# Human Resource Analysis and Employee Churn Prediction

V. Naga Pranava Shashank
*School of Computer Science and Engineering*
*Vellore Institute of Technology*
Chennai, India

Sathiyanarayan GS
*School of Computer Science and Engineering*
*Vellore Institute of Technology*
Chennai, India

Keerthana V
*School of Computer Science and Engineering*
*Vellore Institute of Technology*
Chennai, India

Sajal Verma
*School of Computer Science and Engineering*
*Vellore Institute of Technology*
Chennai, India

***Abstract*—Employee retention or attrition causes a major cost to any organization which on the long run effects the overall efficiency. This not only increases the Human Resource cost, but also impacts the market value of an organization. Many businesses around the world are trying to find a solution to this serious issue. In this study, we have used supervised learning models which describe, demonstrate, and predict the employee attrition within an organization. Supervised learning models like Linear Regression, Decision Tree, Random Forest, K-NN, SVM and an ensemble model of Decision Tree and Random Forest has been applied which will predict whether an employee will leave the organization or not based on the metrics like number of projects, satisfaction rate, average monthly hours, effectiveness of employee appraisal, etc. and also to predict various other related features. The performance of all these models is evaluated and established using statistical methods. Using these results, various reasons for employee attrition can be identified and used by the management to take preventive measures for each employee individually.**

***Keywords—Predictive analytics, Employee Attrition, Decision Tree, Random Forest, Ensemble model.***

## I. Introduction

Human resource analysis is the examination of an organization's current workforce, analysis of its strengths and weaknesses, and development of strategies to maximize the use of human resources. Human resource analysis entails gathering and evaluating data on the abilities, experience, and performance of employees, as well as the organizational culture and structure. Predicting employee turnover is the process of using data and statistical analysis to determine which employees are most likely to depart a firm in the near future. This forecast is based on a review of multiple criteria, including employee performance, job satisfaction, prospects for professional advancement, remuneration, and benefits.

Employee attrition rate is the rate at which employees leave the organization due to environmental changes or due to the rules of the company. This can happen for several reasons. These include dissatisfaction with the employee perks or compensation structure, lack of opportunities for professional growth, and even terrible working environment. Let's see what is the need for employee churn prediction for the organizations.

- If managers or HR learned that some employees were considering leaving the company, they may speak with those individuals to persuade them to stay or manage the workforce by recruiting a replacement for those individuals.

- If the HR for a particular project learns of a worker who wants to leave the company, he or she can control the number of hirings and secure the important.

- The project pipeline will be fluid if all of the employees are working continuously on it, but the workflow won't be as smooth if, for example, one of the project's efficient employees abruptly leaves the organization.

In the past, the "rates" have received most of the attention rather than specific terminations. In an effort to forecast future turnover rates, we calculate historical rates of turnover. And it is crucial that you do it and keep doing it.

### A. *Different types of attrition:*

- If only two or three employees retired from the organization this year, statistically speaking, this employee group is too tiny to be included in attrition. However, attrition may occur if a substantial portion of the team departs at the same time.
- The senior professionals may decide to retire early or become independent consultants for reasons unrelated to age, thus attrition due to retirement shouldn't be ignored.
- High-value talent voluntarily attrition should be aggressively avoided as it might gradually reduce productivity. For instance, it is obviously a cause for concern if a corporation notices that its marketing professionals are leaving various business units.
- Employees in this situation are leaving one department to work in another. Internal attrition can be advantageous in some circumstances since it directs talent toward more lucrative fields. Additionally, it guarantees greater job-employee fit.
- However, it is worth looking into if a certain department has experienced a high rate of turnover in a single year. Does the job still need something? Is the management lacking in abilities? HR should investigate the replies to these queries.

In this research, we are going to apply known models like Linear Regression, Decision Tree, Random Forest, K-NN, etc. which will predict whether an employee will leave the organization based on the metrics like number of projects, satisfaction rate, average monthly hours, the effectiveness of employee appraisal, etc. The findings will enable businesses to anticipate employee turnover and as a result, lower their human resource expenses.

## II. Related Work

Employee churn is a critical issue for all businesses, since it negatively impacts their overall revenue and brand image. Several solutions based on machine learning (ML) have been developed to handle the ECn problem. Ineffectively handling the ECn problem, but crucial issues such as staff categorization, category-wise turnover forecast, and retention policy are neglected.

A multi-attribute decision making (MADM) based approach combined with ML techniques has been proposed in this research. Initially, an accomplishment-based employee importance model that employs a two-stage MADM technique to categorize individuals is built. In order to allocate relative weights to employee successes an entropy weight method is used by the authors. To measure the significance of the employees' class-based categorization performance order preference by similarity to ideal solution approach is adopted. The CatBoost algorithm is then used to predict employee attrition by class. The authors suggest an employee retention policy based on the observations in the end. The proposed Employee churn prediction and retention scheme was evaluated on a benchmark dataset from the human resource information system, and the findings were compared to those of other ML techniques using a number of performance criteria. It has been concluded that the CatBoost based system outperforms any other ML technique. For two reasons, ML approaches are applied to the problem of employee turnover. In the past few years, machine learning techniques have not been applied to address staff turnover issues. Second, machine learning techniques outperform the problem of employee turnover. From data collecting to determining the cause of employee attrition, the current research has been done using a prediction model as described below. The difficulty that many firms encounter is employee attrition, as valuable and experienced individuals leave the organization regularly. Several businesses around the world are attempting to eliminate this severe issue. The main purpose of this study was to construct a model that can predict whether an employee will leave the organization.

There may also be instances in which human resource believes an employee will leave the organization within a short period of time, but the individual does not. These blunders could be wealthy and bothersome for both employees and human resources but are a better bargain for relational advancement. On the other hand, there could be a false negative if a human resource does not provide the staff with encouragement/a raise and they quit the firm. The primary objective is to evaluate the effectiveness of employee evaluation and satisfaction rates inside the organization, which can help to prevent employee turnover. In this research, a novel approach centered on machine learning was applied to improve various employee retention strategies. This research also attempted to shed light on the various elements influencing the attrition rate of workers and their potential solutions.

The performance of each of the supervised machine learning approaches for forecasting employee turnover was statistically determined after a rigorous and exhaustive evaluation process. The inferences are as follows: tiny HR datasets may contain a high degree of randomness and unpredictability. This indicates that more time should be allocated to data quality evaluations and data augmentation in this instance. The selection of classifiers for small datasets should be made using a heuristic technique. For medium and large HR data sets, the data variance decreases and a more trustworthy model can be constructed.

Using tree-based ensemble approaches, such as extreme gradient boosting and gradient boosted trees, are best practice. Competent workers are a scarce commodity for great businesses. The problem of retaining talented, experienced employees poses a threat to business owners. The issue of employee turnover can be costly to employers, as it is difficult to replace their expertise and productivity. In this study, an automated model capable of predicting employee turnover based on various predictive analytic techniques has been presented. These techniques were applied to various pipeline architectures in order to select the best champion model. In addition, an autotuning strategy was implemented to determine the optimal hyper parameter combination for the champion model. Finally, an ensemble model for selecting the most effective model based on various evaluation metrics was proposed.

The results of the proposed model show that no model up until now could be considered ideal and perfect for each case of business context. Yet, the model proposed was pretty much optimal as per the considered requirements and adequately satisfied the intended goal. People analytics help organizations and HR professionals reduce attrition by changing recruitment and retention strategies in the age of data science and big data analytics. Employee turnover threatens productivity and planning in this climate. This research's main contributions were: a people analytics strategy for predicting employee turnover that emphasizes data quality over quantity and moves from big data to deep data was proposed.

This data-driven technique uses a hybrid methodology to create a meaningful employee attrition model and identify key employee attributes that affect attrition. Second, this attrition prediction technique uses machine, deep, and ensemble learning models and was tested on a big and medium simulated human resources dataset, followed by a real small dataset with 450 replies. The term "turnover intention" refers to an employee's reported willingness to leave her organization within a specified time frame, and it is frequently employed in the analysis of actual employee turnover. Since employee turnover can have negative effects on businesses and the labor market as a whole, it is essential to comprehend the factors that influence such a decision. A one-of-a-kind European-wide employee turnover survey was described and analyzed. Many baseline and cutting-edge categorization methods were examined in terms of their prediction capabilities.

Logistic regression and LightGBM proved to be the two most effective models. The evaluation was done based on the significance of the predictive features for these two models as a means of ranking the turnover intention factors. HR departments are studying employee turnover to understand its causes and motivations. This study uses an upgraded random forest algorithm to predict employee turnover and meet this need. For high-dimensional, unbalanced employee turnover data, the weighted quadratic random forest approach is recommended. First, the random forest method ranks and reduces feature dimensions. Second, the selected features are used using the random forest approach, and the F-measure values for each decision tree are calculated as employee turnover prediction model weights. The recommended algorithm for predicting employee turnover surpasses the random forest, C4.5, Logistic, BP, and others in recall and F-measure. The study hence provides a novel analytic tool that can assist human resource departments in more precisely predicting employee turnover, and its experimental results provide additional insights into how to reduce employee turnover intent. The increased capability for collecting, storing, and analyzing the huge volume of data as a result of the rapid development of information technology has altered the way in which organizational decision-makers approach their work. Human Resource Information Systems provide a wealth of employee-related data, but there are still a few best practices for utilizing this data to improve Human Resources decision-making.

Each organization has its own method for treating its employees and ensuring their pleasure. Infrequently, however, is the customer satisfaction rating measured. As a result, employees frequently quit their jobs abruptly and for no apparent reason. The Random Forest algorithm had the strongest ability to predict staff attrition, according to experimental results. The best accuracy of prediction was 85.12, which is regarded to be good accuracy. The research on predictive analytics for employee turnover remains appealing to both academic and business professionals. In general, Support Vector Machine delivers a trustworthy result for properly predicting staff attrition. Despite the fact that some methods now provide a high degree of accuracy for predicting employee churn, there is still a great deal to learn from all of them. With additional research on predictive analytics in employee turnover, it is anticipated that the optimal strategy for predicting employee churn in a particular firm will be identified.

## III. METHODOLOGY

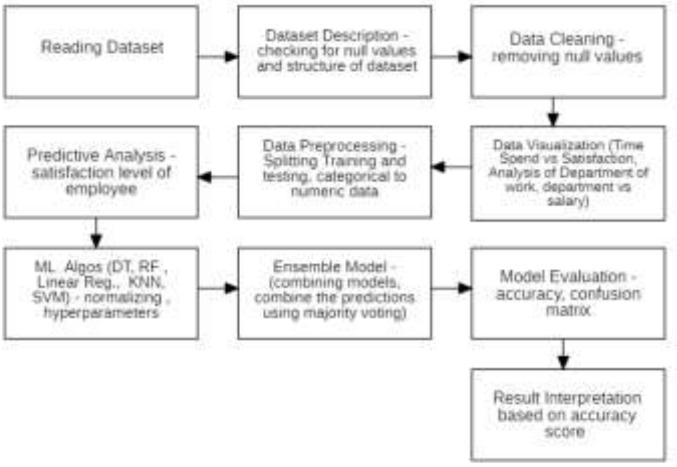

Reading Dataset
Dataset Description - checking for null values and structure of dataset
Data Cleaning - removing null values
Data Visualization (Time Spend vs Satisfaction, Analysis of Department of work, department vs salary)
Data Preprocessing - Splitting Training and testing, categorical to numeric data
Predictive Analysis - satisfaction level of employee
ML Algos (DT, RF , Linear Reg., KNN, SVM) - normalizing , hyperparameters
Ensemble Model - (combining models, combine the predictions using majority voting)
Model Evaluation - accuracy, confusion matrix
Result Interpretation based on accuracy score

Fig 1. Proposed Methodology

### A. Reason to choose proposed work

Predicting employee churn is an important responsibility for any organization, since excessive employee turnover can have numerous negative effects on the firm, including decreased productivity, increased recruitment and training costs, and decreasing morale among the existing workforce. By forecasting which employees are at risk of leaving, firms can take proactive steps to retain them, such as providing opportunities for training and development, higher remuneration packages, and a more favorable work atmosphere. Identifying the underlying causes of employee turnover can also provide useful insights into the organization's vulnerabilities and aid in the development of initiatives to solve these issues. Consequently, predicting staff attrition can result in enhanced employee happiness, retention, and overall organizational performance.

Human Resource analysis is crucial for businesses to comprehend and maximise their personnel. By this method, firms can collect and analyse information on their employees, including performance indicators, compensation, and benefits. This information can then be used to make educated decisions regarding the recruitment, retention, and development of people. Also, HR analysis can highlight areas where HR policies and procedures can be enhanced, such as diversity and inclusion efforts, employee engagement tactics, and talent management practices. In the end, HR analysis can assist firms in developing a more productive and engaged workforce, resulting in greater employee satisfaction and improved business outcomes.

### B. Novelty

Despite the widespread use of hyperparameter tuning and ensemble techniques to improve machine learning model performance, our study has shown that these approaches may not always result in better outcomes. Specifically, we employed hyperparameter tuning and developed an ensemble model using random forest and decision tree algorithms, but found that this approach did not outperform the regular models. This finding challenges the commonly held belief that hyperparameter tuning and ensemble techniques are always effective in improving model accuracy. Our results suggest that the effectiveness of these methods may depend on the specific dataset and the characteristics of the models used. Therefore, further research is needed to explore the conditions under which hyperparameter tuning and ensemble techniques are most effective, and to identify alternative methods for improving model performance when these approaches do not yield better results.

### C. Dataset Description

Datasets used for the project is taken online from kaggle.com in which it is available as “HR Analytics: Employee Churn Prediction” and “IBM HR dataset”. The HR Analytics dataset contains 14,999 records with 10 different attributes and IBR HR dataset contains 1470 records with 35 different attributes. The attributes present in the HR Analytics dataset are satisfaction level, last evaluation, number of projects, average monthly hours, time spent in the company, work accident, promotion in the last 5 years, Domain, Salary, Left. Basic summary of the datasets is given in table. 1, table 2 and table 3.

Table 1. Dataset Summary of IBM HR

| Attribute | min | 1Q | med | mean | 3Q | max |
|---|---|---|---|---|---|---|
| Age | 18 | 30 | 36 | 37 | 43 | 60 |
| Attrition | 0 | 0 | 0 | 0.16 | 0 | 1 |
| BusnessTravel | 0 | 1 | 1 | 1.09 | 1 | 2 |
| DailyRate | 102 | 465 | 802 | 803 | 1157 | 1499 |
| Department | 0 | 0 | 1 | 0.78 | 1 | 3 |
| HomeDist. | 1 | 2 | 7 | 9.19 | 14 | 29 |
| Education | 1 | 2 | 3 | 2.91 | 4 | 5 |
| Edn.Field | 0 | 1 | 2 | 1.98 | 2 | 5 |
| EmplCount | 1 | 1 | 1 | 1 | 1 | 1 |
| EmpNum | 1 | 491 | 1021 | 1025 | 1556 | 2068 |
| EnvSatisfactn | 1 | 2 | 3 | 2.72 | 4 | 4 |
| Gender | 0 | 0 | 1 | 0.6 | 1 | 1 |
| HourlyRate | 30 | 48 | 66 | 66 | 88 | 100 |
| JobInvolve | 1 | 2 | 3 | 2.73 | 3 | 4 |
| Joblevel | 1 | 1 | 2 | 2.06 | 3 | 5 |

Table 2. Dataset Summary of IBM HR

| | | | | | | |
|---|---|---|---|---|---|---|
| JobRole | 0 | 1 | 2 | 2.38 | 4 | 7 |
| JobSatisfaction | 1 | 2 | 3 | 2.73 | 4 | 4 |
| MaritalStatus | 0 | 1 | 1 | 1.24 | 2 | 2 |
| MonthlyInc. | 1009 | 2911 | 4919 | 6503 | 8379 | 19999 |
| MonthRate | 2094 | 8047 | 14236 | 14313 | 20462 | 26999 |
| NumComp | 0 | 1 | 2 | 2.69 | 4 | 9 |
| Over8 | 1 | 1 | 1 | 1 | 1 | 1 |
| Overtime | 0 | 0 | 0 | 0.28 | 1 | 1 |
| SalaryHike | 11 | 12 | 14 | 15.2 | 18 | 25 |
| Performance | 3 | 3 | 3 | 3.15 | 3 | 4 |
| Relation | 1 | 2 | 3 | 2.71 | 4 | 4 |
| StandardHrs | 80 | 80 | 80 | 80 | 80 | 80 |
| StockOpt | 0 | 0 | 1 | 0.79 | 1 | 3 |
| Totalwork | 0 | 6 | 10 | 11.3 | 15 | 40 |
| Training | 0 | 2 | 3 | 2.8 | 3 | 6 |
| worklifebal | 1 | 2 | 3 | 2.76 | 3 | 4 |
| yrscompany | 0 | 3 | 5 | 7.01 | 9 | 40 |
| yrscurrRole | 0 | 2 | 3 | 4.23 | 7 | 18 |
| Lastpromotion | 0 | 0 | 1 | 2.19 | 3 | 15 |
| yearswithcurr5 | 0 | 2 | 3 | 4.12 | 7 | 17 |

Table 3. Dataset Summary of HR Analytics

| Attribute | min | 1Q | med | mean | 3Q | max |
|---|---|---|---|---|---|---|
| Satisf.level | 0.09 | 0.44 | 0.64 | 0.62 | 0.82 | 1 |
| last_eval | 0.36 | 0.56 | 0.72 | 0.71 | 0.87 | 1 |
| numproject | 2 | 3 | 4 | 3.8 | 5 | 7 |
| Monthlyhrs | 96 | 156 | 200 | 201.1 | 245 | 310 |
| promotion | 0 | 0 | 0 | 0.02 | 0 | 1 |
| dept | 14999 | 0 | 0 | 0 | 0 | 0 |
| salary | 14999 | 0 | 0 | 0 | 0 | 0 |
| time_spent | 2 | 3 | 3 | 3.49 | 4 | 10 |
| workacc | 0 | 0 | 0 | 0.14 | 0 | 1 |

## D. Parameter Setting

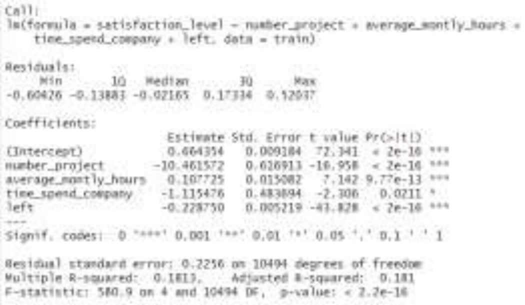
Choosing the optimal hyperparameters for machine learning models is a crucial step in building accurate and reliable models. This involves adjusting the values of certain parameters that are set before training and cannot be learned from the data. The aim is to find the values that result in the best model performance, as measured by a chosen evaluation metric. For different models such as k-Nearest Neighbors, decision trees, random forests, and support vector machines, there are specific hyperparameters that can be adjusted using the train function from the caret package in R. Tuning the hyperparameters requires testing different combinations of values and evaluating the resulting models until the best hyperparameters are identified. For KNN, we tuned the hyperparameter K, which determines the number of neighbors to consider when making a prediction. Increasing K results in a smoother decision boundary but may lead to underfitting. Decreasing K may lead to overfitting. For Decision Trees, we tuned the hyperparameter CP, which determines the complexity of the tree. A smaller CP results in a more complex tree with more splits, which may lead to overfitting. A larger CP results in a simpler tree with fewer splits, which may lead to underfitting. For Random Forests, we tuned the hyperparameter Mtry, which determines the number of features to consider when splitting each node in the tree. A larger Mtry may result in better performance but may also increase the runtime. For SVM, we tuned the hyperparameters sigma and C. Sigma determines the width of the kernel used in the SVM algorithm, while C determines the tradeoff between maximizing the margin and minimizing the misclassification rate. A smaller C results in a wider margin but may lead to misclassification, while a larger C results in a smaller margin but may lead to overfitting. Once the optimal hyperparameters have been found, the final model is trained using the training dataset. The process of selecting the best hyperparameters is an iterative one that aims to achieve the best possible model performance. According to table 4. the best hyperparameters have been identified, and the final model has been trained using the training dataset. After tuning the hyperparameters, we compared the performance of each model using two different datasets. We fixed the hyperparameters that gave the best results for both datasets. Overall, we found that hyperparameter tuning significantly improved the performance of each model.

Table 4. Hyperparameters for different models

| Model | Hyperparameters | Accuracy for dataset 1 | Accuracy for dataset 2 |
|---|---|---|---|
| K-NN | K= 5 | 0.9110905 | 0.8230881 |
| | K= 7 | 0.9102900 | 0.8299126 |
| | K= 9 | 0.9065898 | 0.8357475 |
| | K= 11 | 0.8990893 | 0.8337871 |
| | K= 15 | 0.8970892 | 0.8425724 |
| Decision Tree | Cp= 0.001 | 0.9787983 | 0.8299061 |
| | Cp= 0.012 | 0.9729974 | 0.8376827 |
| | Cp= 0.023 | 0.9711972 | 0.8464208 |
| | Cp= 0.034 | 0.9669972 | 0.8493476 |
| | Cp= 0.045 | 0.9603966 | 0.8464773 |
| Random Forest | Mtry= 1 | 0.9499956 | 0.8425790 |
| | Mtry= 2 | 0.9814979 | 0.8493848 |
| | Mtry= 3 | 0.9824982 | 0.8532778 |
| | Mtry= 4 | 0.9825982 | 0.8590844 |
| | Mtry= 5 | 0.9824982 | 0.8629680 |
| Support Vector Machine | Sigma= 0.3, C= 1 | 0.9598091 | 0.8425701 |
| | Sigma= 0.3, C=10 | 0.9650463 | 0.8425701 |
| | Sigma= 0.5, C= 1 | 0.9646663 | 0.8425701 |
| | Sigma= 0.5, C=10 | 0.9664750 | 0.8425701 |
| | Sigma= 0.5, C=100 | 0.9642847 | 0.8425701 |

## E. Statistical Analysis and Modeling

The goal of statistical modeling is to use the model to make predictions or test hypotheses about the relationship between variables. This involves fitting the model to the data, evaluating the goodness of fit of the model, and using the fitted model to generate predictions or conduct hypothesis tests. Statistical modeling is widely used in many fields, including finance, healthcare, marketing, and social sciences. It is an important tool for understanding complex relationships between variables and making data-driven decisions. We have performed Linear Regression which is a statistical modeling technique that is used to investigate the relationship between a dependent variable (response variable) and one or more independent variables (predictors). The summary of the model is depicted below:



```
Call:
lm(formula = satisfaction_level ~ number_project + average_montly_hours +
    time_spend_company + left, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60426 -0.13883 -0.02165  0.17334  0.52037

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.664354   0.009184  72.341  < 2e-16 ***
number_project      -10.461572   0.626913 -16.958  < 2e-16 ***
average_montly_hours  0.107725   0.015082   7.142 9.77e-13 ***
time_spend_company   -1.115476   0.483894  -2.306   0.0211 *
left                 -0.228750   0.005219 -43.828  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2256 on 10494 degrees of freedom
Multiple R-squared:  0.1813,    Adjusted R-squared:  0.181
F-statistic: 580.9 on 4 and 10494 DF,  p-value: < 2.2e-16
```

Fig 2. Summary of Model

The intercept (the expected value of sales when all predictor variables are 0) is estimated to be 0.664354, with a standard error of 0.009184. The t-value of 72.341 indicates that the intercept is significantly different from 0 ($p < 0.001$). The residual standard error is 0.2256, which represents the standard deviation of the residuals. This indicates that the model has a moderate amount of error in predicting sales. The multiple R-squared value of 0.1813 indicates that 18.13% of the variation in sales can be explained by the predictor variables. The adjusted R-squared value of 0.181 takes into account the number of predictor variables and is slightly lower. The F-statistic of 580.9 and p-value of < 2.2e-16 indicate that the overall model is significant and that at least one of the predictor variables is significantly associated with the output(left).

We have identified relationships between various variables through visualization as given below. Fig. 3 shows the correlation between various features from the HR Analytics dataset. It is noticed that the satisfaction level and left are negatively correlated and similarly there is also some negative correlation between work accident and left and also for time spent in the company and left feature. There is a good positive correlation between average monthly hours and number of projects, average monthly hours and number of projects and average monthly hours and last evaluation.

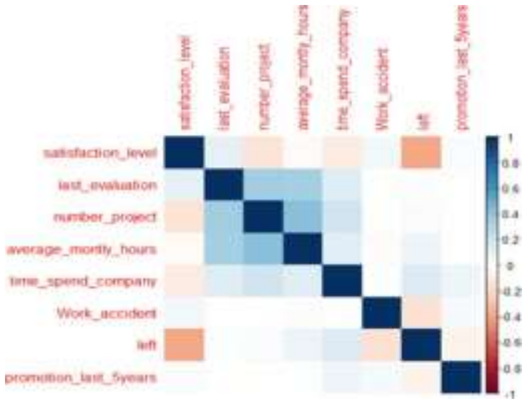


Fig 3. Correlation matrix of features in the dataset 1

Fig. 4 shows the relationship between the salary range and the people left in HR Analytics dataset. It was observed that there has been a significant number of people with the low and high salary who have left the company.

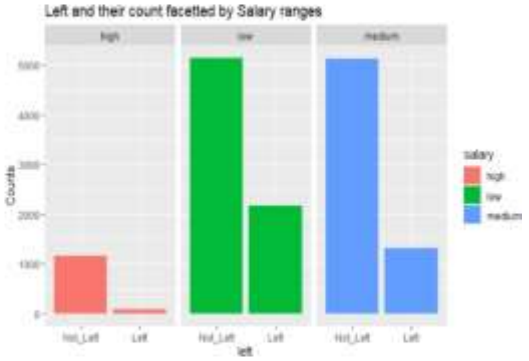


Fig 4. Count of people left vs the salary range

Fig. 5 represents the relationship between satisfaction levels and the count of people who are still working and left in HR Analytics dataset. The number of people who left the company is more when their satisfaction level is less than 0.5. From here, we can also infer that satisfaction level can also be taken as a serious measure in analyzing whether the person will leave or not beforehand and can take necessary actions.

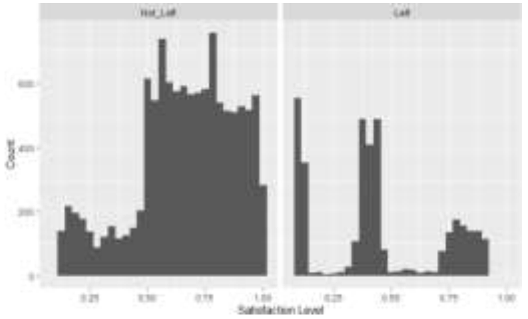


Fig 5. Satisfaction Level vs Left and Not Left

Fig 6. shows the percentage of people who left each department in the HR Analytics dataset. It is observed that the highest percentage of people left is noted in HR and accounting department. The lowest percentage of people leaving is from management and R & D department.

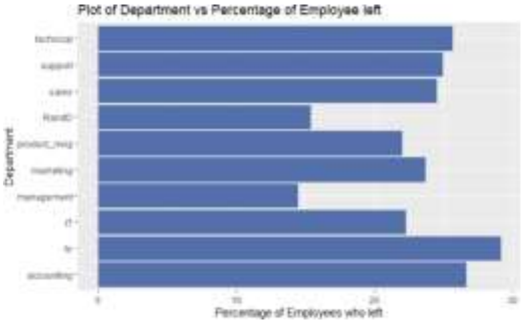


Fig 6. Department vs Employee’s Left

### *F. Data Preprocessing*

Data pre-processing is one of the most crucial and important steps that is required before applying any data analytics model. Any data analytics model uses machine learning algorithms where we basically feed the data from which it gets trained and makes predictions. It is necessary to take care that the data we feed is of suitable and valid format in order to get the desired output and proper working of the model. Having pre-processed and clean data helps in increasing the accuracy as well as the efficiency of the model.

The dataset is split into 70% training set and 30% testing set. Testing data is used to train the model according to the output desired and to make necessary predictions whereas testing data is used to evaluate the performance of the model and make all the necessary changes if needed to improve the efficiency of the model.

Data normalization is done in order to make all the variables to a similar scale for better stability of the model which generally uses gradient descent algorithms. Min - Max normalization is used for attributes like the number of projects, average monthly hours and time spent in the company in order to scale it and maximise the performance. (1) represents the mix – max normalization formula where xmin and xmax are the minimum and maximum absolute value of x. x is the old value and $x_{scaled}$ is the new value after normalization.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

## IV. Data Analytics Model

In this project, four classification models were used to predict, whether an employee will leave the company or not. The models that we are using are Linear Regression, Decision Tree, Random Forest Classifier, K nearest neighbors. The dataset was split into train and test in the 70:30 ratio.

### A. *Linear Regression*

Linear regression is a supervised learning algorithm that performs a regression task. It performs the task to predict a dependent variable value which is ‘y’ based on independent variable which is ‘x’. Here y is the target variable which is ‘left’ here. The hypothesis function for linear regression is given in (2) where theta1 is the intercept and theta2 is the coefficient of x. In our model, linear regression is used to predict the satisfaction level of the employees.

$$h_\theta(x) = \theta_0 + \theta_1 x \qquad (2)$$

### B. *Decision tree*

A decision tree is a supervised machine learning algorithm that produces a non-parametric model. It is mostly preferred for solving classification problems. The supervised part means that the decision tree is built in situations where the values of both the independent and dependent variables are known. Figure [9] shows the working of the decision tree. Decision trees are a great addition to our HR analytics toolbox. They easily find and leverage complex non-linear effects in our HR data and do so almost without the analyst’s involvement.

The fundamental problem that emerges while developing a decision tree is how to choose the optimal attribute for the root node and for sub-nodes. So, a method known as attribute selection measure, or ASM, may be used to tackle these issues. There are two widely used ASM approaches, which are as follows:

a) Information Gain:

$$Information\ Gain = Entropy_S - [(Weighted\ average * Entropy_{each\ feature})]$$

b) Gini Index:

$$Gini\ index = 1 - \Sigma_j P_j^2$$

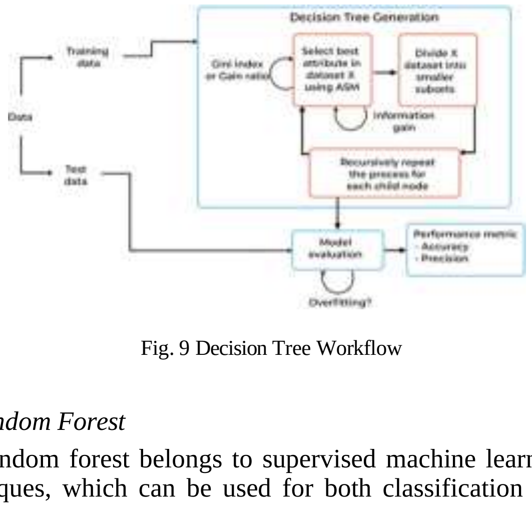


Fig. 9 Decision Tree Workflow

### C. *Random Forest*

Random forest belongs to supervised machine learning techniques, which can be used for both classification and regression problems. It is based on the concept of ensemble learning. It is the process of combining multiple decision trees to solve a problem. This is to improve the performance of the model. The model will get more accuracy when it has more decision trees. The algorithm for random forest is described below.

- In Random Forest, n records are randomly selected from a data collection of k records.
- Unique decision trees are built for every sample.
- An output will be created by each decision tree.
- For Classification and Regression, the final result is based on Majority Voting or Averaging, accordingly.

### D. *K-nearest neighbor*

KNN is a non-parametric and supervised learning algorithm, which makes classifications about the grouping of individual data points. It is typically used as a classification algorithm, where similar points in the dataset can be found near one another. The most common distance metrics used for KNN are Euclidean distance and Manhattan distance

We obtain smoother, more distinct borders between various classifications as K increases. Additionally, when we include more data points in the training set, the accuracy of the above classifier improves.

### *E.* SVM

A supervised machine learning approach called Support Vector Machine (SVM) is used for both classification and regression. Although we also refer to regression issues, categorization is the most appropriate term. Finding a hyperplane in an N-dimensional space that clearly classifies the data points is the goal of the SVM method. The number of features determines the hyperplane's size. The hyperplane is essentially a line if there are just two input features. The hyperplane turns into a 2-D plane if there are three input features. Imagining something with more than three features gets challenging.

SVM classification models are developed in four steps: choosing the input variables, preprocessing and splitting the data, selecting the model parameters, and putting the model into practice.

## V. Results And Discussion

### A. *Linear Regression*

Linear Regression was used to predict the satisfaction level of the employee based on the number of projects, average monthly hours and time spent in the company. The correlation between the actual and the predicted value was 0.417624. The mean squared error and standard error obtained were 5.13% and 22.65% respectively. Since the correlation coefficient is above 0.4, it is considered relatively strong. Outliers are highly detectable using linear regression. Therefore, before using linear regression on the dataset, outliers should be detected and eliminated. Another common issue noted was over-fitting which has to be prevented by employing various techniques like cross-validation and regularizations.

### B. *K-nearest neighbor*

KNN regression was used to predict the satisfaction level of the employee which can be a useful factor in predicting whether a person will leave or not. We took the K value as 10, the standard error is 0.2115 and the mean square error is 0.041. Residual values are the difference between the actual value and the predicted value. Fig. [10] shows that, the residual values for different inputs. As we increase K, we start to notice that errors are happening more often. Since it saves all the training data, the approach is also computationally costly. Compared to other supervised learning techniques, a lot of Memory is needed. When N is huge, prediction is slow. It is very sensitive to the scale of data as well as irrelevant features.

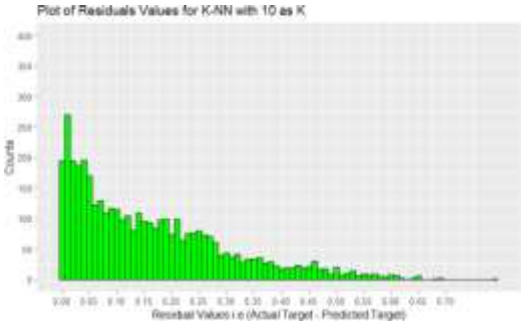


Fig 10. Residual Values of KNN for K = 9

The algorithm for the K-NN model has been given below:

```
Step 1: Load the necessary libraries required for running the code.

Step 2: Load the dataset into R.

Step 3: Split the dataset into training and testing sets.

Step 4: Select the best hyperparameters for the KNN model using the train () function from the caret package.

a. Set the formula for the model to be Attrition ~. where '.' represents all the predictor variables.

b. Set the method to "knn".

c. Set the train control method to "cv" with the number of folds set to 10.

d. Set the tune length to 10 and the metric to "Accuracy".

e. Print the best hyperparameters for the model and the results of all the hyperparameter combinations tried.

Step 5: Build the KNN model with the optimal hyperparameters found in Step 4.

a. Set the formula for the model to be Attrition ~ . where '.' represents all the predictor variables.
b. Set the method to "knn".

c. Set the train control method to "cv" with the number of folds set to 10.

d. Set the hyperparameters for k to 5.

Step 6: Make predictions on the test set using the predict () function with the KNN model built in Step 5.

Step 7: Calculate the accuracy of the KNN model using the confusion matrix and the sum of diagonal elements in the matrix.

Step 8: Print the accuracy of the KNN model.

Step 9: End.
```

### C. *Decision tree*

Using decision tree classifier to predict whether an employee will leave the company or not gave an accuracy of 96%. Decision Tree Classification is easy to grasp since it uses the same reasoning process that people use to decide what to do in everyday life. Each decision is represented by a node in the tree, with branches leading to potential outcomes. Decision trees can be interpreted easily and are useful for explaining the rationale behind a decision. It may be quite helpful for resolving issues with decisions. When compared to other methods, less data cleaning is necessary for this method.

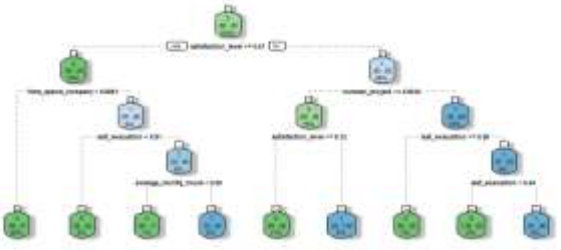


Fig 11. Decision Tree Plot

Fig. [11] shows the decision tree obtained after training the model for HR Analytics dataset. The satisfaction level is the root node, thus, plays a major role in predicting whether a person will leave the company or not. The decision tree is complicated as well since it has several tiers. It could have an overfitting problem, which the Random Forest method can fix. The computing complexity of the decision tree may rise with additional class labels.

Fig. [12] shows the decision tree obtained after training the model for IBM HR dataset. The monthly income is the root node, thus plays a major role in predicting whether a person will leave the company or not. The decision tree is complicated as well since it has several tiers. It could have an overfitting problem, which the Random Forest method can fix.

Fig 12. Decision Tree Plot

The algorithm for the Decision Tree model used is given below:

```
Step 1: Load the necessary libraries required for running the code.

Step 2: Load the dataset into R.

Step 3: Split the dataset into training and testing sets.

Step 4: Select the best hyperparameters for the rpart model using the train() function from the caret package.

    a. Set the formula for the model to be Attrition ~ . where '.' represents all the predictor variables.

    b. Set the method to "rpart".

    c. Set the train control method to "cv" with the number of folds set to 5.
    d. Set the hyperparameters to be tuned using the expand.grid() function. In this case, the complexity parameter 'cp' is tuned over a range of values from 0.001 to 0.1 with a length of 10.

    e. Set the metric to be "RMSE" and print the best hyperparameters for the model and the results of all the hyperparameter combinations tried.

Step 5: Build the rpart model with the optimal hyperparameters found in Step 4.

    a. Set the formula for the model to be Attrition ~ . where '.' represents all the predictor variables.

    b. Set the method to "class".

    c. Set the complexity parameter to the best value found in Step 4.

    d. Set the minimum number of observations required to split a node to be 2 and the minimum number of observations required in a terminal node to be 1.

Step 6: Visualize the decision tree using the fancyRpartPlot() function from the rpart.plot package.

Step 7: Make predictions on the test set using the predict() function with the rpart model built in Step 5.

Step 8: Calculate the accuracy of the rpart model by comparing the predicted values with the actual values in the test set.

Step 9: Print the accuracy of the rpart model.

Step 10: End.
```

### D. *Random Forest*

Random forest classification gave an accuracy of 98.25% for HR Analytics dataset and 85.33% for IBM HR dataset. It is considered one of the best accuracies obtained compared to other classification models. ROC curve in Fig. 13 shows the error rate of random forest at training and testing phase versus the number of trees used in the model for HR Analytics and Fig. 14 shows the curve for IBM HR dataset. It is observed that the error rate decreases as the number of decision tree used is increased. Missing values can also be handled easily using imputation.

Model is good at handling outliers and also is resilient to them which reduces the work of feature scaling. Complexity and longer training data can be the only disadvantages when compared to its efficiency and robustness.

The algorithm for the Random Forest model used is as follows:

```
Step 1: Load the necessary libraries required for running the code.

Step 2: Load the dataset into R.

Step 3: Split the dataset into training and testing sets.

Step 4: Select the best hyperparameters for the random forest model using the train() function from the caret package.

    a. Set the formula for the model to be as.factor(Attrition) ~ ., where '.' represents all the predictor variables.

    b. Set the method to "rf".

    c. Set the train control method to "cv" with the number of folds set to 10.

    d. Set the tuning grid using the expand.grid() function. In this case, the number of variables randomly sampled as candidates at each split (mtry) is tuned over a range of values from 1 to 5.

Step 5: Apply feature scaling to the dataset using the preProcess() function from the caret package with method = "range".

Step 6: Train the random forest model with the best hyperparameters found in Step 4 using the randomForest() function from the randomForest package.

    a. Set the predictor variables to all columns except the 'Attrition' column.

    b. Set the response variable to the 'Attrition' column.

    c. Set the number of trees in the forest to 100 and the number of variables randomly sampled as candidates at each split (mtry) to 3.

Step 7: Make predictions on the test set using the predict() function with the random forest model built in Step 6.

Step 8: Transform the predicted values into binary values by setting a threshold of 0.5.

Step 9: Calculate the confusion matrix using the table() function by comparing the predicted values with the actual values in the test set.

Step 10: Plot the random forest model using the plot() function from the randomForest package.

Step 11: Create an importance plot using the varImpPlot() function from the randomForest package.

Step 12: Calculate the accuracy of the random forest model by comparing the predicted values with the actual values in the test set.

Step 13: Print the accuracy of the random forest model.

Step 14: End.
```

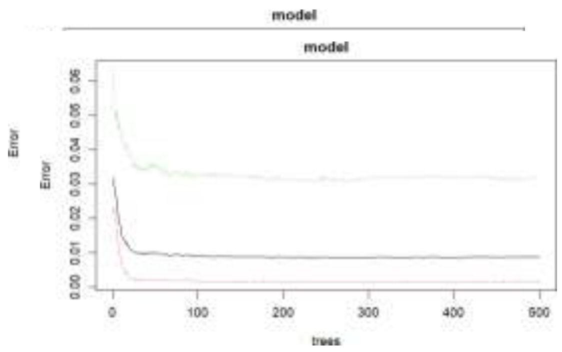


Fig. 13 ROC Curve of Random Forest

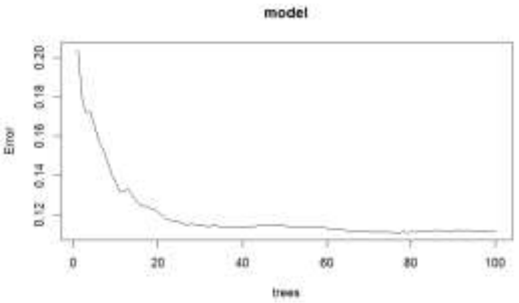


Fig. 14 ROC Curve of Random Forest

### E. *Ensemble Model*

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class. We have constructed the decision tree model and random forest model and combined it for producing an ensemble result. The performance was evaluated using accuracy. The accuracy we got is 76.2%. It means that the ensemble model correctly predicted the class labels for 76.2% of the samples in the test set. Ensemble models can often achieve higher accuracy and better generalization performance compared to single models. This is because they combine the predictions of multiple models, each of which may have different strengths and weaknesses, to arrive at a more robust and accurate prediction. By combining the predictions of multiple models, ensemble models can reduce the risk of overfitting, which occurs when a model is overly complex and fits the noise in the training data rather than the underlying patterns.

### F. *Support Vector Machine*

We performed SVM using radial, polynomial, linear and sigmoid kernels on both the datasets and evaluated the models using accuracy metrics. For the HR Analytics dataset the linear model resulted in an accuracy of 78.53%, while the polynomial kernel achieved an accuracy of 94.93%, the radial basis kernel resulted in an accuracy of 95.27% and the sigmoid kernel achieved 56.97% accuracy. On the other hand, the SVM model with the same hyperparameters when trained using the IBM HR dataset, the results achieved are as follows: for radial kernel the accuracy was 82.99%, for linear kernel the accuracy was 82.99%, for polynomial kernel the accuracy was 61.67%. For sigmoid kernel the accuracy was 82.99%. the results indicate that the radial basis kernel performed the best across all the kernels used to train the SVM model. This could be attributed to its ability to adapt to a wide range of different data types, including continuous, categorical and ordinal data. The radial basis kernel has a hyperparameter called sigma that controls the width of the kernel, and tuning this hyperparameter can be challenging but can lead to improved accuracy when done correctly. We have tuned this hyperparameter and after tuning we got the improved accuracy with sigma = 0.3 and cost = 10, the accuracy we got is 96.42%. Fig [11] shows the SVM model plot for satisfaction level s last evaluation for HR Analytics.

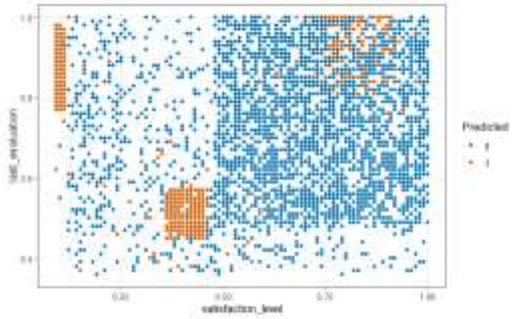


Fig. 11 Support Vector Machine

The algorithm for the SVM model used is given below:

Step 1: Set seed to ensure reproducibility of results.

Step 2: Split the data into training and testing sets using a 7 0:30 split ratio.

Step 3: Convert the target variable to a factor in the training set.

Step 4: Train four SVM models with different kernels (linea r, polynomial, radial, and sigmoid) using the svm() function with the training data.

Step 5: Make predictions on the testing data for each of the f our SVM models using the predict() function.

Step 6: Compute the confusion matrix and accuracy for eac h of the four SVM models using the table() function and su mming the diagonal elements of the confusion matrix.

Step 7: Print the accuracy of each SVM model using the pas te() function.

Step 8: Create a data frame combining the testing data and p redictions from the SVM model with radial basis kernel.

Step 9: Print the resulting data frame.

## VI. Conclusion

Organizations can get an idea of whether their employee will leave or not from this proposed model. The company can also predict the satisfaction level of the company which will give them an idea about their employees and helps them in taking prior initiatives in understanding the needs of the employee and take necessary actions to resolve them. Since the predictions are not 100% accurate, enough care must be taken to prevent trouble for both human resources and employees. The accuracy and efficiency of the prediction can also be increased by adding new attributes to the dataset in the future and also by using exploring various other machine learning models like neural networks for improved efficiency. After training both datasets and comparing their accuracy, we found that the models trained using the IBM HR analytics dataset had significantly lower accuracy than those trained on the HR analysis dataset that was used initially. One possible explanation for this discrepancy is the presence of several unimportant attributes in the IBM HR analytics dataset. To identify these attributes, we plotted a correlation matrix and examined which attributes had the least impact on accuracy. We inferred that in the dataset the attributes Marital Status, Education, yearsWithCurr5, Business Travel, StockOptionLevel, and Overtime are the features that have the most impact on the accuracy of the model and could possibly be a reason as to why the accuracy decreases in the models.

## VII. References

[1] N. Jain, A. Tomar, and P. K. Jana, “A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning,” Journal of Intelligent Information Systems, Sep. 2020, doi: https://doi.org/10.1007/s10844-020-00614-9.

[2] P. K. Jain, M. Jain, and R. Pamula, “Explaining and predicting employees’ attrition: a machine learning approach,” SN Applied Sciences, vol. 2, no. 4, Mar. 2020, doi: https://doi.org/10.1007/s42452-020-2519-4.

[3] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, “Employee Turnover Prediction with Machine Learning: A Reliable Approach,” Advances in Intelligent Systems and Computing, pp. 737–758, Nov. 2018, doi: https://doi.org/10.1007/978-3-030-01057-7_56.

[4] F. K. Alsheref, I. E. Fattoh, and W. M.Ead, “Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms,” Computational Intelligence and Neuroscience, vol. 2022, pp. 1–9, Jun. 2022, doi: https://doi.org/10.1155/2022/7728668.

[5] N. b. Yahia, J. Hlel, and R. Colomo-Palacios, “From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction,” IEEE Access, Access, IEEE, vol. 9, pp. 60447–60458, Jan. 2021, doi: https://doi.org/10.1109/ACCESS.2021.3074559.

[6] M. Lazzari, J. M. Alvarez, and S. Ruggieri, “Predicting and explaining employee turnover intention,” International Journal of Data Science and Analytics, vol. 14, May 2022, doi: https://doi.org/10.1007/s41060-022-00329-w.

[7] P. R. Srivastava and P. Eachempati, “Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction,” Journal of Global Information Management, vol. 29, no. 6, pp. 1–29, Nov. 2021, doi: https://doi.org/10.4018/jgim.20211101.oa23.

[8] X. Gao, J. Wen, and C. Zhang, “An Improved Random Forest Algorithm for Predicting Employee Turnover,” Mathematical Problems in Engineering, vol. 2019, pp. 1–12, Apr. 2019, doi: https://doi.org/10.1155/2019/4140707.

[9] M. Pratt, M. Boudhane, and S. Cakula, “Employee Attrition Estimation Using Random Forest Algorithm,” Baltic Journal of Modern Computing, vol. 9, no. 1, 2021, doi: https://doi.org/10.22364/bjmc.2021.9.1.04.

[10] “Employee Churn Rate Prediction and Performance Using Machine Learning,” International Journal of Recent Technology and Engineering, vol. 8, no. 2S11, pp. 824–826, Nov. 2019, doi: https://doi.org/10.35940/ijrte.b1134.0982s1119.