**Abstract--Companies overlook their employees’ happiness and satisfaction which leads to the employees moving to another company that allows them to showcase their talents and grow in their careers. The mental health of an employee, the perks and incentives that are given to them and the work hours assigned to them should be constantly monitored by the Human Resources department to ensure that an employee is not facing any difficulties in their company. Several studies state that losing an employee causes a company much more loss, compared to the annual salary of the employee. This project aims to avoid that. The primary goal is to help companies find whether an employee will leave their organization, based on various factors that were decided using machine learning. These results can help the HR department mark which employees need assistance of any sort.**

**Keywords--Employee Attrition,Human Re-Source,Attrition Prediction,Dataset,ML**

1. **Introduction**

Employee Attrition is a very influential factor in deciding the annual profit earned by an organization. Loss of talented employees is a major issue faced by business leaders within such organizations. Retaining a good employee can boost business in many ways. The work is done efficiently, and the quality of work is not compromised, having a good employee as a company representative leaves a good impression on clients, and major projects are completed according to client needs. This is very profitable for a company in the long run and gives returns that are much higher than the employee’s annual salary.

Looking at the factors mentioned above, losing an employee due to reasons like dissatisfaction in the workplace is undesirable. Some factors that cause dissatisfaction are not getting credit for work done by them, feeling underappreciated, heavy workload or lack of incentives or bonuses. In the later stages of the project, suggestions can be given to the HR department of the companies that use our model to avoid losing employees for the reasons mentioned above.

1. **Literature Survey**

Numerous studies have been conducted in the field of employee churn in the past, however the use of machine learning has been explored only recently in this field.

**[1]**In a study by Yadav et. al (2019), it was concluded that incurred by the HR department in recruiting and training new employees is much higher than an employee’s annual salary. The study stated various challenges faced by hiring managers and talked about the various categories of employee attrition. The prediction was done by comparison of various Machine Learning models’ performance when it came to the reliable features in this prediction. These features were identified by RFECV (Recursive Feature Elimination with Cross Validation). Models like Logistic Regression, SVM, Random Forest, Decision Tree and AdaBoost were analyzed, where AdaBoost and Random Forest gave the best results. The features that were majorly analyzed were average monthly hours, satisfaction level, number of projects and last evaluation. The results showed how this trend can be prevented by increasing employee satisfaction levels and other factors, and how preventing it is very beneficial for the company’s future.

1. In 2020, Jain et. al stated that employee retention could be achieved only when employee appraisal and satisfaction rates were higher. The results showed that features like satisfaction level, number of projects and work accidents contributed most to an employee’s attrition. To the processed data, the support vector machine (SVM), decision trees (DT), and random forest (RF) algorithms were applied. Random Forest gave the best results, an accuracy of 99% was seen and it was checked through the standard confusion matrix.
2. In the study by Fallucchi et. Al, a preliminary exploratory analysis of the application of machine learning methodologies for employee attrition prediction was proposed. Several classification models, like, Gaussian Naive Bayes, Naive Bayes classifier for multivariate Bernoulli models,Logistic Regression classifier, K-nearest neighbors (K-NN), Decision tree classifier, Random Forest classifier, Support Vector Machines (SVM) classification and Linear Support Vector Machines (LSVM) classification, were used with the goal of finding the best one. Among the proposed methods, Logistic Regression performed the best, with an accuracy of 88% and an AUC-ROC of 85%. Results obtained by the proposed automatic predictor demonstrated that the main attrition variables are monthly income, age, overtime, and distance from home.
3. The primary objective of the study by Jain et. al was to predict employee attrition and the XGBoost model was used for the prediction of Employee Attrition. After analysis, the study concluded with the features that influence the turnover rate of an organization. These features include age, gender, distance from home, department, Job involvement, job satisfaction, marital status, monthly income and years since the last promotion. The XGBoost model is the best algorithm in this scenario as it is efficient in terms of efficient memory utilization, high accuracy and low

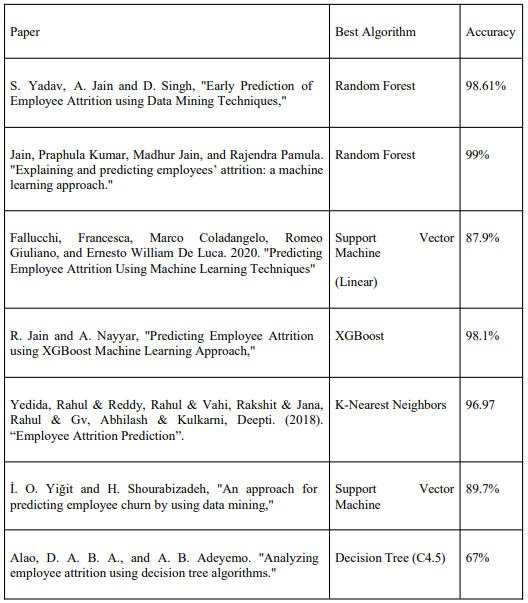
running times.

**[5]**In a study by Yedida et. al (2018), the attrition problem was listed and listed how it could be solved and then 4 machine learning models were applied and compared on the pre-processed data in terms of the features used. The KNN classifier gave the best results on a dataset pulled from Kaggle. AUC and ROC curves were used to check the general predictiveness of the model. The methods used here were Naive-Bayes, Logistic Regression, Multi-layer Perceptron Classifier and K-Nearest Neighbours (KNN). The dataset was pulled from Kaggle, pre-processing was done on it and then the training-test split used here was 70-30. When comparing the results, the AUC and ROC curves were used to check the general predictiveness of the model. This comparison showed that the KNN classifier showed good ROC-AUC and accuracy. The study suggested using the KNN classifier to accurately predict employee attrition to enable HR to take necessary action to avoid that.

**[6]**In 2017, Yiğit et. al conducted a study that demonstrated that data mining algorithms can be used to build reliable and accurate predictive models for employee churn. The study used the Employee Attrition data set provided by IBM which contained employee information such as demographics, experience, skills, nature of work or unit, position etc. They applied well-known classification methods including, Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes methods on the dataset. The final finding was that SVM gave better results than the other methods in terms of accuracy, precision and F-measure .

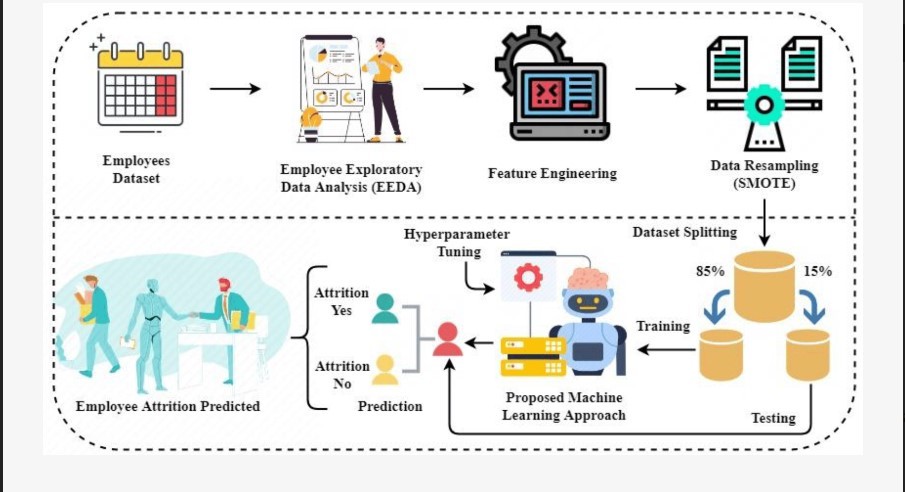
1. Alao et. al (2013) discussed the types of voluntary turnovers, which are functional and dysfunctional. Their work aims to avoid dysfunctional turnover as it can be very harmful to an organization. Their study aims to use Decision Trees for classification and regression, and the primary analyzer used was CART (Classification and Regression Trees) analysis. Other decision tree algorithms used were ID3, C4.5 and CHAID (Chi-square automatic interaction detection). C4.5 was the best performing Decision Tree algorithm .

**Table 1.** Result Comparison of Various Studies



1. **Methodology**

When preparing to implement the project development of a machine learning-based employee attrition system, several steps need to be taken.

**Fig.1** - Methodological analysis of our proposed research study for employee attrition prediction.

**[1]Data Gathering and pre-processing**--The fundamental phase in the machine learning pipeline is data gathering for training the ML model. The accuracyof the predictions provided by ML systems is only as good as the training

data. The dataset chosen was created by IBM Data Scientists based on features like Age, Monthly Income, Distance from Home, Job Role etc.

Data preprocessing is the process of preparing and transforming raw data into a format that can be easily analyzed and interpreted . In this stage, the dataset is modified using methods like Data Cleaning which includes removing noise values, incomplete records, outliers and inconsistencies in data. Several methods that can be used here include binning, filling incomplete data with attribute mean and regression.

**[2]Hyperparameter Tuning**--The goal of hyperparameter tuning was to find the hyperparameters that lead to the best performance of the model on the validation set. Hyperparameters are the internal parameters that can change the results of any complex machine learning model because of the sensitivity of the model towards these parameters, tuning methods include forcing a practitioner to evaluate thousands of hyperparameters, increasing the training time of complex models and huge datasets for tuning or managing a large number of hyperparameters and high training times using parallel or distributed computing for hyperparameter optimization.

1. **Method Selection and Training**

Modeling involves examining various machine-learning techniques to find the best possible classifier. Examination of each technique can be done by training each classifier on the feature set and the classifier with the best results can be used for prediction.

**[1]K-Nearest Neighbours (KNN)-**-The K-Nearest Neighbours algorithm, also known as KNN, is a simple non-parametric classification method. In this method, data records can be classified in their respective neighborhoods through majority voting among the data records in the neighborhood . The cost of classifying instances is very high because this method does not involve pre-modelling, prohibiting it to be applied to fields where dynamic classification can be needed for large datasets.

**[2]Logistic Regression--**The concept of logistic regression is based on a mathematical concept known as the logit—the natural logarithm of an odds ratio. This concept is well suited for describing and testing hypotheses about relationships between multiple predictor variables and a categorical outcome variable.

**[3]Decision Tree**--It is a common technique used for developing predictive models or for establishing classification systems based on multiple covariates. A population is classified into a branch-like structure to

construct an inverted tree with root, internal and leaf nodes. The branches between each node represent the outcomes from root or internal nodes that lead into leaf nodes.

**[4]Random Forest--**This technique uses multiple classification and regression trees to overcome the problem of poor accuracy in decision trees. It is an ensemble learning method which uses a collection of classification and regression trees which use binary splits on predictor variables for determining the desired outcomes .

**[5]Multi-Layer Perceptron (MLP)-**-It is the most popular type of neural network and uses a feed-forward architecture. The MLP classifier consists of an input layer, multiple hidden layers and an output layer where each layer is connected to multiple neurons in the next layers through weighted connections .

**[6]Extreme Boosting Tree--**This is an end-to-end gradient tree-boosting algorithm which incorporates a regularized model to prevent overfitting. The gradient boosting method is used in this technique where the weight of the wrongly classified observation is increased, and the weight of the correctly classified observation is reduced. The classifier is trained using the observations whose weights were modified. All the different classifiers obtained here are amalgamated to build a highly accurate classifier .

1. **Method Validation and Optimization**

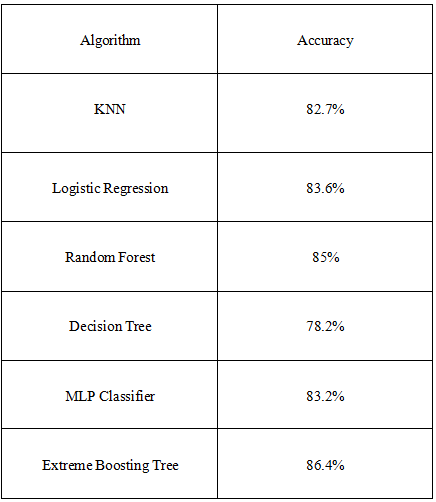
Model validation is the step where a trained machine learning model’s performance is assessed using newly collected data or a separate dataset, other methods suggest applying a Train/Test split on an existing dataset and using the Test data to validate a model . The model's capacity to generalize to new data is assessed using the validation datasets, which differ from the training datasets.

In the process of Model Optimization, the machine learning model’s hyper-parameters are optimized to enhance its performance. In this study, Grid Search was used for optimization. In this technique, the best combination of hyper-parameters is created for optimal results .

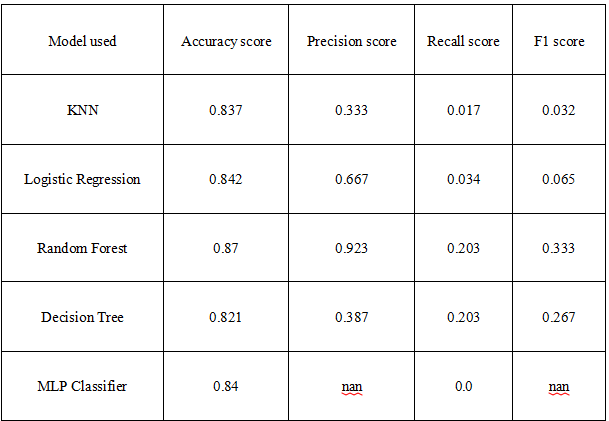
1. **Results**

The dataset was divided in a 75:25 ratio in the train-test split, which was done before the preprocessing step. The test data was used to measure the accuracy of different models and the results are discussed below.The resultant accuracy score, prediction score, recall score and F1 score produced after using two techniques, i.e., correlation matrix and chi-square distribution are as follows:

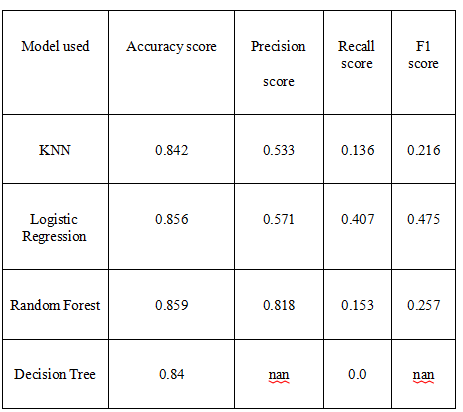
**Table 2.** Results before Hyperparameter Tuning



**Table 3.** Correlation Matrix Model Results

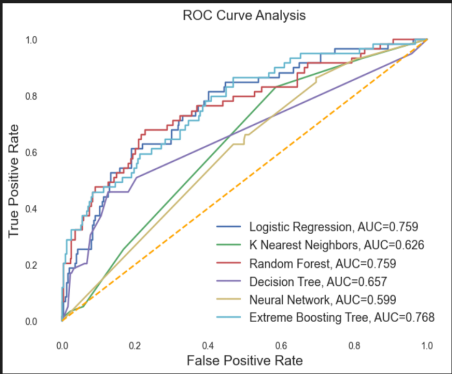


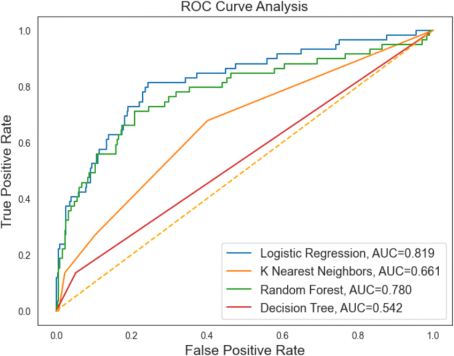
**Table 4.** Chi-Square Model Results



The ROC-AUC metric was also used for evaluating the binary classifications of each model, where the ROC curve has recall of the model (true positive rate) at the x-axis and the false positive rate is on the y-axis. The results for the models trained under each feature selection technique are given in the figures below

**Fig. 2 -** Correlation Model ROC Curve

**Fig.3-**-Chi-Square Model ROC Curve



**7.Discussion and Future Scope**

Two features i.e., Performance Rating and Business Travel were dropped from the dataset after which the feature selection process was used using two methods, i.e., Correlation Matrix and Chi-Square Distribution.According to the scores obtained by the various models used in the above two tables, we come to the conclusion that when we used a correlation matrix for feature selection, we trained six models and the model that gave the best performance was Random Forest, with an accuracy of 87% which is a quite high accuracy score. The second best-trained model according to the dataset was the Extreme Boosting Tree, which achieved an accuracy of 86.1%. The least-performing model was Decision Tree with an accuracy of 82.1%. The ROC-AUC score of Extreme Boosting Tree was the highest with an AUC score of 0.768(see figure 2).When we used chi-square distribution for feature selection, we trained 4 models and the model that gave the best performance here was also Random Forest, with an accuracy of 85.9%. The second best-trained model according to the dataset was Logistic Regression, which achieved an accuracy of 85.6%, which was only slightly lesser than Random Forest. The least-performing model in this method too was the Decision Tree with an accuracy of 84%. The ROC-AUC score of Logistic Regression was highest with an AUC score of 0.819(see figure 3).

Future Work might include the important features and best-performing model could be used to create a web application where a user can give data specific to an employee and the model can predict their attrition. Secondly, these results can be used by HR to counsel the valuable employees who are facing some issues in the company and help them out. This can be done by giving incentives, reducing the workload or just a token of appreciation. This study can change the profitability of a business to a huge extent.

**8.Conclusion**

In conclusion, the model designed performs well when Correlation Matrix is used for feature selection, compared to the Chi-Square test. When using Correlation Matrix, the Random Forest algorithm’s performance is the best with an accuracy of 87%. In terms of the F1 score, Logistic Regression under the Chi-Square test gives the best F1 score of 0.475 and its accuracy is 85.6%, which is the best among the models used in that technique. Therefore, both models are good for us in their respective use cases. This project has a lot of scope for improvement, the first step being rigorous training of the model on actual datasets and the next one will be giving these results to businesses to avoid losing valuable employees.

**Acknowledgment**

We are profoundly grateful to **Dr. Amiya Ranjan Panda** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion.

**References**

**[1]**S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques," 2018 IEEE 8th International Advance Computing Conference (IACC), Greater Noida, India, 2018, pp. 349-354, doi: 10.1109/IADCC.2018.8692137.

1. Jain, Praphula Kumar, Madhur Jain, and Rajendra Pamula. "Explaining and predicting employees’ attrition: a machine learning approach." SN Applied Sciences 2 (2020): 1-11.
2. Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. 2020. "Predicting Employee Attrition Using Machine Learning Techniques" Computers 9,

no. 4: 86.

1. R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2018, pp. 113-120, doi: 10.1109/SYSMART.2018.8746940.
2. Yedida, Rahul & Reddy, Rahul & Vahi, Rakshit & Jana, Rahul & Gv, Abhilash & Kulkarni, Deepti. (2018). Employee Attrition Prediction.
3. İ. O. Yiğit and H. Shourabizadeh, "An approach for predicting employee churn by using data mining," 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2017, pp. 1-4, doi: 10.1109/IDAP.2017.8090324.
4. Alao, D. A. B. A., and A. B. Adeyemo. "Analyzing employee attrition using decision tree algorithms." Computing, Information Systems, Development Informatics and Allied Research Journal 4, no. 1 (2013): 17-28.
5. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset 10/03/2023
6. Alasadi, Suad A., and Wesam S. Bhaya. "Review of data preprocessing techniques in data mining." Journal of Engineering and Applied Sciences 12, no. 16 (2017): 4102-4107.
7. ]Alavi, Mousa, Denis C. Visentin, Deependra K. Thapa, Glenn E. Hunt, Roger Watson, and Michelle L. Cleary. "Chi-square for model fit in confirmatory factor analysis." (2020).
8. Pham-Gia, Thu & Choulakian, Vartan. (2014). Distribution of the Sample Correlation Matrix and Applications. Open Journal of Statistics.04.330-344. 10.4236/ojs.2014.45033.
9. Li, Liam, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. "Massively parallel hyperparameter tuning." arXiv preprint arXiv:1810.05934 5 (2018).