Machine Learning Techniques for Aspect Analysis of Employee Attrition

Anamika Hooda[1(✉)], Purva Garg[1], Nonita Sharma[2], and Monika Mangla[3]

[1] Department of Electronics & Communication Engineering (AI), Indira Gandhi Delhi Technical University for Women, Delhi, India
{anamika107bteceai21,purva117bteceai21}@igdtuw.ac.in

[2] Department of Information Technology, Indira Gandhi Delhi Technical University for Women, Delhi, India
nonitasharma@igdtuw.ac.in

[3] Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

Abstract. Employee attrition is the reduction in the employee workforce, which can be defined as the rate of employees leaving the company faster than the rate they are hired. Attrition may be for the whole establishment but sometimes it might be particular for a business field. This happens when there is intervention of technology that contribute in replacing the human workforce. There are several factors contributing to employee attrition, a few being age, number of years in the company, manager, technology change, etc. It is vital to understand the impact of these factors on employee attrition so that necessary action can be taken to avoid this. Thus, Machine learning technique is being used nowadays to inspect and predict the data of several real-life applications. After employing the models, authors performed the analysis on each of them using confusion matrix, F-1 score, recall, precision, etc., and found that the best model is SVM with an accuracy of 85.60%.

Keywords: Attrition · Data Analytics · SVM · Logistic Regression · Heatmap · Correlation Matrix

1 Introduction

Employees are one of the most significant asset of any organization. The longevity of employees' retention in any organization advocates a healthy work environment. On the contrary, frequent resignations of the employees can be attributed to various reasons. Thus the employees' retention can be directly considered as a measure of satisfaction of employees towards work place. Therefore, it is needed to understand the employees' attrition. Here, employee attrition is the decrease in number of employees [1, 2]. This means that employees leave their organizations before the new staff is hired. Employee attrition can be for the whole company but it can also be particular to a business field. It may be also because the skill sets that are needed change constantly [3].

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 286–296, 2023.
https://doi.org/10.1007/978-3-031-35081-8_23

More the employee attrition, the more no of workers we need to hire but recruiting employees can be a difficult job. Getting able employees and training them for the job, and after some time they resign is frustrating for the employer. Hence, human resource (HR) managers are there to recognize the factors leading to attrition and take the preventive and corrective measures for reducing attrition. The main focus of this study is to present the aspect analysis of attrition in terms of:

1) Factors leading to employee attrition
2) How it can be decreased
3) Accuracy of four machine learning models used employed in the prediction of attrition.

The models in this project are;

a) Logistic Regression
b) Naïve Bayes
c) SVM Model
d) Random Forest

Logistic Regression is a one of the statistical models often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring; the next model being the Naïve Bayes algorithm, which is a supervised learning algorithm, based on the Bayes theorem aiming to solve classification problems.

Support vector machines are a set of supervised learning methods used for classification, regression, and outlier detection. Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems [4, 5].

Authors have employed the dataset of employee attrition from Kaggle [6] to perform various analyses on it. This dataset consists 1470 rows and 35 columns out of which 34 factors are showing their dependency on attrition. Attrition is denoted as Yes/No. There is no null value in our dataset. It has string and integer as data types.

Before employing machine learning models, the research work utilized a heatmap for finding correlations between factors and removing the appropriate ones. After employing the models, the analysis is performed on each of them using confusion matrix, F-1 score, recall, precision, etc. and the results concluded that the best model is SVM with an accuracy of 85.60%.

The current work is organized into various sections. Section 1 introduces the concept of workforce attrition. Similar problems have been undertaken by various researchers. The noticeable research by few researchers has been presented in Sect. 2. The methodology proposed by authors has been elaborated in Sect. 3. Results are discussed in Sect. 4 while the conclusion is presented in Sect. 5.

2 Related Work

The significance of employee attrition has been realized since many years and resultantly several authors have tried their hand to understand the association behind. For instance, authors in [7] considered a sample of 297 employees in a private firm. The authors considered various factors such as firm characteristics, location, employee benefits, and

work culture etc. Authors performed hierarchical regression analyses and concluded that firms with high benefits and high benefits packages observed slight employee attrition.

The authors in [3] also proposed a model to find association among service and employee attrition. For the same, authors considered data from 64 business units. Authors also determined that a high attrition (specifically for customer-facing employee) will adversely impact the relations with existing customers. This may also impact the revenue generation of the business units. Aligning with this research, authors in [1] also believe that the employees who are leaving from an organization are carrying some tacit knowledge with them which could be advantageous to the competitors. Hence, the organizations must strive to retain the employees and minimize employees' attrition. In order to validate the belief authors considered 309 employees who had left their organization during 1978 to 2006. This dataset was classified into some predefined attrition classes using decision tree models and rule-sets. The classification produced by these 2 classifiers were used for a predictive model so as to predict upcoming employee attrition.

Considering the interest to understand the pattern of employee attrition, machine learning (ML) was also employed for the same considering the demonstrated supremacy of machine learning algorithms. Hence, authors in [2] performed a study to understand employee attrition using ML models. For the same, authors used a synthetic data created by IBM Watson of employees' attrition, an imbalanced dataset. This IBM Watson dataset is processed using support vector machine (SVM), random forest and K-nearest neighbor (KNN). Further, authors also used adaptive synthetic (ADASYN) to address class imbalance. During all experimental setup, the machine learning models demonstrated its efficiency and effectiveness to predict the employee attrition.

Further, authors in [8] also aimed to determine the association between manager's interpersonal skills and employees' attrition. As per the finding by authors in [8], management skills have a strong impact on employee turnover.

Thus it is still to debate that what all factors are there which closely impact the employee attrition and hence needs an extensive research in this domain.

3 Methodology

The methodology is depicted in Fig. 1. As evident from Fig. 1, authors have used the IBM Watson dataset of employee attrition from Kaggle [6]. The considered dataset is preprocessing to drop null values and unwanted attributes. Authors converted categorical data with the help of a label encoder [9, 10]. Four models were employed i.e., naïve Bayes, random forest, SVM, and logistic regression, and a heatmap was plotted to determine the correlation between the factors. Analysis of models was done by confusion matrix and then they were evaluated by classification report and accuracy score [11]. Graphs and tables were plotted to get the result and determine the most important attributes for attrition. The following steps are employed in the methodology:

1. Data Understanding: The dataset contains the data of 1470 employees. There are 35 attributes present in the database some are – Job level, Percent hiking rate, Years with current manager, Years since promotion etc. There is no null value in our dataset. It has string and integer as data types.

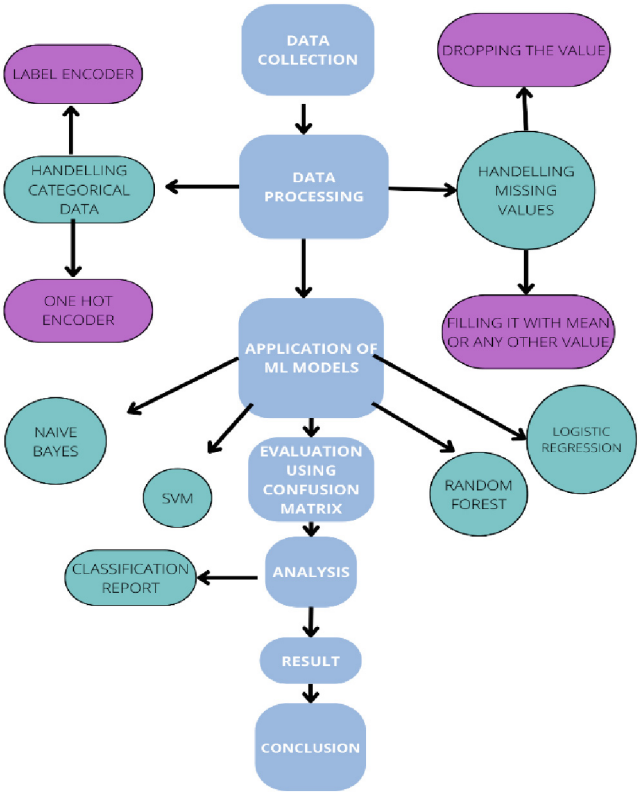


Fig. 1. Flowchart depicting the Methodology

2. Data Pre-processing: The attributes with string datatype were converted to integer datatype with the help of a label encoder. A heatmap is plotted to check the correlation between attributes and the attributes with higher correlation were removed as shown in Fig. 2. Attributes with no correlation to attrition were also dropped. These attributes were Employee number, employee count, over 18, standard hours, percent salary hike, years in a current role, and total working years. With the help of heatmap, we found the Attributes with no correlation to attrition and the ones which are highly correlated and modified the dataset accordingly.

3. Modelling: The Naïve Bayes, Logistic regression, SVM, and Random Forest algorithms are employed.

4. Analysis: The confusion matrix is used to check how many times the prediction was these models [12, 13]. We also used a classification report to check the statistical results of the models used. Following metrics were used to make the analysis, Here TP

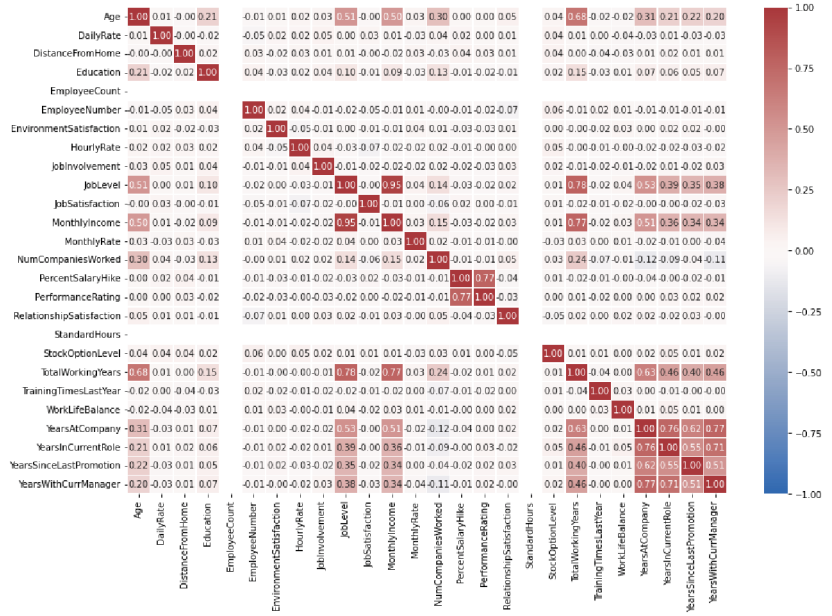


Fig. 2. Heatmap representing the correlation of Attributes

represents the true positive values, TN represents the true negative values, FP represents False Positive values and FN represents False Negative values;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

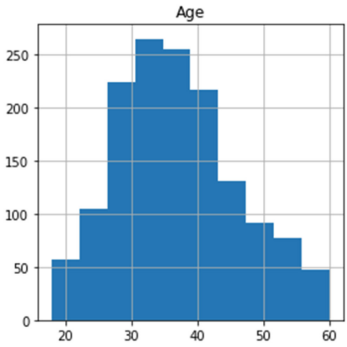
$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

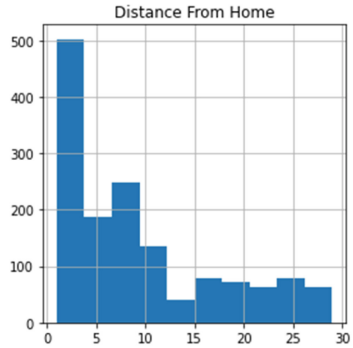
5. Result and conclusion: After performing the analysis, we get results and conclusions regarding the problem statement.

4 Result and Discussion

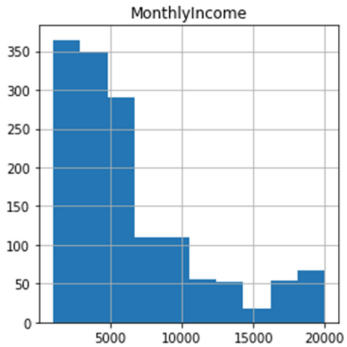
Ahead of employing the ML models, correlation among various attributes is evaluated. This correlation is illustrated in the heat map shown in Fig. 2. After employing the models, we performed the analysis on each of them using the confusion matrix, F-1 score, recall, precision, etc. The section is divided into two subsections namely Aspect analysis using visualization and Comparative analysis of ML models.



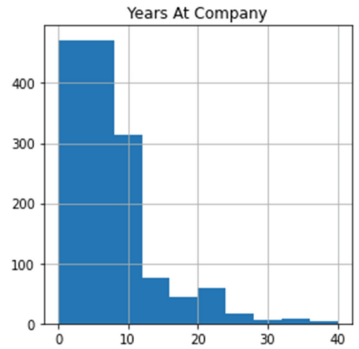
(a) Attrition vs Age



(b) Attrition vs Distance from Home

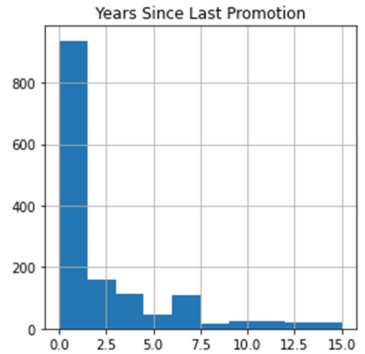


(c) Attrition vs Monthly Income



(d) Attrition vs Years at Company

(e) Attrition vs Years with current Manager



(f) Attrition vs Years since Promotion

Fig. 3. Aspect Analysis of Attrition with various factors

4.1 Aspect Analysis Using Visualization

Aspect analysis means analyzing the data with the help of various plots and graphs [14–16]. We plotted various graphs to analyze attributes and understand our data.

Impact Analysis of Factors on Attrition
Figure 3 represents the factors affecting Employee Attrition. From the graphs below it is conclude that attrition is maximum among people aged between 30 to 40, living within 10 km, and with a monthly income of less than 6k. People working at the company for less than 12 years, under the same manager for 8 years, and who haven't been promoted for two years are most likely to leave the company. Figure 4 visualizes attrition concerning gender. Here, it can be safely concluded that the attrition among males is 25.6% more

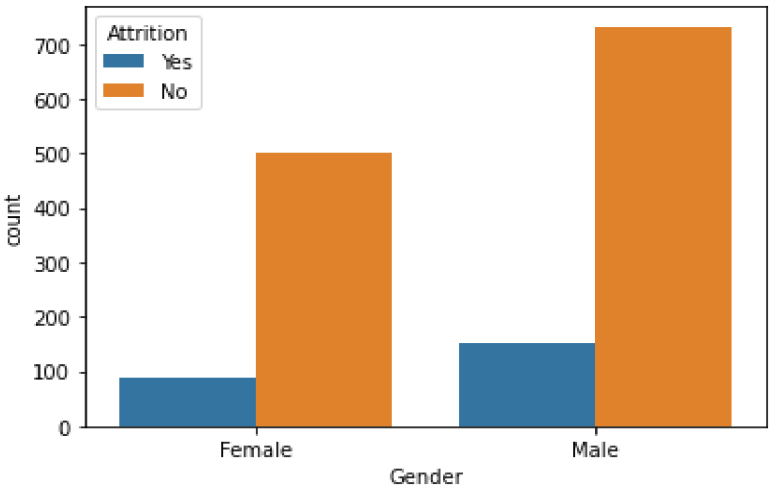


Fig. 4. Aspect Analysis of Attrition with respect to Gender

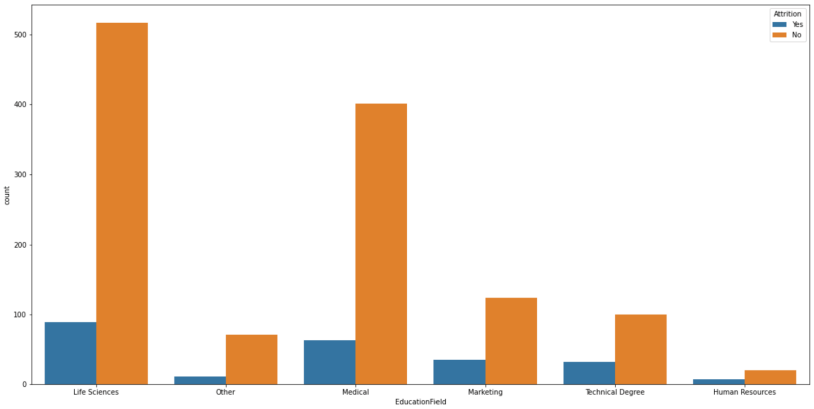


Fig. 5. Aspect Analysis of Attrition with respect to Education Field

than among females. Further, Fig. 5 presents the impact of the education field on attrition. Here, we can establish the fact that attrition is most in Lifesciences as compared to other fields and HR has the least attrition. Furthermore, Fig. 6 determines the impact of salary hikes on attrition. Percent salary hike between 11 to 14% leads to more attrition and between 19 to 25%, it's the least.

4.2 Comparative Analysis of Various Machine Learning Models

Table 1 displays us the confusion matrix for four models. There are more true positives than false positives and fewer false negatives than false positives [17, 18]. Table 2 displays us the contrast in the classification report of all the models. So, from Table 2 and the bar graph representation in Fig. 7, we can see that with the Logistic Regression method the performance score as the objective variable was 84.98% accuracy, naïve Bayes method the performance score as the objective variable was 82.04% accuracy, SVM method the performance score as the objective variable was with 85.60% accuracy and Random Forest method the performance score as the objective variable was with 85.19%. So, from the above results, we conclude that the best model is SVM with an accuracy of 85.60%.

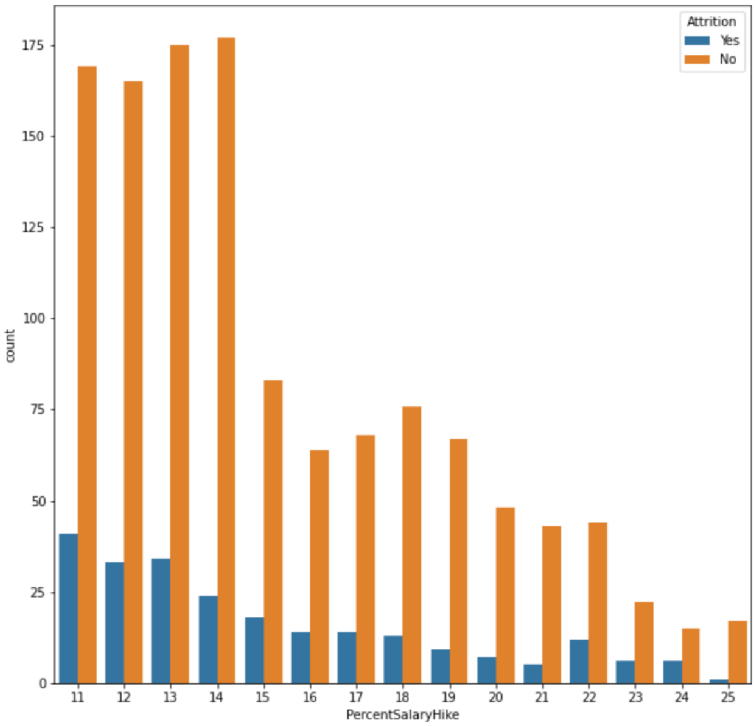


Fig. 6. Aspect Analysis of Attrition with respect to Present Salary Hike

Table 1. Results in terms of the Confusion Matrix of the models

Models		Positive Values	Negative Values
Logistic regression	True	412	4
	False	69	1
Naïve Bayes	True	370	46
	False	40	30
SVM	True	416	0
	False	70	0
Random Forest	True	407	9
	False	63	7

Table 2. Classification Report of various models

Models		Precision	Recall	F1 Score	Support
Logistic regression	0	0.86	0.99	0.92	416
	1	0.20	0.01	0.03	70
	Macro avg	0.53	0.50	0.47	486
	Weighted avg	0.76	0.85	0.79	486
Naïve Bayes	0	0.90	0.89	0.90	416
	1	0.39	0.43	0.41	70
	Macro avg	0.65	0.66	0.65	486
	Weighted avg	0.83	0.82	0.83	486
SVM model	0	0.86	1.00	0.92	416
	1	0.00	0.00	0.00	70
	Macro Avg	0.43	0.50	0.46	486
	Weighted avg	0.73	0.86	0.79	486
Random Forest	0	0.87	0.98	0.92	416
	1	0.44	0.10	0.16	70
	Macro avg	0.65	0.54	0.54	486
	Weighted avg	0.80	0.85	0.81	486

Fig. 7. Comparative Analysis of Machine Learning Models concerning Accuracy

5 Conclusion

Attrition is unavoidable; its presence is always there. Though, we can find ways to reduce it. Turnover is a costly drain on company resources. Internal attributes are equally responsible, if not more important than external attributes in case of attrition. Constructive leadership can help in the betterment of attrition. From the evaluation, the accuracy of the best-proposed model, SVM is 85.60%. It indicates that the SVM technique is very good at predicting attrition. The SVM model is significant to support the decision-making process and can be efficiently used for improving employee attrition.

References

1. Alao, D.A., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. **4**(1) 17–28 (2013)
2. Alduayj, S.S., Rajpoot, K.: Predicting employee attrition using machine learning. In: 2018 International Conference on Innovations in Information Technology (it), pp. 93–98. IEEE (2018)
3. Subramony, M., Holtom, B.C.: The long-term influence of service employee attrition on customer outcomes and profits. J. Serv. Res. **15**(4), 460–473 (2012)
4. Yadav, S., Sharma, N.: homogenous ensemble of time-series models for Indian stock market. In: Mondal, A., Gupta, H., Srivastava, J., Reddy, P.K., Somayajulu, D.V.L.N. (eds.) BDA 2018. LNCS, vol. 11297, pp. 100–114. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04780-1_7
5. Verma, U., Garg, C., Bhushan, M., Samant, P., Kumar, A., Negi, A.: Prediction of students' academic performance using machine learning techniques. In: 2022 International Mobile and Embedded Technology Conference (MECON), pp. 151–156. IEEE (2022)
6. https://www.kaggle.com/datasets/patelprashant/employee-attrition
7. Bennett, N., Blum, T.C., Long, R.G., Roman, P.M.: A firm-level analysis of employee attrition. Group Org. Manag. **18**(4), 482–499 (1993)

8. Hoffman, M., Tadelis, S.: People management skills, employee attrition, and manager rewards: an empirical analysis. J. Polit. Econ. **129**(1), 243–285 (2021)
9. Sharma, N., Yadav, S., Mangla, M., et al.: Multivariate analysis of COVID-19 on stock, commodity & purchase manager indices: a global perspective (2020). Preprint (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-68388/v1
10. Mangla, M., Sharma, N., Mohanty, S.N.: A sequential ensemble model for software fault prediction. Innov. Syst. Softw. Eng. **18**, 301–308 (2022). https://doi.org/10.1007/s11334-021-00390-x
11. Sharma, N., Mangla, M., Mohanty, S.N., Pattanaik, C.R.: Employing stacked ensemble approach for time series forecasting. Int. J. Inf. Technol. **13**(5), 2075–2080 (2021). https://doi.org/10.1007/s41870-021-00765-0
12. Banik, S., Sharma, N., Mangla, M., Mohanty, S.N., Shitharth, S.: LSTM-based decision support system for swing trading in the stock market. Knowl. Based Syst. **239**, 107994 (2022)
13. Singh, N., Sharma, N., Sharma, A.K., Juneja, A.: Sentiment score analysis and topic modelling for GST implementation in India. In: Bansal, J.C., Das, K.N., Nagar, A., Deep, K., Ojha, A.K. (eds.) Soft Computing for Problem Solving. AISC, vol. 817, pp. 243–254. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1595-4_19
14. Sadiku, M., Share, A.E., Musa, S.M., Akujuobi, C.M., Perry, R.: Data visualization. Int. J. Eng. Res. Adv. Technol. (IJERAT) **2**(12), 11–16 (2016)
15. Grinstein, U.M., Wise, A.: Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, Burlington (2002)
16. Chen, C.H., Härdle, W.K., Unwin, A.: Handbook of Data Visualization. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-33037-0
17. Sharma, N., Juneja, A.: Combining of random forest estimates using LSboost for stock market index prediction. In: 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 1199–1202. IEEE (2017)
18. Mangla, M., Shinde, S.K., Mehta, V., Sharma, N., Mohanty, S.N.: Handbook of Research on Machine Learning: Foundations and Applications. CRC Press, Boca Raton (2022)