Shape

Description automatically generated with low confidencemestrado

métodos quantitativos para a decisão económica e empresarial

trabalho final de mestrado

dissertação elaborada 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

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 o modelo preditivo mais adequado. 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 que 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*; mortes; PSI; casos; vacinação; temperatura; mobilidade; *machine learning*; Portugal; *python*

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; cases; vaccination; temperature; mobility; machine learning; Portugal; python

Índice

[Agradecimentos I](#_Toc114960848)

[Resumo II](#_Toc114960849)

[Abstract III](#_Toc114960850)

[Índice IV](#_Toc114960851)

[Índice de Figuras V](#_Toc114960852)

[Índice de Tabelas V](#_Toc114960853)

[Lista de Siglas e Abreviaturas V](#_Toc114960854)

[1. Introdução 1](#_Toc114960855)

[2. Revisão da Literatura 3](#_Toc114960856)

[2.1. Métodos de Previsão – Aplicação na *COVID-19* 3](#_Toc114960857)

[2.2. Tipos de aprendizagem de *Machine Learning* 3](#_Toc114960858)

[2.3. Algoritmos de *Machine Learning* 4](#_Toc114960859)

[2.3.1. Ordinary Least Squares (OLS) 4](#_Toc114960860)

[2.3.2. Ridge 5](#_Toc114960861)

[2.3.3. Least Absolute Shrinkage and Selection Operator (LASSO) 5](#_Toc114960862)

[2.3.4. Gradient Boosting (GB) 6](#_Toc114960863)

[2.3.5. Multilayer Perception (MLP) 7](#_Toc114960864)

[2.3.6. Random Forest (RF) 8](#_Toc114960865)

[2.3.7. *Principal Component Analysis (PCA)* 9](#_Toc114960866)

[2.3.8. *Cross Validation - K Fold* 10](#_Toc114960867)

[2.4. Métricas de Precisão 10](#_Toc114960868)

[2.4.1. Mean Absolute Error (MAE) 11](#_Toc114960869)

[2.4.2. *Mean Squared Error (MSE)* 11](#_Toc114960870)

[2.4.3. *Median Absolute Error (MdAE)* 11](#_Toc114960871)

[2.4.4. *R2 Score* 12](#_Toc114960872)

[2.4.5. *Explained Variance Score* 12](#_Toc114960873)

[3. Metodologia 13](#_Toc114960874)

[3.1. Modelo 1 – Previsão do número de mortes associadas à *COVID-19* 13](#_Toc114960875)

[4. Resultados 17](#_Toc114960876)

[5. Conclusão e Trabalhos Futuros 20](#_Toc114960877)

[Referências Bibliográficas 21](#_Toc114960878)

Índice de Figuras

Índice de Tabelas

Lista de Siglas e Abreviaturas

*COVID-19* - *Coronavirus Disease 2019*

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

GB – *Gradient Boosting*

LASSO - *Least Absolute Shrinkage and Selection Operator*

MLP - *Multi-Layer Perceptron*

OLS - Ordinary Least Squares

PSI – *Portuguese Stock Index*

RF – *Random Forest*

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

TFM – Trabalho Final de Mestrado

1. Introdução

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. [1] 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). [2] A pandemia mais impactante deste século é a pandemia de *COVID-19*. *COVID-19* é uma doença respiratória causada pelo vírus *SARS-CoV-2*, [1] que afeta todas as faixas etárias, mas tem consequências mais graves em indivíduos mais velhos e/ou pessoas com condições médicas pré-existentes. [3] Os primeiros casos registados datam de 31 de dezembro de 2019 na cidade de Wuhan, China. [4] Esta doença espalhou-se rapidamente por todo o mundo, em Portugal, o primeiro caso foi registado a 2 de março de 2020. [5] 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. [2]

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 [39]* quatro vezes em dez dias, algo que em toda a história apenas tinha sido ativado uma vez em 1997 [38]. 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 [38]. 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]](#footnote-1)*, teve uma descida de aproximadamente 15% desde o início da pandemia (considera-se a data registada da primeira infeção, 2 de março de 2020) até ao fim desse mesmo mês, segundo dados de [40].

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 [41]. 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. [42] Isso levou a uma diminuição das idas ao escritório e por consequência da utilização dos transportes públicos também.

O estudo do impacto demográfico e financeiro que a pandemia teve tem sido bastantes frequente ao longo destes últimos dois anos, por exemplo [2] e [43].O objetivo deste TFM é encontrar dois modelos adequado 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 melhor poder preditivo, de acordo com as métricas escolhidas para esse efeito.

1. Revisão da Literatura
   1. Métodos de Previsão – Aplicação na *COVID-19*

Os dados da mortalidade por *COVID-19* podem ser previstos por vários métodos de previsão, como algoritmos de *machine learning* ou *statistical forecast*. [1] Além dos algoritmos de *machine learing*, diversos estudos utilizaram os modelos *ARIMA* e *SARIMA*, considerando o comportamento sazonal presente na série da mortalidade. [7] e [54]

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 [17], ou da Regressão Linear e Polinomial [2]. Também o *Gradient Boosting* já foi utilizado por [55] para modelar e prever os dados da mortalidade resultante da *COVID-19*, provando ser um algoritmo bastante eficiente.

Neste TFM, apenas foram usados algoritmos de *machine learning* ​​na modelagem e previsão dos dados de mortes por consequência da COVID-19.

* 1. Métodos de Previsão – Aplicação nos *Stock Index’s*

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 [49], ou de algoritmos de *ensemble* como o *Gradient Boosting* [50] ou o *Random Forest* [51], 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* [52]. Vários artigos usam modelos híbridos entre os ARIMA e vários modelos de *machine learning,* como por exemplo [53] devido à eficácia de ambos em diferentes situações.

Neste TFM apenas foram abordados algoritmos de *machine learning* para a previsão dos dados do *PSI*.

* 1. *Machine Learning – Introdução*

De acordo com o artigo [6], “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, [8] 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. [8]

* 1. Tipos de aprendizagem de *Machine Learning*

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. [9] 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. [10] 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”*, segundo [11]. 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”, segundo [11]. Um exemplo deste tipo de algoritmos é o *K-means*.

* 1. Algoritmos de *Machine Learning*

Diagram

Description automatically generatedEste 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

* + 1. Ordinary Least Squares (OLS)

*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).[12] O modelo estrutural da Regressão Linear pode ser representado por:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

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

* + 1. Ridge

A regressão de *Ridge* é um algoritmo usado quando se enfrenta problemas de multicolinearidade[[2]](#footnote-2) entre as variáveis ​​preditivas do modelo. [12] Este tipo de regressão linear é bastante similar com o *OLS*, com a diferença que na estimação dos *β*, é adicionado um termo aos elementos diagonais da matriz de correlação. [44] 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. [44]

A picture containing diagram

Description automatically generatedFigura 2: Comparação da estimação dos coeficientes (OLS v.s Ridge)

Fonte: Elaboração Própria

Através da fórmula da estimação do parâmetro de *Ridge* da Figura 2, pode-se observar que se o valor de for igual a zero, então o estimador vai ser igual ao estimador do *OLS*.

* + 1. Least Absolute Shrinkage and Selection Operator (LASSO)

*Least Absolute Shrinkage and Selection Operator* ou *LASSO* é um algoritmo que melhora 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.[12] O estimador de *LASSO* pode ser estimado através do seguinte problema de minimização:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Ao contrário da regularização feita na regressão de *Ridge*, a de *LASSO* gera coeficientes exatamente iguais a zero, quando o parâmetro é demasiado grande, melhorando o modelo em termos de interpretação. [45]

* + 1. Gradient Boosting (GB)

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. [14] Este método tem sido muito utilizado em variados estudos sobre a pandemia de *COVID-19*. [13] 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. [13] O modelo aditivo do GB pode ser representado da seguinte forma:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

onde representa o modelo anterior, e é a taxa de aprendizagem usada para diminuir os erros da previsão. [14] é um multiplicador que pode ser representado da seguinte forma:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Diagram, schematic

Description automatically generatedonde é a classificação da classe de destino. [14]

Figura 3: Arquitetura do algoritmo Gradient Boosting

Fonte: [46]

* + 1. Multilayer Perception (MLP)

*Multilayer Perception* (*MLP*) é um método de *machine learning* que usa redes neuronais artificiais. Como refere o artigo [15], *“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. [15] 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*. [47]. 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. [47]

Diagram

Description automatically generatedFigura 4: Arquitetura do algoritmo MLP

Fonte: [47]

* + 1. Random Forest (RF)

Diagram

Description automatically generated*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. [17] 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. [16]

Figura 5: Arquitetura do algoritmo Random Forest

Fonte: [48]

* + 1. *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 [30]:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

em que representa a matriz de covariâncias da matriz dos dados , centrada através das colunas, e a condição é indicativa que o vetor tem norma unitária. A solução deste problema de maximização será o maior valor próprio (*eigenvalue*) λ da matriz **,** o que significa que o primeiro componente principal é o vetor próprio (*eigenvector*) de norma unitária associado ao maior valor próprio , o segundo componente principal será o vetor próprio associado ao segundo maior valor próprio 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 [31]:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

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. [32]

* + 1. *Cross Validation - K Fold*

Diagram

Description automatically generatedO método de validação *K-Fold* é um método de *Cross Validation* que produz uma divisão do *dataset* em vários subconjuntos 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 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. [33]

Figura : Processo K-Fold Cross Validation

Fonte: [34]

* 1. 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 (MdAE), Explained Variance Score* e *R2 Score.*

* + 1. Mean Absolute Error (MAE)

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. [35] Pode ser representado pela seguinte fórmula:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

em que representa o número de observações e os resíduos da observação .

* + 1. *Mean Squared Error (MSE)*

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. [35] Esta medida também pertence ao grupo das medidas de erro absoluto e pode ser representada da seguinte forma:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

* + 1. *Median Absolute Error (MdAE)*

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. [35] A sua representação pode ser feita através de:

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

* + 1. *R2 Score*

O R2 ou Coeficiente de Determinação é uma das medidas de precisão mais utilizadas em *estatística*. [36] 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 R2 representa a proporção da variância explicada na variável dependente que é previsível a partir das variáveis independentes [37], podendo estar representado na seguinte forma:

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

onde representa as observações da variável dependente e as observações da variável independente. Esta métrica é escalada entre . [37]

* + 1. *Explained Variance Score*

1. Metodologia

FALAR SOBRE O CRISP DM

* 1. Modelo 1 – Previsão do número de mortes associadas à *COVID-19*

Chart, histogram

Description automatically generatedPara 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* [18]. 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. [28] As figuras seguintes mostram os gráficos de todas as variáveis ​​de 2 de março de 2020 a 28 de fevereiro de 2022:

Chart, histogram

Description automatically generatedFigura 2: Número diário de mortes relacionadas com o vírus SARS-CoV-2 em Portugal

Figura 3: Número de infeções e número de infeções ao quadrado em Portugal

Chart, line chart, scatter chart

Description automatically generatedChart, line chart, histogram

Description automatically generatedFigura 4: Total de pessoas totalmente vacinadas e número diário de vacinações em Portugal

Figura 5: Temperatura média diária em Portugal

O processo descrito abaixo foi executado de acordo com o método *CRISP-DM*. [27] Foram escolhidos dados da vacinação 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* [19], 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 aumentar bastante, o número de mortes não seguiu o mesmo comportamento, em termos de quantidade, como antes do processo de vacinação, como se pode observar nas Figuras 2 e 3 e por último foi utilizada a temperatura média diária devido ao padrão sazonal presente nos dados como é possível observar na Figura 6.

Figura 6: Decomposição Sazonal da séria das Temperaturas média diárias em Portugal

Graphical user interface

Description automatically generatedForam removidos todos os fins de semana (Sábado e Domingo) dos dados tendo em conta a inconsistência dos dados nesses mesmos períodos. Por fim, para que todas as variáveis ​​estejam na mesma escala e para 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*. [22]

FALAR SOBRE O VIF ENTRE O NÚMERO DE CASOS E O NÚMERO DE CASOS AO QUADRADO. Por fim, o *p-value* do teste *t* das variáveis [24] foi observado e verificou-se que não existiam variáveis não significantes, ficando o modelo final com as seguintes variáveis: *new\_deaths*, *people\_fully\_vaccinated*, *new\_vaccinations\_lag21d*, *new\_cases\_square\_lag21d*, *new\_cases\_lag21d*, *temperature\_lag21d*.

O próximo passo foi dividir os dados em amostras de treino e teste, de forma aleatória, sendo que amostra de treino corresponde a 80% da amostra e a amostra de teste a 20% da amostra. Foi feita também uma hiper parametrização dos algoritmos com o objetivo de estimar os parâmetros ótimos para cada um. A parametrização de *Ridge*, *LASSO*, *Gradient Boosting*, *MLP* e *Random Forest* foi realizada, inserindo dados aleatórios para os parâmetros dos algoritmos e realizando um grande número de iterações, através de vários métodos de *cross-validation*, [22] até atingir a convergência. Após terem sido estimados os dados da previsão para a fase de teste, realizou-se um teste de *Durbin Watson* [25] 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. [26] Por último, foi avaliada a eficiência da previsão, através de uma comparação de algumas medidas de validação como o Erro Absoluto Médio (EAM), Erro Quadrado Médio (EQM), Erro Absoluto Mediano, Variância Explicada (VE) e o R2 para todos os algoritmos. [22]

1. Resultados

Como acima mencionado, o primeiro passo foi a estimativa do modelo de um modelo de regressão linear (*OLS*) conforme mostra a Figura 6.

Table

Description automatically generatedFigura 6: Modelo OLS

A partir da saída da Figura 6, o número diário de casos de *COVID-19* tem um impacto positivo, tanto antes quanto após o processo de vacinação, com maior impacto antes da vacinação (coeficiente mais alto no modelo). Todas as variáveis ​​de vacinação têm impacto positivo, exceto a defasagem de um mês do número diário de vacinas administradas, resultado que vai contra o que seria esperado. Por fim, pode-se inferir que as temperaturas têm um impacto negativo no número de óbitos, o que está de acordo com os dados que podemos observar nos gráficos de ambas as variáveis. Para a estimação dos demais modelos, utilizou-se uma hiperparametrização dos modelos. Passando agora para a identificação do modelo com melhor poder preditivo, podemos observar a tabela abaixo com as informações referentes a cada modelo. (ARRANJAR TODO O TEXTO)

Tabela 1: Medidas de precisão dos algoritmos

Observando a Tabela 1, podemos inferir que o Gradient Boosting foi o melhor algoritmo preditivo, obtendo os melhores escores em todas as medidas. Random Forest e MLP também obtiveram bons resultados, sendo RF superior a MLP em todas as medidas de pontuação. Isso indica que esses três algoritmos podem ser candidatos para fazer uma boa previsão futura dos dados diários de mortalidade por *COVID-19*. (ARRANJAR TODO O TEXTO)

Chart, bar chart

Description automatically generatedFigura 7: Gradient Boosting – Importância dos Preditores

Na Figura 7 podemos observar a importância da importância de cada um dos preditores, dada pelo algoritmo Gradient Boosting. As pessoas vacinadas e a temperatura são as variáveis ​​mais importantes para a previsão de óbitos por *COVID-19*, ao contrário do que aconteceu no OLS, em que o número de casos diários foi a variável com maior coeficiente. Curiosamente, o GB atribui menos peso ao número de casos antes da vacinação em comparação com o número de casos após a vacinação. Uma curiosidade relevante também é que as variáveis ​​que não foram significativas no MQO são as duas com menor importância no GB. 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 por SARS-CoV2. (ARRANJAR TODO O TEXTO)

Por fim, podemos observar na Tabela 2 abaixo os resultados do teste de Durbin-Watson e a média dos resíduos, para testar sua qualidade. (ARRANJAR TODO O TEXTO)

Tabela 2: Resultados do teste Durbin-Watson e média dos resíduos

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, sendo os piores resultados nos três primeiros. [25] Podemos dizer que todos os modelos capturam adequadamente as informações presentes nos dados. [26] (ARRANJAR TODO O TEXTO)

1. Conclusão e Trabalhos Futuros

O objetivo deste trabalho foi inferir sobre o impacto da vacinação, temperatura e número de casos na mortalidade por *SARS-CoV-2* em Portugal. Vários dados de vacinação e as defasagens desses dados foram usados, como uma "divisão" do número de casos diários registados antes e depois da vacinação e a temperatura média diária. O modelo inicial começou a ser construído pelo método OLS, e depois replicado em outros algoritmos. Houve uma correlação positiva entre a variável dependente e o número de casos, como esperado, mas a diferença no coeficiente antes e depois da vacinação foi muito clara, enquanto quase todos os dados de vacinação presentes no modelo tiveram um coeficiente negativo, como já era esperado, exceto o número diário de vacinações defasadas em um mês. Os resultados em Gradient Boosting, MLP e Random Forest foram satisfatórios, enquanto em OLS, Ridge e LASSO, os valores de ajuste do modelo ficaram abaixo do esperado, o que pode significar que a relação entre os preditores e a variável dependente não é linear. Os objetivos do trabalho foram alcançados, pois foi identificado o algoritmo com maior poder preditivo, que consiste em um algoritmo em conjunto, Gradient Boosting, e ficou comprovado que a vacinação é uma boa medida preventiva contra mortes por *SARS-CoV-2* e a temperatura tem um impacto negativo no número de mortes. (ARRANJAR TODO O TEXTO)

Referências Bibliográficas

1. Almalki, A., Gokaraju, B., Acquaah, Y., Turlapaty, A.: Regression Analysis for *COVID-19* Infections and Deaths Based on Food Access and Health Issues. Healthcare. 10, 324 (2022).
2. Rustagi, V., Bajaj, M., Tanvi, Singh, P., Aggarwal, R., AlAjmi, M.F., Hussain, A., Hassan, Md.I., Singh, A., Singh, I.K.: Analyzing the Effect of Vaccination Over COVID Cases and Deaths in Asian Countries Using Machine Learning Models. Frontiers in Cellular and Infection Microbiology. 11, (2022).
3. Sarirete, A.: A Bibliometric Analysis of *COVID-19* Vaccines and Sentiment Analysis. Procedia Computer Science. 194, 280–287 (2021).
4. Sohrabi, C., Alsafi, Z., O’Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, R.: World Health Organization declares global emergency: A review of the 2019 novel coronavirus (*COVID-19*). International Journal of Surgery. 76, 71–76 (2020).
5. Milhinhos, A., Costa, P.M.: On the Progression of *COVID-19* in Portugal: A Comparative Analysis of Active Cases Using Non-linear Regression. Frontiers in Public Health. 8, (2020).
6. Aparicio, J.T., Romao, M., Costa, C.J.: 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 (2022)
7. Perone, G.: 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, 18 (2022).
8. Mahesh, B.: Machine Learning Algorithms -A Review. (2019).
9. Cord, M., Cunningham, P.: Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Springer Science & Business Media (2008).
10. Zhu, X. (Jerry): Semi-Supervised Learning Literature Survey. University of Wisconsin-Madison Department of Computer Sciences (2005).
11. Mendelson, S., Smola, A.J. eds: Advanced lectures on machine learning: Machine Learning Summer School 2002, Canberra, Australia, February 11-22, 2002: revised lectures. Springer, Berlin; New York (2003).
12. Saleh, H., Layous, J.: Machine Learning -Regression, (2022).
13. Shrivastav, L.K., Jha, S.K.: A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of *COVID-19* in India. Appl Intell. 51, 2727–2739 (2021).
14. Gumaei, A., Al-Rakhami, M., Mahmoud Al Rahhal, M., Raddah H Albogamy, F., Al Maghayreh, E., AlSalman, H.: Prediction of *COVID-19* Confirmed Cases Using Gradient Boosting Regression Method. Computers, Materials & Continua. 66, 315–329 (2020).
15. Borghi, P.H., Zakordonets, O., Teixeira, J.P.: A *COVID-19* time series forecasting model based on MLP ANN. Procedia Computer Science. 181, 940–947 (2021).
16. Yeşilkanat, C.M.: Spatio-temporal estimation of the daily cases of *COVID-19* in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals. 140, 110210 (2020).
17. Gupta, V.K., Gupta, A., Kumar, D., Sardana, A.: Prediction of *COVID-19* confirmed, death, and cured cases in India using random forest model. Big Data Mining and Analytics. 4, 116–123 (2021).
18. *COVID-19* Data Explorer, <https://ourworldindata.org/coronavirus-data-explorer>, last accessed 2022/07/05.
19. 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.: 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, 357–366 (2022).
20. Dyer, O.: *COVID-19*: Moderna and Pfizer vaccines prevent infections as well as symptoms, CDC study finds. BMJ. n888 (2021).
21. Albon, C.: Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning. O’Reilly Media, Inc. (2018).
22. Avila, J., Hauck, T.: scikit-learn Cookbook: Over 80 recipes for machine learning in Python with scikit-learn. Packt Publishing Ltd (2017).
23. Seabold, S., Perktold, J.: Statsmodels: Econometric and Statistical Modeling with Python. Presented at the Python in Science Conference, Austin, Texas (2010).
24. Kim, T.K.: T test as a parametric statistic. Korean J Anesthesiol. 68, 540–546 (2015).
25. Mckinney, W., Perktold, J., Seabold, S.: Time Series Analysis in Python with statsmodels. (2011).
26. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts (2018).
27. Costa, C., Aparício, J.T.: POST-DS: A Methodology to Boost Data Science. (2020).
28. Menne, M.J., Durre, I., Korzeniewski, B., McNeill, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R.S., Gleason, B.E., Houston, T.G.: Global Historical Climatology Network - Daily (GHCN-Daily), Version 3, https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html, (2012).
29. Akossou, A., R., P.: Impact of data structure on the estimators R-square and adjusted R-square in linear regression. International Journal of Mathematics and Computation. 20, 84–93 (2013).
30. 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.
31. How to compute PCA loadings and the loading matrix with scikit-learn. (2020, janeiro 27). Simone Centellegher, PhD - Data Scientist and Researcher. PROCURAR REFERÊNCIA MELHOR
32. Graham, M. H. (2003). Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, *84*(11), 2809–2815.
33. Berrar, D. (2019). Cross-Validation. Em *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). Elsevier.
34. Kubben, P., Dumontier, M., & Dekker, A. (Eds.). (2019). *Fundamentals of Clinical Data Science*. Springer International Publishing.
35. Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. (2013). *A Survey of Forecast Error Measures*. 7.
36. 36Redell, N. (2019). Shapley Decomposition of R-Squared in Machine Learning Models. 9.
37. 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.
38. *Financial markets under the global pandemic of COVID-19 | Elsevier Enhanced Reader*. (2020).
39. Chen, H., Petukhov, A., & Wang, J. (2018). *The Dark Side of Circuit Breakers*. 57.
40. *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](https://live.euronext.com/pt/product/indices/PTING0200002-XLIS)
41. 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.
42. 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.
43. The relationship between mobility and COVID-19 pandemic: Daily evidence from an emerging country by causality analysis | Elsevier Enhanced Reader. (2021).
44. *Ridge Regression*. (sem data). 20.
45. Melkumova, L. E., & Shatskikh, S. Ya. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746–755.
46. 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.
47. Park, Y.-S., & Lek, S. (2016). Artificial Neural Networks. Em *Developments in Environmental Modelling* (Vol. 28, pp. 123–140). Elsevier.
48. Ornella, L., Kruseman, G., & Crossa, J. (2020). Satellite Data and Supervised Learning to Prevent Impact of Drought on Crop Production: Meteorological Drought. Em G. Ondrasek (Ed.), *Drought—Detection and Solutions*. IntechOpen.
49. 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.
50. 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.
51. Research Scholar, Dept of CSE, University college of Engineering, JNTU Kakinada, East Godavari, AP, INDIA..., Polamuri\*, S. R., Srinivasi, Dr. K., Professor, Dept Of CSE, V R Siddhartha Engineering College, Vijayawada, INDIA, Mohan, Dr. A. K., & Professor, Dept Of CSE, University college of Engineering, JNTU Kakinada, East Godavari, AP., INDIA. (2019). Stock Market Prices Prediction using Random Forest and Extra Tree Regression. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(3), 1224–1228.
52. 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.
53. 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.
54. 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.
55. 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.

1. ***free float*** *– A capitalização é calculada multiplicando po preço das ações pelo número de ações disponíveis no mercado.* ***Fonte:*** [Free-Float Methodology Definition (investopedia.com)](https://www.investopedia.com/terms/f/freefloatmethodology.asp) [↑](#footnote-ref-1)
2. ***multicolineariedade:*** *Condição que sucede quando uma ou mais variáveis ​​independentes do modelo podem prever outra variável independente de uma forma eficiente e linear.* ***Fonte: [44]*** [↑](#footnote-ref-2)