Assessing Machine Learning Linear Models to Predict Egyptian Stock Market Prices

Ismail M. Hagag
El Madina Higher Institute of
Administration and Technology
Egypt

ABSTRACT
Despite significantly related difficulties, the search for models to predict the prices of financial markets remains a highly researched subject. Financial time series are challenging to forecast because of the non-linear, chaotic, and dynamic nature of the prices of financial assets. Given their capacity to recognize complex patterns in various applications, machine learning models are among the most researched of the most recent techniques. The objective of this research is to identify the most effective method (among the selected methods) for forecasting stock markets by evaluating the accuracy of these models using the EGX100 market indicator. By applying nine different algorithms on the EGX100 daily prices we, namely Decision Tree (DT), Support Vector Machine (SVM), extreme Gradient Boost (XGBoost), AdaBoost, Multiple layer perceptron (MLP), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and K Nearest Neighbour (KNN). we have found that Linear regression is by far the best machine-learning algorithm for this type of problem. This result is reached through the usage of four different metrics after changing the problem into a classification problem.

Keywords
Stock Market, EGX100, ARIMA, Machine Learning, KNN, MSE

1. INTRODUCTION
The stock market is a vital part of any economy in today's world, it's where investors can direct their personal savings to maximize their wealth, which will also help the nation's economy to grow, and as stated, the financial markets can be advantageous for any economy [1].

The Egyptian securities market is regarded as one of the most important and rapidly expanding markets in the Middle East due to a large number of investors and trading activity, as well as a large number of listed securities [2].

However, selecting a company that will be profitable or the ability to forecast an index is considered a difficult task for both researchers and investors, as it is One of the most significant and difficult problems involving time series is stock market forecasting [3], as it follows a random pathway and therefore they are unpredictable [4]. The idea of creating a successful model was the attention of academia despite the efficient market hypothesis (EMH) [5], but, a predictive model that can predict the market indices overtime and generate profits would prove the falseness of the EMH and also make it possible to make significant profits from financial operations [6, 7, 8, 9].

To forecast stock prices in the future there are two factors are used in such an approach, fundamental and technical analysis. The fundamental analysis consists of the company's parameters, and the technical analysis of the stock price history is significant according to Dow theory [10]. Recently, a lot of data scientists have focused on creating stock price prediction models based on complex, nonlinear machine learning techniques like artificial neural networks (ANN) and regression techniques. In recent years, a variety of research projects have been carried out in the area of using various neural network types for stock price prediction. Therefore, research into risk management is a crucial subject.

Another aspect of the problem of forecasting the future of stock prices is that it is affected by other aspects such as the whole economy, industry characteristics, politics, and investor psychology [11]. That's why most of the literature regarding stock prices were using different methods and real-world applications that are based on historical data, for instance, this application may contain, support and resistance levels [12], economic factors that influence the market trends [13], moving average, autoregressive models, discriminating analyses, and correlations [14].

Recently, the focus of the new techniques has been on chaotic data, randomness, and non-linearity [15], The advancement in technologies have made it possible to discover patterns in a huge dataset with a computational system [16]. Using different models and techniques such as machine learning which is a term that is frequently used to describe the usage of computational intelligence to create a predicting model, and the data that is obtained from the financial market is time series analysis [17], for predicting when stock market volatility will return, there are numerous models available.

The objective of this research is to identify the most effective method (among the selected methods) for forecasting stock markets by evaluating the accuracy of these models using the EGX100 market indicator. To help investors and future academics which algorithm is better, doing that using an extensive analysis to determine the most suitable classification algorithm to better predict stock prices. This paper uses the daily data of EGX100 that have been extracted from investing.com.

To assess different techniques we have compared set of algorithms, and they are: Decision Tree (DT), Support Vector Machine (SVM), extreme Gradient Boost (XGBoost), AdaBoost, Multiple layer perceptron (MLP), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and K Nearest Neighbour (KNN).and we will determine the most effective algorithm to solve the problem through comparing the accuracy of each approach.

2. RELATED WORK
2.1 Data Mining Algorithms
Recent work has been focusing on using different techniques in different fields such as the medical field such as [19], as they used a regression tree to predict congestive heart failure

(CHF) and classify the high risk and low risk with 93.3% and 63.5% respectively.

We also have researchers that are focused on the stock market prices such as [20], where they used a few R packages of different algorithms to predict the EGX30 index and their finding was that ARIMA (0,0,0) is the best model for stationary time series and ARIMA (0,1,0) is the best for non-stationary. [21], have found that using principal components analysis (PCA) on EGX30 could summarize 83% of data variability, furthermore, the result of applied cross-validation shows that the first two components are more than enough to capture the observed value for this index.in [22] the author was investigating the machine learning techniques that are applied to predict the financial market, they used bibliographic survey techniques to highlight the important texts of research, and they investigated fifty-seven texts of research on the North American market to classify market, assets, methods, and variables, and their finding the best models was support vector machines and neural networks, and that data from developing markets was a research opportunity. In [23] the author attempted to combine neural networks and ARIMA models to predict EGX30, and the results showed that the combination of these two models was more appropriate, because it gives more accurate results as one of these models compensates the shortage of the other model.

In [24] was using a similar approach to assess the deep learning methods that are used to predict EGX30, EGX50, EGX70, EGX100, and NILE, using different deep learning such as Bayesian regularization (BE), Levenberg–Marquardt (LM), and scaled conjugate gradient (SCG), on different time frames daily, 3 days, 5 days, 7 days, and 30 days, and their results suggest that BR was the best for the short term, especially for 3 days, and LM was the best when it comes to long term prediction, especially for the 7 days prediction. These results are based on Mean Squared Error (MSE).

Machine learning algorithms, extract patterns learned from historical data to predict the future pattern of such data [25] such a process contains two phases, training which require the extracting of the variables and splitting the data into two parts one for training and the other for assessing the accuracy of the model. [26] this research tried to predict price trends in the stock market using oscillators and indicators, such as the moving Average Convergence Divergence (MACD), the Relative Strength Index (RSI), the Stochastic Oscillator (KDJ), and Bollinger Band (BB). and predicting short-term and long-term stock trends using the decision tree technique. [27] used the support vector machine (SVM) technique to predict the prices of the stock market, where they enhanced it through using GASVM which stands for Genetic algorithm SVM to select better features and improve the SVM and avoid overfitting, underfitting, and such model, they reached an accuracy of 93.7%. [28] tested the XGBoost technique for stock market price predictions and found that XGBoost can predict the general trend of stock prices. [29] have used AdaBoost-based long short-term memory as an approach to predict financial time series of daily change of the data, and finding that such approach was promising for such type of data. [30] was approaching the prediction of stock market prices using Multi-Layer Perceptron (MLP) of Bombay Stock Exchange (BSE), and using MSE to assess the accuracy of the model as they have reached an average of 95% accuracy as different companies and different dates have shown different accuracy. In [31] the author used a random forest algorithm to predict trends in the stock market and tried to minimize the error by treating the prediction problem as a classification problem, with such approval they reached an accuracy of 80% all the way up to 92%. [32] approached the prediction of the stock prices using Xtreme Gradient Boosting, and reached an accuracy of 87%, for 60 days and 90 days periods, and again they also considered it as a classification problem instead of predicting problem. In [33] used the Logistic regression model to predict the stock performance of Pakistan's Stock Exchange, the accuracy of the prediction was 89.77% and their variable did not include any Macroeconomic variables. [34] used K Nearest Neighbour (KNN) based on EEMD to forecast the financial time series, and specifically to predict closing prices of the stock market, such model when integrated with ensemble empirical mode decomposition reached a high score for short-term prediction, they compared it with different algorithms and have found that KNN and ARIM was the highest score of them all as the MASE was 0.97 for the KNN and was 1.0 for the ARIMA, using different accuracy measures such as MAPE and NMSE, the KNN and ARIMA were the top two models.

2.2 Methodology

The main goal of this research is to compare multiple machine learning algorithms and their ability to predict EGX100 closing prices over time. The methodology will consist of four main steps, as shown in Fig. 1.

1. Dataset collection
2. Data preprocessing
3. Machine learning algorithm
4. Evaluation of the prediction of model

2.3 Dataset Collection

We have collected our dataset from investing.com website, it contains daily EGX100 data starting from 01/01/2020 till 31/12/2021, as we tried to stay away from the COVID-19 effect on the Egyptian stock market [35], the data set also contains six features, open, close, high, low, volume, and change and the description of each one is in the Table 1

Table (1):

Field	Description
Date	The date of the day data were recorded
Open	The price at which the financial security opens in the market when trading begins
High	The highest price the index has reached on this day
Low	The lowest price the index has reached on this day
Vol	The volume traded on this day
Price	Is the price of the index on this day

The goal of our algorithm is to predict the price using the other 5 features (open, close, high, low, volume, and change)

2.4 Data Preprocessing

After collecting the data we need to look through the data and find any missing or duplicated values and scale the data to be consistent to ensure the smoothness of the model and ensure no biased results occur starting by removing all the duplicate data and duplicated data, and scaling the volume as the volume is calculated in millions, and any inconsistent data is corrected, and then all the data are saved in CSV file to be ready for the model to be prepared for training. The training position of the data will be 80% and 20% of the data was treated as testing data while using cross-validating to train the model.

2.5 Machine Learning Algorithms

In this section we start our analyses using different techniques that we mentioned before are(Decision Tree (DT), Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), AdaBoost, Multiple layer perceptron (MLP), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and K Nearest Neighbour (KNN)), depending on the tool that we are going to use the data will be affected and significantly [36, 37].

We will be using python, an easy syntax scripting language that will make it easy to maintain, using python packages for the algorithms that we need will provide security and reusability of the script for a different set of data, and can be used by organizations, or individuals [38].

As mentioned in the previous phase we split the data into two data sets one for training with 80% of the data and a testing data set with 20% of the data. Even though we are using the same dataset but changing the algorithms changes the results as well, as we mentioned before we are using Decision Tree (DT), Support Vector Machine (SVM), XGBoost, AdaBoost, Multiple layer perceptron (MLP), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and K Nearest Neighbour (KNN)[38].

Decision Tree (DT) is a decision support tool, that can be used to support a decision-making process based on a tree-like model that is built upon the possible consequences It changes according to the features that will be used in the training set for an instance equation (1) represents the using of one attribute and equation (2) represents the using of two attributes

Equation (1)

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Equation (2)

$$E(T, X) = \sum_{c \varepsilon X} p(c)E(c)$$

As it works on the Entropy, where the model is built top-down from a root node and it shatters the data into subsets that have instances with the value.[26]

Support Vector Machines (SVM) is a machine learning algorithm that is mainly built for classification, however, it can work for regression and classification problems, and it is a supervised machine learning and its goal is to find a hyperplane that best separates the two classes, doing so by finding the maximum margin such hyperplane is produced through equation (3) [27].

Equation (3)

w.x+b = 0

Where (w) is a vector and (b) is an offset

Extreme Gradient Boost (XGBoost) is a machine learning algorithm that is built on a decision tree as an extent and designed to be highly extensible, according to this the main focus of XGBoost is to focus on decision tree as basic classifier but using various methods to improve and speed up the process and handling the complexity of the tree as shown in equation (4).[40]

Equation (4)

$$l^{(t)} = \sum_{i=1}^{n} \iota(yi, \hat{y}i^{(t-1)} + f_t(X_i)) + \Omega(f_t)$$

As presents it's easy to see that it is a function of functions AdaBoost is a machine learning algorithm that has some advantages such as simplicity, as it doesn't require much adjustment in its parameters, and faster calculation [41] and the equation of this algorithm is presented in equation (5), it also contains a weak classifier that is trained to provide a set of theoretical learning guarantees equation (6)

Equation (5)

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

Equation (6)

$$D_{t+1}(i) = \frac{D_t(i) \epsilon exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Multiple layer perceptron (MLP) is a neural network model that work approximators as universal for any continuous functions [42] and the equation of the model is presented in equation (7) [42],

Equation (7)

$$w_1 x_1 + w_2 x_2 + \ldots + w_n x_n - \theta = 0$$

Where it is the equation of the hyperplane

Random Forest (RF) is a combination of decision trees that is very successful when handling a large dataset yet it is a bit slower but can handle multiple issues in the dataset such as the missing values [43], and it is used to solve classification and regression problems.

Gradient Boosting (GB) is an algorithm that builds a new base learner that correlates to the negative slope and maximum of the loss function through the ensemble, which means that if the error is the squared error loss, then it depends on the learning step (equation (8)), and it's all up to the user[37], and its equation is shown in equation (9).

Equation (8)

Y = ax + b + e

Equation (9)

$$y_i^p = y_i^p + \alpha * \sigma \sum (y_i - y_i^p)^2 \mid \sigma y_i^p$$

Logistic Regression (LR) is named after the core function of the method, the logistic function as it also called the sigmoid function and its equation is presented in equation (10), it also represents an S-shaped curve that map values between but not exactly 0 and 1 [44].

Equation (10)

$$y = \frac{e^{(b_0+b_1*x)}}{1 + e^{(b_0+b_1*x)}}$$

K Nearest Neighbour (KNN) is considered a lazy learning method, and it is one of the simplest machine learning algorithms as it relies on distance metrics as shown in equation (11), the better these metrics reflect similarity the better the classification will be, and it can run on multiple choices, the most common one is Minkowski distance, and it is an algorithm that does not require an explicit training [45]

Equation (11)

$$dist(x,z) = \left(\sum_{r=1}^{d} |x_r - z_r|^p\right)^{\frac{1}{p}}$$

2.6 Evaluation Metrics

After using different machine learning algorithms we need a way to assess the model efficiency in predicting, we have picked few methods that will help assess the model's prediction. In this section, we will go over all the measurements that we have picked. Our picks depend on the availability to be executed on the model.

These are the main components for any evaluation metrics, because depending on accuracy only is not a valid metric to be considered [46].

- Precision is one of the evaluations that we are going to use and it is calculated as the number of true positives divided by the total number of true and false positives as shown in equation (12), it calculates a ratio of the correctly predicted values.

  Equation (12)

  $$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall is another evaluation measurement that will be used in this research to determine the model correctness, as it represents the total number of correct positive predictions made out of all positive predictions that could have been made, and it differs from precision as it takes in consideration all positive predictions as it also an indication of missed positive predictions, as represented in equation (13).

  Equation (13)

  $$\text{Recall} = \frac{TP}{TP + FN}$$

- F-Measure some also call it F-Score or F1-score and it works on evaluating binary classification system to find the accuracy of a model's accuracy on a dataset as it classifies examples into positive and negative and it depends on Precision and Recall to calculate it as presented in equation (14).

  Equation (14)

  $$\text{F-Measure} = \frac{2 * Precision * Recall}{Precision + Recall}$$

- Accuracy which is called a proportionate classification as well, and it measures the percentage of the data that classified correctly its equation is in equation (15). It represents how much the proposed model is to the ground truth [47].

  Equation (15)

  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

3. EXPERIMENTS

As mentioned in the previous section we have chosen some machine learning algorithms to implement on the EGX100 dataset and after that, we compare the results through the evaluation metrics that we have mentioned in the previous section, these studies have been carried out on a PC with windows 10 operating system and with 1 processor 1 with Intel® Core ™ i5-10400F CPU @ 2.90GHz 2.90 GHz Processor, and 32 GB RAM.

Before diving into the experiment results and analysis, we will be looking at the data, starting with the coefficient matrix of our variables as it is represented in fig2.

Most of the coefficients are near the one which suggests the importance of the variables that we have picked and they are all positively related to the price and each other however the vol is the only one that is negative and that is because of the huge difference between vol and the other variables, which requires scaling the data [48].

We will be using data normalize with the equation (16), so there is not much difference for the model and the data could be fitted into our models without any issues.

Equation (16)

$$Y = \frac{(x - min)}{(max - min)}$$

We also have drawn the daily price of EGX100 all over the year in fig.3 to represent the chaotic nature of the data and to make sure there is no seasonality effect or any other types of effects such as COVID-19 [35].

3.1 Experimental Result

Running the models and assessing the accuracy over the data we have split the data into a training and test dataset with 80% for training and 20% for the test, and using the evaluation metrics that have been mentioned before we have got the results that represented in Table 2

Table 2 Results of the machine learning algorithms

classifier	DT	SVM	XGBoost	AdaBoost	MLP	RF	GB	LR	KNN
Precision	0.40816	0.26030	0.75552	0.26030	0.23990	0.41836	0.56150	0.94533	0.45102
Recall	0.42857	0.51020	0.53061	0.51020	0.48979	0.48979	0.55102	0.93877	0.48979
F1-	0.4	0.5	0.5	0.5	0.4	0.4	0.5	0.9	0.4

score	2857	1020	3061	1020	8979	8979	5102	3877	8979
Accuracy	0.42857	0.51020	0.53061	0.51020	0.48979	0.48979	0.55102	0.93877	0.48979

For this results to be shown we have changed the prediction problem to be a classification problem just following most of the literature review such as [31], [32]. Our classification was following a daily prediction to suit the type of data so if we predict the next day will be a higher index value we will buy otherwise we sell.

In table 2 we can see clearly that the LR algorithm is the best for such types of data as it got the highest score in all of our metrics as it scored 0.945 in precision and 0.938 in the recall, 0.938 in f1-score, and an accuracy of 0.938 it followed by the XGB with 0.755 which is a huge difference in precision, but GB is the following in all other metrics but huge gap as well.

B. Identify the Headings

Using another package the Yellowbrick [49] for reporting the results of the experiment we got the following results, which can be represented in figs, 4, 5, 6, 7, 8, 9, 10, 11, 12

As mentioned in the previous section LR is the best by far in all of our metrics as it bypasses all the other machine learning algorithms by huge far when using only these features.

Most of the models aside from DT and KNN have gotten a score of around 50% for all metrics with one exception of the XGBoost, which is even higher than the neural network model.

3.2 Conclusion & Further Work

In this paper we have investigated different machine learning algorithms on the EGX100 dataset, the machine learning are in Python Language. We have used nine machine learning algorithms and evaluated the results with four different metrics.

We started by cleaning the dataset and looking through the literature and related work we find that they change the prediction from regression to a classification which we followed, doing that by changing the data to 1 and 0 for buying or not if the prediction to predict a higher price we by and if the lower price we sell.

We investigated the variable in most related work and found the usage of Open, Low, High, and Volume is the main variable, and as we looked in the coefficient matrix the variables are related but suggested that we scale the Volume as it comes in very high amount compared with other variables.

Our results suggest that the Logistic regression (LR) algorithm is by far is the best one to be used compared to the other suggested 8 different algorithms as it scored 93% accuracy in prediction, and maybe the data for such a problem should have been more data to make sure that our results apply but if we extend the data we would have to go through the COVID-19 effect [35], so we just wanted to check the algorithms on the data to determine the best and then in the future, we will be taking into consideration such effect and add more data for better results.

4. FIGURES/CAPTIONS

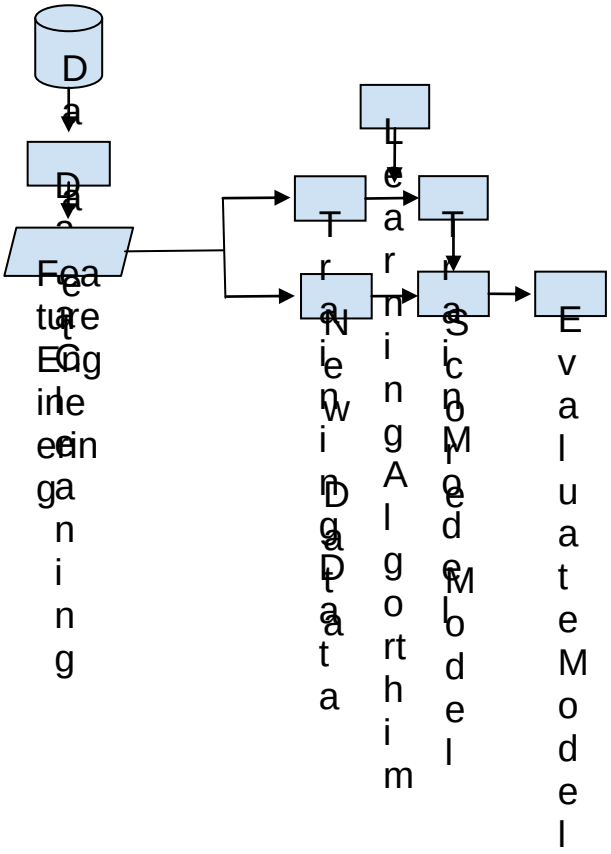

Fig (1):

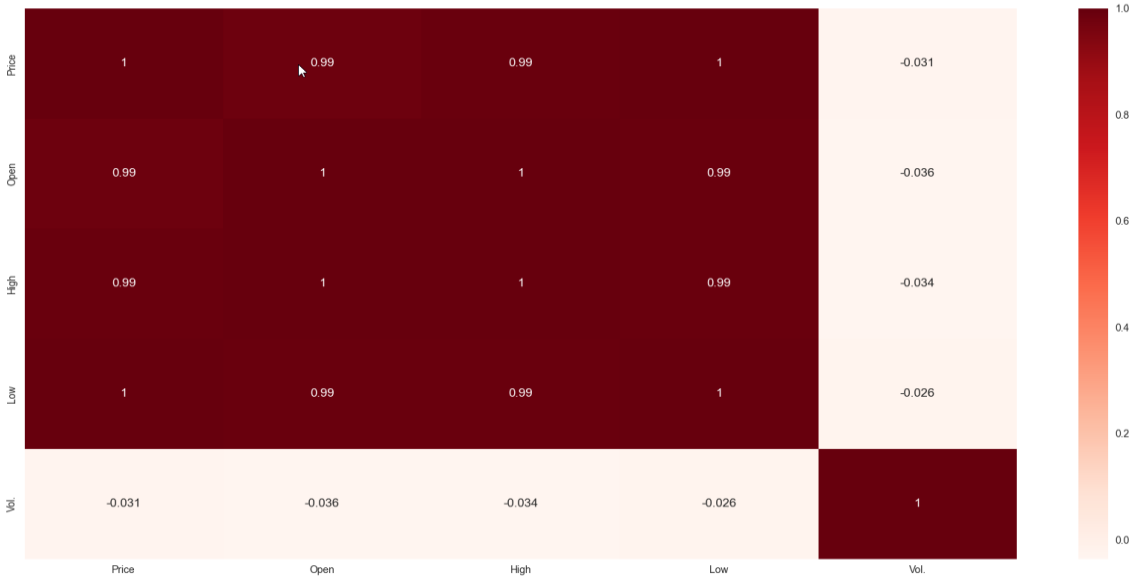

Fig (2):

Fig (3):

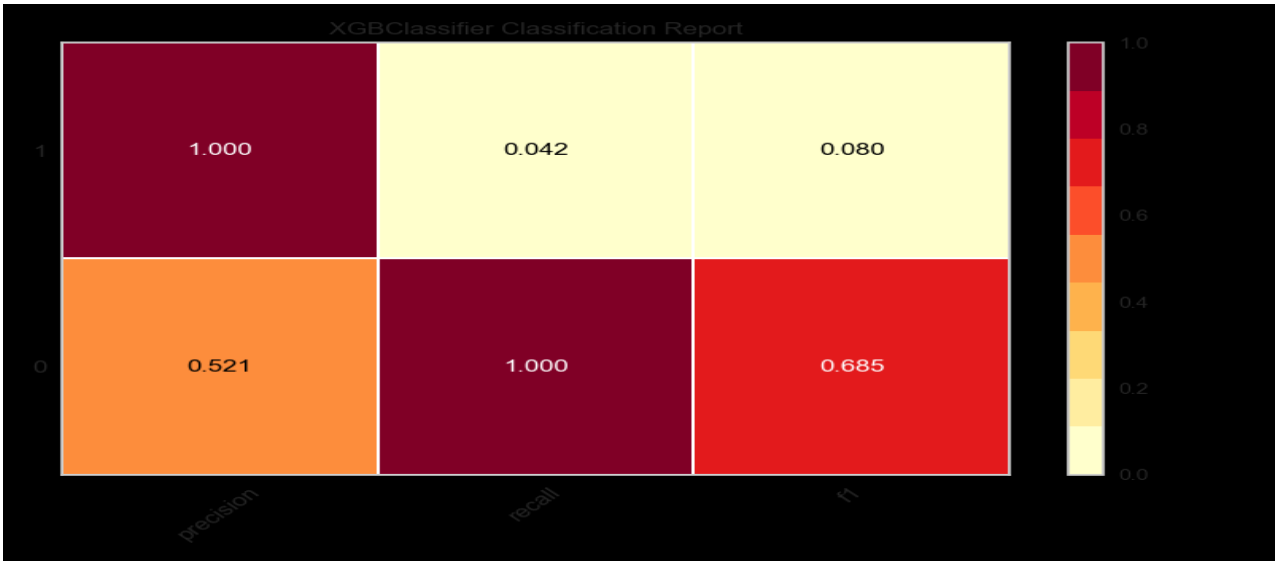

Fig (4): XGBoost Classification report

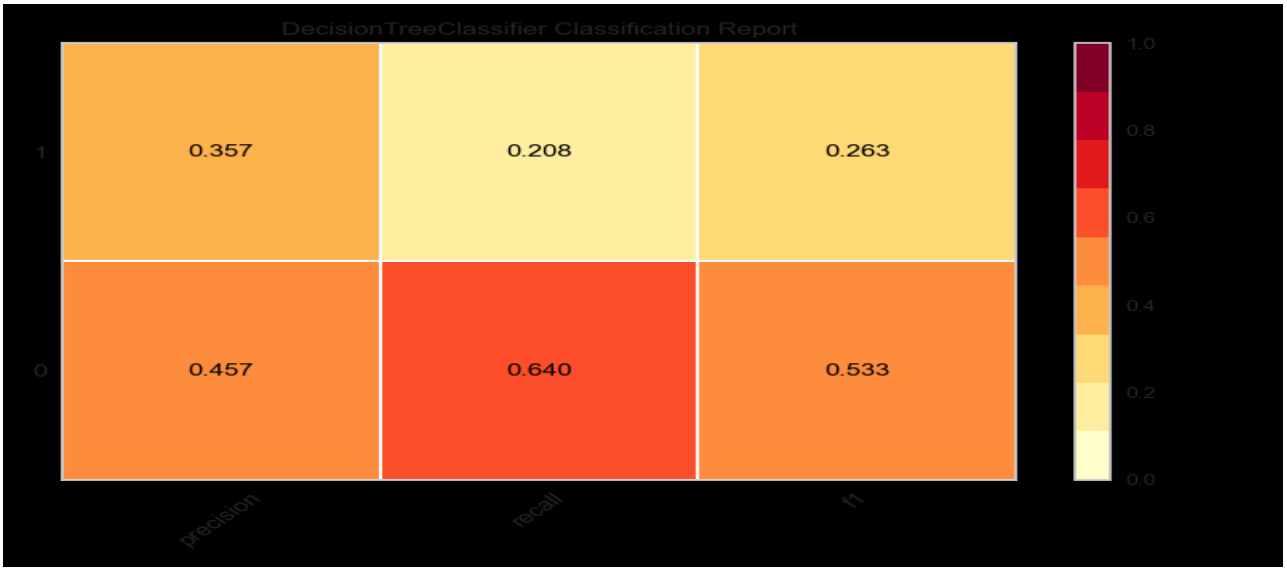

Fig (5): DT Classification report

Fig (6): GB Classification report

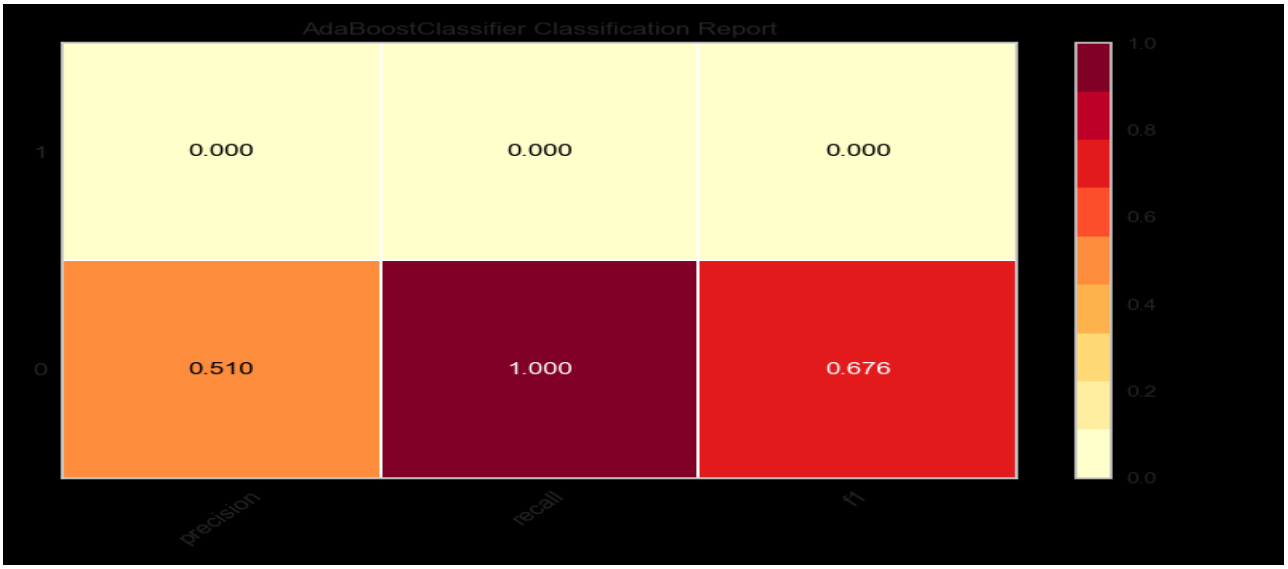

Fig (7): Ada Classification report

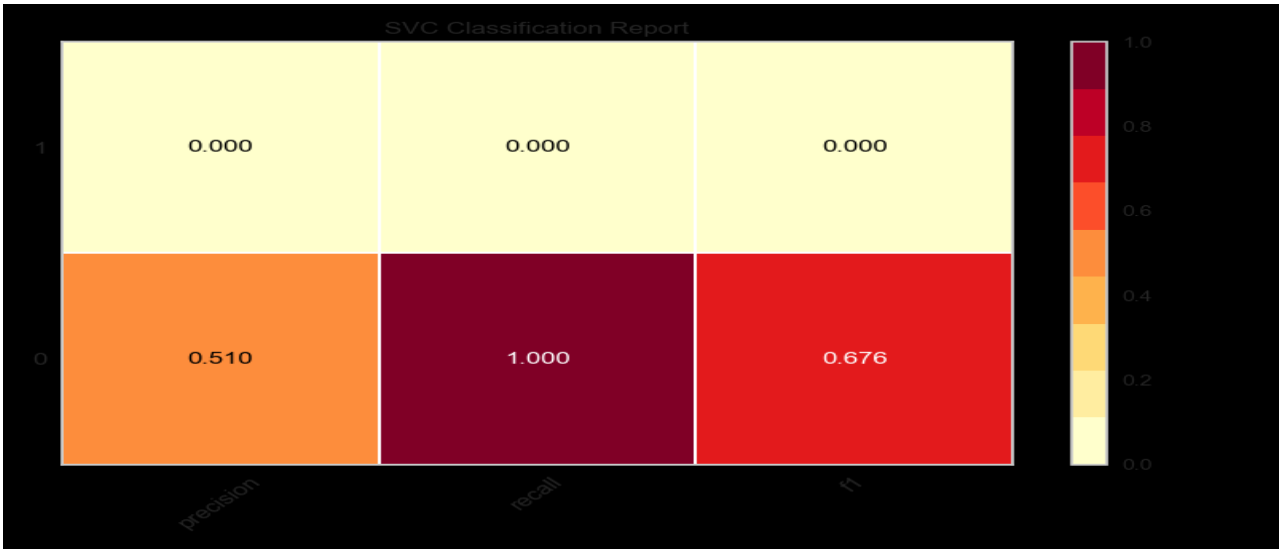

Fig (8): SVC Classification report

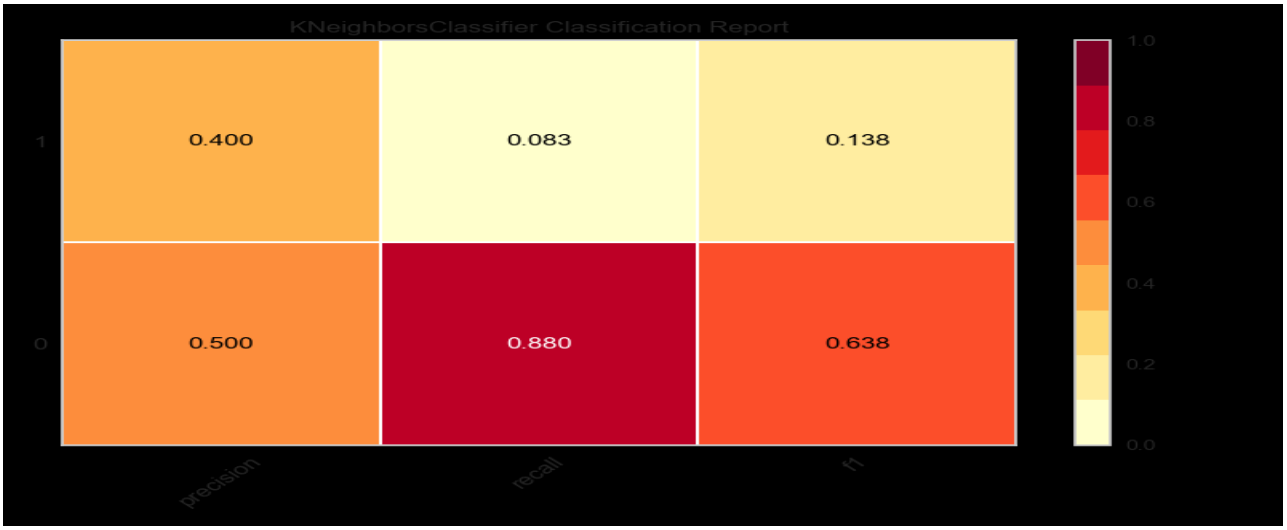

Fig (9): KNN Classification report

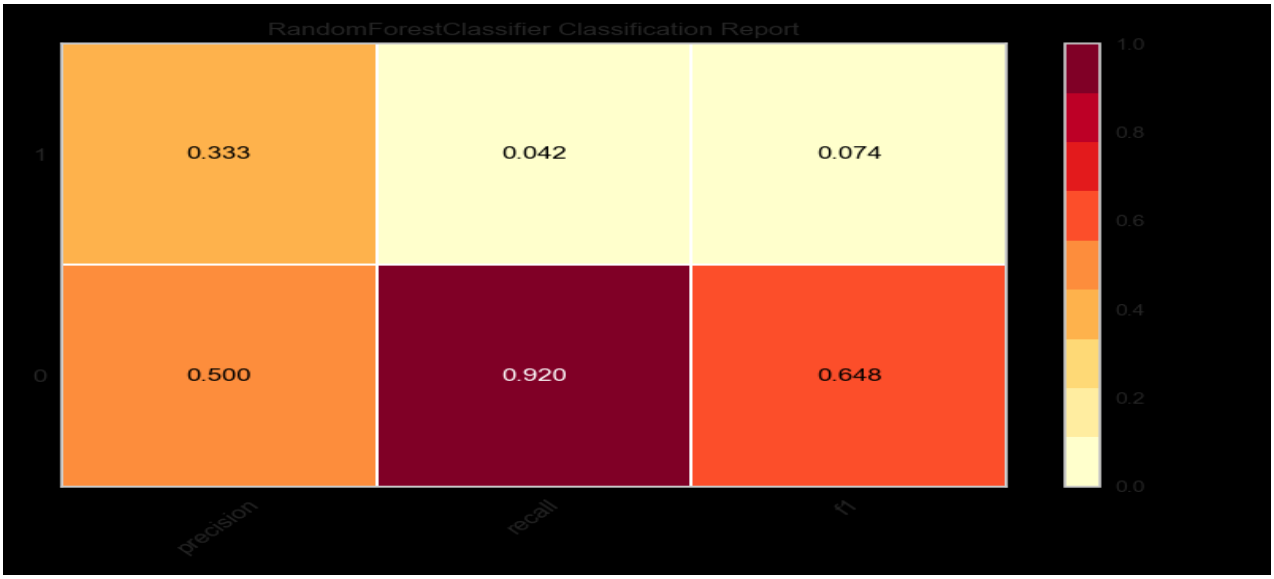

Fig (10): RF Classification report

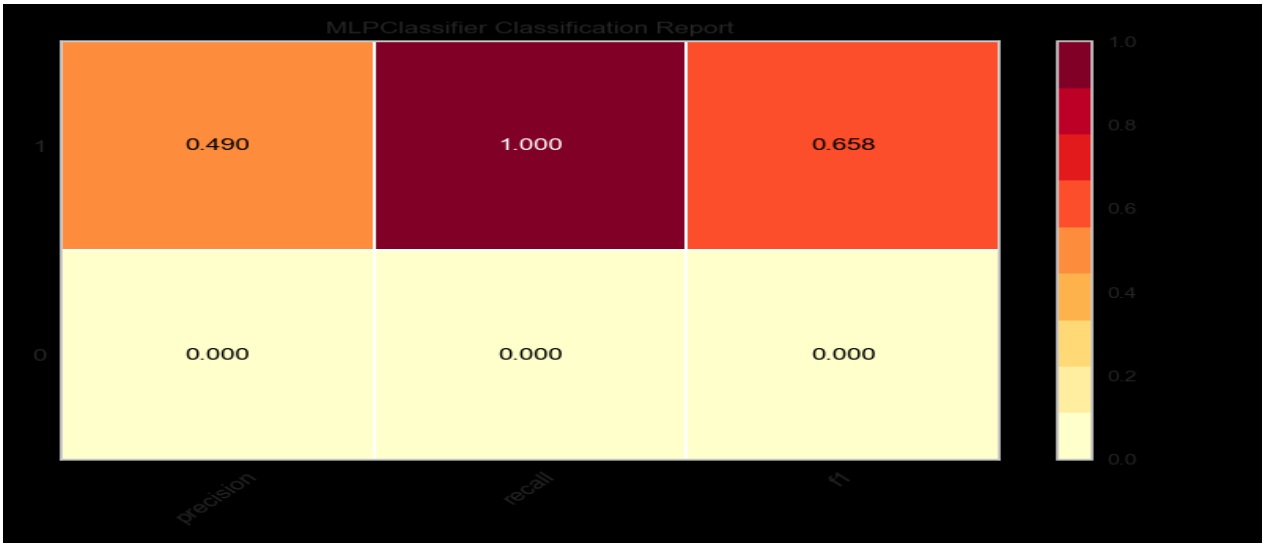

Fig (11): MLP Classification report

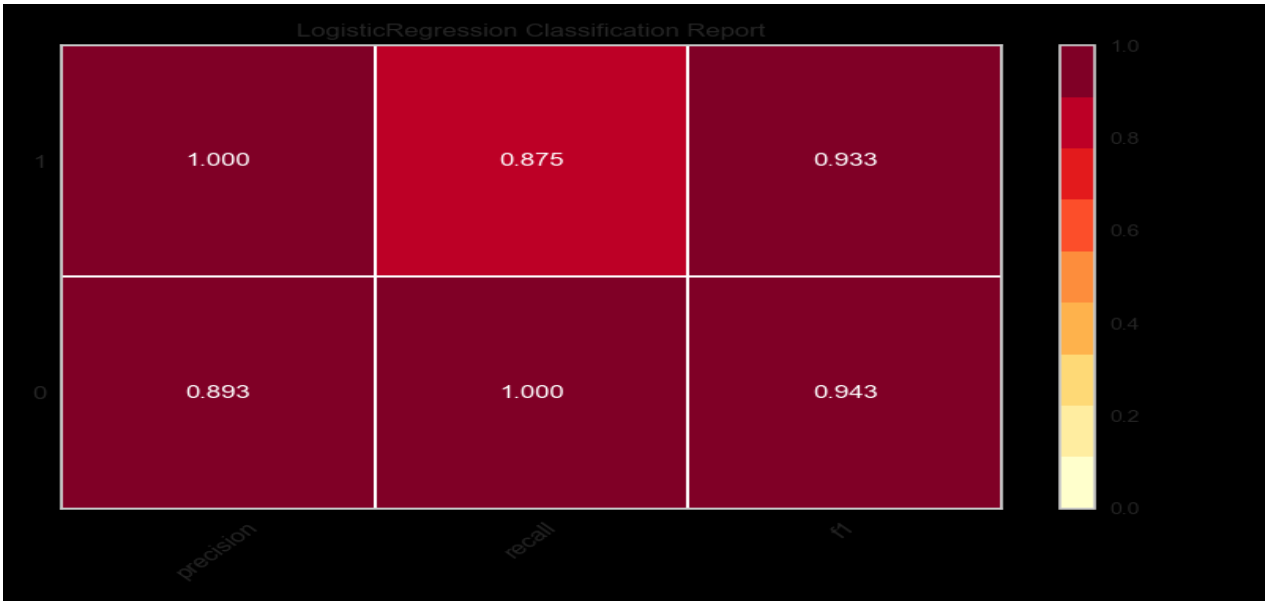

Fig (12): LR Classification report

5. REFERENCES

[1] Boivin, J., Kiley, M. T., & Mishkin, F. S. (2010). How has the monetary transmission mechanism evolved over time?. In *Handbook of monetary economics* (Vol. 3, pp. 369-422). Elsevier.

[2] Ahmed, A. A., Abd El-Baky, M. M. H., Zaki, M. F., & Abd El-Aal, F. S. (2011). Effect of foliar application of active yeast extract and zinc on growth, yield and quality of potato plant (Solanum tuberosum L.). *Journal of Applied Sciences Research*, *7*(12), 2479-2488.

[3] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251.

[4] Malkiel, B. G., & Fama, E. F. J. T. (1970). Efficient capital markets: A review of theory and empirical work" nn, 25 (2), 383-417. *DOI: https://doi. org/10.1111/j*, 1540-6261.

[5] Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, *79*, 153-163.

[6] Kumar, D., Meghwani, S. S., & Thakur, M. (2016). Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. *Journal of Computational Science*, *17*, 1-13.

[7] Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with applications*, *36*(3), 5932-5941.

[8] Malkiel, B. G. (2003). Passive investment strategies and efficient markets. *European Financial Management*, *9*(1), 1-10.

[9] Fama, E. F. (1991). Efficient capital markets: II. *The journal of finance*, *46*(5), 1575-1617.

[10] Benjamin, H. S. (1942). The dow theory of stock prices. *Social Research*, 204-224.

[11] (Chen et al., 2017; Zhong and Enke, 2017)

[12] (Chen, Cheng, and Tsai, 2014, pp. 329-330)

[13] (Cavalcante, Brasileiro, Souza, Nobrega, and Oliveira, 2016, p. 194),

[14] (Kumar and Thenmozhi, 2014; Wang, Wang, Zhang, and Guo, 2012, p. 285; p. 758)

[15] (Chen et al., 2017, pp. 340–341)

[16] Chiang, Enke, Wu, and Wang (2016)

[17] Hsu, Lessmann, Sung, Ma, and Johnson (2016, p. 215).

[18] L. K. Morrison, et al., "Utility of a rapid Bnatriuretic peptide assay in differentiating congestive heart failure from lung disease in patients presenting with dyspnea," Journal of the American College of Cardiology, vol. 39, pp. 202-209, 2002.

[19] Forecasting EGX30 index time series using vector autoregressive models VARS

[20] Ezzat, H. M. (2021). Principal component regression for egyptian stock market prediction. *The International Journal of Informatics, Media and Communication Technology*, *3*(1), 23-39.

[21] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251.

[22] Elwasify, A. I. (2015). A combined model between Artificial Neural Networks and ARIMA Models. *Int J Recent Res Commer Econ Manag*, *2*(2), 134-140.

[23] Houssein, E. H., Dirar, M., Hussain, K., & Mohamed, W. M. (2021). Assess deep learning models for Egyptian exchange prediction using nonlinear artificial neural networks. *Neural Computing and Applications*, *33*(11), 5965-5987.

[24] Xiao, Y., Xiao, J., Lu, F., & Wang, S. (2014). Ensemble ANNs-PSO-GA approach for day-ahead stock e-exchange prices forecasting. *International Journal of Computational Intelligence Systems*, *7*(2), 272-290.

[25] Kamble, R. A. (2017, June). Short and long term stock trend prediction using decision tree. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1371-1375). IEEE.

[26] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). Efficient stock-market prediction using ensemble support vector machine. *Open Computer Science*, *10*(1), 153-163.

[27] Er, X., & Sun, Y. (2021, July). Visualization Analysis of Stock Data and Intelligent Time Series Stock Price Prediction Based on Extreme Gradient Boosting. In *2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)* (pp. 272-279). IEEE.

[28] Wu, Y., & Gao, J. (2018). AdaBoost-based long short-term memory ensemble learning approach for financial time series forecasting. *Current Science*, *115*(1), 159-165.

[29] Devadoss, A. V., & Ligori, T. A. A. (2013). Forecasting of stock prices using multi layer perceptron. *International journal of computing algorithm*, *2*(1), 440-449.

[30] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

[31] Dey, S., Kumar, Y., Saha, S., & Basak, S. (2016). Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting. *PESIT South Campus*.

[32] Ali, S. S., Mubeen, M., Lal, I., & Hussain, A. (2018). Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX). *Asian Journal of Empirical Research*, *8*(7), 247-258.

[33] Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics and its Applications*, *477*, 161-173.

[34] Ezzat, H. M. (2021). The effect of COVID-19 on the Egyptian exchange using principal component analysis. *Journal of Humanities and Applied Social Sciences*.

[35] B. Ratner, Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data: CRC Press, 2017

[36] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Frontiers in neurorobotics, vol. 7, p. 21, 2013.

[37] H. Hu, et al., "A comparative study of classification methods for microarray data analysis," in Proceedings of the 5th Australasian Data Mining Conference (AusDM 2006): Data Mining and Analytics 2006, 2006, pp. 33-37.

[38] S. Sangeetha and S. Saradhambekai, "Python Libraries and Packages for Data Mining-A Survey," 2019.

[39] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[40] T. Chengsheng, et al., "AdaBoost typical Algorithm and its application research," in MATEC Web of Conferences, 2017, p. 00222.

[41] Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. In *Advances in computers* (Vol. 117, No. 1, pp. 339-368). Elsevier.

[42] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," 2005

[43] M. Maalouf, "Logistic regression in data analysis: an overview," International Journal of Data Analysis Techniques and Strategies, vol. 3, pp. 281-299, 2011.

[44] D. Cheng, et al., "kNN algorithm with datadriven k value," in International Conference on Advanced Data Mining and Applications, 2014, pp. 499-512.

[45] M. Sokolova, et al., "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in AI 2006: Advances in Artificial Intelligence, ed: Springer, 2006, pp. 1015-1021.

[46] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," Communications Surveys & Tutorials, IEEE, vol. 10, pp. 56-76, 2008.

[47] Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

[48] Bengfort, B., & Bilbro, R. (2019). Yellowbrick: Visualizing the scikit-learn model selection process. *Journal of Open Source Software*, *4*(35), 1075.