DEVELOPING A FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

[1]Thammisetty Venkata Naga Radha Parameswari, [2]K.Chandra Prasad

[1]PG Scholar, Department of CSE, Srinivasa Institute of technology and science, Kadapa.
[2]Assistant professor, Department of CSE, Srinivasa Institute of technology and science, Kadapa.
[1]radhaparameswari2020@gmail.com, [2]chandu.sp2009@gmail.com

ABSTRACT

Flight delay is a major problem in the aviation sector. During the last two decades, the growth of the aviation sector has caused air traffic congestion, which has caused flight delays. Flight delays result not only in the loss of fortune also negatively impact the environment. Flight delays also cause significant losses for airlines operating commercial flights. Therefore, they do everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In this paper, using machine learning models such as ad boost classifier, decision tree classifier, grid search classifier, voting classifier, xgb classifier, naive bayes, decision tree classifier we predict whether the arrival of a particular flight will be delayed or not.

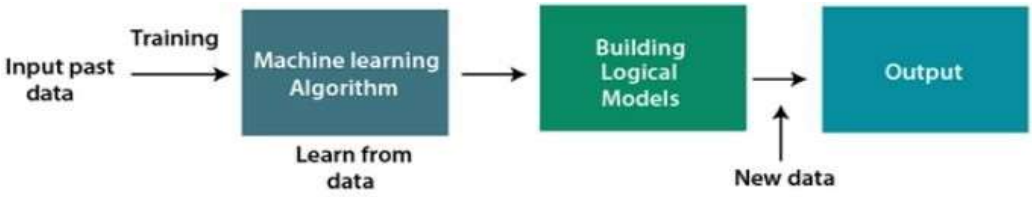
Keywords: Flight Prediction, Machine Learning, Error Calculation, Logistic Regression, Decision Tree, Bayesian Ridge, Random Forest, Gradient Boosting, Logistic Regression.

I. INTRODUCTION

Flight delay is studied vigorously in various research in recent years. The growing demand for air travel has led to an increase in flight delays. According to the Federal Aviation Administration (FAA), the aviation industry loses more than $3 billion in a year due to flight delays and, according to the BTS data, 1,386,699 flights were delayed nationwide in 2023, out of more than 6.8 million. That's about 20.2% of all scheduled domestic flights in the country. Meanwhile, 87,943 flights were canceled, about 1.3% of the total. The reasons for the delay of commercial scheduled flights are air traffic congestion, passengers increasing per year, maintenance and safety problems, adverse weather conditions, the late arrival of plane to be used for next flight. In the United States, the FAA believes that a flight is delayed when the scheduled and actual arrival times differs by more than 15 minutes. Since it becomes a serious problem in the United States, analysis and prediction of flight delays are being studied to reduce large costs.

i) Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more. Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the



output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm.

Figure: Processes of Machine Learning

II. BACKGROUND WORK

[1] SobhanAsian – 2019, Flight delay prediction for commercial air transport: A deep learning approach. This study analyzes high-dimensional data from Beijing International Airport and presents a practical flight delay prediction model. Following a multifactor approach, a novel deep belief network method is employed to mine the inner patterns of flight delays. Support vector regression is embedded in the developed model to perform a supervised fine-tuning within the presented predictive architecture.

[2] Xiaotong Dou – 2020: Flight Arrival Delay Prediction And Analysis Using Ensemble Learning. With the development of the civil aviation transportation industry in recent years, the volume of civil aviation transportation has increased rapidly. Increased carrier costs and reduced airport operating efficiency caused by flight delays have become issues that need to be addressed. How to improve the accuracy of predicting flight arrival delay time is of great significance for improving airport transportation efficiency, rationally scheduling flights and improving passenger comfort. In this paper, the Cat-boost model is utilized on the U.S Domestic airline on- time performance data from U.S.

[3] Suvojit Manna , Sanket Biswas , Riyanka Kundu , Somnath Rakshit 2021: A statistical approach to predict flight delay using gradient boosted decision tree. Supervised machine learning algorithms have been used extensively in different domains of machine learning like pattern recognition, data mining and machine translation. Similarly, there has been several attempts to apply the various supervised or unsupervised machine learning algorithms to the analysis of air

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

traffic data. However, no attempts have been made to apply Gradient Boosted Decision Tree, one of the famous machine learning tools to analyses those air traffic data.

III. SYSTEM ANALYSIS

i) Existing System

The Existing system proposed that, the expected growth in air travel demand and the positive correlation with the economic factors highlight the significant contribution of the aviation community to the U.S. economy. On-time operations play a key role in airline performance and passenger satisfaction. Thus, an accurate investigation of the variables that cause delays is of major importance. The application of machine learning techniques in data mining has seen explosive growth in recent years and has garnered interest from a broadening variety of research domains including aviation. This study employed a support vector machine (SVM) model to explore the non-linear relationship between flight delay outcomes. These findings provide insight for better understanding of the causes of departure delays and the impacts of various explanatory factors on flight delay patterns.

The primary contribution of Existing study is to investigate the possibility of using SVM models for analysis of the causes of flight delay and investigation of flight delay patterns. The maximum precision achieved was 79.7% with gradient booster as a classifier with a limited data set.

ii) Proposed System

In proposed system for detecting flight delays data set is collected from Kaggle website and machine learning algorithms like adaboost classifier, decision tree classifier, grid search classifier, voting classifier, xgb classifier, naive bayes, decision tree classifier. Model is trained and saved as pickle and used in web application to predict flight delay and accuracy is calculated for each model. Our proposed model does everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In this model, using machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression, we predict whether the arrival of a particular flight will be delayed or not. We develop a system that predicts for a delay in flight departure based on certain parameters. We train our model for forecasting using various attributes of a particular flight, such as arrival performances, flight summaries, origin/destination, etc.

IV. SYSTEM DESIGN

The System Design describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

i) Modules

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

a) Data Collection

To predict flight delays to train models, we have collected data accumulated by the Bureau of Transportation; U.S. Statistics of all the domestic flights taken in 2015 was used. The US Bureau of Transport Statistics provides statistics of arrival and departure that includes actual departure time, scheduled departure time, and scheduled elapsed time, wheels-off time, departure delay and taxi-out time per airport. Cancellation and Rerouting by the airport and the airline with the date and time and flight labelling along with airline airborne time are also provided.

b) Pre-Processing

Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

c) Feature Extraction

We have studied from various sources to find out which parameters will be most appropriate to predict the departure and arrival delays. After several searches, we conclude the following parameters:

- Day Departure
- Delay Airline
- Flight Number
- Destination Airport
- Origin Airport
- Day of Week
- Taxi out

d) Evaluation

After pre-processing and feature extraction of our dataset, 60% of the dataset was selected for training and 40% of the dataset was selected for testing. For error calculation, we are using scikit-learn metrices. Results are divided between two sections, Departure Delay(A) and Arrival Delay(B).

- Departure Delay

our results for departure delay which compares different Machine Learning models, i.e. Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor and Gradient

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Boosting Regressor, based on various evaluation metrics. Further, we compare each model concerning one evaluation metric at a time.

- Arrival Delay

our results for arrival delay which compares different Machine Learning models, i.e. Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor and Gradient Boosting Regressor, based on various evaluation metrics. Further, we compare each model concerning one evaluation metric at a time.

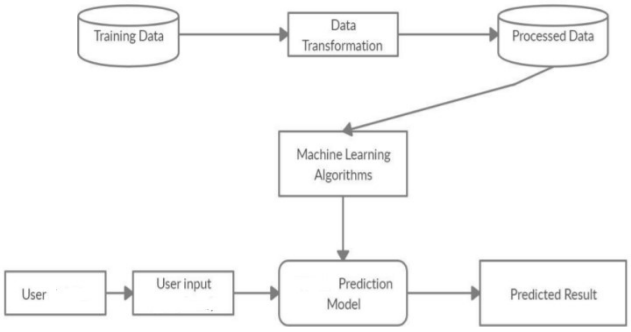


Figure: System Architecture

ii) Algorithms

a) Random Forest

The Random Forest classification model is made up of several decision trees. In simple terms, it combines the results from numerous decision trees to reach a single result. The main difference between decision trees and random forests is that decision trees consider all the possible feature splits, however, random forests will only select a subset of those features.

Random Forest algorithm was developed by Breiman, L. [60]. This is an ensemble learning algorithm made up of several DT classifiers, and the output category is determined collectively by these individual trees. When the number of trees in the forest increases, the fallacy in generalization error for forests converges. There are also important benefits of the RF. For example, it can manage high-dimensional data without choosing a feature; trees are independent of each other during the training process, and implementation is fairly simple; however, the training speed is generally fast and, at the same time, the generalization functionality is good enough [4]. Random forest algorithm for machine learning has tree predictions, and based on tree predictions, the RF provides random forest predictions [61]. The Random Forest algorithm model is visualized in Figure.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

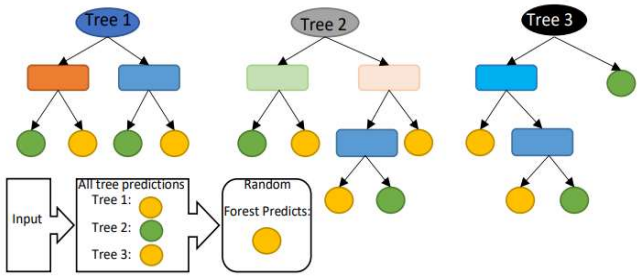


Figure: Random Forest algorithm

b) Logistic Regression (LR)

The adjusted r square by each key contributor was established using hypothesis testing. The factors that were found to be statistically significant after cross tabulations were used to train the multiplex logit model, which was then used to create the equation. In their study, they utilized unstructured historical machine data to train the ML classification algorithms including RF, XGBoost, and LR in predicting the machine failures. Various metrics were analyzed to determine the goodness of fit of the models. These metrics include empirical cross-entropy, area under the receiver operating characteristic curve (AUC), receiver operating characteristic curve itself (ROC), precision-recall curve (PRC), number of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) at various decision thresholds, and calibration curves of the estimated probabilities. Based on the results obtained, in terms of ROC, all the algorithms performed significantly better and almost similar. But in terms of decision thresholds, RF and XGBoost perform better than LR. Using a given set of independent variables, linear regression is used to estimate the continuous dependent variations. However, using a given set of independent variables, logistic regression is used to estimate the categorical contingent variations [68]. Graph of the linear regression model and logistics regression model are shown in Figure.

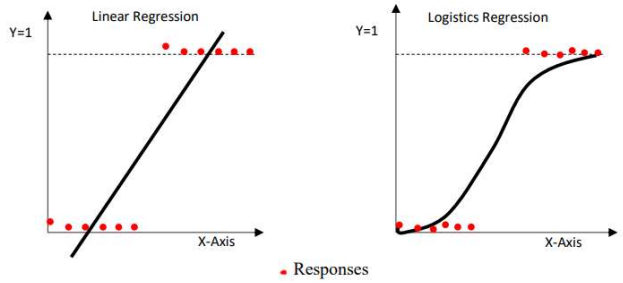


Figure: Logistic Regression

c) Decision Tree (DT)

Decision Tree is a network system composed primarily of nodes and branches, and nodes comprising root nodes and intermediate nodes. The intermediate nodes are used to represent a feature, and the leaf nodes are used to represent a class label [52]. DT can be used for feature

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

selection [57]. DT algorithm is presented in Figure. DT classifiers have gained considerable popularity in a number of areas, such as character identification, medical diagnosis, and voice recognition. More notably, the DT model has the potential to decompose a complicated decision-making mechanism into a series of simplified decisions by recursively splitting covariate space into subspaces, thereby offering a solution that is sensitive to interpretation.

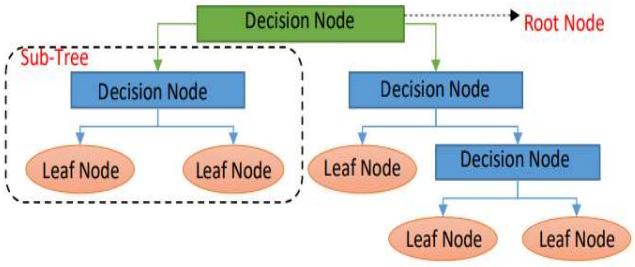


Figure: Decision Tree algorithm

V. TESTING

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Stating formally, we can say, testing is a process of executing a program with intent of finding an error.

1. A successful test is one that uncovers an as yet undiscovered error.
2. A good test case is one that has probability of finding an error, if it exists.
3. The test is inadequate to detect possibly present errors.
4. The software more or less confirms to the quality and reliable standards.

i) Code testing

This examines the logic of the program. For example, the logic for updating various sample data and with the sample files and directories were tested and verified.

ii) Specification Testing

Executing this specification starting what the program should do and how it should perform under various conditions. Test cases for various situation and combination of conditions in all the modules are tested.

iii) Unit testing

In the unit testing we test each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried

out during programming stage itself. In the testing step each module is found to work satisfactorily as regard to expected output from the module. There are some validation checks for fields also. For example, the validation check is done for varying the user input given by the user which validity of the data entered. It is very easy to find error debut the system.

Test Case Id	Test Case Name	Test Case Desc.	Test Steps			Test Case Status	Test Priority
			Step	Expected	Actual		
01	Upload the tasks dataset	Verify either file is loaded or not	If dataset is not uploaded	It cannot display the file loaded message	File is loaded which displays task waiting time.	High	High
02	Upload dataset and preprocess	Verify either dataset loaded or not and preprocessed	If dataset is not uploaded and preprocessed	It cannot display dataset reading process completed	It can display dataset reading process completed	low	High
03	Preprocessing	Whether preprocessing on the dataset applied or not	If not applied	It cannot display the necessary data for further process	It can display the necessary data for further process	Medium	High
04	Prediction model	Whether Prediction algorithm applied on the data or not	If not applied	Algorithm model is created	Algorithm model is created	High	High

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

05	Prediction	Whether predicted data is displayed or not	If not displayed	It cannot view prediction containing flight delay	It can view prediction containing flight delay	High	High
06	Noisy Records Chart	Whether the graph is displayed or not	If graph is not displayed	It does not show the variations in between clean and noisy records	It shows the variations in between clean and noisy records	Low	Medium

VI. RESULTS

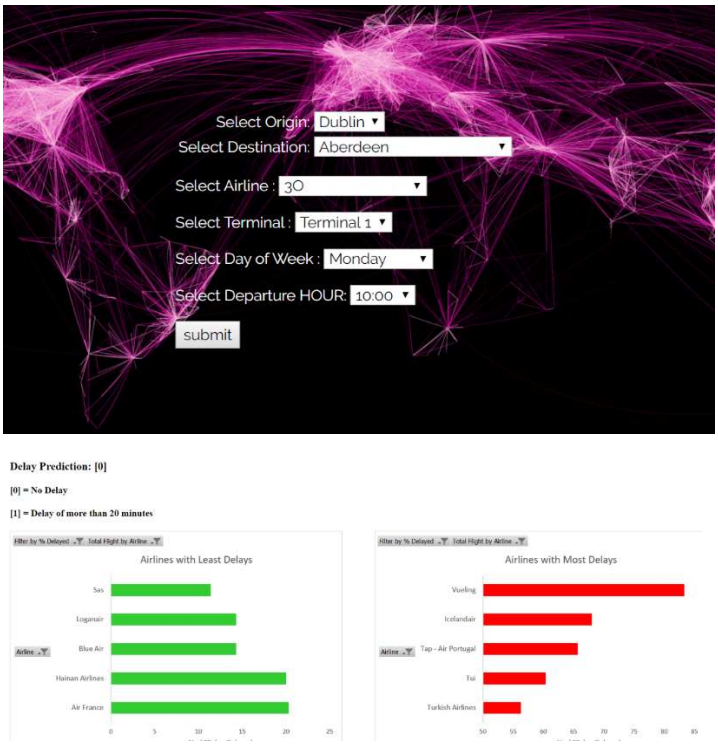


Figure: Delay Prediction

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

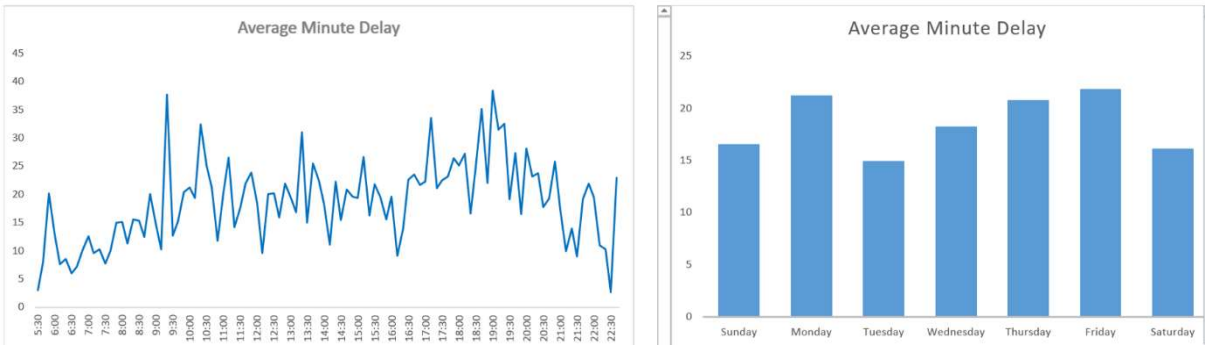


Figure: Average Minute Delay

VII. CONCLUSION

Machine learning algorithms were applied progressively and successively to predict flight arrival & delay. We built five models out of this. We saw for each evaluation metric considered the values of the models and compared them. We found out that: - In Departure Delay, Random Forest Regressor was observed as the best model with Mean Squared Error 2261.8 and Mean Absolute Error 24.1, which are the minimum value found in these respective metrics. In Arrival Delay, Random Forest Regressor was the best model observed with Mean Squared Error 3019.3 and Mean Absolute Error 30.8, which are the minimum value found in these respective metrics. In the rest of the metrics, the value of the error of Random Forest Regressor although is not minimum but still gives a low value comparatively. In maximum metrics, we found out that Random Forest Regressor gives us the best value and thus should be the model selected.

REFERENCES

[1] N. G. Rupp, "Further Investigation into the Causes of Flight Delays," in Department of Economics, East Carolina University, 2007.

[2] "Bureau of Transportation Statistics (BTS) Databases and Statistics," [Online]. Available: http://www.transtats.bts.gov.

[3] "Airports Council International, World Airport Traffic Report," 2015,2016.

[4] E. Cinar, F. Aybek, A. Caycar, C. Cetek, "Capacity and delay analysis for airport manoeuvring areas using simulation," Aircraft Engineering and Aerospace Technology, vol. 86, no. No. 1,pp. 43-55, 2013.

[5] Navoneel, et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.

[6] Y. J. Kim, S. Briceno, D. Mavris, Sun Choi, "Prediction of weatherinduced airline delays based on machine learning algorithms," in 35th Digital Avionics Systems Conference (DASC), 2016.

[7] W.-d. Cao. a. X.-y. Lin, "Flight turnaround time analysis and delay prediction based on Bayesian Network," Computer Engineering and Design, vol. 5, pp. 1770-1772, 2011.

[8] J. J. Robollo, Hamsa, Balakrishnan, "Characterization and Prediction of Air Traffic Delays".

[9] S. Sharma, H. Sangoi, R. Raut, V. C. Kotak, S. Oza, "Flight Delay Prediction System Using Weighted Multiple Linear Regression," International Journal of Engineering and Computer Science, vol. 4, no. 4, pp. 11668 - 11677, April 2015.

[10] Hatıpoğlu, I., Tosun, Ö. & Tosun, N. Flight delay prediction based with machine learning. LogForum 18, 1 (2022).

[11] Wang, F., Bi, J., Xie, D. & Zhao, X. Flight delay forecasting and analysis of direct and indirect factors. IET Intell. Transp. Syst. 16(7), 890–907 (2022).

[12] Wang, Z. et al. Distribution prediction of strategic flight delays via machine learning methods. Sustainability 14(22), 15180 (2022).

[13] Xu, H., Shi, J. & Wang, T. Departure flight delay prediction model based on deep fully connected neural network. J. Comput. Appl. 42(10), 3283 (2022).

[14] Paramita, C., Supriyanto, C., Syarifuddin, L. A. & Rafrastara, F. A. The use of cluster computing and random forest algorithm for f light delay prediction. Int. J. Comput. Sci. Inf. Secur. 20, 2 (2022).

[15] Li, Q., Jing, R. & Dong, Z. S. Flight delay prediction with priority information of weather and non-weather features. IEEE Trans. Intell. Transp. Syst. 1, 1 (2023).

[16] Kaiquan, C. A. I. et al. A geographical and operational deep graph convolutional approach for flight delay prediction. Chin. J. Aeronaut. 36(3), 357–367 (2023).

[17] Qu, J., Chen, B., Liu, C. & Wang, J. Flight delay prediction model based on lightweight network ECA-MobileNetV3. Electronics 12(6), 1434 (2023).

[18] Qu, J., Wu, S. & Zhang, J. Flight delay propagation prediction based on deep learning. Mathematics 11(3), 494 (2023).

[19] Wu, Y., Yang, H., Lin, Y. & Liu, H. Spatiotemporal propagation learning for network-wide flight delay prediction. IEEE Trans. Knowl. Data Eng. 1, 1 (2023).

[20] Chen, H., Tu, S. & Xu, H. The application of improved grasshopper optimization algorithm to flight delay prediction-based on spark. In Complex, Intelligent and Software Intensive Systems: Proceedings of the 15th International Conference on Complex, Intel ligent and Software Intensive Systems (CISIS-2021) 80–89 (Springer, 2021).

[21] Yang, H., Zhang, X., Li, Z. & Cui, J. Region-level traffic prediction based on temporal multi-spatial dependence graph convolutional network from GPS data. Remote Sens. 14(2), 303 (2022).

[22] Chen, J. et al. A flow feedback traffic prediction based on visual quantified features. IEEE Trans. Intell. Transp. Syst. 24(9), 10067 10075 (2023).

[23] Jiang, Y., Yang, Y., Xu, Y. & Wang, E. Spatial-temporal interval aware individual future trajectory prediction. IEEE Trans. Knowl. Data Eng. https://doi.org/10.1109/TKDE.2023.3332929 (2023).

[24] Yang, M., Wang, Y., Liang, Y. & Wang, C. A new approach to system design optimization of underwater gliders. IEEE/ASME Trans. Mechatron. 27(5), 3494–3505 (2022).

[25] Singh, D. & Singh, B. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 97, 105524 (2020).

[26] Chatzi, A. & Doody, O. The one-way ANOVA test explained. Nurse Res. 31, 2 (2023).

[27] Venkatesh, B. & Anuradha, J. A review of feature selection and its methods. Cybern. Inf. Technol. 19(1), 3–26 (2019).

[28] Deng, D. DBSCAN clustering algorithm based on density. In 2020 7th International Forum on Electrical Engineering and Automa tion (IFEEA) 949–953 (IEEE, 2020).

[29] Pierezan, J. & Coelho, L. D. S. Coyote optimization algorithm: A new metaheuristic for global optimization problems. In 2018 IEEE Congress on Evolutionary Computation (CEC) 1–8 (IEEE, 2018).

[30] Goutte, C. & Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval 345–359 (Springer, 2005).

[31] Dudek, A. Silhouette index as clustering evaluation tool. In Classification and Data Analysis: Theory and Applications Vol. 28 (ed. Dudek, A.) 19–33 (Springer, 2020).