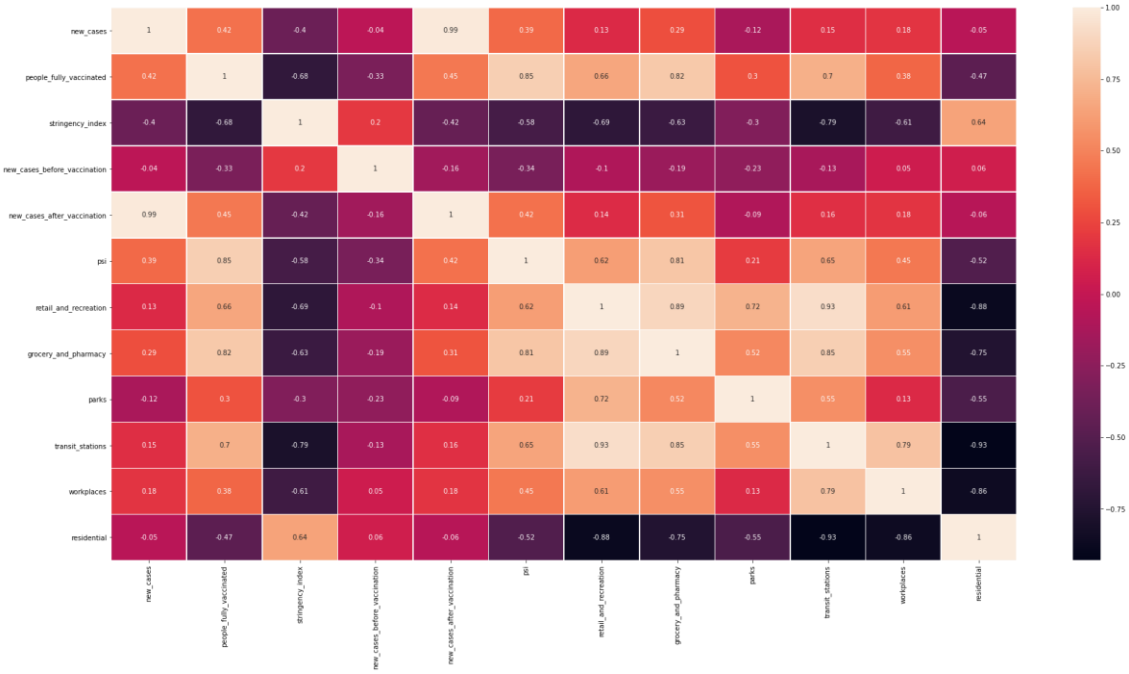



***Anexo 12:*** *Convergência dos coeficientes de LASSO com o aumento do parâmetro* ***k*** *– Modelo 1*

**VIF (Variance Inflation Factors):**

| | feature | VIF |
|---|---|---|
| 0 | const | 1.000000 |
| 1 | retail_and_recreation | 27.238731 |
| 2 | grocery_and_pharmacy | 11.677101 |
| 3 | parks | 5.965109 |
| 4 | transit_stations | 50.270912 |
| 5 | workplaces | 14.417332 |
| 6 | residential | 36.323946 |
| 7 | people_fully_vaccinated | 8.171560 |
| 8 | stringency_index | 5.268427 |
| 9 | new_cases_before_vaccination | 1.558996 |
| 10 | new_cases_after_vaccination | 2.200914 |

***Anexo 13:*** *VIF's Modelo 2*




***Anexo 14:*** Matriz de correlações do Modelo 2

**Optimal Alpha (Ridge):** 0.01

***Anexo 15:*** *Modelo 2 – Hyperparameter Optimization: Ridge*

**Optimal Alpha (Lasso):** 0.0018092989327430178

***Anexo 16:*** *Modelo 2 – Hyperparameter Optimization: LASSO*

| Predictors | Best Values |
|---|---|
| learning_rate | 0.015659 |
| max_depth | 5.000000 |
| max_features | 2.000000 |
| min_samples_leaf | 14.000000 |
| min_samples_split | 6.000000 |
| n_estimators | 714.000000 |
| subsample | 0.674415 |

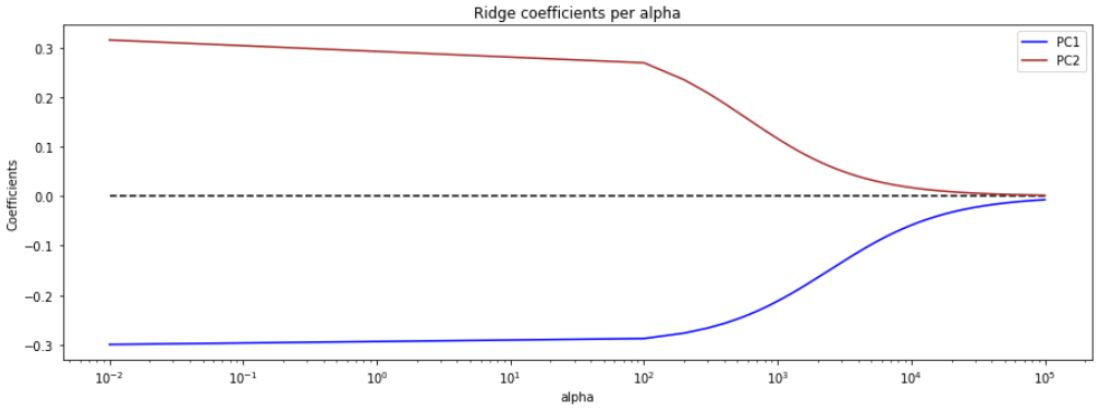
***Anexo 17:*** *Modelo 2 – Hyperparameter Optimization: Gradient Boosting*

| Predictors | Best Values |
|---|---|
| activation | relu |
| alpha | 0.017141 |
| hidden_layer_sizes | 91 |
| solver | lbfgs |

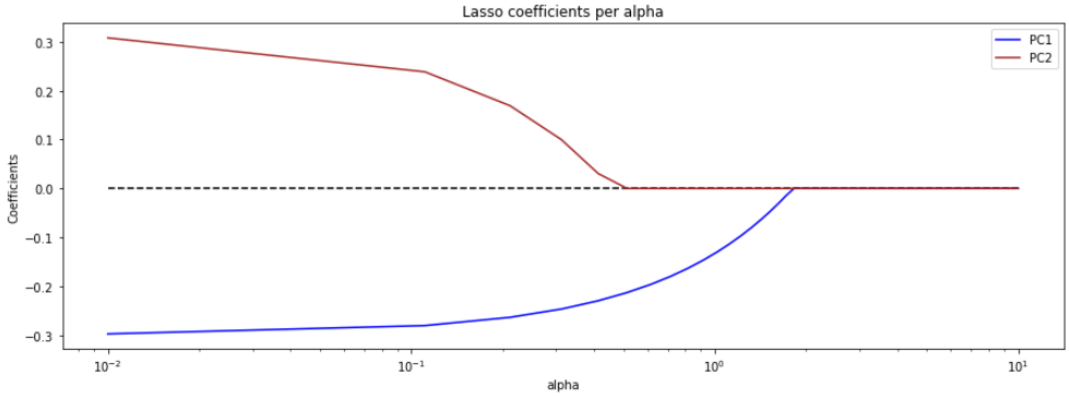
***Anexo 18:*** *Modelo 2 – Hyperparameter Optimization: MLP*

| Predictors | Best Values |
|---|---|
| max_depth | 6 |
| max_features | 2 |
| min_samples_leaf | 5 |
| min_samples_split | 6 |
| n_estimators | 284 |

***Anexo 19:*** *Modelo 2 – Hyperparameter Optimization: Random Forest*




***Anexo 20:*** *Convergência dos coeficientes de Ridge com o aumento do parâmetro* ***k*** *– Modelo 2*




***Anexo 21:*** *Convergência dos coeficientes de LASSO com o aumento do parâmetro* ***k*** *– Modelo 2*

O teste *t* é um teste individual dos parâmetros de um modelo econométrico. (Ferreira, 2010) A variante do teste utilizada neste TFM apresenta as seguintes hipóteses:

$$H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \tag{15}$$

E a estatística de teste e o *p-value*[9] do teste podem ser representados, respetivamente, por:

$$t = \frac{\hat{\beta}_j - \theta}{s.e(\hat{\beta}_j)} \sim t_{n-p} \tag{16}$$

$$p - value(t) = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \sim t_{n-p} \tag{17}$$

Através destas duas medidas referidas anteriormente podemos aceitar ou rejeitar a hipótese nula, de que o coeficiente da variável não é estatisticamente significativo para um determinado nível de significância (o valor mais utilizado é o de 5%) (Ferreira, 2010). Se o *p-value* for maior que o nível de significância a hipótese nula não é rejeitada logo, considera-se que a variável não é estatisticamente significativa para o modelo, para esse nível de significância. Se o contrário acontecer, ou seja, se o *p-value* for menor ou igual ao nível de significância, considera-se que a variável é estatisticamente significativa para o modelo.(Sampi e Jooste, 2020)

***Anexo 22:*** *Teste t – Significância Individual das Variáveis*

[9] ***p-value:*** *probabilidade de observar o valor da estatística de teste, ou superior, sob a hipótese nula.* ***Fonte:*** *(Sampi e Jooste, 2020)*