Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online)
2024, Vol. 4, No. 1, 30-34

Predicting Employee Attrition using an Ensemble Method

Akash GC
Nepal College of Information Technology,
Pokhara University
akash.202902@ncit.edu.np

Roshan Chitrakar
Nepal College of Information Technology,
Pokhara University
roshanchi@ncit.edu.np

Article History:
Received: 25 February 2024
Revised: 28 April 2024
Accepted: 27 May 2024

Keywords— Employee Attrition, Ensemble Model, Employee retention efforts, Employee Turnover.

***Abstract*— Employee attrition poses significant challenges for organizations, impacting productivity, performance, and financial stability. Existing research on predicting attrition suffers from limitations such as accuracy constraints, insensitivity, and a lack of finding the important features contributing to employee attrition. This study aims to address these gaps by developing an ensemble model for predicting attrition and enhancing employee retention rates. The objectives include improving prediction accuracy, identifying informative factors, dealing with imbalanced data, and incorporating hyperparameter tuning. The proposed ensemble model, augmented with hyperparameter tuning, achieved impressive performance metrics, including an accuracy rate of 92.75%, precision score of 98%, recall rate of 88.83%, specificity of 97.68%, and F-beta score of 93.25%. These results indicate the model's effectiveness in identifying employees at risk of attrition and its potential for aiding organizations in retention efforts.**

I. Introduction

Employees play an essential role within the organization, but more than half of businesses struggle to keep the best performers or those who are most marketable [1]. Employee attrition is the term used to describe the loss of human capital experienced by an organization owing to either voluntary or involuntary factors. Unlike voluntary attrition, which happens when employees leave their jobs or retire, involuntary attrition happens when employees are fired or laid off[2]. Employee attrition can be caused by a variety of circumstances. Employees depart the company quicker than they are hired. When an employee departs an organization, the vacancies go unfilled, resulting in a loss for the company. The employee attrition rate provides insight into an organization's development. The high attrition rate indicates that staff leave frequently. The loss of organizational benefits is a result of the high rate of attrition [3].

Therefore, it is crucial for businesses to recognize the causes of attrition and implement preventative measures to retain important employees. Organizations may use predictive modelling approaches to identify employees who are possibly departing the company and can take appropriate steps to retain them. The purpose of this study was to investigate organizational characteristics that contributed to employee attrition and to analyse the prediction of employee attrition using ensemble approaches.

II. Related Work

Many previous studies that use various machine learning algorithms have limitations in terms of prediction accuracy, sensitivity, and complete recommendations for improving employee retention rates[1][4]. This study attempts to provide significant insights and recommendations to firms in efficiently managing their human resources and taking proactive steps to retain their valued staff by filling the research gaps. it.

The prediction of employee attrition rate using machine learning-based classification algorithms was proposed[4]. Machine learning models like Naïve Bayes, Support Vector, and Decision Tree have been adopted for employee attrition prediction. The different steps were applied to obtain high accuracy among them Random Forest classifier delivered the highest accuracy of 83.3%.

The study[1] proposes automated prediction of employee attrition based on numerous machine learning algorithms. The IBM HR employee dataset was used in the learning model construction and model assessment procedure. For the prediction challenge, the Ad Boost Model, Random Forest Model, Decision Tree, Logistic Regression, and Gradient Boosting Classifiers were used. The accuracy of the Decision Tree with Logistic Regression was 86%.

All the work done to develop a model which can early predict the employee attrition was helpful for the organization to retain the valuable employee. But the accuracy of these models is not much in satisfactory level so in this Research a new model, ensemble methos for employee attrition prediction is proposed. Random Forest, Gradient Boost, Logistic Regression, and Neural Network were used to build the model which aims to predict employee attrition more accurately and helps to find the most critical factor in employee attrition. The following diagram shows the research gap in the existing system.

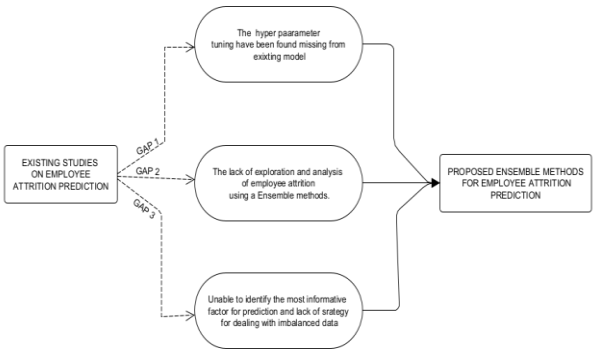

Fig. 1. Limitation in Existing Studies of Employee Attrition Prediction

III. Methodology

The block diagram of the proposed model is shown in fig 2. The employee dataset used in this model is an IBM HR dataset which is publicly available. At first, a detailed literature review has been conducted on the employee attrition prediction algorithm and different data analysis methods. Literature study helps us to refine the structure of the proposed model Employee attrition prediction to overcome the weakness of the existing models.

In the proposed model Random Forest, gradient boost, Neural and logistic regression algorithms will be used. These algorithms are chosen out of many machines learning algorithm because of their better results[1][5][6]. The result of the mentioned algorithms can be increased by the parameter tuning[7] so parameter will be tuned in this proposed model. After all these max-voting based ensemble method is implemented, and results are tested on dataset. The result will be examined based on the accuracy, precision, ROC, and f-measure. justified.

The system flow diagram as shown in Fig 2 shows the actual flow of the study. The following step describes the flow of the system.

A. *Random Forest*

The Random Forest is a classifier that utilizes numerous decision trees on distinct subsets of a given dataset and averages them to improve the projected accuracy of that dataset. The random forest algorithm consists of a collection of decision trees, each of which is generated using a bootstrap sample, which is a data sample acquired from the training set with replacement[8].

B. *Gradient Boosting*

Gradient Boosting is another effective ensemble strategy that generates an ensemble of weak prediction models, such as decision trees, step by step. It seeks to reduce the mistakes generated by earlier models by giving more weight to misclassified events. This repeated procedure yields a powerful prediction model capable of managing a wide range of data types and capturing complex relationships within the data[9].

C. *Logistic Regression*

Logistic regression is a classification algorithm. It is used to predict a binary result based on a set of independent factors. To anticipate the result of a categorical dependent variable, logistic regression is used. As a result, the outcome must be a categorical or discrete value, such as Yes or No, 0 or 1, true or false, and so on.

D. *Ensemble Method*

Ensemble techniques are strategies for creating several models and combining them to achieve better results. In majority voting ensemble models, each model predicts for all test cases, and the final output prediction is the one that receives the majority of votes chooses the critical characteristics, which aids in more accurate result prediction[10].

IV. Experimental Evaluation

After successfully modelling a system, it is trained using the training data set supplied by the standard datasets, and the model is validated using the testing dataset. The system's performance will be assessed by calculating accuracy, recall, precision, and F-beta.

A. *Dataset*

The IBM HR dataset, obtained from Kaggle was utilized for processing in this study. There were 1470 records and 35 features in the collection. The data columns included factors such as 'Age,' 'Gender,' 'Department,' 'Distance from Home,' and others. 'Attrition' is the dependent variable in this work, which is comprised of two class labels - 'Yes' or 'No'. The 'Attrition' rate in the organization was.16%. This Employee attrition dataset is defined by the following attributes.

B. *Data Preprocessing*

Preprocessing describes the procedures and methods used to convert unprocessed data into an appropriate format for analysis and modelling. Preprocessing is an essential phase in the data science pipeline because it facilitates data organization and cleaning, handles missing values, reduces noise, and converts data into a more intelligible form.

1) Imputation of missing values:

An important stage in data analysis is to investigate and rectify missing values to guarantee data completeness and quality. The existence of missing values was explored in the context of the IBM HR dataset. After examining the dataset, it was found that there are no missing values in the IBM HR dataset.

2) Data Type Conversion:

Categorical variables were translated into numerical representation using one-hot encoding to make the dataset compatible with machine learning methods such as logistic regression. In our dataset Attrition, Business Travel, Over Time, Department, Gender, Education Field, Marital Status, Job Role and Over18 have the datatype categorical. Among them the "Over18" column only has one type of value, 'Y,' suggesting that this column provides no substantial variance and is unlikely to impact the model's training outcomes.

On the other hand, three columns, namely "Attrition," "Gender," and "Overtime," have two sorts of values. We may encode these columns by giving labels 0 and 1 to represent the various categories.

3) Dealing with duplicate data:

It was discovered that there are no duplicate values in our dataset after running an operation to explore them. This means that we may continue with the dataset's processing and analysis without having to deal with or address any duplicate values.

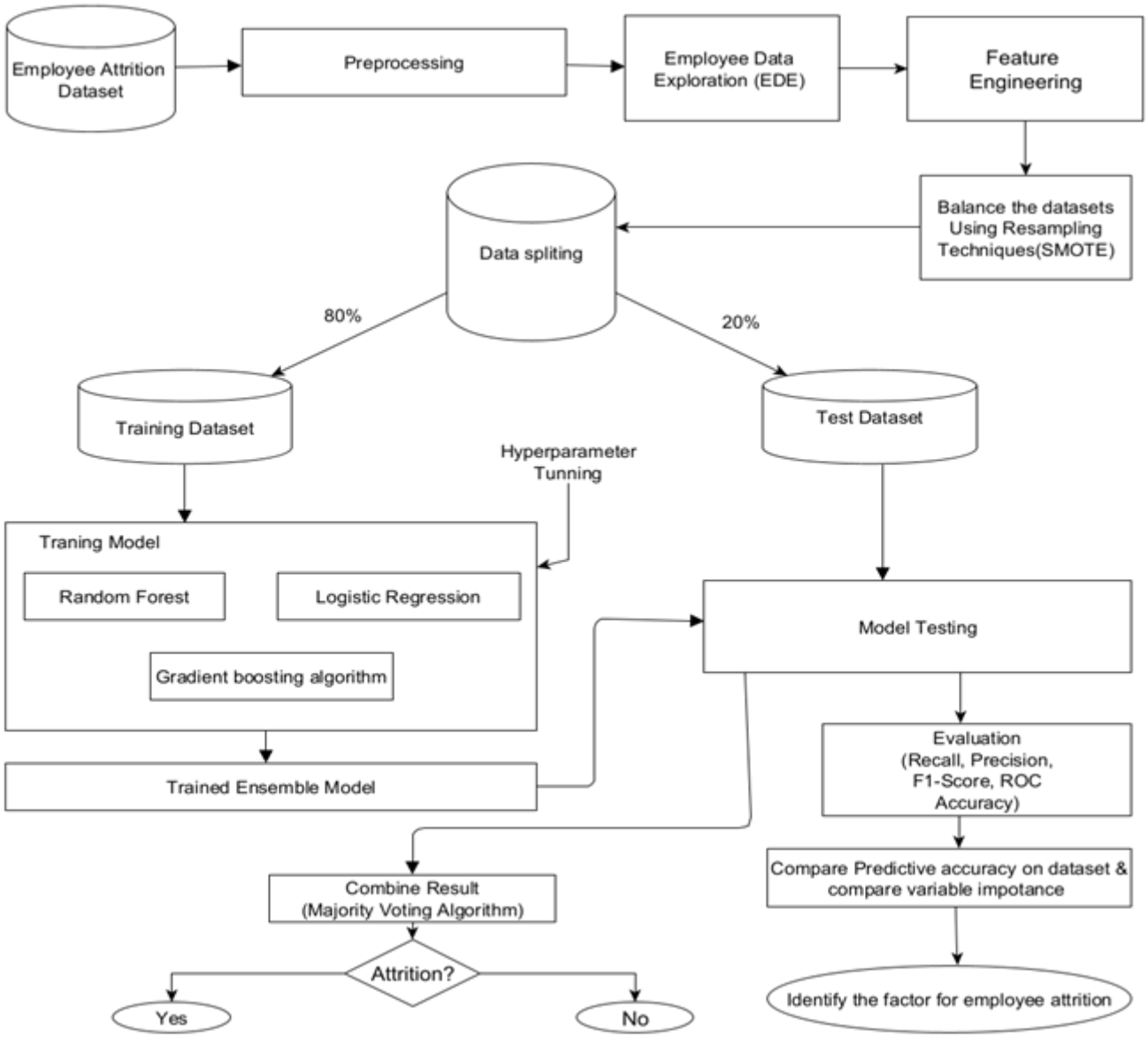

Fig. 2. Block Diagram of the Methodology

C. *Employee Data Exploration (EDE)*

The Employee Data Exploration (EDE) includes evaluating the HR dataset to identify useful insights. To investigate the distribution of numerical columns, histograms were used. This research allowed for a more in-depth knowledge of the dataset, allowing for the identification of important trends and areas of interest in the HR data.

D. *Correlation Based Feature Engineering*

Feature correlation is a statistical metric that illustrates how variables in a dataset are connected. It aids in determining whether variables tend to change together (positive correlation) or in opposing ways (negative correlation). Correlation coefficients vary from -1 to 1, reflecting the strength and direction of the link. Fig. 3 illustrates the correlation between various feature in Dataset. After the correlation and EDE analyses, the following features were removed from our dataset: Total Working Years', 'Job Level', 'Percent Salary Hike', 'Years in Current Role',' Years with Current Manager', 'Department Human Resources', 'Department Sales', 'Employee Count', and 'Over18'.

E. *Data Resampling*

To balance the dataset, the SMOTE (Synthetic Minority Oversampling approach) dataset resampling approach was used. The model's complexity was lowered by balancing the dataset. This was accomplished by training the model with an equal amount of target distributions, resulting in a higher model accuracy score.

F. *Data Splitting*

The model dataset was separated into two separate datasets, one for training and one for testing. For this model, the dataset will be divided into a testing set (20%) and a tanning set (20%). The training set is used to train the model, while the testing set is used to evaluate its performance.

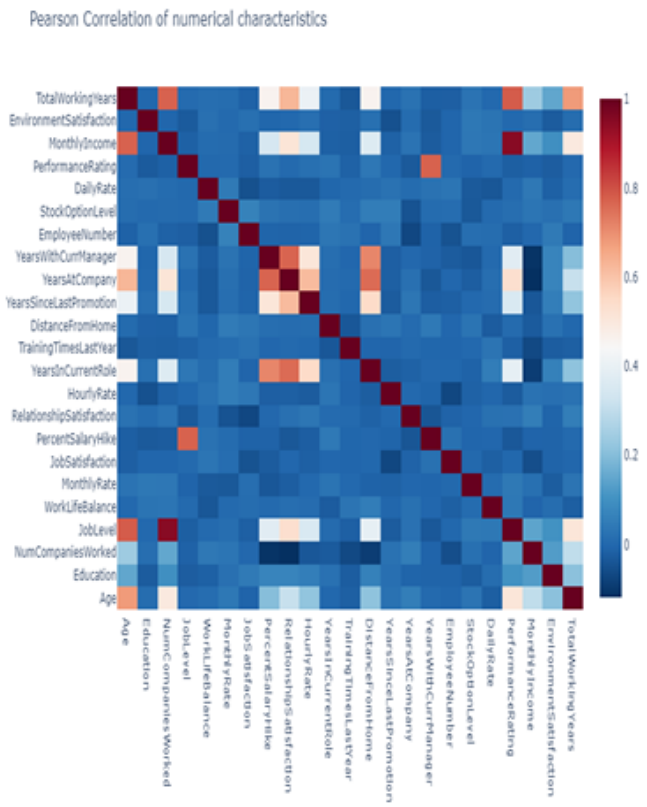

Fig. 3. Person correlation of numerical characteristics

G. *Hyperparameter Tunning*

Hyperparameter tuning is an important step in machine learning research that tries to improve a model's performance by fine-tuning its parameters and configurations. Researchers can determine the ideal values for a certain job or dataset by experimenting with different combinations of hyperparameters such as learning rate, batch size, and regularization intensity. The Optimized C value for Logistic Regression is 100 and for the others two models following are the Optimized hyperparameters values.

TABLE I. OPTIMIZED HYPERPARAMETER VALUES

S. N	Regular	Value for Gradient Boost	Value for Random Forest
1	Learning_rate	0.5	-
2	Min_samples_leaf	2	1
3	Random_state	400	-
4	Subsample	0.80	-
5	Min_samples_split	2	2
6	N_estimators	500	500
7	Max_depth	-	20

V. RESULT AND DISCUSSION

All In this study the proposed model is an ensemble model combining three different classification algorithms.

Logistic Regression, Gradient Boosting, and Random Forest are among the machine learning techniques used. To combine these algorithms, the max voting approach is used. The model's performance is compared to other existing models when the final predictions are received.

TABLE II. CONFUSION METRIX OF PROPOSED MODEL

		Predicted	
		Positive (1)	**Negative (0)**
Actual	**Positive (1)**	222	9
	Negative (0)	35	228

The outcomes of the experiment are assessed using many performance criteria, including Accuracy, Precision, Recall, Specificity, and F-beta score.

By studying these numbers, we can calculate several assessment metrics such as accuracy, precision, recall, and F1-score to further examine the model's effectiveness in predicting employee attrition. The following diagram shows the overall results of the proposed Employee prediction model.

The ensemble model, which was improved with hyperparameter tunning, performed well in predicting employee attrition. The model performed well in categorizing occurrences, with an accuracy rate of 92.75%. Its precision score of 98% revealed that the most anticipated attired employees were, in fact, attired. The model also had a commendable recall rate of 88.83%, recognizing a sizable fraction of genuinely attired personnel. Furthermore, the model had a high specificity of 97.68%, indicating that it correctly identified non-attrition personnel. Overall, the model's F-beta score of 93.25%, which emphasizes accuracy, showed its ability to predict employee's attrition. These findings demonstrate the model's strong performance and potential use in aiding firms in identifying employees in danger of leaving.

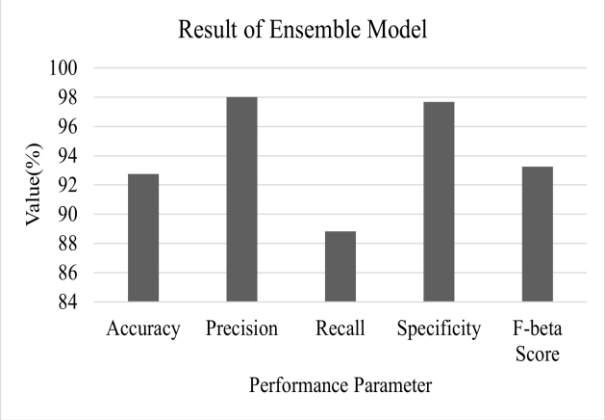

Fig. 4. Pearson correlation of numerical characteristics

The ROC curve in fig 4 illustrates the performance of our binary classification model. It demonstrates how well the model differentiates between positive and negative circumstances at different levels. The rapid early ascent of the curve reflects the model's excellent capacity to detect positive events while limiting false positives. The model's total performance is summarized by the Area Under the Curve (AUC) score, which is 98%. A higher AUC value

indicates greater performance, and in our situation, 98% indicates performed well in predicting employee attrition.

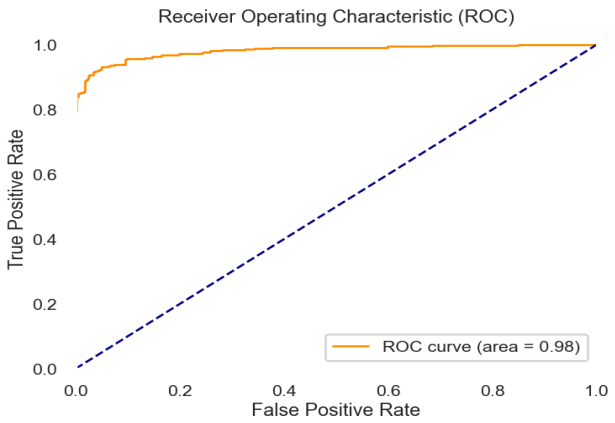

Fig. 5. Pearson correlation of numerical characteristics

A. *Comparison between proposed model and existing model*

After comparing our suggested model to current models in terms of accuracy, we determined that our model outperforms them. We fine-tuned the hyperparameters thorough review and testing to obtain higher accuracy than the previous models.

TABLE III. CONFUSION METRIX OF PROPOSED MODEL

S.N.	Paper –Year	Technology Used	Accuracy
1	[1]- 2019	Ensemble learning.	86.39%
2	[5]-2023	logistic regression (LR) found best	87.96%
3	[11]-2021	Support Vector found best	80.00%
4	[12]-2021	Random Forest	85.12%
5	Proposed	Ensemble Model	92.75%

VI. CONCLUSION

The average result is taken from multiple runs to obtain the more stable result based on this, the proposed model gives the result as accuracy of 92.75%, Recall of 88.83%, Precision of 98%, Specificity of 97.68% and F-beta Score of 93.25% which is the better result as compared to the existing model. Among the various features of the dataset used in this model it is found that Job satisfaction, Relationship Satisfaction, Job Involvement, Work Life Balance, Age, Gender, and Education are the most important features in employee attrition.

The accuracy and other performance parameters of the proposed model are well but in this model real time data was not used which may vary the performance of the model so in future real time data can be used to build the model which can helps to improve the performance of model. And we can add the recommendation system on this study to minimize the employee attrition rate.

REFERENCES

[1] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods," *Int. J. Mach. Learn. Comput.*, vol. 11, no. 2, pp. 110–114, Mar. 2021, doi: 10.18178/ijmlc.2021.11.2.1022.

[2] A. P. Dilip Singh Sisodia, Somdutta Vishwakarma, *Evaluation of Machine Learning Models for Employee Churn Prediction*. IEEE, 2017.

[3] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Appl. Sci.*, vol. 12, no. 13, Jul. 2022, doi: 10.3390/app12136424.

[4] N. Bhartiya, P. Shukla, S. Jannu, and R. Chapaneri, *Employee Attrition Prediction Using Classification Models*. IEEE, 2019.

[5] F. Guerranti and G. M. Dimitri, "A Comparison of Machine Learning Approaches for Predicting Employee Attrition," *Appl. Sci.*, vol. 13, no. 1, Jan. 2023, doi: 10.3390/app13010267.

[6] F. K. Alsheref, I. E. Fattoh, and W. Mead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7728668.

[7] H. Herodotou, Y. Chen, J. L.-A. C. S. (CSUR), and undefined 2020, "A survey on automatic parameter tuning for big data processing systems," *dl.acm.org*, vol. 53, no. 2, Jun. 2020, doi: 10.1145/3381027.

[8] "What is Random Forest? IBM", https://www.ibm.com/topics/random-forest (accessed Sep. 16, 2023).

[9] "Introduction to the Gradient Boosting Algorithm | by Anjani Kumar | Analytics Vidhya | Medium." https://medium.com/analytics-vidhya/introduction-to-the-gradient-boosting-algorithm-c25c653f826b (accessed Sep. 16, 2023).

[10] M. Subhashini and R. Gopinath, "Employee Attrition Prediction in Industry Using Machine Learning Techniques," *Int. J. Adv. Res. Eng. Technol.*, vol. 11, no. 12, pp. 3329–3341, 2020, doi: 10.34218/IJARET.11.12.2020.313.

[11] İ. ERSÖZ KAYA and O. KORKMAZ, "Machine Learning Approach for Predicting Employee Attrition and Factors Leading to Attrition," *Çukurova Üniversitesi Mühendislik Fakültesi Derg.*, vol. 36, no. December, pp 913–928, 2021, doi: 10.21605/ cukurovaumfd. 1040487.

[12] M. Pratt, M. Boudhane, and S. Cakula, "Employee attrition estimation using random forest algorithm," *Balt. J. Mod. Comput.*, vol. 9, no. 1, pp. 49–66, 2021, doi: 10.22364/BJMC.2021.9.1.04.