A PROJECT REPORT

on

"Employee Attrition Prediction"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN

COMPUTER SCIENCE

BY

Aindrila Roy	2005358
Shinjini Banerjee	2005964
Syed Farhan Ali	2005209
Samik Ranjan Das	2005605
Avik Ranjan Das	2005794
Arghya Hazra	2005440

UNDER THE GUIDANCE OF

KUMAR DEVADUTTA

SCHOOL OF COMPUTER ENGINEERING

December ,2023

A PROJECT REPORT

on

"Employee Attrition Prediction"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN

COMPUTER SCIENCE

BY

Aindrila Roy 2005358

Shinjini Banerjee 2005964

Syed Farhan Ali 2005209

Samik Ranjan Das 2005605

Avik Ranjan Das 2005794

Arghya Hazra 2005440

CERTIFICATE

This is certify that the project entitled

"**Employee Attrition Prediction**"

submitted by

Aindrila Roy	2005358
Shinjini Banerjee	2005964
Syed Farhan Ali	2005209
Samik Ranjan Das	2005605
Avik Ranjan Das	2005794
Arghya Hazra	2005440

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Computer Sci-ence Engineering at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2023-2024, under our guidance.

Date: 5/12/2023

Kumar Devadutta
Project Guide

ACKNOWLEDGEMENT

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

Aindrila Roy
Shinjini Banerjee
Syed Farhan Ali
Samik Ranjan Das
Avik Ranjan Das
Arghya Hazra

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: Machine Learning, Human Resources, Employee Attrition, Attrition Prediction, Dataset

CONTENTS

LIST OF FIGURES

Chapter 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.

Chapter 2

Literature Review

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.

2.1 Related Work :

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.

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.

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.

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.

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.

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 .

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

Paper	Best Algorithm	Accuracy
S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques,"	Random Forest	98.61%
Jain, Praphula Kumar, Madhur Jain, and Rajendra Pamula. "Explaining and predicting employees' attrition: a machine learning approach."	Random Forest	99%
Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. 2020. "Predicting Employee Attrition Using Machine Learning Techniques"	Support Vector Machine (Linear)	87.9%
R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach,"	XGBoost	98.1%
Yedida, Rahul & Reddy, Rahul & Vahi, Rakshit & Jana, Rahul & Gv, Abhilash & Kulkarni, Deepti. (2018). "Employee Attrition Prediction".	K-Nearest Neighbors	96.97
İ. O. Yiğit and H. Shourabizadeh, "An approach for predicting employee churn by using data mining,"	Support Vector Machine	89.7%
Alao, D. A. B. A., and A. B. Adeyemo. "Analyzing employee attrition using decision tree algorithms."	Decision Tree (C4.5)	67%

Chapter 3

Problem Statement

High employee attrition poses a significant challenge for organizations, leading to talent loss and financial repercussions. Recognizing the detrimental impact of overlooking employee satisfaction, this project aims to address the pressing need for proactive measures. By employing machine learning to predict potential employee departures, the goal is to empower Human Resources departments to intervene early, ensuring a supportive work environment and minimizing the adverse effects of talent turnover on organizational success.

3.1 Project Planning

When planning to execute a project development, it's crucial to define the requirements and features that will guide the development process. Here's a list of steps and considerations:

1. Project Objectives:
   - Clearly outline project goals and objectives.
2. User and Functional Requirements:
   - Identify user needs and define specific features.
3. Technology Stack:
   - Choose appropriate technologies and tools.
4. Scope and Timeline:
   - Define project boundaries and establish milestones.
5. Resource Allocation and Budget:
   - Allocate human and technological resources.
   - Determine the project budget.
6. Risk Assessment:
   - Identify risks and develop mitigation strategies.
7. Documentation:
   - Create comprehensive project documentation.
8. Testing and Quality Assurance:
   - Plan for testing throughout development.
9. Scalability and Future Expansion:
   - Consider scalability and plan for future updates.
10. Stakeholder Communication:
    - Establish a communication plan with stakeholders.

By addressing these key points, you set a foundation for successful project development. Adjustments can be made as needed, ensuring flexibility in execution.

3.2 Project Analysis

Ensure that there are no conflicting or contradictory statements.

Prioritization:

Prioritize requirements based on their importance and impact.
Establish a roadmap for implementation.

Traceability Matrix:

Create a traceability matrix to link requirements back to project objectives.
Ensure that each requirement contributes to fulfilling a specific goal.

Communication with Stakeholders:

Communicate analysis findings to stakeholders for validation.
Seek approval or feedback for any necessary adjustments.

Conflict Resolution:

Address and resolve any conflicts or discrepancies found during the analysis.Project analysis after collecting requirements is a crucial step to ensure clarity and accuracy. Here's an overview of the key aspects to consider during the project analysis phase:

Requirement Validation:

Review and validate collected requirements to ensure they align with project objectives.
Seek clarification from stakeholders on any ambiguous or unclear points.

Feasibility Assessment:

Evaluate the feasibility of implementing the identified requirements.
Consider technical, financial, and operational aspects.

Risk Identification:

Identify potential risks associated with the project.
Analyze the impact and likelihood of each risk.

Scope Refinement:

Refine the project scope based on a more detailed understanding.
Ensure that all necessary features are included while avoiding scope creep.

Consistency Check:

Check for consistency among different requirements and project documents.
Ensure that the project team has a shared understanding.

Quality Assurance:

Establish quality assurance measures for ongoing development.
Integrate feedback loops for continuous improvement.

The project analysis phase serves as a bridge between requirements gathering and actual implementation. It ensures that the project team has a clear and accurate understanding of what needs to be achieved, setting the stage for a successful development process.

3.3 System Design

3.3.1 Design Constraints:

Working Environment:

Platform: Google Colab (Colaboratory)
IDE: Jupyter Notebooks in Google Colab
Python Version: Colab typically supports the latest version of Python.

Hardware:

CPU: Colab provides access to a virtual machine with various CPU options.
GPU: Colab offers free GPU resources (e.g., NVIDIA K80, T4, P100) for faster training. Note: Availability is not guaranteed.
TPU: Colab supports Tensor Processing Units (TPUs) for even faster training.
Experimental Setup:

1. **Data Loading:**

Load datasets from cloud storage or provide a link to external datasets.
Utilize Colab's file upload feature if working with smaller datasets.

2. **Library Installation:**

Install necessary libraries using the !pip install command.
python.

3. **Environment Configuration:**

Check and configure the runtime type in Colab, selecting between CPU, GPU, or TPU. Set up any required environment variables or configurations.

4. **Model Development:** Develop the machine learning model using popular libraries like TensorFlow or PyTorch. Utilize GPU acceleration if available for faster training.

5. **Hyperparameter Tuning:**

Experiment with different hyperparameters using multiple training runs.
Use Colab's resources efficiently, considering session timeouts.

6. **Results Visualization:**

Use Matplotlib or Seaborn for visualizing training/validation curves and model performance.

7. **Collaboration:**

Share the Colab notebook with collaborators for real-time collaboration.
Use version control (e.g., Git) to track changes if collaborating outside of Colab.

8. **Monitoring and Debugging:**

Utilize Colab's built-in monitoring tools for memory and GPU usage.
Include debugging statements or use external tools for debugging if necessary.

3.3.2 System Architecture:

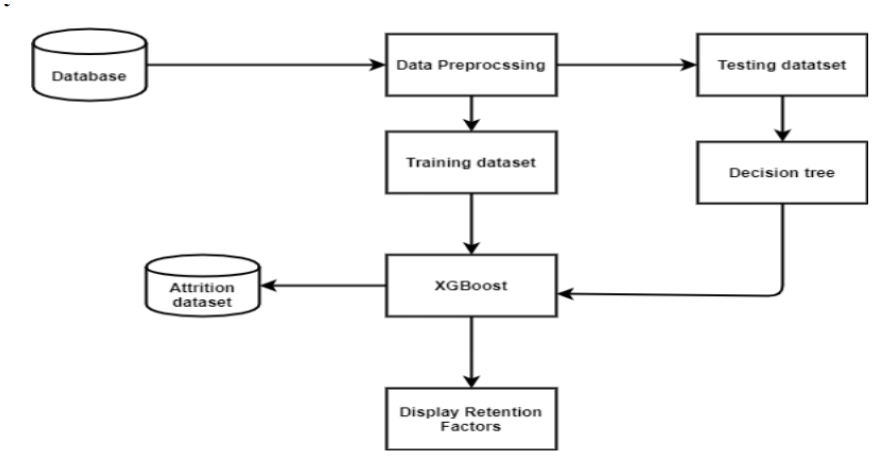

Fig 1 - Demonstration of the system architecture

Explaining the architecture of the system design:

3.3.2.1. Data set: Dataset is a group of data. Most commonly a dataset agrees to the contents of a single database, where every column of the table signifies a particular variable, and each row agrees to a member of the dataset. For our project we take employee statistics from IBM which contains 1470 records and 35 fields including categorical and numeric features. Each record in the employee dataset signifies a single employee information and each field in the record signifies a feature of that particular employee.

3.3.2.2 Data pre-processing: From the IBM employee dataset we implement a feature assortment method to select the most important features of the dataset and divide total dataset into two sub datasets. One is the test dataset, the one is the training dataset. That is if suppose any feature value in the record contain any worthless value or undefined or irrelevant value then separate that entire record from the unique dataset and place that record into training dataset, else if the record contains faultless data with all features then place that into the test dataset. Test dataset contain all important features to predict employee attrition or employee attrition and training dataset contain immaterial data.

3.3.2.3 Correlation of Attributes : The data shows that we have had a large number of attributes, but we have used some major attributes in finding out the turnover rate. We have found out many interesting relationships among these attributes that led us to our goal of finding the turnover rate and in which year the turnover rate touched its peak. In our data, we have shown a correlation between attributes such as how many years an employee spent in a company, how many years an employee spent in a company with the current manager and how many years spent in the company since the last promotion. We have also shown the correlation between the level of job or service an employee is doing and monthly income of the employee.

3.3.2.4 Test dataset and training dataset: Extrication data into test datasets and training datasets is an important part of evaluating data mining models. By parting the total data set into two data sets we can minimize the effects of data inconsistency and better understand the characteristics of the model.

3.3.2.5 XGBOOST: XGBoost belongs to a boosted tree algorithm and works on the principle of gradient boosting. As compared to others, practices a more regularized model reinforcement to regulate overfitting and thus improvises performance.

Advantages of the system design :
A. To guess employee attrition.
B. Helpful to their financial growth by reducing their human resource cost.

Chapter 4

Implementation

4.1 Methodology

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

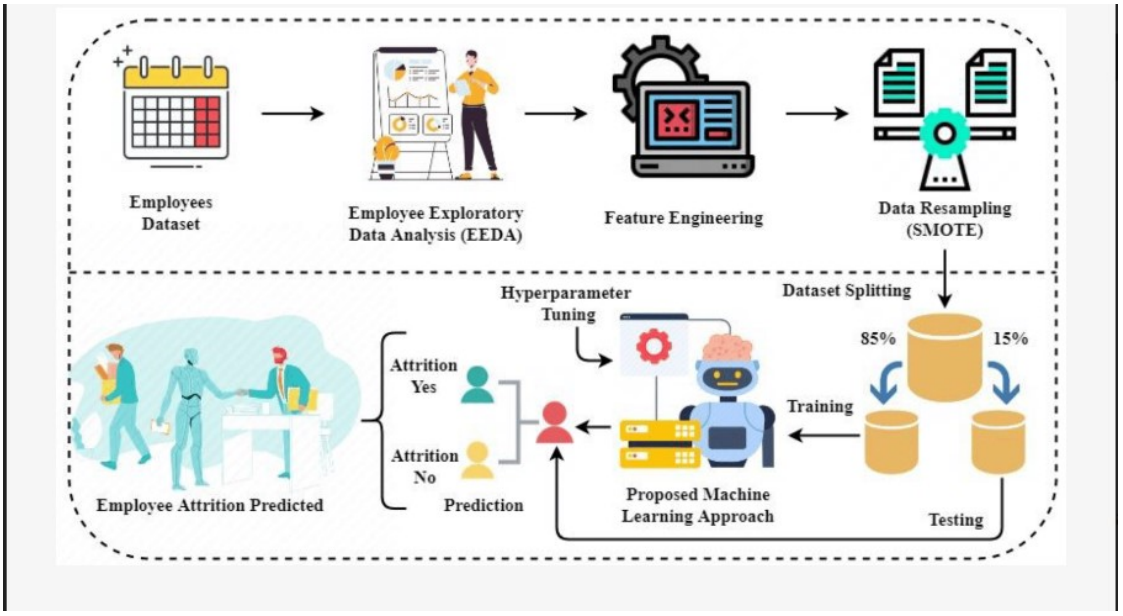

Fig.2 - Methodological analysis of our proposed research study for employee attrition prediction.

Data Gathering and Pre-processing

The fundamental phase in the machine learning pipeline is data gathering for training the ML model. The accuracy of 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.

The next step involves feature selection, Chi-Square Test and Correlation Matrix were used for the same. Chi-squared is a numerical test that evaluates the deviation from the expected distribution using metrics like true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The Correlation Matrix evaluates the relationship between multiple variables and plays an important role in multivariate analysis by capturing pairwise degrees of relationship between different components , which are features like the Age and Gender of an employee in this case.

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.

4.1.1 Model 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.

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.

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.

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 .

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 .

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 .

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 .

4.1.2 Model 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 .

4.2 Result Analysis

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

Algorithm	Accuracy
KNN	82.7%
Logistic Regression	83.6%
Random Forest	85%
Decision Tree	78.2%
MLP Classifier	83.2%
Extreme Boosting Tree	86.4%

Table 3. Correlation Matrix Model Results

Model used	Accuracy score	Precision score	Recall score	F1 score
KNN	0.837	0.333	0.017	0.032
Logistic Regression	0.842	0.667	0.034	0.065
Random Forest	0.87	0.923	0.203	0.333
Decision Tree	0.821	0.387	0.203	0.267
MLP Classifier	0.84	nan	0.0	nan

Table 4. Chi-Square Model Results

Model used	Accuracy score	Precision score	Recall score	F1 score
KNN	0.842	0.533	0.136	0.216
Logistic Regression	0.856	0.571	0.407	0.475
Random Forest	0.859	0.818	0.153	0.257
Decision Tree	0.84	nan	0.0	nan

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.

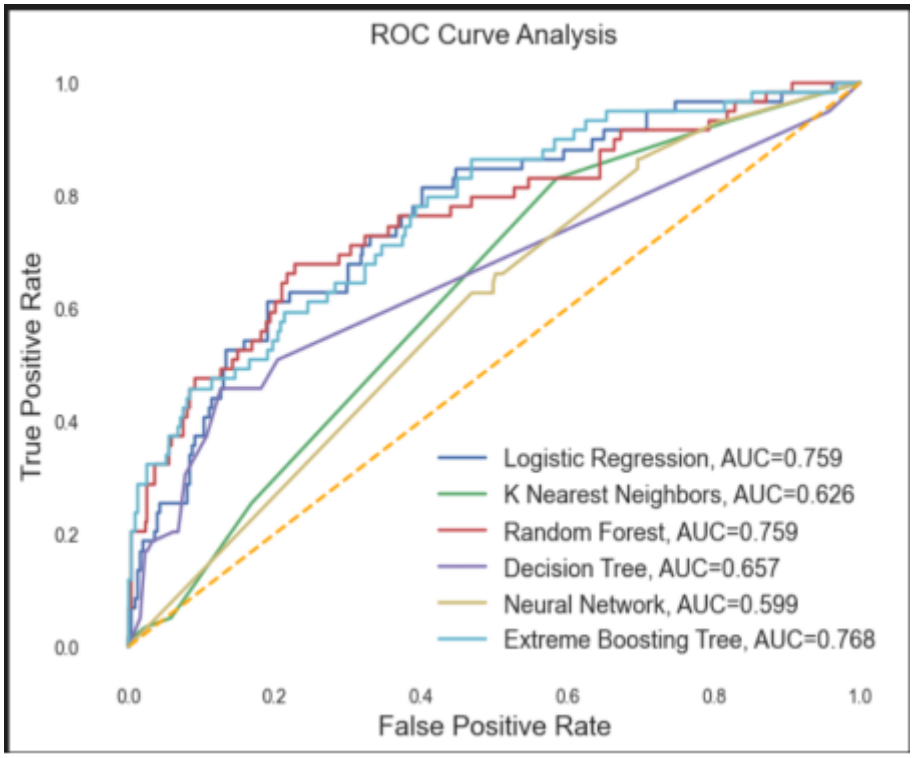

Fig. 3 - Correlation Model ROC Curve

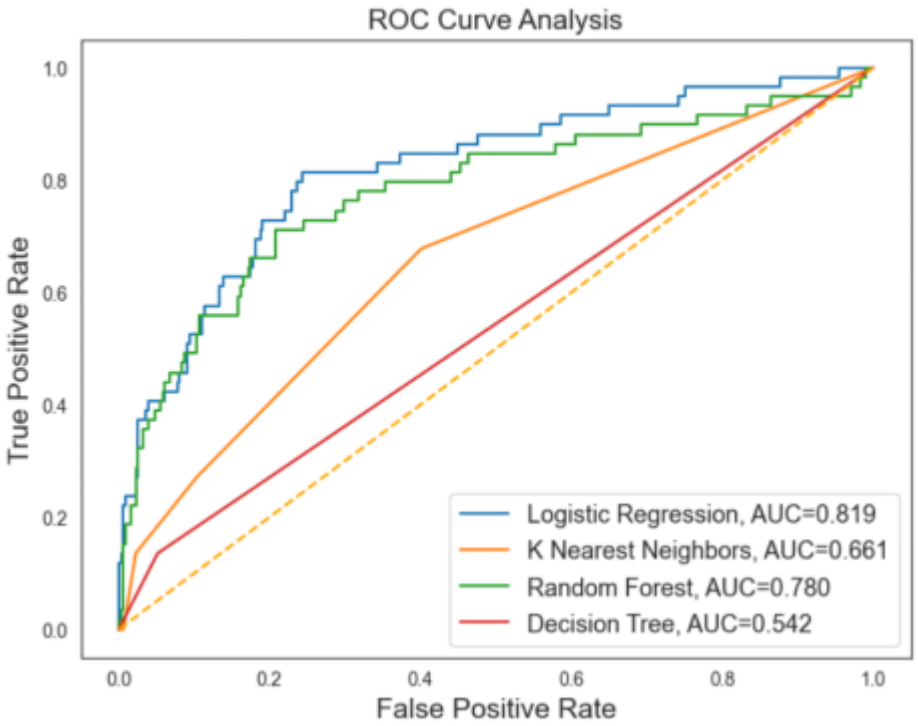

Fig. 4 - Chi-Squared Model ROC Curve

4.3 Discussions

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 3).

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 4).

Chapter 5

Standards Adopted

5.1 Design Standards

In all the engineering streams, predefined design standards are present such as IEEE, ISO etc. List all the recommended practices for project design. In software the UML diagrams or database design standards also can be followed.

5.2 Coding Standards

Coding standards are collections of coding rules, guidelines, and best practices. Few of the coding standards are:

1.Write as few lines as possible.
2.Use appropriate naming conventions.
3.Segment blocks of code in the same section into paragraphs.
4.Use indentation to mark the beginning and end of control structures. Clearly specify the code between them.
5.Don't use lengthy functions. Ideally, a single function should carry out a single task.

5.3 Testing Standards

ISO (International Organization for Standardization) and IEEE (Institute of Electrical and Electronics Engineers) provide standards that guide quality assurance and testing processes for our software product. Here are some relevant standards:

ISO Standards:

1. **ISO/IEC 25010:2011 - Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models:**

- Defines quality models for software products and systems.

2. **ISO/IEC 9126:2001 - Software engineering - Product quality:**

- Specifies quality characteristics and metrics for software products.

3. **ISO/IEC 29119 - Software and systems engineering - Software testing:**

- Comprises a series of standards covering various aspects of software testing, including test processes, documentation, and test techniques.

IEEE Standards:

1. **IEEE 829 - IEEE Standard for Software Test Documentation:**

- Provides a standard for test documentation, including test plans, test cases, and test procedures.

2. **IEEE 1008 - IEEE Standard for Software Unit Testing:**

- Focuses on guidelines for unit testing, outlining the process of designing and conducting tests for individual software units or components.

3. **IEEE 1012 - IEEE Standard for System and Software Verification and Validation:**

- Offers guidelines for verification and validation activities throughout the software life cycle.

4. **IEEE 1061 - IEEE Standard for Software Metrics:**

- Defines a set of software metrics to be used for the evaluation of software processes and products.

5. **IEEE 12207 - IEEE Standard for Information Technology - Systems and Software Life Cycle Processes:**

- Describes life cycle processes, including testing and quality assurance, in the development of software.

6. **IEEE 829 - IEEE Standard for Software and System Test Documentation:**

- Provides guidelines for the preparation of test documentation, covering various testing phases.

7. **IEEE 730 - IEEE Standard for Software Quality Assurance Processes:**

- Outlines the processes involved in software quality assurance, ensuring that quality is systematically planned and executed throughout the software development life cycle.

8. **IEEE 610.12 - IEEE Standard Glossary of Software Engineering Terminology:**

- Offers a glossary of terms commonly used in software engineering, including those related to software quality and testing.

Chapter 6

Conclusion and Future Scope

6.1 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.

6.2 Future Scope

This study has immense scope. First of all, 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.

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.

[2] Jain, Praphula Kumar, Madhur Jain, and Rajendra Pamula. "Explaining and predicting employees' attrition: a machine learning approach." SN Applied Sciences 2 (2020): 1-11.

[3] Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. 2020. "Predicting Employee Attrition Using Machine Learning Techniques" Computers 9, no. 4: 86.

[4] 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.

[5] Yedida, Rahul & Reddy, Rahul & Vahi, Rakshit & Jana, Rahul & Gv, Abhilash & Kulkarni, Deepti. (2018). Employee Attrition Prediction.

[6] 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.

[7] 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.

[8] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset 10/03/2023

[9] 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.

[10] 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).

[11] 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.

[12] 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).

INDIVIDUAL CONTRIBUTION REPORT:

EMPLOYEE ATTRITION PREDICTION SYSTEM

SYED FARHAN ALI
2005209

Abstract: 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.

Individual contribution and findings: In the project, the member's contribution mostly lies around the introduction and literature review. It involves framing the project's significance, emphasizing the critical role of attrition prediction in organizational strategy. In the Literature Review, there is an existing research on attrition prediction models, exploring methodologies, features, and data sources. Key findings include the evaluation of diverse machine learning models and the identification of crucial features like job satisfaction. Emphasis is placed on addressing gaps in current literature and ensuring ethical data handling. This comprehensive review informs the project's direction, aligning it with state-of-the-art practices in HR analytics while establishing the unique contributions our approach brings to the domain.

Individual contribution to project report preparation: For the report, the member contributed mostly to the introduction and related work of the project , where there were researches on different attrition prediction models discussed.

Individual contribution for project presentation and demonstration: In the project presentation, the member focuses on delivering a compelling introduction, outlining key objectives. In the Related Work section, the member concisely presents the existing papers, their algorithms and accuracy, emphasizing our project's unique contributions.

Full Signature of Supervisor:

…….

Full signature of student:

Syed Farhan Ali

……..

INDIVIDUAL CONTRIBUTION REPORT:

EMPLOYEE ATTRITION PREDICTION SYSTEM

AINDRILA ROY
2005358

Abstract: 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.

Individual contribution and findings: The project planning and analysis for employee attrition prediction were skillfully executed. A comprehensive project plan was developed, encompassing goals, scope, and deliverables. The data analysis phase was led with a meticulous approach, ensuring the identification of key factors influencing attrition. The impact of our member's contribution is reflected in the project's well-organized structure and the insightful analysis that laid the groundwork for subsequent stages.

Individual contribution to project report preparation: The project report preparation was significantly advanced by meticulous attention to detail and organizational skills played a pivotal role in compiling and structuring the comprehensive project documentation. Our member ensured the inclusion of technical specifications, data summaries, and relevant visualizations, contributing to the report's clarity and coherence. Their commitment to producing a thorough and well-documented report greatly facilitated communication of the project's methodologies, findings, and recommendations.

Individual contribution for project presentation and demonstration: In the project presentation, the member focuses on delivering a compelling conclusion, outlining the future scopes. In the Results section, the member concisely analyzes the existing results, their algorithms and accuracy, emphasizing our project's unique contributions

Full Signature of Supervisor:

…….

Full signature of student:

Aindrila Roy

……..

INDIVIDUAL CONTRIBUTION REPORT:

EMPLOYEE ATTRITION PREDICTION SYSTEM

SHINJINI BANERJEE
2005964

Abstract: 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.

Individual contribution and findings: A substantial contribution to the project report was made by our member, with a focus on the methodology in Employee Attrition Prediction. Expertly detailing the chosen approach, their meticulous work ensured transparency. Active involvement in refining and structuring the methodology section enhanced coherence. The commitment to presenting a robust methodology significantly contributed to effective communication, enriching the overall quality of the project report.

Individual contribution to project report preparation: Our member made a significant contribution to the project report, focusing on the methodology in Employee Attrition Prediction. Demonstrating expertise and meticulous attention, they crafted a comprehensive methodology section, ensuring clarity and transparency. Their active involvement in refining and structuring enhanced the report's coherence. The commitment to presenting a robust methodology greatly contributed to effective communication, enriching the overall quality of the project report.

Individual contribution for project presentation and demonstration: In the project presentation, the member focuses on delivering the methodology, outlining the data preprocessing algorithms. In this section, the member concisely presents the existing methods, emphasizing our project's unique contributions.

Full Signature of Supervisor:

…….

Full signature of student:

Shinjini Banerjee

……..

INDIVIDUAL CONTRIBUTION REPORT:

EMPLOYEE ATTRITION PREDICTION SYSTEM

SAMIK RANJAN DAS
2005605

Abstract: 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.

Individual contribution and findings: A significant contribution to the project's result analysis and discussion was made by our member. Expertly detailing the findings, a meticulous approach was applied to ensure thorough analysis and interpretation. The presentation of results and discussions was enriched by their active involvement in refining and structuring. The commitment to delivering insightful analysis significantly contributed to the overall quality of the project in Employee Attrition Prediction.

Individual contribution to project report preparation: Our member, specializing in Result Analysis and Discussion, played a key role in the project report preparation for Employee Attrition Prediction. Their meticulous refinement and structuring significantly enhanced the quality and clarity of findings, ensuring a comprehensive and insightful presentation.

Individual contribution for project presentation and demonstration: In the project presentation, the member focuses on delivering the process of feature selection and the different model training techniques, outlining them. In the section, the member concisely presents the existing project code, their accuracy, emphasizing our project's unique contributions.

Full Signature of Supervisor:

…….

Full signature of student:

Samik Ranjan Das

……..

INDIVIDUAL CONTRIBUTION REPORT:

EMPLOYEE ATTRITION PREDICTION SYSTEM

ARGHYA HAZRA
2005440

Abstract: 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.

Individual contribution and findings: A significant contribution to the project's adoption of standards was made by our member. Their meticulous approach ensured the integration of industry standards, enhancing the project's credibility and overall quality.

Individual contribution to project report preparation: A substantial contribution to the project report preparation was made by our member, focusing on standards adoption in Employee Attrition Prediction. Their meticulous approach ensured the integration of industry standards, enhancing report quality and credibility..

Individual contribution for project presentation and demonstration: In the project presentation, the member focuses on delivering the result analysis.

Full Signature of Supervisor:

…….

Full signature of student:

Arghya Hazra

……..

INDIVIDUAL CONTRIBUTION REPORT:

EMPLOYEE ATTRITION PREDICTION SYSTEM

AVIK RANJAN DAS
2005794

Abstract: 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.

Individual contribution and findings: A significant contribution to the project's conclusion and future scope was made by our member. Their insightful findings enriched the conclusions, and a forward-looking perspective was provided, contributing to the project's success.
.

Individual contribution to project report preparation: A substantial contribution to the project report preparation was made by our member, focusing on the conclusion and future scope in Employee Attrition Prediction. Their insights enriched the conclusions, and a forward-looking perspective was provided, contributing to the report's comprehensive outlook.
.

Individual contribution for project presentation and demonstration: In the project presentation, the member focuses on delivering the hyperparameter tuning and model fitting , outlining key objectives.

Full Signature of Supervisor:

…….

Full signature of student:

Avik Ranjan Das

……..

PLAGIARISM REPORT: