ISSN 2320-9984

Volume 12 Issue 3 | September 2024 ISSN: 2320-9984 (Online)

International Journal of Modern Engineering & Management Research

Website: www.ijmemr.org

A Predictive Approach to Employee Turnover Through Machine Learning

Monika Sahu
Research Scholar, M.Tech.
Computer Science and Engineering
Takshshial Institute of Engineering and Technology
Jabalpur (M.P.), India
Email: monikasahu00009@gmail.com

Swati Soni
Assistant Professor
Department of Computer Science and Engineering
Takshshial Institute of Engineering and Technology
Jabalpur (M.P.), India
Email: swatisoni@takshshila.org

Abstract**—***Employee turnover affects team relationships and operational effectiveness, making it a major concern for organisations. In this paper, we investigate machine learning approaches for the job of employee turnover prediction. 14,999 rows and 10 columns make up the dataset that was downloaded from Kaggle. Each record in the dataset represents a single employee and includes various attributes such as the employee's satisfaction level, last evaluation, number of projects, average monthly hours, time spent in the company, work accident, left (i.e., a binary column that indicates whether the employee has left the company (1) or is still employed (0)), promotion in the last five years, department, and salary level. Using ML models such as Decision Tree, Support Vector Classifier (SVC), Gradient Boosting, Random Forest, KNN, Logistic Regression, and Artificial Neural Network (ANN), the study uses binary classification over the dataset. A train-test split is used to evaluate the model initially, and then performance is validated using k-fold cross-validation (k=10). Gridsearch CV and Random search CV are used to tune hyperparameters for further optimisation. Among the models, the KNN Classifier demonstrated its efficacy in identifying employee turnover with an astounding accuracy of 100% on the test data. The Random Forest model performed robustly, following closely behind with an accuracy of 99.95%. The accuracy of the Gradient Boosting Classifier was 98.79%, whilst the Decision Tree Classifier achieved 99.04%. This study emphasises how crucial hyperparameter tweaking is to the optimisation of employee turnover prediction models, with KNN, RF, and DT demonstrating good prediction accuracy.*

Keywords:**—***Employee Turnover, ANN, Logistic Regression, Random Forest, KNN, Decision Tree, SVC, GBC*

1. INTRODUCTION

Successful human resource management and organisational success depend on having a solid understanding of the ability to predict employee turnover. Employee turnover is the rate at which current employees leave an organisation and are replaced by new ones. It is an important measure that businesses should monitor because it affects a number of company characteristics, such as financial performance, morale, and productivity. Employee departures can be either voluntary or involuntary, depending on the situation. Voluntary departures occur when an employee is fired or laid off by the organisation. Here are several reasons why predicting staff turnover is important, as well as how AI and machine learning might help with the process:

Early Identification of At-Risk Employees: Predictive models can find patterns linked to employee turnover by analysing a variety of data, including work history, job satisfaction, employee performance, and external market trends.

Cost Savings and Resource Allocation: The costs associated with hiring, onboarding, and training new employees can make employee turnover expensive. By identifying employees who are likely to leave, organisations can more efficiently allocate resources by concentrating on retaining important personnel and avoiding at-risk employees. This is made possible by the ability to predict turnover. To enable a seamless transition in the event of an employee's departure, this involves early recruiting activities.

Better Workforce Planning: Better workforce planning is made possible by knowing which departments or occupations are more likely to experience employee turnover. Customised Customer Engagement: Churn prediction aids in the customisation of communication tactics, enabling businesses to interact with clients in ways that are more likely to strike a chord and hold their attention.

Client Segmentation: Churn prediction makes it easier to segment customers, which enables businesses to allocate resources and efforts according to the particular requirements of various client segments.

Machine learning, a branch of artificial intelligence (AI), allows computers to learn from data and provide predictions or opinions without the need for explicit programming. While there are many different approaches used, the three primary types are as follows:

Supervised Learning: Using labelled datasets that link input data to intended results, algorithms are trained in supervised learning. The objective is to find an input-to-output mapping so that new, undiscovered data may be accurately predicted by the model.

Unsupervised Learning: Unsupervised learning looks for patterns, correlations, or structures in datasets that lack labels. Common assignments involve grouping together similar data points or reducing the dimensionality of the data.

Reinforcement Learning: Reinforcement learning aims to educate computers how to make decisions about actions that will ultimately maximise the total amount of rewards. The model picks up new abilities from its environment and receives feedback in the form of incentives or punishments based on its behaviour.

Machine learning helps artificial intelligence (AI) because it enables systems to get better over time through experience. Tasks in machine learning can be broadly classified into two categories:

1. ***Classification:*** Sorting incoming data points into discrete groups or classes is the first step in the supervised learning process. The model can be applied to projects where the output variable is categorical because it is designed to map input data to predefined output classes.

2. ***Regression***: Regression is a type of supervised learning in which continuous numerical values are predicted using input data. Because the model learns to establish a relationship between the input features and the output variable through regression problems, it becomes useful for tasks where the output variable is continuous.

The suggested work uses the following machine learning classification algorithms:

- Logistic Regression
- Random Forest Classifier
- KNeighbors Classifier
- Decision Tree Classifier

- Support Vector Classifier
- Gradient Boosting Classifier

There are serious drawbacks to train-test splitting using K-fold Cross-Validation. First off, there could be a lot of variation in performance evaluation due to the arbitrary nature of the data point selection for the training and testing sets. A model's performance may vary significantly across train-test splits due to this randomness, which makes it challenging to draw valid conclusions about how well a model generalises to new data. Additionally, train-test splitting might lead to data waste, especially if the dataset is small. Since some of the data is reserved for testing, there is less data available for training. When dealing with limited quantities of data, models that have fewer training data may be less robust or more prone to overfitting, which can be problematic.

To address these issues, we employ k-Fold Cross-Validation. The data is split into k folds, or subsets, using this procedure, and each fold is utilised as a training and testing set sequentially. It provides a more consistent and representative evaluation of a model's generalisation ability by averaging the performance over multiple iterations. Furthermore, by providing each data point with an equal chance of being included in both the training and testing sets, cross-validation ensures efficient use of the data. When there are insufficient data, this is particularly crucial. Cross-validation is an effective technique for assessing and improving machine learning models because it facilitates reliable comparisons between different models or parameter configurations and aids in model selection and hyperparameter adjustment.

We use k-Fold Cross-Validation to solve these problems. With this method, the data is divided into k folds, or subsets, and each fold is used iteratively as a training and testing set. An estimate of a model's capacity for generalisation is produced that is more reliable and representative by averaging the performance over these rounds. Additionally, cross-validation guarantees effective data use by giving every data point a chance to be included in both the training and testing sets. This is particularly crucial when there is insufficient data. Cross-validation is a solid method for evaluating and enhancing machine learning models since it helps with model selection and hyperparameter tweaking and allows for trustworthy comparisons across various models or parameter settings.

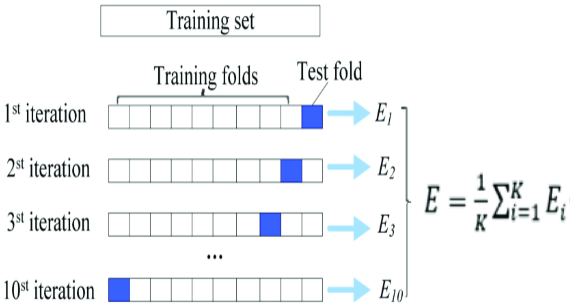


Figure 1: Ten-fold-cross-validation

Hyperparameter tuning - The rigorous process of identifying the optimal set of hyperparameters to maximise a machine learning model's performance is known as hyperparameter tuning, commonly known as hyperparameter optimisation. The aim of hyperparameter tuning is to identify the hyperparameter values that yield the best model performance, which can be defined as maximising prediction accuracy, minimising error, or reaching a certain performance measure. A validation dataset is typically used to evaluate the model's performance after it has been trained using each set of hyperparameters. Typically, this process involves testing with different hyperparameter values or ranges. Depending on the computer power and task complexity, several techniques, including random search, grid search, and Bayesian optimisation, can be used for hyperparameter tuning. The model may produce more accurate predictions on unknown data, prevent overfitting, and grow more accurate with the help of hyperparameter optimisation. Two popular techniques in machine learning for adjusting hyperparameters are RandomizedSearchCV and GridsearchCV. Their methods for navigating the hyperparameter space vary, even though they both aim to find the optimal collection of hyperparameters for a model.

The process of systematically searching for the perfect hyperparameter combination is known as hyperparameter tuning, often referred to as hyperparameter optimisation for a machine learning model to achieve optimal performance.

GridsearchCV is a thorough search technique that evaluates the model's performance for each and every possible combination of hyperparameters that fall inside a preset range. It creates a grid of hyperparameter values and trains and assesses the model for each combination. This approach is systematic and ensures that every possible combination of hyperparameters is examined. However, it could be computationally expensive, particularly if a wide range of values or a lot of hyperparameters are being employed.

GridSearchCV (Grid Search Cross-Validation)

Technique that evaluates the model's performance for all possible combinations of hyperparameters within a predefined range or set. It creates a grid of hyperparameter values and trains and validates the model for each combination.

Example: Random Forest

param_grid = {'n_estimators': [10, 50, 100, 200], # Number of Trees in Random Forest

'max_depth': [None , 10, 20, 30] # Maximum Depth of the Trees

}

On the other hand, RandomizedSearchCV adopts a more haphazard strategy. Rather of looking at every possible combination, it chooses a predetermined number of hyperparameter possibilities at random from the hyperparameter space. This randomness makes it more computationally efficient, especially when the hyperparameter space is huge. In comparison to GridsearchCV, RandomizedSearchCV may quickly study a wide range of hyperparameters, potentially leading to the discovery of optimal configurations faster. However, due to its randomness, there's a possibility that it won't identify the ideal pairing.

RandomSearchCV (Randomized Search Cross Validation)

It randomly selects a predetermined number of hyperparameter configurations from the hyperparameter space rather than analysing every possible combination.

RandomizedSearchCV can quickly explore a wide range of hyperparameters, potentially finding good configurations faster than Grid search CV. However, there's a chance it may miss the optimal combination due to its randomized nature.

2. BACKGROUND AND RELATED WORK

Mehul Jhaver et al. [1, 2019], proposed a system where the dataset has 14,999 records and 10 distinct attributes. The dataset's properties, a description, and potential values are displayed in the list that follows.

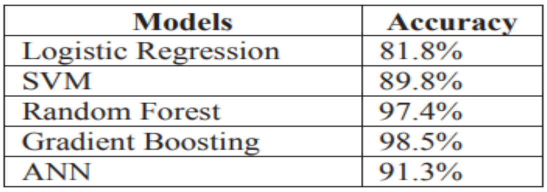
1. Level of employee satisfaction (0–1).
2. An employer's performance appraisal of an employee (0–1).
3. The quantity of projects the worker has finished.
4. The average number of hours an employee works each month.
5. The length of time an employee has worked for that company.
6. Any workplace accidents have ***included*** employees or not.
7. Was there a promotion for the staff member?

8. Which department does the employee work for?

9. Employee salary (Low, Medium, and High)

10. Did the worker depart from the company?

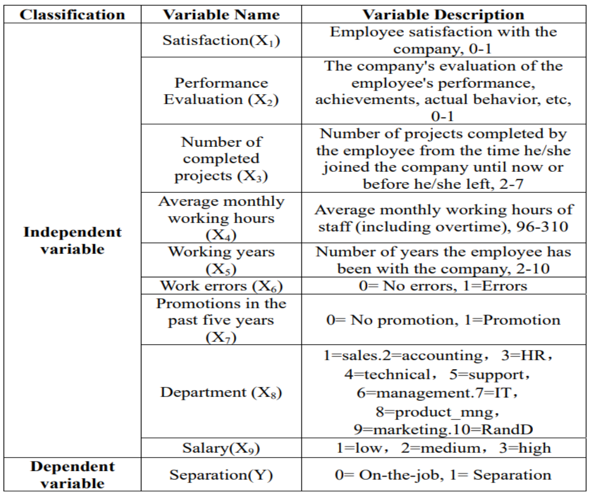
Data Preprocessing – handling missing values, before using machine learning algorithms, a few categorical attributes, such as salary and departments, needed to have their labels encoded. All of the columns have also been renamed for easier and better comprehension. Exploratory Data Analysis is done for data visualization, in this study the following classification models and boosting algorithms were implemented - Logistic Regression, SVM, Random Forest, GBC, Decision Tree Classifier, ANN.



Models	Accuracy
Logistic Regression	81.8%
SVM	89.8%
Random Forest	97.4%
Gradient Boosting	98.5%
ANN	91.3%

Figure 2: Accuracy of Models

Yongkang Duan et al. [2, 2022], Performing Predictive analysis on 15,000 sample data. Where accuracy rate of the Logistic regression model prediction is 79.23%, and accuracy of the XGBoost model prediction is 98,17%.



Classification	Variable Name	Variable Description
Independent variable	Satisfaction(X_1)	Employee satisfaction with the company, 0-1
	Performance Evaluation (X_2)	The company's evaluation of the employee's performance, achievements, actual behavior, etc, 0-1
	Number of completed projects (X_3)	Number of projects completed by the employee from the time he/she joined the company until now or before he/she left, 2-7
	Average monthly working hours (X_4)	Average monthly working hours of staff (including overtime), 96-310
	Working years (X_5)	Number of years the employee has been with the company, 2-10
	Work errors (X_6)	0= No errors, 1=Errors
	Promotions in the past five years (X_7)	0= No promotion, 1=Promotion
	Department (X_8)	1=sales.2=accounting， 3=HR， 4=technical， 5=support， 6=management.7=IT， 8=product_mng， 9=marketing.10=RandD
	Salary(X_9)	1=low， 2=medium， 3=high
Dependent variable	Separation(Y)	0= On-the-job, 1= Separation

Figure 3: Definition of the Variables

Vengai Musanga et al. [3, 2022],A sampling of the IBM dataset was used to create the study's data. 30 features remained in the dataset after 5 redundant features were deleted.

1. Open the Jupyter Notebook and import the IBM dataset.

2. Analysis of Exploratory Data.

3. Excessive sampling of the asymmetric data.

4. The data is split into 70% training data and 30% testing data for the purpose of adopting machine learning. Empirical research has demonstrated that optimal results are obtained when 20–30% of the data is used for testing and the remaining 70–80% for training, which is why the 70/30 ratio was selected.

5. Use the feature selection techniques of Recursive Feature Elimination, Information Gain, and Pearson Correlation to identify the most important variables for the prediction.

6. Use DT, KNN, GBM, LR, and RF for every feature selection technique.

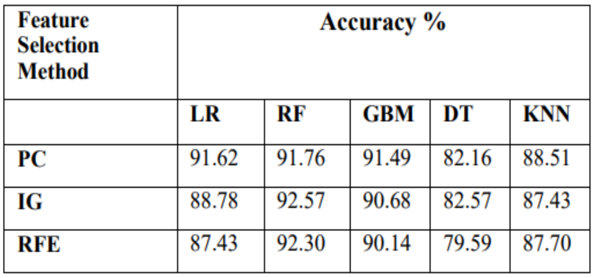
7. Using test data, the machine learning algorithms are compared in terms of accuracy, precision, recall, and F-Score.

8. The data is reanalyzed using the machine learning algorithms; however, feature selection techniques are not applied.

9. A comparison is made between the classification results obtained before and after feature selection techniques were applied.

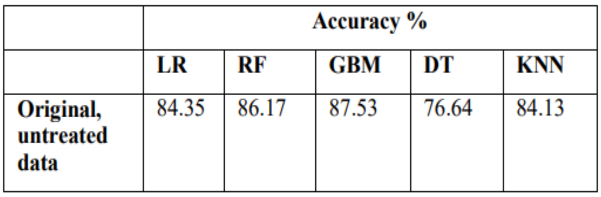
10. High accuracy and precision feature selection and classification methods are used.

11. An examination of the key elements that affect worker's decisions to quit a company is provided.



Feature Selection Method	Accuracy %				
	LR	RF	GBM	DT	KNN
PC	91.62	91.76	91.49	82.16	88.51
IG	88.78	92.57	90.68	82.57	87.43
RFE	87.43	92.30	90.14	79.59	87.70

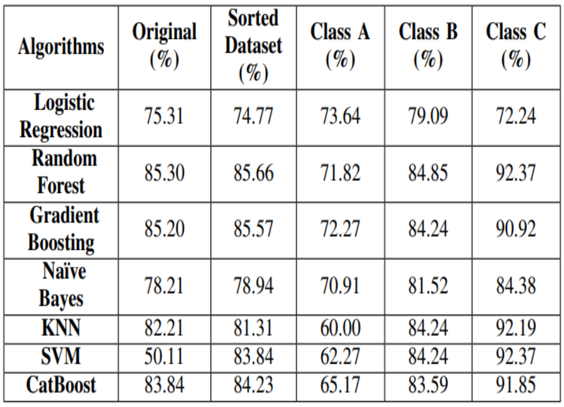
Figure 4: Classification Accuracy Scores Results with Feature Selection



	Accuracy %				
	LR	RF	GBM	DT	KNN
Original, untreated data	84.35	86.17	87.53	76.64	84.13

Figure 5: Classification Accuracy Scores Results Without Feature Selection

Shefayatuj Johara Chowdhury et al. [4, 2023], A total of 1471 employee datasets with 35 attributes are gathered from Kaggle for this approach. In this study data collection, feature selection, dataset development, MCDM and ML method application, and result comparison are all included in this work. Additionally, related ML algorithms and MCDM (Multi Criteria Decision Making) techniques are explained. All datasets in this study are divided into three categories: A, B, and C, using an integration of the AHP (Analytical Hierarchy Process) and TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) methodologies. Ten criteria are used in the model's application to a Kaggle1 human resources dataset. Initially, the dataset was subjected to a series of machine learning methods, including Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, KNN, SVM, and CatBoost, to determine its accuracy, precision, and recall. Subsequently, the most crucial characteristics that contribute to employee turnover are quantified. Six of the most significant criteria were selected, and the dataset was categorised using AHP, which assigned weight to the relatively simple criterion. The acquired weights are then entered into the TOPSIS procedure to rank the employees according to how likely they are to resign.



Algorithms	Original (%)	Sorted Dataset (%)	Class A (%)	Class B (%)	Class C (%)
Logistic Regression	75.31	74.77	73.64	79.09	72.24
Random Forest	85.30	85.66	71.82	84.85	92.37
Gradient Boosting	85.20	85.57	72.27	84.24	90.92
Naïve Bayes	78.21	78.94	70.91	81.52	84.38
KNN	82.21	81.31	60.00	84.24	92.19
SVM	50.11	83.84	62.27	84.24	92.37
CatBoost	83.84	84.23	65.17	83.59	91.85

Figure 6: Comparative Analysis

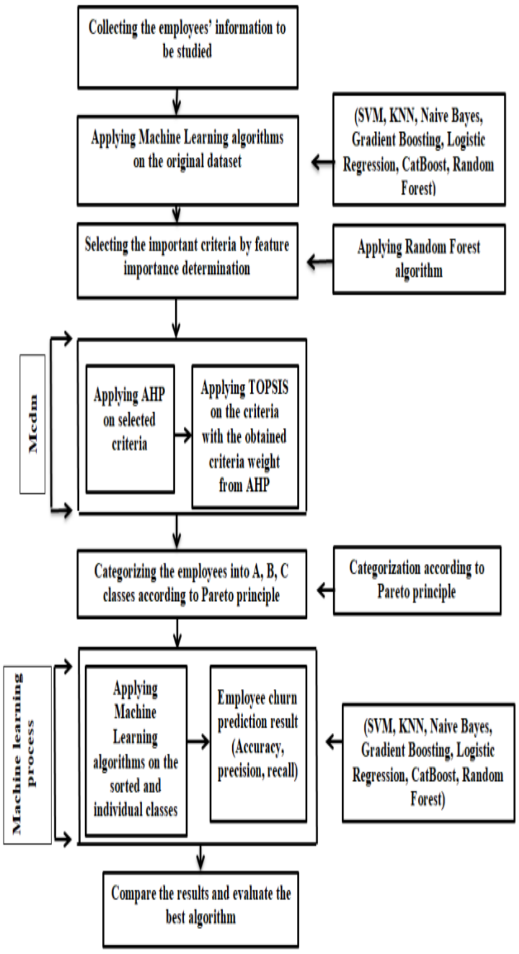

Collecting the employees' information to be studied
Applying Machine Learning algorithms on the original dataset
(SVM, KNN, Naive Bayes, Gradient Boosting, Logistic Regression, CatBoost, Random Forest)
Selecting the important criteria by feature importance determination
Applying Random Forest algorithm
Mcdm
Applying AHP on selected criteria
Applying TOPSIS on the criteria with the obtained criteria weight from AHP
Categorizing the employees into A, B, C classes according to Pareto principle
Categorization according to Pareto principle
Machine learning process
Applying Machine Learning algorithms on the sorted and individual classes
Employee churn prediction result (Accuracy, precision, recall)
(SVM, KNN, Naive Bayes, Gradient Boosting, Logistic Regression, CatBoost, Random Forest)
Compare the results and evaluate the best algorithm

Figure 7: Work flow diagram of the proposed method

Train-test splitting has its disadvantages, Firstly, it can result in high variability in performance evaluation because the choice of data points in the training and testing sets is random. This randomness means that the performance of a model can vary significantly between different train-test splits, making it challenging to draw consistent conclusions about how well a model generalizes to unseen data. Result may Lead to Overfitting or Underfitting Situation.

Accurate prediction of Employee Turnover is required, the traditional methods like train-test splitting have shown limitations in delivering precise results. The Primary Objective is to identify the most effective combination of Data Preprocessing Steps and Machine Learning Algorithms to achieve the Highest Prediction Accuracy.

Research Gap Found:

Cross Validation: Apply k-fold cross-validation on the (x, y) data to evaluate model performance across diverse validation sets.

Hyperparameter Tuning: Identify the best-performing models based on cross-validation accuracy. Fine-tune selected models through GridsearchCV and RandomsearchCV to optimize hyperparameters.

Model Testing: Test the models with optimized hyperparameters on the testing dataset to assess their predictive capabilities.

Model Evaluation: Evaluate the final models' performance using confusion matrices to measure accuracy, precision, recall, and F1-score.

3. PROPOSED FRAMEWORK

Data Collection -https://www.kaggle.com/code/serkanp/employee-turnover-prediction/input

Dataset Shape: 14999 rows × 10 columns, File Size: 553 KB

Each row represents an individual employee with various attributes.

IDE - Google Collaborator / Python– Python 3

Applied Binary Classification Problem on Data

1. Importing Libraries

Drive Mount & Data Collection, reading csv file (turnover.csv) using pandas

2. Data Preprocessing

- Exploratory Data Analysis
- Check for Null Values
- Drop Duplicate Rows
- Label Encoding
- Separating the Dataset Features into features (x) and target (y).
- Use train_test_split, Train Data 80%, Test Data 20%.

3. Model Creation:

Classification models are created on Training – Testing Data, the models are:

- Logistic Regression
- Random ForestClassification
- K Neighbors Classifier
- Decision Tree Classifier
- Support Vector Classifier
- Gradient Booster Classifier

4. k-Fold Cross Validation over (x,y) data is Applied on the Models

5. Hyperparameter Tuning

- Based on Cross Validation Accuracy, Selected Best Models then applied GridsearchCV
- And RandomsearchCV for Hyperparameter Tuning

6 Model Evaluation with Best Parameters

7. Predictive Modeling

- That uses input features to predict an outcome, whether the employee is likely to leave the job or not.

8. ANN Model Creation (Train Data 70%, Validation Data 15%, Test Data 15%)

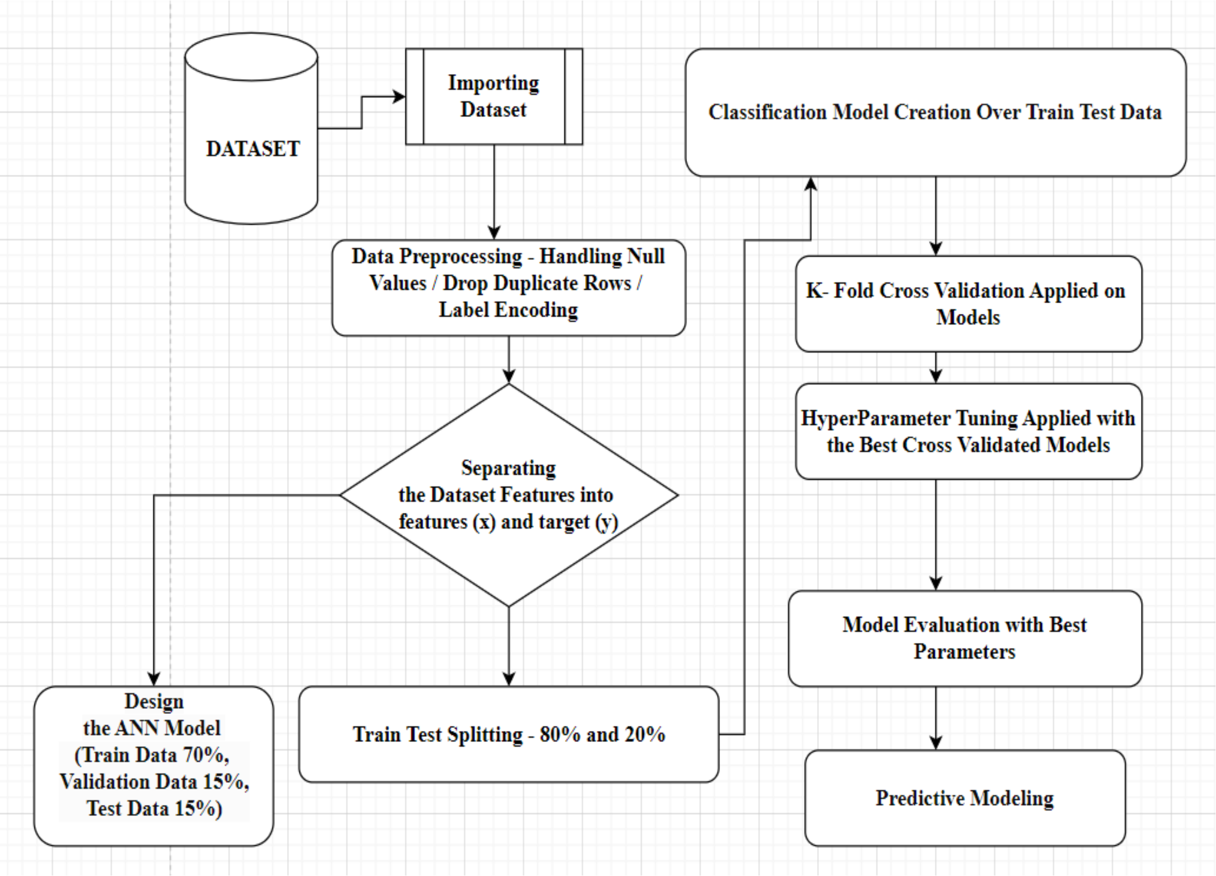

DATASET
Importing Dataset
Data Preprocessing - Handling Null Values / Drop Duplicate Rows / Label Encoding
Separating the Dataset Features into features (x) and target (y)
Design the ANN Model (Train Data 70%, Validation Data 15%, Test Data 15%)
Train Test Splitting - 80% and 20%
Classification Model Creation Over Train Test Data
K- Fold Cross Validation Applied on Models
HyperParameter Tuning Applied with the Best Cross Validated Models
Model Evaluation with Best Parameters
Predictive Modeling

Figure 8. Proposed Work Flow

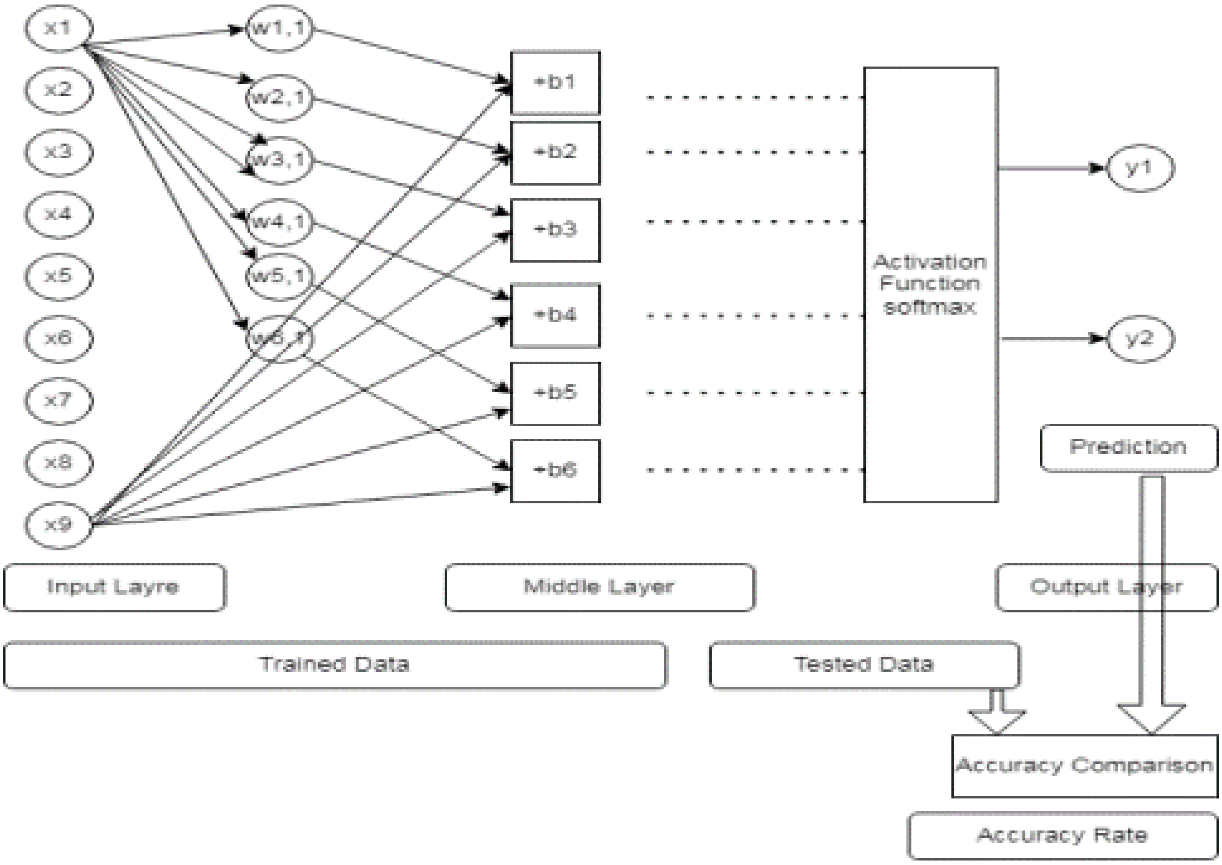


Figure 9. Proposed – ANN Model

Table 1. Dataset Attributes

#	Feature	Description	Value Range
1	satisfaction_level	The level of job satisfaction reported by employees.	0 – 1 0 indicates low satisfaction 1 indicates high satisfaction.
2	last_evaluation	A numerical value indicates the result of the most recent performance evaluation for each employee	0 – 1 0 low Performance 1HighestPerformance
3	number_project:	The number of projects an employee is currently or has been involved in. It could provide insight into the workload or the variety of tasks an employee handles.	2-7 Projects
4	average_montly_hours	The average number of hours worked by employees per month	100 –300 Hrs
5	time_spend_company	The number of years an employee has spent working for the company.	2- 10 Years
6	Work_accident	Whether an employee has had a work-related accident. Incidents that result in injuries or harm to an employee while performing job duties	1- Indicates that the employee has experienced a work accident. 0- Indicates that the employee has not experienced a work accident.
7	left	A binary column	1 – Employee Left the company 0 – Still employed
8	promotion_last_5years	This binary column may indicate whether an employee has received a promotion in the last five years. Promotions can influence job satisfaction	1 – Promoted 0 – Not Promoted
9	dept	The department or team to which an employee belongs. It's a categorical variable that may influence turnover predictions	IT - 0 RandD - 1 accounting -2 hr - 3 management - 4 marketing -5 product_mng - 6 sales - 7 support - 8 technical - 9
10	salary_level	This column probably represents the salary level of employees. It could be categorized into different levels, such as low, medium, and high. Salary is often a significant factor in employee satisfaction and retention.	Low - 0 medium - 1 high - 2

4. RESULTS

This section presents the results and evaluation of the different classifierswith k-fold cross validation and after hyperparameter tuning when final models created with best parameters.

Table 2: Results of the Different Classifiers

#	Classifiers	Train Test Split	Cross Validation CV=10		GridsearchCV / RandomsearchCV
1	Logistic Regression	Train Accuracy 0.8323603002502085 Test Accuracy – 0.8399333055439766 Model Execution Time in Sec: 0.13254022598266602			

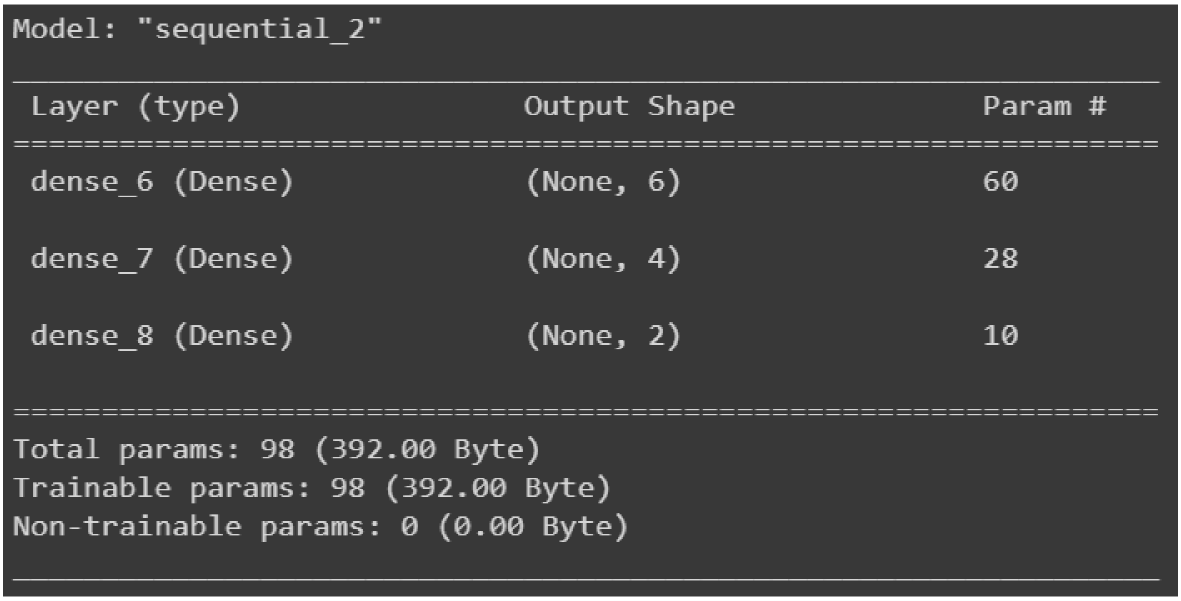
2	Random Forest	Train Accuracy - 1.0
		Test Accuracy - 0.9874947894956232
		Model Execution Time in Sec: 0.9058670997619629
			[0.98583333
			0.99082569
			0.98331943
			0.97998332
			0.98582152
			0.98582152
			0.98582152
			0.98331943
			0.98331943
			0.97998332]
				mean_accuracy_ RF 98.4405	Best Hyperparameters from Grid Search: {'max_depth': 30, 'n_estimators': 100}
					Best Accuracy: 0.984571656936336
					Best Hyperparameters from Random Search: {'n_estimators': 100, 'max_depth': 30}
					Best Accuracy: 0.9847385321100918

3	KNeighbour	Train Accuracy – 0.9641367806505421
		Test Accuracy - 0.9383076281784076
		Model Execution Time in Sec: 0.01872110366821289
			[0.93916667
			0.9557965
			0.93494579
			0.93577982
			0.93911593
			0.94078399
			0.93911593
			0.93911593
			0.94495413
			0.91075897]
				mean_accuracy_ KNC 93.7953	Best Hyperparameters from Grid Search: {'n_neighbors': 9, 'p': 1, 'weights': 'distance'}
					Best Accuracy: 0.9465429541986671
					Best Hyperparameters from Random Search: {'weights': 'distance', 'p': 1, 'n_neighbors': 11}
					Best Accuracy: 0.9470433712128455

4	Decision Tree	Train Accuracy - 1.0
		Test Accuracy - 0.9733222175906627
		Model Execution Time in Sec: 0.11080384254455566
			[0.965
			0.97414512
			0.96747289
			0.96080067
			0.95829858
			0.96914095
			0.97331109
			0.97247706
			0.96747289
			0.9499583]
				mean_accuracy_ DT 96.5808	Best Hyperparameters from Grid Search: {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'}
					Best Accuracy: 0.9818192943195333
					Best Hyperparameters from Random Search: {'splitter': 'best', 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_depth': 7, 'criterion': 'gini'}
					Best Accuracy: 0.9814855439140787

#	Classifiers	Train Test Split	Cross Validation CV=10		GridsearchCV / RandomsearchCV
5	Support Vector	kernel='linear' Train Accuracy - 0.8339241034195163 Test Accuracy - 0.8340975406419341 Model Execution Time in Sec: 187.4665083885193 kernel='poly' Train Accuracy - 0.8339241034195163 Test Accuracy - 0.8340975406419341 Model Execution Time in Sec: 8.67737865447998			

6	Gradient Booster Classifier	Train Accuracy - 0.98185988323603
		Test Accuracy - 0.9808253438932889
		Model Execution Time in Sec: 0.9779415130615234
			[0.98333333
			0.98915763
			0.97998332
			0.97080901
			0.98331943
			0.98081735
			0.98415346
			0.97914929
			0.98081735
			0.97164304]
				mean_accuracy_ GBC 98.0318	Best Hyperparameters from Grid Search: {'learning_rate': 0.01, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
					Best Accuracy: 0.9826531836138285
					Best Hyperparameters from Random Search: {'n_estimators': 70, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_depth': 6, 'learning_rate': 0.01}
					Best Accuracy: 0.9831536006280069



```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_6 (Dense)             (None, 6)                 60

 dense_7 (Dense)             (None, 4)                 28

 dense_8 (Dense)             (None, 2)                 10

=================================================================
Total params: 98 (392.00 Byte)
Trainable params: 98 (392.00 Byte)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 10. ANN Designed

Table 3. Proposed ANN- Accuracy, Loss

#	Hyperparameter	Setting
1	Input Size	9
2	Epochs	30
3	Batch Size	16
4	Activation Function	relu
5	Number of Hidden Layers	2
6	Output Activation Function	softmax
7	Optimizer	Adam
8	Loss	SparseCategoricalCrossentropy
9	Accuracy in %	0.8221
10	Loss in %	0.4074

Table 4: Final Model Accuracies with Best Parameters

#	Classifiers	Accuracy % on Test Data
1	RandomForestClassifier	0.9995831596498541
2	KNN Classification	1.0
3	DecisionTreeClassifier	0.9904126719466444
4	Gradient Booster Classifier	0.9879116298457691

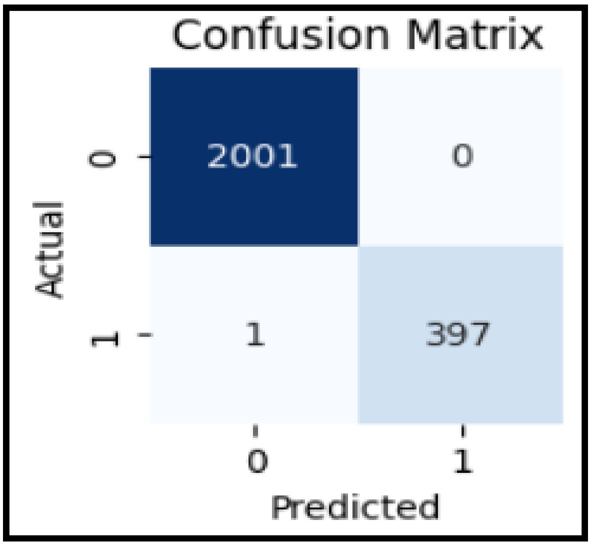


Figure 11. Confusion Matrix for Random Forest Classifier Model with Best Parameters

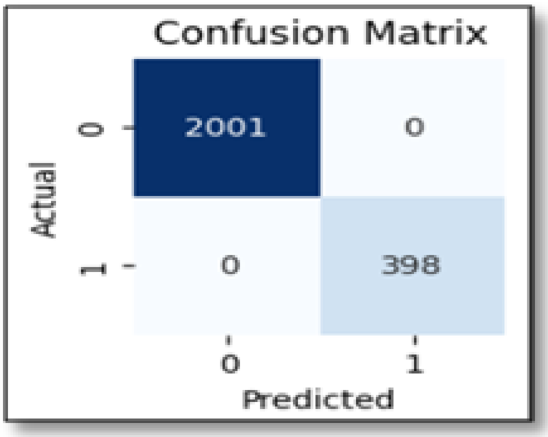


Figure 12. Confusion Matrix for KNN Classifier Model with Best Parameters

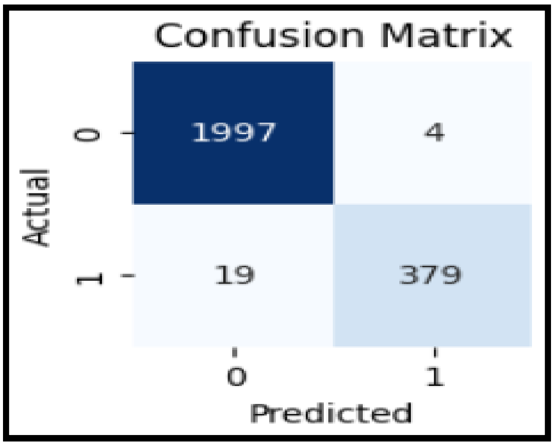


Figure 13. Confusion Matrix for Decision Tree Classifier Model with Best Parameters

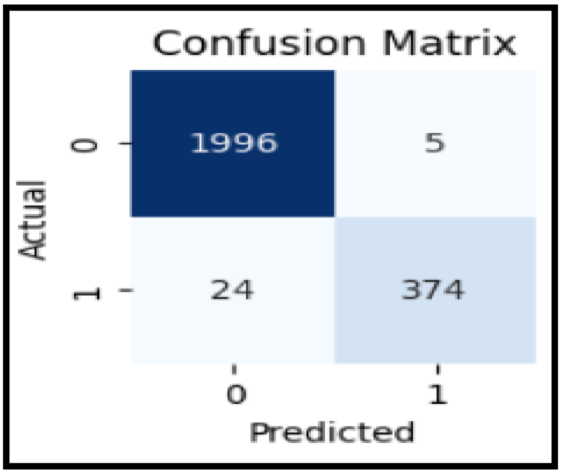


Figure 14. Confusion Matrix for Gradient Booster Classifier Model with Best Parameters

5. CONCLUSION AND FUTURE WORK

Employee turnover is a big concern for organizations because it has an impact on team dynamics and operational efficacy. In this work, we examine machine learning techniques for the task of forecasting employee turnover. The downloaded dataset has 10, columns and 14,999 rows from Kaggle. A single employee is represented by each record in the dataset, which contains a variety of information about them, including their degree of satisfaction, their most recent evaluation, the number of projects they have worked on, their average monthly hours, how long they have been with the company, any work-related accidents, whether they have left the company (a binary column that indicates whether an employee has left the company or is still employed) and whether they have received promotions in the previous five years.

The study employs binary classification across the dataset using machine learning (ML) models as Decision Tree, Support Vector Classifier (SVC), Gradient Boosting, Random Forest, KNN, Logistic Regression, and Artificial Neural Network (ANN). The model is first evaluated using a train-test split, and its performance is then confirmed using k -fold cross-validation (k=10). Hyperparameters are adjusted using GridsearchCV and RandomsearchCV in order to maximize optimization. Using test data, the KNN Classifier outperformed the other models in identifying employee turnover with an amazing 100% accuracy. With an accuracy of 99.95%, the Random Forest model trailed closely behind in performance. While the Decision Tree Classifier achieved 99.04% accuracy, the Gradient Boosting Classifier achieved 98.79%. With KNN, RF, and DT showing strong prediction accuracy, this study highlights how important hyperparameter tweaking is to the optimization of employee turnover prediction models.

Look into advanced feature engineering techniques for upcoming work in order to find fresh and pertinent features that could have a big influence on turnover prediction. Examine other elements and characteristics, such as employee feedback, that could affect employee turnover but were left out of the current study. Utilise NLP tools to examine employee evaluations, comments, and feedback. This can reveal insightful information on the mood of the workforce and point out possible causes of turnover. Investigate the use of ensemble models to integrate the advantages of various techniques, which may increase the overall robustness and accuracy of predictions. Examine more complex deep learning architectures, such as Transformer models or Long Short-Term Memory (LSTM), as they may be able to capture the intricate correlations and temporal patterns between employee behaviours and turnover. Provide a system that predicts employee turnover in real time. This would enable organisations to keep a close eye on the risk of employee attrition and take prompt action in response to evolving circumstances or updated data.To further increase accuracy and robustness, look into the use of ensemble models that combine the advantages of several methods (such as Random Forest, Gradient Boosting, and Neural Networks).

REFERENCES:

[1] Mehul Jhaver, Yogesh Gupta, Amit Kumar Mishra, “Employee Turnover Prediction System”, 2019 4th International Conference on Information Systems and Computer Networks (ISCON), GLA University, Mathura, UP, India. Nov 21-22, 2019, 978-1-7281-3651-6/19/$31.00 ©2019 IEEE.

[2] Yongkang Duan,” Statistical Analysis and Prediction of Employee Turnover Propensity Based on Data Mining”, 2022 International Conference on Big Data, Information and Computer Network (BDICN) | 978-1-6654-8476 -3/22/$31.00©2022 IEEE | DOI: 10.1109/ BDICN55575.2022.00052.

[3] Vengai Musanga, Edmore Tarambiwa, Kudakwashe Zvarevashe, “A Supervised Machine Learning Model to Optimize Human Resources Analytics for Employee

Churn", 2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT) | 978-1-6654-7576-1/22/ $31.00 ©2022 IEEE | DOI: 10.1109/ ZCICT55726.2022.10045987.

[4] Shefayatuj Johara Chowdhury, Mainul Islam Mahi, Sadiqul Alam Saimon, Aynur Nahar Urme, Rashidul Hasan Nabil, "An Integrated Approach of MCDM Methods and Machine Learning Algorithms for Employees' Churn Prediction", 2023 3rd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) | 979-8-3503-4643-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ ICREST57604.2023.10070079.

[5] Raj Chakraborty, Krishna Mridha, Rabindra Nath Shaw, Ankush Ghosh, "Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches", 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) | 978-1-7281-9951-1/21/ $31.00 ©2021 IEEE | DOI: 10.1109/ GUCON50781.2021.9573759.

[6] Namrata Bhartiya, Sheetal Jannu, Purvika Shukla, Radhika Chapaneri, "Employee Attrition Prediction Using Classification Models", 2019 5th International Conference for Convergence in Technology (I2CT), Pune, India. Mar 29-31, 2019, 978-1-5386-8075-9/19/$31.00 ©2019 IEEE.

[7] Sarah S. Alduayj, Kashif Rajpoot, "Predicting Employee Attrition using Machine Learning", 2018 13th International Conference on Innovations in Information Technology (IIT), 978-1-5386-6673-9/18/$31.00 ©2018 IEEE.

[8] Sandeep Yadav, Aman Jain, Deepti Singh, "Early Prediction of Employee Attrition using Data Mining Techniques", 978-1-5386-6678-4/18/ $31.00 c 2018 IEEE.

[9] R Shiva Shankar, J Rajanikanth, V.V. Sivaramaraju, K VSSR Murthy, "Prediction of Employee Attrition Using Data mining", DOI:10.1109/ ICSCAN.2018.8541242.

[10] Andry Alamsyah1 , Nisrina Salma2, "A Comparative Study of Employee Churn Prediction Model", 2018 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 978-1-5386-5813-0/18/ $31.00 ©2018 IEEE.

[11] Ibrahim Onuralp Yigit, Hamed Shourabizadeh, "An Approach for Predicting Employee Churn by Using Data Mining", 978-1-5386-1880-6/17/$31.00 ©2017 IEEE.

[12] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari, "Evaluation of Machine Learning Models for Employee Churn Prediction", Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017), IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9/17/$31.00 ©2017 IEE.

[13] Sepideh Hassankhani Dolatabadi, Farshid Keynia, "Designing of Customer and Employee Churn Prediction Model Based on Data Mining Method and Neural Predictor", The 2nd International Conference on Computer and Communication Systems, 978-1-5386 -0539-4/17/$31.00 ©2017 IEEE.