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](#_Toc114443692)

[Resumo II](#_Toc114443693)

[Abstract III](#_Toc114443694)

[Índice IV](#_Toc114443695)

[Índice de Figuras V](#_Toc114443696)

[Índice de Tabelas V](#_Toc114443697)

[Lista de Siglas e Abreviaturas V](#_Toc114443698)

[1. Introdução 1](#_Toc114443699)

[2. Background 2](#_Toc114443700)

[3. Métodos 6](#_Toc114443701)

[4. Resultados 9](#_Toc114443702)

[5. Conclusão e Trabalhos Futuros 11](#_Toc114443703)

[Referências Bibliográficas 12](#_Toc114443704)

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 2*

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]

O estudo do impacto da vacinação, número de casos registados e temperaturas no número de mortes causadas pelo vírus SARS-CoV-2 tem sido bastante frequente nos últimos tempos. O objetivo deste trabalho é encontrar o modelo adequado para estimar o número de mortes diárias causadas pelo vírus SARS-CoV-2 e posteriormente encontrar o algoritmo com melhor poder preditivo, de acordo com as métricas escolhidas para esse efeito. O principal objetivo é usar vários algoritmos de aprendizado de máquina para prever a mortalidade diária.

1. Background

Os dados de mortalidade por COVID-19 podem ser previstos por vários métodos, 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] Neste artigo, apenas foram usados algoritmos de *machine learning* ​​na modelagem e previsão dos dados. De acordo com o artigo [6], “The premise of machine learning is that a computer program can learn and adapt to new data without the need for human intervention”. 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]

Existem vários tipos de algoritmos de *machine learning*, como os não supervisionados, supervisionados, semi-supervisionados e aprendizagem por reforço. Os do tipo supervisionado realizam um mapeamento das variáveis ​​dependentes e independentes, para prever dados futuros desconhecidos da variável dependente. [9] Os semi-supervisionados usam dados não rotulados (não precisam de intervenção humana) conjuntamente com dados rotulados (precisam de intervenção humana) para prever dados futuros. Esses tipos de algoritmos podem ser mais eficientes, pois precisam de muito menos intervenção humana na construção dos modelos. [10] Algoritmos de aprendizagem por reforço produzem uma série de ações considerando o ambiente onde estão inseridos para maximizar “the future rewards it receives (or minimizes the punishments) over its lifetime”. [11] Por último, mas não menos importantes, em algoritmos de aprendizagem não supervisionada, os dados de entrada são inseridos “but obtains neither supervised target outputs, nor rewards from its environment”. [11] Um exemplo deste tipo de algoritmo é o *K-means*.

Diagram

Description automatically generatedNeste artigo, foram discutidos apenas algoritmos de aprendizagem supervisionada, que são o *OLS*, *LASSO*, *Ridge*, *Gradient Boosting*, *MLP* e *Random Forest* (Figura 1).

Figura 1: Algoritmos de machine learning (tipo supervisionado)

*OLS* ou Regressão Linear é um dos algoritmos de *machine learning* mais simples de compreender. A Regressão Linear pode ser simples (quando apenas uma variável independente é usada no modelo) ou múltipla (quando duas ou mais variáveis ​​são usadas para prever a variável dependente).[12] O modelo estrutural da Regressão Linear é:

(1)

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

A regressão de *Ridge* é um algoritmo usado quando se enfrenta problemas de multicolinearidade entre as variáveis ​​preditoras do modelo. [12] A multicolinearidade é uma condição que sucede quando uma ou mais variáveis ​​independentes do modelo podem prever outra variável independente de uma forma eficiente.

LASSO é um algoritmo que melhora a precisão do modelo por meio de 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 *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 estudos sobre a pandemia de COVID-19. [13] 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:

(2)

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:

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

onde é o rótulo da classe de destino. [14] Os algoritmos de *machine learning* referidos acima já foram utilizados em diversos artigos que abordam o tema central deste artigo, como [13] and [14].

*Multilayer Perception* (MLP) é um método de *machine learning* que usa redes neuronais artificiais. Como refere o artigo [15], “The experience of the network is stored by the synaptic weights between neurons and its performance is evaluated, for example, by the ability to generalize behaviors, recognize patterns, fix errors or execute predictions”. Este algoritmo associa vários neurónios, formando redes neurais que irão realizar diversas funções para melhorar a previsão. [15] O MLP pode ser usado através de aprendizagem supervisionada ou não supervisionada, nesta dissertação apenas será usado através do primeiro tipo de aprendizagem.

*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]

A escolha destes seis algoritmos foi feita para que fosse possível entender como os algoritmos lineares, de *ensemble* e de redes neuronais se comportam na previsão dos dados referenciados, medindo a precisão de cada um deles, através de diversas medidas de precisão. (ESCREVER SOBRE AS MEDIDAS DE PRECISÃO)

1. Metodologia

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:

Figura 2: Número diário de mortes relacionadas com o vírus SARS-CoV-2 em Portugal

Chart, histogram

Description automatically generatedFigura 3: Número de infeções e número de infeções ao quadrado em Portugal

Chart, line chart, histogram

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

Chart, line chart, scatter chart

Description automatically generatedFigura 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 (METER FIGURA DA DECOMPOSIÇÃO SAZONAL).

FALAR DO PRÉ PROCESSAMENTO DAS VARIÁVEIS.

Foram 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. FALAR SOBRE A NÃO OBTENÇÃO DE CONVERGÊNCIA EM ALGUNS DOS ALGORITMOS 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

Table

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

Figura 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

Chart, bar chart

Description automatically generatedObservando 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)

Figura 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).