Stock Prediction and Analysis using Twitter Sentiments

[1]Hardik Lad, [2]Jairaj Saraf, [3]Siddhi Pawar, [4]Vidya Chitre

[1,2,3]UG Student, [4]Professor, Vidyalankar Institute of Technology, Mumbai, India,

[1]hardik.lad773@gmail.com, [2]jairaj.saraf01@gmail.com, [3]siddharth0914@gmail.com, [4]vidya.chitre@vit.edu.in

Abstract- **Stock market prediction and its analysis are one of the most important topics of research in today's world. Historical stock market data has proved to be insufficient in predicting upcoming stock performance. It is hard to predict stock price movement on basis of a single key factor. Historical data when amalgamated with public sentiments has proved to achieve more accurate predictions. In this paper, we propose a unified system that incorporates historical stock data as well as public twitter sentiments for predicting stock price movements.**

Keywords* — *sentiment analysis, stock market, accuracy, machine learning, regression, prediction, sentiments

I.INTRODUCTION

In various researches, it has been proven that news and Twitter sentiments play a pivotal role in shaping and predicting the future stock of a company. The fundamental idea behind this is to recognize patterns in these tweets and find a correlation between them and predict the future behavior of various stock prices.

Stock market prediction is an important problem, hence we propose an enhanced learning-based method for stock prediction. In decision-making, the opinions of others have a significant effect on customers. With a successful model for stock analysis and prediction, we can gain insight into market behavior, finding trends that would otherwise not have been possible. With the increasing computational power, machine learning will be an efficient method to solve this problem. However the public dataset is too limited to work with, the motivating idea is that if we know all information about today's stock trading including the market sentiment, the price is adequately predictable. With the growth of the internet, social networks, and online social interactions, getting daily user sentiments and opinions is a feasible job. Manually analyzing public sentiments and quantifying them is a difficult task. For this reason, we used the Twitter API Tweepy for mining public sentiments. In this paper, we proposed a system that performs sentiment analysis of the fetched tweets by using Natural Language Processing and collaborates them with historical stock data to provide better and more accurate predictions as compared to conventional methods.

II.LITERATURE SURVEY

We analyzed and studied some related works based on Stock movement prediction. The survey of all 10 papers that have been studied are described below.

Pranjal Chakraborty, Ummay Sani Pria, M. R. A. H. Rony, and M. A. Majumdar in [2] tried to predict stocks by analyzing twitter sentiments. According to them, tweets are sort of less informative and are grammatically incorrect so it gets difficult to classify them with traditional classifiers. They observed that SVM works better for classifying twitter sentiments. They concluded that with the Boosted Regression Tree it gets difficult to predict stocks because of extremely high and low differences in stock indexes. S. Naveen Balaji, P. Victer Paul, R. Saravanan in [3] did a survey on Big Data, Data Analytics, and its types. The survey stated the techniques which were related to big Data and prediction techniques. According to their study they cleared that Predictive analytics is better than Analytical technique to predict stock in advance.They added that historical as well as sentimental analysis techniques improves the accuracy of stock prediction. Venkata Sasank Paglou, Kamal Nayan Reddy Challa, Ganapati Panda in [4] classified tweets in three categories i.e positive, neutral and negative. According to them, positive sentiments or emotions of the public would have a huge effect on companies' stock prices. They showed a strong relation exists between rise and fall stock prices of a company due to public opinions. A Sarkar, A K Sahoo, S Sah, C Pradhan in [5] stated that a single model is yet to be developed to predict stock fluctuations. They collected technical and fundamental analysis for the market.They used the LSTM model to predict stocks which resulted in the shape of predicted prices being seen to conform to real prices. In Stock Price Prediction Through the Sentimental Analysis of News Articles, Jaeyoon Kim, Jangwon Seo, Minhyeok Lee, Junhee Seok in [6] predicted stock through emotional analysis using news articles. They achieved the relation between positive indexes for each date. Approximately 0.304 correlation value was checked based on this. Because of Korean language it was highly difficult

DOI : 10.35291/2454-9150.2022.0108

to achieve high accuracy using NLP methods. Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C. Anastasiu in [7] predicted stocks using neural networks and a combination of neural networks. They concluded that there was a strong correlation between financial news articles and stock prices. They achieved better results with RNN and found a relation between textual information and stock price direction. But the models didn't perform better since stock prices were low or highly volatile. Masoud Makrehchi, Sameena Shah, and Wenhui Liao in [8] generated training data based on the stock market. Their training strategy was able to beat S&P 500 by 20%. . They stated that for an external calculation the best approach would be a transferring domain using test data and for that they labeled needed test data that wasn't available. Rubi Gupta and Min Chen in [9] investigated a large-scale amalgam of tweets. They did a sentimental analysis on StockTwits data by three methods (Naive Bayes, SVM, and logistic regression). Because these models had a positive effect on stock prediction.Sunil Kumar Khatri, Himanshu Singhal, and Prashant Johri in [10] tried to collect e-data from various sites and analyzed it with ANN. They tried to minimize error up to the least possible value however e-data proved to be more perfect. Yichuan Xu and Vlado Kesel in [11] tried DNN with a Data set and combined it with stock price history and technical indicators. They used the LSTM model and observed improvement over conventional LSTM on the aggregate dataset. The result was 65%.

III.PROPOSED SYSTEM

We are proposing a system that helps determine the importance of Sentiment analysis of peoples' emotions as a determining factor for Stock price movement. For this, we are using the stock data of United Airlines(UAL) as well as UAL twitter data for prediction. The stock data is retrieved from Yahoo Finance while the UAL tweets are fetched from the Tweepy API. Our system helps determine the correlation of actual Stock prices to that of predicted stock prices using both user sentiments as well as historical stock data.

For this, we use readily available public sentiments from Twitter and stock price data. This data is then cleaned and processed and then combined into a single dataset for further analysis. We then perform sentiment analysis on the Twitter sentiment data and give appropriate polarities and sentiment scores for the tweets. We then split this data into training and testing datasets respectively. Further, we pass this data to the appropriate Machine learning model considering both stocks as well as sentiment scores for prediction. This is then followed by the prediction accuracy comparison of various models and visualization of stock data with the predicted stock data that we got from the ML prediction model.

IV.METHODOLOGY

Following steps were performed while building the system.

Data Collection: The stock's price data was downloaded from Yahoo Finance, whereas the respective United Airlines(UAL) tweets were fetched using Tweepy API. The data contains various metrics of the requested stock from the start to end dates the user has specified. Over 400 tweets get fetched from the API per week and the stock data contains 6 attributes for each date. As the Tweets fetched from Tweepy API have a limitation of fetching tweets only for the past week, we collected tweet data of UAL from various sources and combined it with UAL stock data into a single dataset ranging from the year 2007 to the year 2016.

Data Cleaning: As some days are holidays, not every date has an Adjusted Closing Price which is an important factor for a certain stock. So the blank values are taken care of by calculating the mean of the closing price. Also, we have removed a few unwanted columns such as 'Open', 'High', 'Low', and 'Volume' from the dataset as they will not be useful for prediction. For the tweets fetched, we removed special characters, dots, and spaces from the tweets fetched.Multiple tweets for a certain day were combined into a single tuple so we can match the tweet data and stock data date-wise. We then merged the Stock price data with the tweets fetched into a single Dataset. As the adj close price attribute is a range of stock price and not an arbitrary number, we did not need feature scaling for it.

VARIABLE	DESCRIPTION
Date	Stock price date
Open	Open stock price value for the day
High	Highest stock price value of the day
Low	Lowest stock price value of the day
Close	Closing share price for the day
Volume	Number of shares traded for the day
Tweets	Short descriptive tweet
Compound	A compound overall sentiment value
Negative	Negative sentiment value
Neutral	Neutral sentiment value
Positive	Positive sentiment value

Table 1: Description of the attributes

Data Preprocessing: As for preprocessing the data, we have arranged the tweets from the starting date till the ending date, clubbing multiple tweets for the same day together. Later, a 'price' column is added to the data frame to assign the correct

DOI : 10.35291/2454-9150.2022.0108

closing price to each date. The columns ‘compound’, ‘positive’, ‘negative’, and ‘neutral’ will represent the numeric value of the polarity of the tweet or basically a sentiment score. The adjusted closing price is considered from the stock data and all the attributes were then rearranged accordingly.

Sentiment Analysis: We used NLP for doing Sentiment analysis on the Twitter data. For this, we used the NLTK library Vader Sentiment analyzer to give polarity to the tweets. We further classified these polarities into 4 attributes: positive, negative, neutral, and compound sentiment scores.

Data Splitting: We split the data into a training and testing set in the ratio of 80:20, using the most recent 20% of the data for testing purposes.

Generating the model: The models have been generated using Linear regression and Random Forest regression models. The model which provides the highest accuracy is going to be considered for the final prediction.

Presenting the Final prediction: Both models were evaluated and the final prediction and visualization were done on the basis of evaluation.

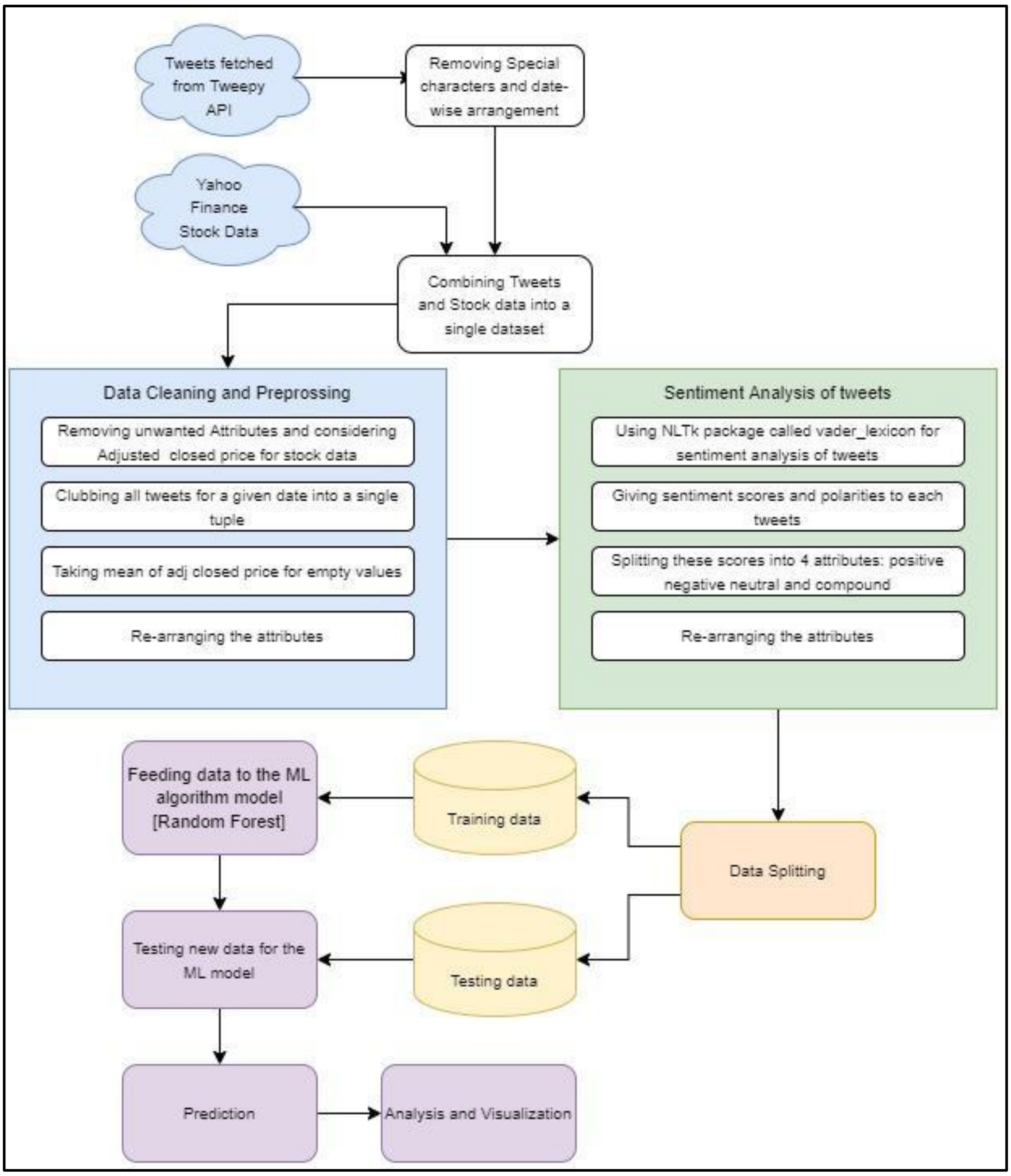
Tweets fetched from Tweepy API
Removing Special characters and date-wise arrangement
Yahoo Finance Stock Data
Combining Tweets and Stock data into a single dataset
Data Cleaning and Preprossing
Removing unwanted Attributes and considering Adjusted closed price for stock data
Clubbing all tweets for a given date into a single tuple
Taking mean of adj closed price for empty values
Re-arranging the attributes
Sentiment Analysis of tweets
Using NLTk package called vader_lexicon for sentiment analysis of tweets
Giving sentiment scores and polarities to each tweets
Splitting these scores into 4 attributes: positive negative neutral and compound
Re-arranging the attributes
Feeding data to the ML algorithm model [Random Forest]
Training data
Data Splitting
Testing new data for the ML model
Testing data
Prediction
Analysis and Visualization

Fig 1: System Methodology Flowchart

V.IMPLEMENTATION

- **Models:**

a) Linear Regression:

Linear Regression is a Supervised Machine learning algorithm that depends on historical data for prediction. It attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

In the case of this project, the value of the close price is the dependent variable(y) and the input factors used for prediction are the independent variables. [12]

b) Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. [13]

VI.RESULTS

We are using evaluation metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R squared Value for evaluating both models.

➢ Mean Absolute Error (MAE) is the average of the difference between the actual data value and the predicted data value. [14]

➢ Mean Squared Error is the average squared difference between the estimated values and the actual value.

Where, n = Data set observations

Y_i = Observation values

Y^$_i$ = Predicted Values [15]

❖ Root Mean Squared Error is the root of MSE.

Where, n = Data set observations

S_i = Predicted values

O_i = Observations [16]

❖ R squared value is used for measuring the accuracy of the model.

Where

R^2 = coefficient of determination [17]

Since the Random Forest Regression technique is often used on data containing volatile values such as stock prices, it yielded better accuracy for our system. Following pie charts and graphs is the result of the analysis and visualization of the models.

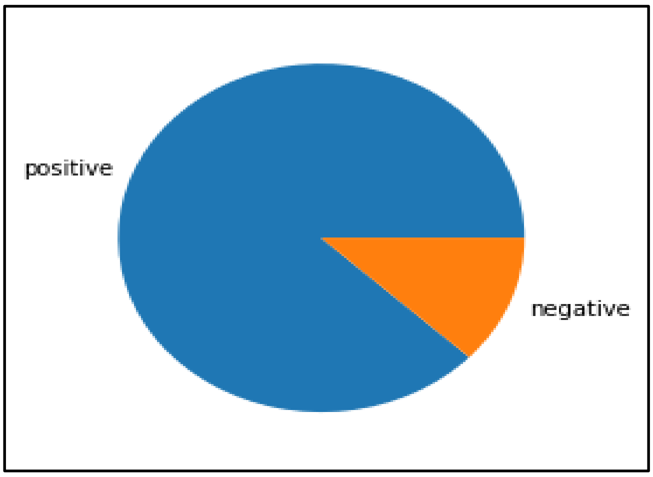

Fig 2. Sentiment scores of recent UAL Tweets(March 2022)

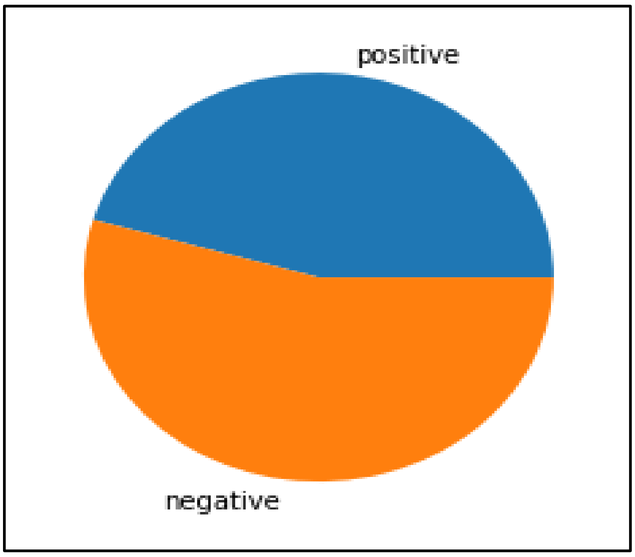

Fig 3. Sentiment scores of Historical UAL Tweets(2007-2016)

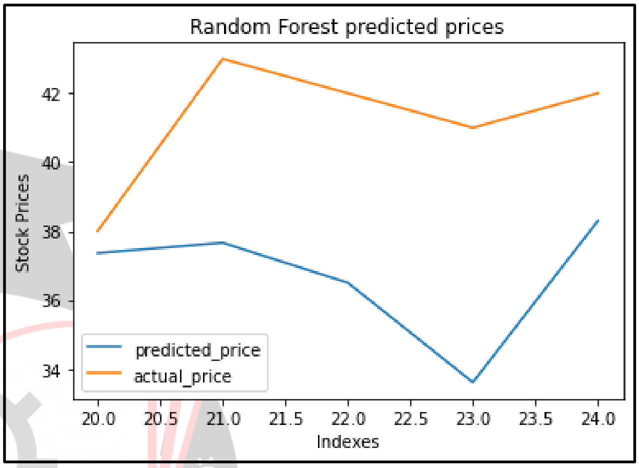

Fig 4. Random forest Regression on Recent UAL Data(March 2022)

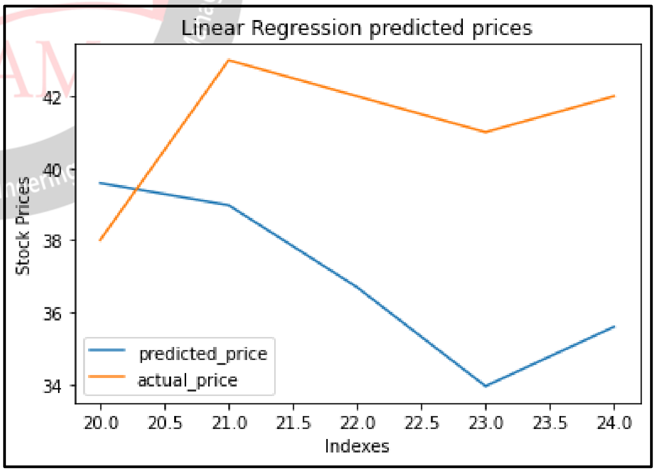

Fig 5. Linear Regression on Recent UAL Data(March 2022)

DOI : 10.35291/2454-9150.2022.0108

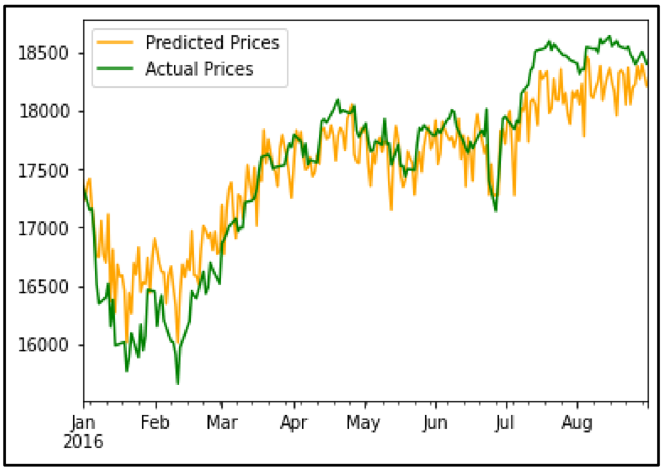

Fig 6. Random Forest Historical Data

The following table is the result of evaluation metrics done on both models. It consists of the respective accuracies and errors of both models.

Table 2: Values for evaluation metrics

Models	RMSE	MSE	MAE	R Square Value	Accuracy
Linear Regression	4.97	24.72	4.64	-8.30	77
Random Forest Regression	4.87	23.76	4.64	-7.6	86.47

VII. Conclusion And Future Scope

We studied and applied two regression techniques in this project. On the basis of the evaluation and visualization metrics, we can conclude that the random forest regression technique fits our hybrid model and yields better accuracy.

From this study, we have learned that people's sentiments are a significant factor in stock price movements. The study results also indicate that performance of stock movement prediction over a longer period of time yields better results than that done over a short period of time. Our proposed model using the Random Forest regression technique achieved significant performance with an 86.47% accuracy for UAL stocks.

This approach can help end-users in analyzing various stocks and also for its forecasting. This methodology can be applied to various stocks such as AAPL, TSLA, and AMZN, and its analysis and prediction can be displayed on a front-end web app for better visualization as the future scope of this study.

References

[1] Hardik Lad, Jairaj Saraf, Siddhi Pawar, Dr. Vidya Chitre. “Stock prediction and Analysis using Twitter Sentiments: A Survey. Issued February 2022. DOI : 10.35291/2454-9150.2022.0045

[2] P. Chakraborty, U. S. Pria, M. R. A. H. Rony and M. A. Majumdar, "Predicting stock movement using sentiment analysis of Twitter feed," 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT), 2017, pp. 1-6, doi: 10.1109/ICIEV.2017.8338584.

[3] S. N. Balaji, P. V. Paul and R. Saravanan, "Survey on sentiment analysis based stock prediction using big data analytics," 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), 2017, pp. 1-5, doi: 10.1109/IPACT.2017.8244943.

[4] Pagolu, Sasank & Reddy, Kamal & Panda, Ganapati & Majhi, Babita. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 1345-1350. 10.1109/SCOPES.2016.7955659.

[5] Pagolu, Sasank & Reddy, Kamal & Panda, Ganapati & Majhi, Babita. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 1345-1350. 10.1109/SCOPES.2016.7955659.

[6] J. Kim, J. Seo, M. Lee and J. Seok, "Stock Price Prediction Through the Sentimental Analysis of News Articles," 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), 2019, pp. 700-702, doi: 10.1109/ICUFN.2019.8806182.

[7] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[8] M. Makrehchi, S. Shah and W. Liao, "Stock Prediction Using Event-Based Sentiment Analysis," 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, pp. 337-342, doi: 10.1109/WI-IAT.2013.48.

[9] R. Gupta and M. Chen, "Sentiment Analysis for Stock Price Prediction," 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2020, pp. 213-218, doi: 10.1109/MIPR49039.2020.00051.

[10] S. K. Khatri, H. Singhal and P. Johri, "Sentiment analysis to predict Bombay stock exchange using artificial neural network," Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, 2014, pp. 1-5, doi: 10.1109/ICRITO.2014.7014714.

[11] Y. Xu and V. Keselj, "Stock Prediction using Deep Learning and Sentiment Analysis," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 5573-5580, doi: 10.1109/BigData47090.2019.9006342.

[12] http://www.stat.yale.edu/Courses/1997-98/101/linreg.html

[13] https://levelup.gitconnected.com/random-forest-regression-209c0f354c84#:~:text=Random%20Forest%20Regression%20is%20a%20supervised%20learning%20algorithm%20that%20uses,prediction%20than%20a%20single%20model

[14] https://www.statology.org/mean-absolute-error-python/

[15] https://www.sciencedirect.com/topics/engineering/root-mean-squared-error/

[16] https://www.sciencedirect.com/topics/engineering/root-mean-squared-error/

[17] https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html

DOI : 10.35291/2454-9150.2022.0108