Employee Attrition Analysis Using XGBoost

1st Isha
Department of Computer Science
Graphic Era Hill University
Dehradun, India
ishu19603@gmail.com

2nd Nitin Thapliyal
Department of Computer Science
Graphic Era Hill University
Dehradun, India
thapliyal.nitin@gmail.com

3th Sheetal Solanki
Department of Computer Science
Graphic Era Hill University
Dehradun, India
sheetalsolanki2207@gmail.com

4rd Sangita Papola
Department of Computer Science
Graphic Era Hill University
Dehradun, India
sangitapapola@gmail.com

***Abstract*—In today's time, employee attrition is a major problem faced by organisations. The most precious resource for a firm is its workforce. They are the ones that bring values both quantitatively and qualitatively. A steady decline in the workforce due to resignation, death or retirement is referred to as attrition. It is a matter of concern when employees leave their jobs of better opportunities but this leads the company to face its consequences. Employee turnover and attrition are equivalent. Leaders can avoid it by implementing better retention techniques by mastering this concept. In this paper, we use dataset to analyse attrition and determine the root cause of employee resignation. We did analysis of the training dataset for effective data exploration. Machine Learning algorithms, such as Random Forest Classifier, Logistic Regression, XGBoost and ensemble stacking technique were used. On data exploration and enabling algorithm we discovered that XGBoost classifier with an accuracy of 88.% and precision 89% was the most appropriate algorithm for this dataset. We may also conclude that monthly income, age, distance from home can be a reason behind employees leaving their job and finding better opportunities.**

***Index Terms*—Attrition, Random Forest, Logistic Regression, XGBoost, Exploration.**

I. Introduction

Employee attrition is the term coined for the process of gradual decline in size of organization's workforce over time due to inescapable causes such as employee resignation for personal or professional reasons or death. The rate of employees leaving the organization is relatively higher than the rate of recruitment and such situations when occur are beyond the reach of the employer. An employee would choose to join or resign from a company considering many factors. Some take into consideration variables like salary, department, working environment, distance from home, gender equity, etc while other takes into account personal reasons like maternity, family, health issues. Employee attrition is the unexpected or unpredictable process either voluntarily or involuntarily. The basic purpose of this analysis is to get understanding of and handling employee attrition. Employee attrition is one of the critical aspects for any company because it has an enormous impact on the company's financial and market performance. It becomes a setback for companies when trained, expertise and important employees quit their jobs for better opportunities of growth in career. Higher attrition rate acts as a setback for the company and negatively impacts its performance, market value and reputation. Continual staff turnover can create uncertainties leading to less employees and financial drawbacks for the organization. Attrition not only affects the inside business but also its image and reputation in the eyes of the outside world. Usually employee attrition over a period of time follows the above mentioned as basics and calculate a numeric value in percentage referred as attrition rate.

To compute the attrition rate dividing the number of employees who quit their jobs by the total number of employees that were recruited at the beginning when the period started. For example, let us count employees to know how many of them were there at the initial year of a company, suppose 1,000.Keep a check on how many of the people left the company throughout the year. Suppose there were 250 employees who left the company intentionally or unintentionally. We also view at how many new employees were hired within the year and conduct a final count for employees at the end of the year. For example, let suppose a company hired 500 new employees in that particular year. This concludes that the final headcount of that company would be 1,500 for that year. Now let's look at the average number of employees for that particular year. It comes out as (1000+1500)/2 = 1,250. As we have the average number of employees with this we calculate the percentage of employees who left. The formula for it is **Average number of employees/number of employees who left*100**. Using the above formula attrition rate can be calculated as: (250/1250)*100. The percentage comes out to be 20%.

It is very essential for an organization to minimize its employee attrition to minimum to standout in the competition market. Therefore, it is important for organisations to take necessary steps and opt proper strategies to improve their

finances and productivity. Google's 2018 Annual Diversity Report dedicates a separate section precisely discussing about employee attrition. It specified that black and Latin employee's attrition was way more than the attrition rate in company average by a wide margin. The bulk of workforce are white and Asian. Black employees contribute to just 2.5% of Google's U.S. employee while Asian people are more with a very slight margin of at 3.6%. This clearly shows why measuring and keeping a check on employee attrition of an organization is important and necessary. This led Google to focus on their employees such that to provide justifiable future. Uber is another example which highlights the importance of keeping a check on employee attrition. While Uber has continuously denied having an above average attrition rate but there are not enough evidence to prove it suggest otherwise. When Uber filed for an IPO in 2019, its public disclosure documents highlighted a number of problems of workspace culture. It included a decrease in employee incentives, poor employer's image which can be a major reason for employees who reportedly left the company in the past few years. From the previously mentioned two case studies, it is evident that Google and Uber posses different perspective on attrition. On the one hand Uber views attrition as an inevitable consequence of its policies at workplace, Google on the other hand opts for a more honest and candid approach, Google openly admits that employee attrition is one issue that needs consideration and addressing. In their 2019 Retention Report, Work Institute discovered that preventable attrition surpasses unpredictable attrition in frequency. 6 out of 100 employees left their jobs due to retirement, a change of career or other such reason which can also be expressed as 22%. The report anticipated that the attrition rate would be so high by the end of this year that 35% of employees would leave there present jobs to work in a different company. For the first time in 2018, it was noticed that the number of vacancies crossed the number of unemployed workers.This helped companies drew a conclusion that job satisfaction was not enough to retain the employees. Work-life development came out to be the number one reason for employees leaving jobs, followed by work-life balance and manager behaviour. Also, attrition cannot always be considered negatively. It can also turn out the other way round for businesses where poor performing or weak employees leave the company and this empty space can be filled with new and promising talent. It is noticed that in tech companies usually males have a higher number of employment which can cause inequality.Here, attrition may assist in the termination of workers who aren't a right fit for the organization. It helps to establish an engaged workforce in the organisation as the same employees having same ideologies won't be operating the company for a longer time. Employee attrition is confidential data of the company and can't be disclosed to the world, so a dataset is randomly selected from Kaggle and it is used to analyze the data and predict the attrition rate of a company. We used some methodologies of machine learning by stacking them for classification of data and prediction of result.

II. Literature Survey

Several machine learning research on employee attrition have been conducted. Taking ideas from the completed work we apply combinations of some of these models and techniques on our selected dataset and analysis it to get proper results. Below are some of the works:
Punnoose et al.[1]suggested Global retailer's HRIS database, BLS(Bureau of Labour Statistics) data and trained and tested supervised algorithms like Random Forest, Logistic Regression, XGBoost, KNN, SVM Naïve Bayes on ROC-AUC metric on the database. Suggested the use of Extreme Gradient Boosting (XGBoost), a more reliable approach. The primary objective of the research is to demonstrate that XGBoost forecasts employee attrition with a high degree of accuracy when compared to different models.Zangeneh et al.[2]presents a methodology for predicting attrition and uses real dataset from IBM HR as a case study.As there are many features in the databases, the "max out" feature selection strategy is used during the pre-processing step. The results show that this technique improves the F1-score performance.Yang et al.[3] analysed IBM Employee Attrition dataset to determine the factors affecting employee's decision to quit their respective jobs. Using a correlation matrix, firstly unimportant or unwanted attributes were eliminated from the dataset and selected the most important features. Then, a Random Forest approach was used to split the population into two clusters utilizing monthly income,age and the count of companies worked for. This was done after dividing the population through K-means clustering and conducting quantitative analysis using binary logistic regression,it was determined that the proportion of employees who frequently traveled was 2.4 times greater than those who traveled infrequently. Additionally, there was observed to be a higher propensity for employees in the human resources department to depart.Setiawan et al.[4] aimed to use logistic regression to evaluate employee attrition. R was applied for data preparation, integration, exploratory analysis, logistic regression, model evaluation, and visualization. Five datasets containing identical, unique employee IDs were created from the data. To understand the dataset, the functions summary() and str() were utilized. The desired value of a variable was transformed from yes/no to levels 0/1 for the purposes of model selection and training. The dataset divided into seventy-thirty percentage in which (30%)for the test set and (70%) the training set. The prediction model's sensitivity, evaluation accuracy, and test accuracy were 75%, 73%, and 75%, respectively. It was shown that an employee's probability of leaving a company rises noticeably with the number of years and companies they had worked for. By accessing the job satisfaction, working environment, and workload with communications between supervisors and staff, the business can enhance its HR division and reduce staff churn.Nesree et al.[5] analyzes records of employees who worked in reputed institutions in Nigeria and left it between 1978 and 2006, were used as a dataset. Mostly, job-related and demographic data were used to categorize workers into established attrition

classifications. See5 for Windows and the Waikato Environment for Knowledge Analysis (WEKA) were used to create decision tree models. The predictive model was then built using the results of the decision tree model and the produced rule-sets. With this technique, predictions about probable cases of employee attrition were made. A solitary decision tree of 15 by 3 sub-trees was produced, with a misclassification rate of 25.2 per cent. Salary was used 100 percent of the time, duration of service was used 49 percent of the time, and rank was used 16 of the time. These findings suggest that employee salary and length of service were the main drivers of whether or not they chose to continue with the organization.

Kamath et al.[6] primary objective of the research was to examine the reasons behind the unreasonable departure of skilled and experienced employees and also forecast which employee would resign. The dataset was selected from Kaggle. The study was done using R and Rattle data mining platform. The dataset was divided into training, testing and validation. Decision trees, random forest, support Vector Machine and linear regression techniques were used to build the model. The model's perfromance were evaluated through error matrix and pseudo R square estimation of error rate. It was predicted that Random Forest was the most suitable model for employee attrition analysis in comparison to other models. "Job satisfaction" was the dominant factor that influenced the attrition rate the most.Raza et al.[7] examined the organizational components that influence employee attrition and used machine learning to foresee attrition. Additionally, Employee Exploratory Data Analysis (EEDA) was used to determine the contributing causes of employee attrition. Support Vector Machines (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC), and Extra Tree Classifier (ETC) are the four machine learning approaches compared in this work for analysis.The proposed Extra Trees Classifier (ETC) approach achieved a 93% accuracy score in employee attrition prediction. It was concluded that age, work level, hourly rate, and monthly income were the primary factors contributing to employee attrition.

III. Proposed Methodology

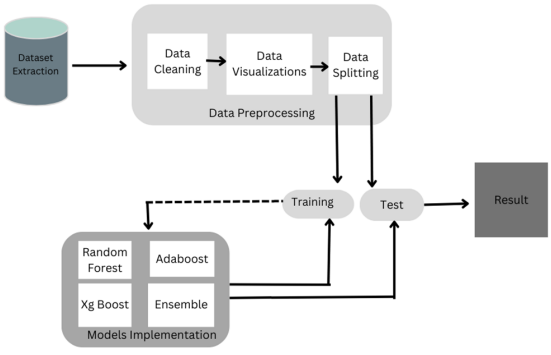

Fig. 1. Architecture of the proposed approach.

The dataset is a set of data represented in the form of table to make it easily visible and understandable. In dataset each column describes a particular feature and each row gives information about a member corresponding to that feature this way it constitutes to data management. Being specific about this project, dataset was selected from Kaggle.
The dataset comprises 1470 records and 35 fields, including definite and numeric features.

The heatmap of the dataset, which represents the coefficients, gives a visualization of the strength of correlation among variables.

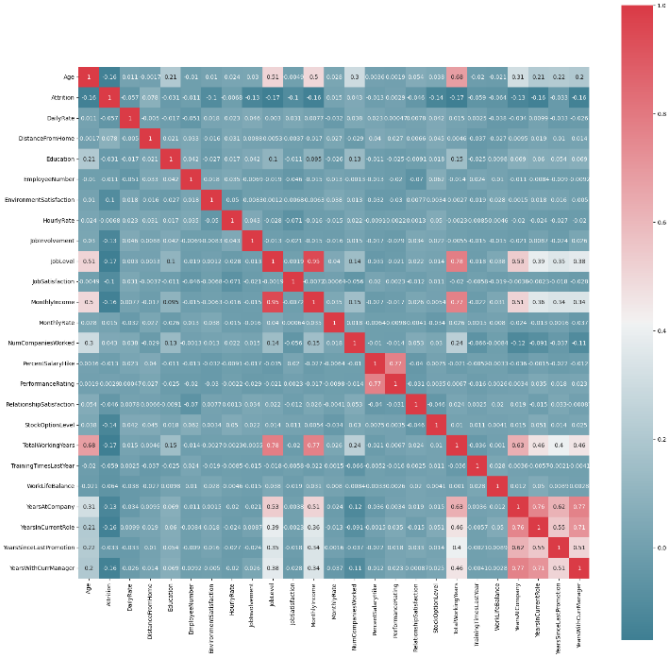

Fig. 2. Heatmap

From fig.6 monthly income and job level are highly correlated, also job level and total working year.
Recursive feature eliminator (RFE) is used for data preprocessing and feature selection. Feature selection can be defined as a technique that chooses a subset of features that are most relevant for the problem statement from the dataset. By using feature selection, unnecessary and irrelevant features are excluded and machine learning algorithms run more efficiently as they have less space and time complexity. If this technique is avoided, our prediction may result in wrong or worse results and we may redo it.
For feature selection algorithm, RFE is a wrapper-type algorithm known to be popular, easy to use and effective in selecting relevant features from the dataset.The process of data exploration and visualization plays a significant role in data science. Data analysts use exploration process to see or identify specific patterns in the dataset. It helps the information to be represented easily and interactive visual elements making the data being understood better.
This step helps users to understand their data statistically and visually through graphical means. Graphs and charts are used

for making complex structures simpler and finding relations within data as the ability to see different visual patterns and colors helps to interpret data more accurately. Breaking down the data into simpler structures will be helpful for professional sectors. We used various graphs for plotting. Having attrition as a common feature, we plotted it with the number of employees, department, age of the employee, job satisfaction etc.

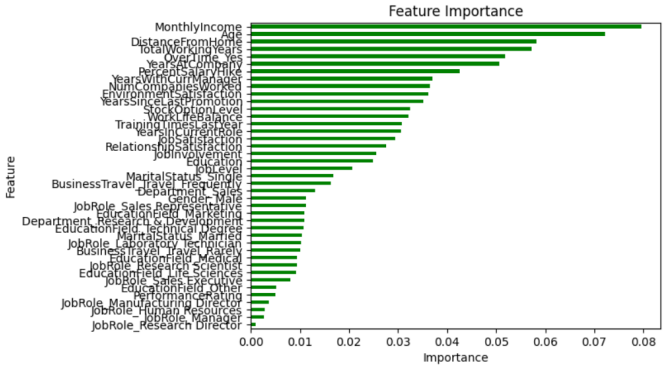

Fig. 3. random forest feature importance

From fig.3, the top three features are monthly income, age and distance from home are dominant, and they indicate whether the employee has a tendency to leave. While job role, marital status indicate employees are less likely to go. Monthly income is the primary reason that employees choose to leave their firm. People getting low salaries will depart the firm more likely and the employees getting high salary package will stay in the organization. No one wanted a low salary package particularly the fresher or young employees who just started their careers. In search of a higher salary package, they find other firms that offer more money than earlier. Additionally, recent graduates are more inclined to explore various job opportunities and eventually discover a fitting career path as they transition from universities.

Distance from home the other main reason is the distance of company(office) and home as commuting for longer time leads to wastage of time and energy which can be physically and mentally draining further can lead to job dissatisfaction. Due to the adverse effects of long commutes on mental health, physical activity, and stress levels, workers are prioritizing their health and seeking for opportunities closer to home. Overall, the distance between home and the workplace can significantly influence the determination of an employee to depart from a company, as it directly impacts their quality of life, job satisfaction, and overall happiness.

Age adds to attrition because of many factors such as health considerations, retirement, career switches and the need to explore new domains and opportunities.

We divide the dataset into training and testing dataset consisting of 1176 and 294(80-20%) records respectively with the same 24 features.

IV. Model Building

The report offers recommendations and insights to assist firms in lowering their rate of employee turnover. Certain machine learning classifiers were put to use to identify significant features and conduct data analysis.Classifier is an algorithm that automatically distinguish or classifies data according to classes. Types of classifiers that are used in machine learning are random forest classifier, logistic regression, naive Bayes, decision tree, support vector machine (SVM) and many more.Here, we have used Random forest classifier, Logistic Regression, XG-boost and ensemble stacking technique to find which suits best on our dataset. Machine learning tasks including both regression and classification, Random Forest is a flexible and effective ensemble learning technique. It constructs multiple decision trees during training, uses a random subset of features at each node and a variable part of the data to train each tree. For classification, where the mode of predictions is chosen, or for regression tasks, where the mean of predictions is derived, an averaging process determines the final prediction in a Random Forest. By combining predictions from several trees, the ensemble method reduces over-fitting and boosts adaptability.

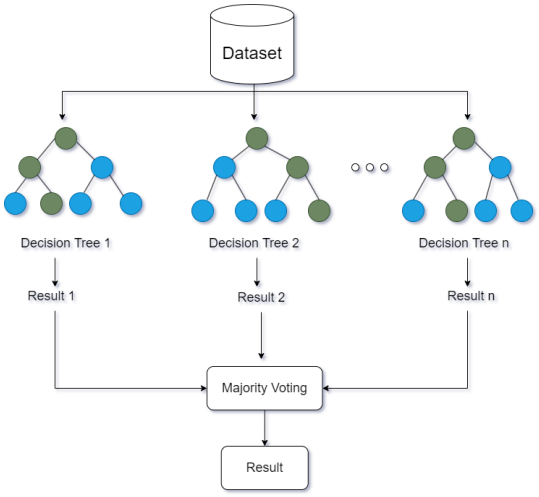

Fig. 4. Random Forest.

For binary classification tasks in machine learning, logistic regression is a basic and popular approach. The likelihood of a binary result dependent on one or more predictor variables is modeled using this linear model, despite its name.Due to its simplicity, interpret-ability, and efficiency in modeling binary outcomes, logistic regression is widely utilized in a variety of disciplines, including marketing (for customer churn prediction), finance (for credit scoring), and healthcare (for illness risk prediction). The linear relationship between the log-odds of the outcome and the predictor variables, which logistic regression depends on, might not always stay true in real-world circumstances.

Extreme Gradient Boosting, or XGBoost for short, is a sophisticated and very effective use of the gradient boosting

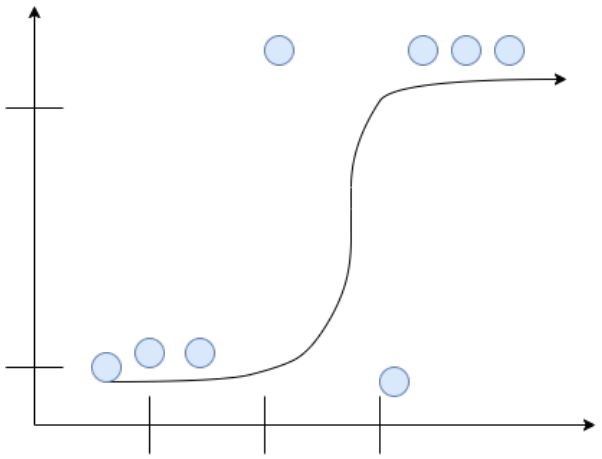

Fig. 5. Logistic Regression.

technique. It is widely used in machine learning competitions and real-world applications due to its exceptional performance and scalability. Feature engineering, structured data, and tabular data are just a few of the machine learning applications that choose XGBoost due to its state-of-the-art performance, scalability, and adaptability.By providing insights into feature relevance, XGBoost enables users to choose which features are most pertinent for usage in prediction.

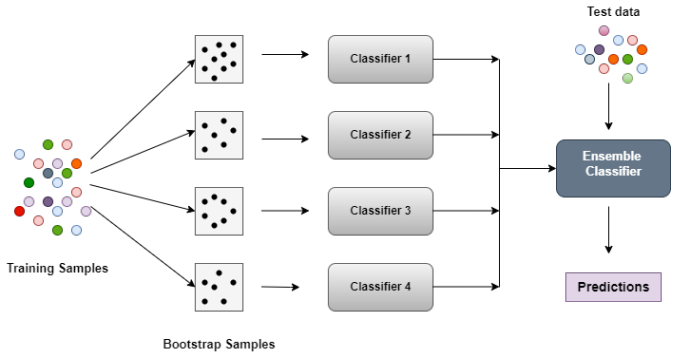

Fig. 6. XGBoost.

Results can be improved and made more accurate by combining several methods of machine learning. This is the concept Ensemble learning. By using multiple models on behalf of single model, the performance and prediction both are improved. Its advantage is that it gives majority vote, Bagging Randomness Injection, Feature selection, Error-correcting output coding. In ensemble classification we used stacking technique. Stacking consists model training for combination of predictions of multiple different algorithms.
Models of Logistic Regression are usually used as combiners. Fig 5 presents a combiner algorithm makes the final prediction after combining all the other algorithms as additional input or using cross-validated prediction. This gives better predictive results. Both supervised and unsupervised learning have been successfully accomplished with ensemble learning. Regression, classification, and distance learning are examples of supervised learning; density estimation is an example of unsupervised learning. It has also proved to be useful for error rate estimation of bagging.

V. Results

After data exploration and applying model algorithms, we equated them to know the best-suited algorithm with the highest accuracy for our dataset. The confusion matrix of the algorithms came out to be as follows.

This list consists of random forest classifier, Xg-boost, Logistic Regression and ensemble classifiers. Data processing is done using RFE(recursive feature elimination).We made some graphs between various features like age, job satisfaction with attrition rate to get a rough idea about feature selection. After data exploration, we use models to find most suitable algorithm with highest accuracy and precision for our dataset.

	precision	recall	f1-score	support
0	0.89	0.98	0.93	242
1	0.85	0.42	0.56	52
accuracy			0.88	294
macro avg	0.87	0.70	0.75	294
weighted avg	0.88	0.88	0.87	294

Fig. 7. Confusion Matrix of XG-Boost

The analysis revealed that Xg-boost proved as the most suitable algorithm for predicting employee attrition, boasting an accuracy of 88% and a precision of 89%. The graphical representation of feature importance indicated that Monthly Income exerted the greatest influence, followed by distance from home and age, prompting employees to seek better opportunities elsewhere. Apart from this using different model we found comparisons that males tend to depart more as compared to female, people who travel frequently leaves more because travelling leads to disruption of work life balance , fatigue and stress and those who work overtime are more likely to leave. Additionally, the Sales Department witnessed the highest turnover rate among employees.

VI. Future Scope

Future studies could enhance the analysis by taking into account factors that are positively correlated with worker intentions to leave, such as inequality, insufficient peer support, unfavorable working conditions, and limited career chances.This paper would help in analysis of employee attrition of an organization and letting it know the reason for the same.

VII. Conclusion

The results obtained by the model indicate that our findings align with observations from the real world and previous studies conducted by other academics. When attrition occurs within a company the workload among existing members of the team increases despite a pay raise. The workload of the existing team members increases when attrition occurs within a company without any pay raise. The HR specialist even deals with this workload. Due to the role being eliminated due to attrition, there may not be a chance for promotion in the workplace. We can conclude that it is an issue which is to be taken under consideration. Job satisfaction, daily rate, employee number and age were some of the factors that led to attrition. It was also observed that sales department had the highest attrition rate. Knowing these aspects will enable the organization to take the required actions to lower attrition rate and prevent a significant number of employees from leaving.

References

[1] Rohit Punnoose and Pankaj Ajit."Prediction of employee Turnover in organization using Machine Learning Algorithms".In: International Journal of Advanced Research in Artificial Intelligence (2016).

[2] Mangal, V., Dhamija, S. Analysing Theoretical Models for Predicting Employee Attrition: A Comparative Study in the FMCG Sector.

[3] Yang, Shenghuan, and Md Tariqul Islam. "IBM employee attrition analysis." arXiv preprint arXiv:2012.01286 (2020).

[4] Setiawan, I. A., et al. "HR analytics: Employee attrition analysis using logistic regression." IOP Conference Series: Materials Science and Engineering. Vol. 830. No. 3. IOP Publishing, 2020.

[5] El-Rayes, Nesreen, et al. "Predicting employee Attrition using Tree-based Models." International Journal of Orgaizational Analysis (2020).

[6] Kamath, Dr RS, Dr SS Jamsandekar, and Dr PG Naik. "Machine learning approach for employee attrition analysis." Int. J. Trend Sci. Res. Dev., vol. Special Is, no. Special Issue-FIIIIPM2019 (2019): 62-67.

[7] Raza, Ali, et al. "Predicting employee attrition using machine learning approaches." Applied Sciences 12.13 (2022): 6424. employee-attrition-and-factors.

[8] Reference for recursive feature elimination https://machinelearningmastery.com/rfe-feature-selection-in- python/

[9] Qutub, Aseel, et al. "Prediction of employee attrition using machine learning and ensemble methods." Int. J. Mach. Learn. Comput 11.2 (2021): 110-114.

[10] Jain, Rachna, and Anand Nayyar. "Predicting employee attrition using xgboost machine learning approach." 2018 international conference on system modeling advancement in research trends (smart). IEEE, 2018.

[11] Link for dataset https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors

[12] Najafi-Zangeneh, Saeed, et al. "An improved machine learning-based employees attrition prediction framework with emphasis on feature selection." Mathematics 9.11 (2021): 1226.

[13] James, M.J. and Faisal, U., 2013. Empirical study on addressing high employee attrition in BPO industry focusing on employee salary and other factors in Karnataka and Kerala states of India. Research Journal of Management Sciences.

[14] Pandey, N. and Kaur, G., 2011. Factors influencing employee attrition in Indian ITeS call centres. International Journal of Indian Culture and Business Management, 4(4), pp.419-435.

[15] Alshiddy, M.S. and Aljaber, B.N., 2023. Employee Attrition Prediction using Nested Ensemble Learning Techniques. International Journal of Advanced Computer Science and Applications, 14(7).

[16] Al-Alawi, A.I. and Ghanem, Y.A., 2024, January. Predicting Employee Attrition Using Machine Learning: A Systematic Literature Review. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 526-530). IEEE.

[17] Gim, Simon, and Eun Tack Im. "A Study on Predicting Employee Attrition Using Machine Learning." In IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science Engineering, pp. 55-69. Cham: Springer International Publishing, 2022.

[18] Sheth, Kalgi, Jaynil Patel, and Jaiprakash Verma. "Machine Learning-Based Investigation of Employee Attrition Prediction and Analysis." In Emerging Technology Trends in Electronics, Communication and Networking: Select Proceedings of the Fourth International Conference, ET2ECN 2021, pp. 221-238. Singapore: Springer Nature Singapore, 2022.