A Machine Learning approach for shared bicycle demand forecasting

Margarida Mergulhao
ISEG (Lisbon School of Economics & Management),
Universidade de Lisboa,
Lisbon, Portugal
mmendes.margarida@outlook.com

Myke Palma
ISEG (Lisbon School of Economics & Management),
Universidade de Lisboa,
Lisbon, Portugal
palmamyke@gmail.com

Carlos J. Costa
Advance/ISEG (Lisbon School of Economics & Management),
Universidade de Lisboa,
Lisbon, Portugal
cjcosta@iseg.ulisboa.pt

Abstract — - **More than 9 million bicycles are shared worldwide through more than 3.000 Bicycle Shared Systems (BSS). Investigating possible behaviours related to the demand for these services will optimize their success. The purpose of this research is to identify the impact of weather conditions, covid and pollution on the usage of BSS. Different machine learning algorithms are studied and used to analyze the different variables. Results were consistent with the literature and theory. In what concerns the algorithms, random forest and multi-layer perceptron regressor performed better, showing a better prediction power.**

Keywords – sustainability; data science; machine learning; bicycle shared usage.

I. Introduction

Between conventions on climate change, protocols and agreements, cities are being pressured to change their ways and habits into sustainable ways to live. Around the globe, there is a special struggle to improve sustainable indicators, due to the increase in pollution emissions, road traffic and high and inefficient energy consumption [1,2]. Cycling in Portugal occupies a lower position when compared to other means of transportation, being also one of the countries in Europe with the highest motor vehicle ownership [3], concerning environmental entities and political parties fighting for a better engagement towards it. Over the years, there were multiple and successful attempts to implement Bicycle Shared Systems (BSS) in some of the main Lisbon areas. We can describe these systems as a spread agglomerate of customized bicycles with automated docking stations secured by user payment card details [4]. They are being used worldwide and count more than 9 million shared bicycles, available in more than 900 cities (https://bikesharingworldmap.com). Implementing BSS in greatly populated cities will have exponential impacts on health, economics, environment, local businesses, transport mode substitution, and travel behaviours [6], as well as in time spent while travelling. Normally, these bike trips are short, taking between 15 to 35 minutes, providing easy mobility on short distances [7]. But several conditions will impact bicycle usage, like weather conditions and air quality [5]. In the last years, pandemic situations resulting from Covid 19 also contribute to behavioural changes related to transportation usage [8]. In this context, it is especially important to understand the impact weather conditions, covid and pollution have on the use and demand of these bicycles. This objective may be decomposed into the following objectives: Identification of the impact of weather conditions, covid, and pollution on bicycle usage and create a model to predict bike usage. To execute this, data was collected, and a model was created. The first model will allow an understanding of the impact of each variable. Then, several algorithms are used to improve the predictive power of the initial model.

The following section includes the background and related work used in this study. Following the previous, methods and results will be described and finally completed with an overall discussion and main conclusions.

II. Literature review

To understand the factors that impact, the demand for BSS, major studies were conducted. It was possible to verify, between numerous related investigations and experiments, the diversity of data and models used for forecasting. A neural network-based machine learning method was created in Shenzhen, China [11] to study the behaviour of bicycle usage in different time and space dimensions and weather conditions, demonstrating the importance of each external component. A similar study was conducted in Jiangsu province, which proved that deep learning approaches, compared to previous traditional models, can better capture nonlinear relationships between data [12]. El-Assi, Salah Mahmoud and Nurul Habib [13] in a related study consider the temperature as a better indicator for the overall feel of the bike user and therefore for the BSS demand forecasting, being used as a factor in multiple other analyses that defend the growth of the daily total number of bicycle rentals when an equal rise on the temperature values is registered [14]. Considering factors like precipitation is proving to be effective to consider when predicting users' demand, due to the strong negative correlation with daily bicycle usage [15]. As was referred to previously, there were changes in the pattern of transportation [8]. Research also shows that there was a growth in the number of cycling trips in 2020 – COVID-19 period, significantly impacting all features of daily travel for all transportation modes [16] [17]. There are some suggestions that the quality of air may impact bicycle usage [5]. It makes sense that its relationship is addressed, and it is done in this present study.

2022 17th Iberian Conference on Information Systems and Technologies (CISTI)
22 – 25 June 2022, Madrid, Spain
ISBN: 978-989-33-3436-2

Ordinary Least Squares (OLS) is a regression algorithm that minimizes the error - the residual sum of the squares [19] - between the training dataset and the values predicted by the linear model. Ridge regression uses a sum-of-squares error function [20] to solve multicollinearity and decrease mean square error [21]. Lasso or Lasso regression algorithm is regularly used in parameter estimation and variable selection [22]. To avoid overfitting, it shrunk the coefficients of determination toward zero [30]. Bayesian Ridge regression will analyze the data affected by multicollinearity [22]. This model is like OLS, aiming to minimize a residual sum of squares. SVM (Support Vector Machine) is a supervised learning and non-probabilistic algorithm that separates data sequentially and non-sequentially [24]. After training a specific set of data, this model will assign it to one category or the other [25]. Multiple-layer perceptron is an object that makes numerical predictions. These predictions are defined through layers – an input layer, hidden layer, and output layer [26] -, training data based on perceptrons. The number of layers and perceptrons are selected arbitrarily and it does not affect its prediction. Random forest is a regression analysis algorithm that calculates the result of a prediction using the average value of all decision trees in a specific set [27]. It is a very attractive model for data study, known for its high precision [28].

III. Method

The main goal is to understand the impact weather conditions, covid and pollution have on the use and demand for these bicycles.

The methodological approach used in this study is based on the CRISP-DM [18],[23]. To execute the objectives mentioned above, we collected data and created a model. To estimate the models, Python language and respective libraries – Pandas and scikit-learn- were used.

To evaluate the quality of the algorithms, the following metrics were used. – e.g Mean absolute error, mean deviation absolute error, mean squared error and the coefficient of determination (R^2) – as depicted in Table IV.

A collection of data was gathered that contained records related to the number of bicycle docks, and the number of bicycles stationed in them reported on a minute basis, sourced from a BSS integrated in Lisbon, Portugal named Gira. Additionally, a collection of data was considered, containing records covid cases, pollution levels, temperature and precipitation. Gira was first implemented in 2017 with the mission of disposing of shared bicycles along the main areas of the capital to promote an easy and sustainable way of mobility among the users from the Emel website of Giras (https://www.gira-bicicletasdelisboa.pt/). The temperature (ºC), precipitation (mm.), pollution (PM2.5) and Covid-19 cases were also gathered through distinct platforms, each one chosen individually for the reliable number of daily records that allowed us to conduct a better and more extensive analysis. The whole data contains 366 rows corresponding to the 366 days in study, between 01/01/2020 and 01/01/2021.

TABLE I. Variables

Variables	**Description**	**Unit**	**Source**
NumEmptyDocs	Number of Empty docs	Average per day	https://www.gira-bicicletasdelisboa.pt/
Total Cases	Number of Covid Cases	number	https://ourworldindata.org/explorers/coronavirus-data-explorer?zoomToSelection=true&time=2020-03-02..2020-12-31&facet=none&hideControls=true&Metric=Confirmed+cases&Interval=New+per+day&Relative+to+Population=false&Align+outbreaks=false&country=PRT~2020+Summer+Olympics+athletes+%26+staff
New Cases	Daily number of Covid Cases	number	
PM2.5	Fine particulate matter (PM2.5) air pollutant	µg/m3	https://aqicn.org/data-platform/register
Precipitation	Raining level	mm	https://meteo.tecnico.ulisboa.pt/obs/history/pp/monthly/2020/12
Avg Temp	Daily average temperature	**ºC**	https://www.wunderground.com/history/monthly/pt/lisbon/LPPT/date/2020-12

The first model allows understanding the impact of each variable.

Then, several algorithms were used to improve the precative power of the initial model. To build the appropriate models, this data was split into train and test "sub" data sets, with a 75% and 25% train-test ratio. Afterwards made the use of pipelines to verify each model's mean absolute error, median absolute error, mean square error and squared error. Between a vast range of diverse models, it was necessary to analyze and compare the results between them to pick the ones with better performance and use it when forecasting. Therefore, the models Ordinary least squares (OLS), Ridge, Lasso, Bayesian Ridge, SVM or SVR, multiple-layer perceptron regressor and random forest regression were analyzed. Depending on the results, the best model will be chosen to carry out the process of forecasting.

IV. Results

By connecting to an app, users can visualize bicycles availability and routes (Fig.1,2) through the city and benefit from a fairly priced mobility service. Lisbon, known as the city of the seven hills, is dominated by an irregular topography [9]. Gira also provides a line of both classic and electric bicycles to facilitate the terrain inconveniences.

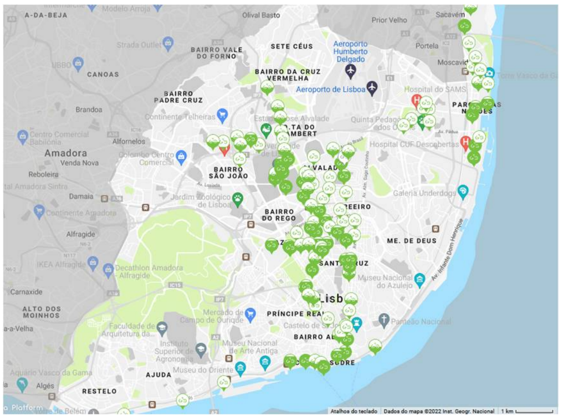
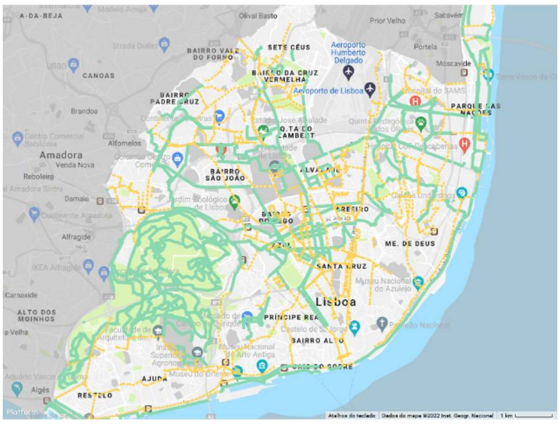


Figure 1. Bikes disposition by type and location, cc. GIRA.



Bicycles stations are also segmented into different zones, between Parque das Nações, Alvalade and Avenidas Novas and Marquês de Pombal and Avenida da Liberdade, located in the centre of the city where transport alternatives are essential due to large amounts of road traffic.

It was possible to visualize the usage of GIRA bicycles in Lisbon when the data was organized by daily records.

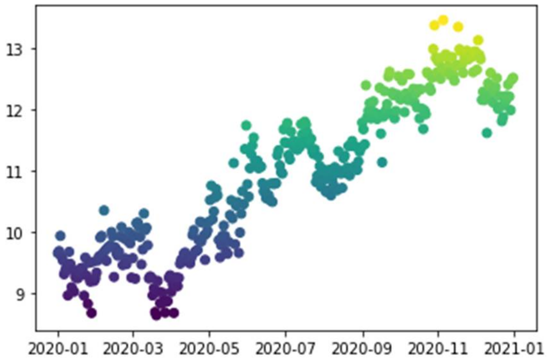


Figure 3. GIRA usage between January of 2020 and January of 2021.

There is a significant growth during the year of 2020. It is also possible to visualize the downfall registered around March of the same year, explained by the first lockdown due to the exponential number of COVID-19 cases in Portugal. Before comparing each model and the respective results, it was necessary to investigate and analyze the significance of the variables, using the values calculated by the OLS model:

TABLE II. REGRESSION USING OLS

	Coef	**Std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	8.2507	0.141	58.709	0.000	7.974	8.527
Total Cases	7.654e-06	5.15e-07	14.854	0.000	6.64e-06	8.67e-06
New Cases	0.0002	3.04e-05	8.141	0.000	0.000	0.000
PM2.5	-0.0054	0.002	-3.207	0.001	-0.009	-0.002
Precipitation (mm.)	-0.0201	0.010	-1.977	0.049	-0.040	-0.000
Avg temp (ºC)	0.1208	0.006	19.272	0.000	0.108	0.133

TABLE III. REGRESSION MODEL USING OLS

Dep. Variable:	NumEmptyDocs	*R-squared:*	0.780
Model:	OLS	*Adj. R-squared:*	0.777
Method:	Least Squares	*F-statistic:*	254.6
Nº Observations:	366	*Prob (F-statistic):*	7.73e-116
Df Residuals:	360	*Log-Likelihood:*	-323.26
Df Model:	5	*AIC:*	658.5
Covariance Type:	nonrobust	*BIC:*	681.9

It was possible to consider all variables relevant (p-value < 0.05).

It was verified, as time passed, an increasing tendency related to the usage of GIRA bicycles. To avoid any bias when developing a model, it was necessary to add 'days' as a new variable. Multiple-layer perceptron regressor and Random Forest regression methods were optimized, calculating the best combinations of each:

- 4 layers were determined with the following perceptron count (8,9,11,11), calculated using an algorithm that iterated through a finite number of ranges of selected layers and perceptrons, and calculated its accuracy, to output the most precise combination for future use.

- 98 trees were calculated using an algorithm that iterated through combinations of the number of trees to output the most precise model.

Analyzing our data correlation, we could verify that we had an acceptable Variance Inflation Factor. Analyzing the performance of several algorithms will help us select the appropriate and the best model for predicting the usage of bicycles considering the external factors also analyzed in this study. The following table exhibits the prediction accuracy of 7 machine learning models through MAE, MDAE, MSE and R2.

TABLE IV. TABLE TYPE STYLES

Model	**MAE**	**MDAE**	**MSE**	**R2**
OLS	0.357	0.310	0.194	0.888
Ridge	0.357	0.310	0.194	0.888
Lasso	0.359	0.314	0.197	0.887
BayesianRidge	0.352	0.312	0.196	0.888
SVM	0.293	0.231	0.140	0.888
MLPRegressor	0.295	0.219	0.146	0.916
Random Forest	0.297	0.222	0.140	0.920

MAE - mean absolute error (lower is better)
MDAE - Median absolute error (lower is better)
MSE - mean squared error (lower is better)
R^2 - squared error (higher is better)

As the table shows, the highest accuracy between all the models corresponds to the Random Forest model, with 0.920 of R2. Followed by a Multi-layer Perceptron regressor, with an R2 of 0.916, MAE of 0.295, MDAE of 0.219 and MSE of 0.146. The remaining agglomerate of models showed similar R2 results between 0.887 and 0.888. Considering that the Random Forest method has such a high value of trees, and to avoid overfitting, it was ideal to use the MLP regressor model as the optimal model. As mentioned, we found an increasing tendency in GIRA usage over the year of 2020, and it had an impact on our predictions. Therefore, the time variable was considered in order to remove this trend growth and verify the existing individual fluctuations of each variable. With this model, it was possible to study the weight of each variable through nonlinear graphs.

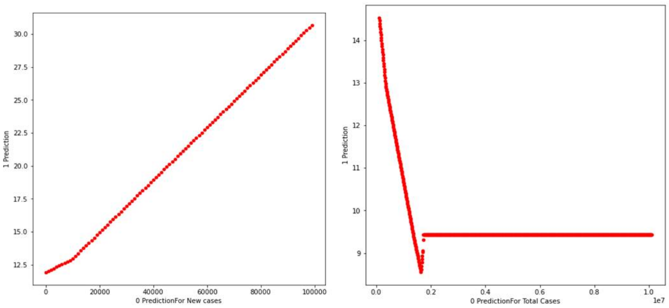


Figure 4. Prediction for new cases and total cases.

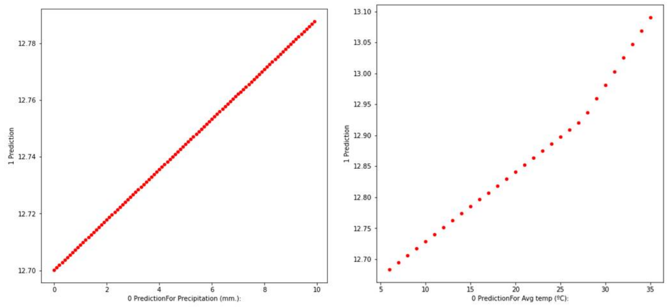


Figure 5. Prediction for precipitation and temperature.

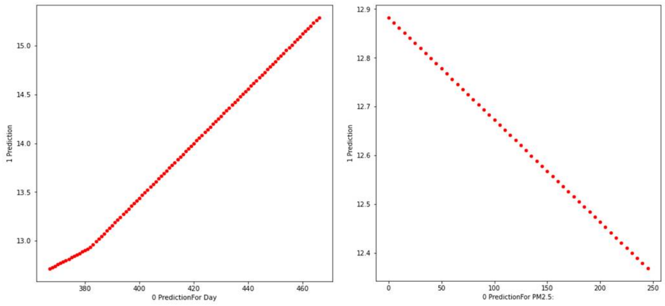


Figure 6. Prediction for day and pollution.

As shown above, there is a positive relationship between the variable's average temperature, days, new cases and precipitation with the bicycle usage, demonstrating an increasing usage guided by the growth of each variable value.

Despite the lack of evidence to support a possible relation between the pollution levels and bicycle usage, we came across a negative slope - bigger pollution levels are followed by smaller bicycle usage.

The last graph provided an outside look of what it is or it was the behaviour of the individuals when confronted with a higher number of COVID-19 cases over time. Despite the clear negative relation from the beginning, it registered a small increase followed by the constant usage of bicycles during the pandemic.

V. DISCUSSION

Studies on BSS usage were conducted all around the world to find a balance between supply and demand that would determine the success of the deployment of BSS stations or bicycles [10]. Similarly, to what has been analyzed here, there is a strong significance in weather-related variables that would fluctuate individuals' behaviour accordingly. Surprisingly, as other authors could visualize in their studies, the pandemic was not a constraint for BSS. On the other hand, it might have been a peculiar solution for a new way of living, replacing the traditional means of transportation with cycling. To better visualize this, it was necessary to choose one algorithm that could help in the forecasting act of these behavioural patterns. The Random Forest algorithm proved to be the most effective and attractive one when studying data [28], followed by a multiple-layer perceptron regressor. This last model can normally perform much better than other complex models, reducing the loss from the prediction [29], therefore our algorithm of choice.

VI. Conclusions

As a result of the OLS model, every variable proved to be necessary to understand the behaviour around shared bicycles usage. Either triggering a positive effect or negative. The multiple-layer perceptron regressor algorithm proved to be the most optimal algorithm to help us understand the behaviour of each variable in its predictions. It is shown that temperature plays a major role to exponentiate the usage when higher values are recorded. The opposite occurs when analyzing precipitation per mm. or pollution indicators; when bigger values are registered the usage decreases significantly. As mentioned in other studies the pandemic was, on the other hand, neither a positive nor negative factor on Bicycle Shared Systems. Over time its usage stabilizes. It is worth reminding of the importance of using sizable amounts of data when forecasting in Machine Learning. It was possible to observe clear patterns in this study, but not entirely in detail. In future studies, analyzing these variables in a wider time interval, with the help of better and more complex algorithms would allow us to visualize fluctuations in our predictions and better understand seasonal changes and the conditions when choosing to use these types of services. It is important to note that directly analyzing bicycle usage and its routes will give us an insight with better accuracy than the actual results reflected by bicycle docks. Bicycle Shared Systems' impact and similar services are still being studied and to what extent the deployment can be enforced in society. Learning about multiple dimensions and other external factors not mentioned in this study and how they condition the BSS usage will facilitate these companies and others with products partially or totally dependent on these dimensions and on the behavioural change in people to allocate or even launch new products.

Acknowledgment

We gratefully acknowledge financial support from FCT - Fundação para a Ciência e a Tecnologia (Portugal), national funding through research grant UIDB/04521/2020.

References

[1] I. Matias, B. Santos, A. Virtudes, "Making Cycling Spaces in Hilly Cities" *KnE Engineering,* 5(5) 2020 pp. 152–165. https://doi.org/10.18502/keg.v5i5.6933

[2] I. Bouzguenda, C. Alalouch, N. Fava, "Towards smart sustainable cities: A review of the role digital citizen participation could play in advancing social sustainability" in Sustainable Cities and Society, Vol. 50, 2019, https://doi.org/10.1016/j.scs.2019.101627.

[3] L. Laker Hilly Lisbon launches electric bike share system in bid to solve congestion, Laura Laker in Lisbon, The Guardian 2017, https://www.theguardian.com/cities/2017/aug/03/hilly-lisbon-portugal-electric-bike-share-congestion

[4] J. Todd, O. Obrien, J. Cheshire, "A global comparison of bicycle sharing systems" in Journal of Transport Geography,, Vol. 94, 2021, https://doi.org/10.1016/j.jtrangeo.2021.103119.

[5] A. Campbell, C. Cherry, M Ryerson, X. Yang. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. Transportation research part C: emerging technologies, 67, 2016, pp. 399-414. https://doi.org/10.1016/j.trc.2016.03.004

[6] M. Ricci, "Bike sharing: A review of evidence on impacts and processes of implementation and operation" in Research in Transportation Business and Management , Vol.15, 2015, Pg. 28-38, ISSN 2210-5395, https://doi.org/10.1016/j.rtbm.2015.03.003.

[7] Deldot Bicycle Program (2016). Wilmington Bike Share: Feasibility Study. Wilmington, DE.

[8] J. T. Aparicio, E. Arsenio, R. Henriques, R. Understanding the Impacts of the COVID-19 Pandemic on Public Transportation Travel Patterns in the City of Lisbon. Sustainability, 13(15), 2021, p.8342. https://doi.org/10.3390/su13158342

[9] United Cities and Local Governments (UCLG) (2016). Committee on Culture, Lisbon CITY PROFILE.

[10] G. Cantelmo, R. Kucharski C. Antoniou. Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data. Transportation Research Record.; 2674(8), 2020, pp. 132-144. doi:10.1177/0361198120932160

[11] Y. Peng, T.Liang, X. Hao., Y. Chen., S. Li., Y. Yi . "CNN-GRU-AM for Shared Bicycles Demand Forecasting" in Computational Intelligence and Neuroscience. 2021. Pp. 1-14. doi: 10.1155/2021/5486328.

[12] X. Chengcheng, J. Junyi, L. Pan, "The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets" in Transportation Research Part C: Emerging Technologies, Volume 95, 2018, pp 47-60, https://doi.org/10.1016/j.trc.2018.07.013.

[13] W. El-Assi, M. Salah Mahmoud, K. Nurul Habib. "Effects of built environment and weather on bike-sharing demand: a station level analysis of commercial bike-sharing in Toronto. Transportation" in Transportation, Springer, vol. 44(3), 2017, pp 589-613, https://doi.org/10.1007/s11116-015-9669-z.

[14] K. Kim. "Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations" in Journal of Transport Geography. Vol. 66., 2018, pp. 309-320, https://doi.org/10.1016/j.jtrangeo.2018.01.001.

[15] Xu X, Ye Z, Li J, Xu M, "Understanding the Usage Patterns of Bicycle-Sharing Systems to Predict Users' Demand: A Case Study in Wenzhou, China" in Computational Intelligence and Neuroscience. 2018, https://doi.org/10.1155/2018/9892134.

[16] Yu C., O'Brien O., DeMaio P., Rabello R., Chou S., Benicchio T. (2021). The Meddin Bike-sharing World Map: Mid-2021 Report. PBSC Urban Solutions.

[17] A Mehdizadeh Dastjerdi, C. Morency "Bike-Sharing Demand Prediction at Community Level under COVID-19 Using Deep Learning" in Sensors, Vol. 22(3), 2022, https://doi.org/10.3390/s22031060.

[18] C. J. Costa, J., Aparício. "POST-DS: A Methodology to Boost Data Science" in 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020. pp. 1-6, IEEE. https://doi.org/10.23919/CISTI49556.2020.9140932

[19] W. H. Lee, M. Antoniades, H.G. Schnack, R. S. Kahn, S. Frangou, "Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter?" in Psychiatry Research: Neuroimaging, Volume 310, 2021, dot: 10.1101/2020.07.28.224931.

[20] T. Li, Y. Zhou, X. Li, J. Wu, T. He, "Forecasting Daily Crude Oil Prices Using Improved CEEMDAN and Ridge Regression-Based Predictors" in Energies 12, 3603, 2019, https://doi.org/10.3390/en12193603

[21] E. Bas, U. Yolcu, E. Egrioglu, “Intuitionistic fuzzy time series functions approach for time series forecasting” in Granul. Comput. 6, 619–629 (2021). https://doi.org/10.1007/s41066-020-00220-8

[22] R. Muthukrishnan, R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning” in 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, 18-20, https://doi.org/10.1109/ICACA.2016.7887916.

[23] C. Costa, J. Aparicio A Methodology to Boost Data Science in the Context of COVID-19. In Advances in Parallel & Distributed Processing, and Applications, Springer, Cham, 2021, pp. 65-75, https://doi.org/10.1007/978-3-030-69984-0_7

[24] M. Bhumika, B. Vimalkumar, “Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis” in International Journal of Computer Applications. 146, 2016, pp 26-30. https://doi.org/10.5120/ijca2016910921.

[25] Hassan S. U., Imran M., Iqbal S., “Deep context of citations using machine-learning models in scholarly full-text article” in. Scientometrics 117, 2018, pp. 1645–1662. https://doi.org/10.1007/s11192-018-2944-y

[26] T. Deepika, P. Prakash., “Power consumption prediction in cloud data center using machine learning” in International Journal of Electrical and Computer Engineering (IJECE), Vol. 10, 2020, pp1524-1532. doi: 10.11591/ijece.v10i2..

[27] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li, ”A parallel random forest algorithm for big data in a spark cloud computing environment,” in IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 4, 2016, pp. 919-933, https://doi.org/ 10.1109/TPDS.2016.2603511

[28] D. Palmer, N. O'Boyle, R. Glen, J. Mitchell J.,. “Random Forest Models To Predict Aqueous Solubility” in Journal of chemical information and modeling, 47(1):150-8, 2007, https://doi.org/10.1021/ci060164k. PMID: 17238260.

[29] J. Hong, S. Lee, J. Bae,J. Lee, W. Park, D. Lee, J. Kim, K. Lim, “Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow” in Water. 12(10):2927. 2020, https://doi.org/10.3390/w12102927.

[30] R. Tibshirani "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88, 1996.