EXISTING LABOR TURNOVER PREDICTION

T.J.A Udayana – IT19199672

Procedure Document

B.Sc. (Honors) Degree in Information Technology

Specializing in Computer Systems & Network Engineering

Department of Computer Systems Engineering

Sri Lanka Institute of Information Technology Sri

Lanka

November 2023

Declaration

I declare that this is my own work, and this policy document does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
Udayana T.J. A	IT19199672	Janith

ABSTRACT

The turnover rate of a corporation implies that its employees have decided to quit. Employee turnover has become increasingly prevalent along with the rapid growth of the economy and businesses in recent years. On the one hand, employees decide to quit the organization for a variety of reasons. On the other hand, employee retention and job security affect the business's usual operation. Companies must comprehend the primary causes of employee turnover and make appropriate efforts to address this issue.

In contrast to physical systems, human resource issues cannot be explained using a scientific formula. Consequently, machine learning and deep learning approaches are the most effective methods for achieving this objective. The purpose of this study is to present a methodology for predicting employee churn by applying classification algorithms to analyze the particular behaviors and qualities of the employee.

In accordance with the standard machine learning pipeline, this paper presents a five-stage framework for predicting employee attrition (data retrieval, data preparation, modeling, model evaluation & tuning, and deployment). Our work was tested using the IBM analytics imbalanced dataset, which contains 35 features for 1470 employees. We created a balanced version from the original using SMOTE and ADASYN oversampling techniques to obtain realistic results.

This study aims to examine the effectiveness of deep learning techniques like Artificial Neural Networks classifiers and machine learning techniques like Logistic Regression and Random Forest classifiers. In this study, cross-validation and parameter-tuning strategies are used for optimization in order to address overfitting problems. The optimized LR model, which had the highest AUC score of 0.74 compared to the other classification models tested, emerged as the best model that can be used to forecast employee attrition, according to the comparative study on the three classifiers. ANN and RF were second and third, respectively. We discovered that employee factors, including monthly income, age, daily rate, total working years, and overtime, significantly impact turnover.

Keywords: Employee Turnover, Employee Attrition, Churn Prediction, Random Forest, Logistic Regression, Hyperparameter Tuning, Oversampling, SMOTE, ADASYN, Machine Learning, Classification, Imbalanced Data, Prediction Model

TABLE OF CONTENTS

1 INTRODUCTION

Employee Attrition is the term used to describe employees who quit an organization for personal, work-related, financial, or environmental reasons. There are two distinct types of employee turnover: voluntary and involuntary. Involuntary attrition is the termination of employees by their employers for reasons such as poor performance or operational demands. In contrast, high-performing workers who choose to leave the company voluntarily do so despite efforts made by the employer to keep them there. For instance, early retirement or employment offers from other companies are two examples of voluntary attrition. Although organizations that value their workers frequently invest in them by providing extensive training and a positive work environment, the organizations also experience voluntary attrition and the loss of brilliant workers. Hiring replacements is another problem that comes at a substantial cost to the business, including the price of recruiting, hiring, and training.[17]

Employee attrition can be prevented or at least lessened by management if it is foreseen before it happens. According to some research, motivated workers are more likely to be innovative, productive, and effective. Based on predictive models that can be created for this purpose, organizations can use their HR data to generate such forecasts. Artificial intelligence (AI) is being used in a wide range of industries, including business, government, healthcare, and education. The use of AI to forecast employee attrition has attracted a lot of recent scientific interest. Additionally, the growing body of information on this subject encourages more research in the area.[8]

This paper mainly focuses on predicting existing labor turnover using machine learning and deep learning techniques. More specifically, using logistic regression, random forest, and artificial neural networks. The IBM HR dataset has been used to train and test the machine learning model. This dataset includes a total of 1470 records and 35 features. The target column "Attrition" mainly consists of two classes. (Yes, for employees who left and No, for employees who are still staying). These samples are highly imbalanced; there are only 237 (16.12%) positive samples (employees left) and 1233 (83.88%) negative samples (employees staying). This highly imbalanced dataset makes it difficult to make predictions.

Following are the main contributions of our research work to predicting employee turnover.

- Two advanced machine learning classifiers and one deep learning classifier were applied for employee turnover prediction, which are logistic regression, random forest, and artificial neural networks, respectively.
- AUC score has been used to evaluate ML models rather than accuracy as we

deal with an imbalanced dataset.

- In order to extract valuable insight from the dataset, exploratory data analysis was applied.
- Compare two different oversampling techniques, which are SMOTE and ADASYN, to balance the imbalanced dataset.
- Used hyperparameter tuning and cross-validation to optime the ML algorithms and reduce overfitting.

1.1 Background Literature

The main purpose of the existing labor turnover prediction is to identify the employees who are about to leave the company and find the most compelling reasons for it. On the subject of churn, attrition, and turnover, there has been a substantial amount of study and research work carried out by a wide variety of organizations and individuals.

The research paper that was published in 2018 by Sanghvi college [4] has taken a sample of IBM USA's employee database and, using the notion of information value; discovered that employee characteristics such as Job Role, overtime, and job level have a significant impact on turnover. They implemented Several classification techniques, including logistic regression, LDA, ridge classification, lasso classification, decision trees, and random forests, were utilized to make prediction of any new employee's attrition probability and tested them simultaneously. LDA provided the highest accuracy, logistic regression provided the highest precision, and ridge provided the highest recall.

In 2021, the Brac University [1] introduced a machine learning approach for employee retention prediction. The method proposed in this paper successfully modeled and analyzed a variety of machine learning classifiers to demonstrate Factors that influence the decision of the targeted candidate and determine the likelihood of candidate retention prior to training.In addition, this article utilizes the oversampling techniques of SMOTE and ADASYN to work with the imbalanced dataset. Classical metrics are applied to express the outcomes of the utilized algorithms, and the random forest classifier was shown to have the highest accuracy percentage.

The research paper "Early Prediction of Employee Attrition using Data Mining Techniques," which was published in 2018 [5], aims to give a system for predicting employee turnover by conducting in-depth analyses of each employee's specific behaviors and characteristics utilizing several classification methods such as logistic regression, support vector machines, random forest, decision tree, and AdaBoost. Moreover, they have used a feature selection method called recursive feature

elimination with cross-validation (RFECV) to find the optimal features that may fit into the machine learning model. According to the research, salaries and other financial factors, such as promotions, are not the only factors that lead to the departure of individuals from their jobs.

The research paper "Employee Attrition Prediction Using Deep Neural Networks," published in 2021 [8], was evaluated with the help of IBM analytics' unbalanced dataset, which consists of 35 attributes for 1470 employees. They started with the first model and worked backward to generate a version more aligned with reality. The final step, the implementation of cross-validation, accurately evaluates their job. They were able to achieve a prediction accuracy of approximately 91% when utilizing the original dataset. However, when using a synthetic dataset, it was around 94%. Moreover, their studies reveal that factors such as job level, monthly salary, and extra hours are the most influential factors that influence employee decisions.

The scholars from the department of Mathematics at Chaudhary Charan Singh University and the University of Delhi [12] help comprise an active study domain on how AI-based intelligence may be interpreted and exploited to assess the input resources required to put in an employee. This research aims to anticipate which employee would choose a job move and which employee would remain in a company. It is suggested to combine natural language processing, opinion mining, fuzzy logic, and a variety of frequently used classifiers. These classifiers include Random Forest (RF), Cat Boost Classifier, Support Vector Machine (SVM), and Naive Bayes (NB). To predict employee turnover inside a corporation, they used a Mamdani-based fuzzy inference system that included nine inputs and nine outputs. The robust XGBoost algorithm delivers the highest accuracy level for their particular application case. They have stated that the utilization of ensemble techniques is the basis for this, citing the strength and beauty that these techniques possess. According to their research findings, the researchers concluded that a company's attrition rate is heavily influenced by a variety of characteristics, including the city development index, gender, and education level. They employed the AUC-ROC metric in order to visualize the performance of this binary classifier that was constructed using XG Boost. This classifier was built to sort data into two categories. It provides a concise summary of the compromise that must be made between the true positive and false positive sets. In addition, they have utilized the SMOTE oversampling approach to address the imbalanced dataset. A compromise must be reached between the sets of true positives and false positives. They were able to make an accurate prediction of the probability by using the ROC (Receiver Operating Characteristic curve) because they had previously used SMOTE to balance the target class.

The research paper that was published in 2020 by the department of mathematics at Southern Illinois University [13] reveals a solution for the primary considerations

for solving classification problems are the problems with Unbalanced data. Since the default assumption of the majority of machine learning algorithms is that all data are balanced, the algorithms do not account for the distribution of the data sample class. The results are typically unsatisfactory and skewed toward the distribution of the majority of the sample class. This indicates that, in both theory and practice, the results of utilizing a model constructed with Imbalanced data without accounting for the imbalance in the data could be misleading. Although the majority of researchers have focused their works on the application of the SMOTE and ADASYN Sampling Approach in handling data imbalance independently and have not sufficiently explain the algorithms behind these methodologies with computed examples, this paper concentrates on both synthetic oversampling techniques and manually computes synthetic datasets to enable efficient awareness of the algorithm. This is because the majority of researchers have failed to adequately explain the algorithms behind these techniques with computed examples. They investigated how these synthetic oversampling approaches may be used to binary classification issues that included imbalanced sample sizes and proportions.

The research paper "Predicting Employee Attrition using Machine Learning," published in 2018 [17] uses machine learning models to research employee turnover. In order to make accurate projections regarding staff turnover, three primary experiments were run using IBM Watson's fabricated data. In the first experiment, the original dataset with an uneven distribution of classes was used to train three different types of machine learning models. Support vector machine (SVM) with several kernel functions, random forest, and K Nearest Neighbor were these models (KNN). The second study focused on utilizing an adaptive synthetic (ADASYN) technique to address class imbalance, followed by retraining on the new dataset with the previous mentioned machine learning models. In the third experiment, a manual under sampling of the data was utilized in order to achieve a balance between the classes. As a consequence of this, training a KNN model on an ADASYN balanced dataset using a KNN model with a K value of 3 achieved the maximum performance, with a score of 0.93 for the F1-score. In conclusion, an F1-score of 0.909 was accomplished through the utilization of feature selection and random forest by utilizing 12 features out of a total of 29 features.

In 2021 ,the Universities Dian Nuswantoro, researchers[6] proposed a feature selection process that includes eliminating duplicate features, correlating features, and using a univariate receiver operating characteristics curve (ROC) in order to minimize the number of features from 35 to 21. After that, once they had compiled the most significant features, they utilized Decision Trees and Random Forest to analyze the data. The research came to the conclusion that Random Forest with feature selection may accurately predict Employee Attrition and Performance by achieving an accuracy of 79.16%, recall of 76%, and precision of 82.6%. This was accomplished by optimizing parameter selection using a parameter grid. As a result

of these results, they are able to draw the conclusion that they are able to acquire a better forecast utilizing 21 features for employee attrition and performance, which assists higher management in the decision-making process.

In 2022 research was carried out Khwaja Fareed University and the University of Hafr Al Batin [18]. This was aimed at investigating the organizational elements that led to employee turnover and utilizing machine learning methods to predict employee turnover in the future. A comparison was carried out using the four different approaches to machine learning. The accuracy score for the suggested improved Extra Trees Classifier (ETC) technique achieved a score of 93% when it was applied to the prediction of employee attrition. The proposed method fared better than recent studies that were considered to be state-of-the-art. In order to determine the factors that led to employee turnover, an Employee Exploratory Data Analysis (EEDA) was carried out. According to their research findings, the primary reasons why employees leave their jobs are changes in their monthly income, hourly rate, work level, and age. In addition, the hyperparameter tuning technique was utilized in this study in order to locate the parameters of the applied machine learning models that provided the greatest match.

In 2021, scholars from Jiangxi University and Syracuse University [23] did an employee attrition analysis. In this study, the authors conducted an analysis of the dataset IBM Employee Attrition to determine the most common factors that lead employees to leave their jobs. First, they used the correlation matrix to identify some features in our dataset that did not have a significant link with any of the other qualities, and then they eliminated those features from our dataset. Secondly, they exploited Random Forest to pick relevant features and discovered that factors such as monthly income, age, and the number of firms worked for substantially impacted employee turnover. After that, they used a K-means clustering technique to divide people into two groups. Finally, they carried out a binary logistic regression quantitative analysis, which led them to the conclusion that the rate of attrition among people who traveled frequently was 2.4 times higher than the rate among people who traveled only occasionally. In addition, we discovered that employees who work in the Human Resources department have a greater propensity to quit their jobs. Researchers trained and tested the dataset to predict employee turnover, split it in half (80% for training, 20% for testing), and documented the accuracy of the test set.. This was done so that they could evaluate the performance of the model. Both Random Forest and Logistics Regression have an accuracy of 0.8456 and 0.8843, respectively. This indicated that the Logistics Regression model was a better fit for their dataset and more appropriate for prediction use.

In 2021, research was conducted by the center for artificial intelligence technology of the University Kebangsaan Malaysia and the quality engineering research center of the University Kuala Lumpur, Johor, Malaysia, on "Machine Learning for

Predicting Employee Attrition" [16]. The objective of this study is to establish which of three machine learning techniques—the Decision Tree (DT) classifier, the Support Vector Machines (SVM) classifier, and the Artificial Neural Networks (ANN) classifier—produces the most precise results.The IBM Human Resource Analytic Employee Attrition and Performance dataset is used to facilitate the comparison of these different machine learning algorithms. The steps of data exploration, data visualization, data cleansing and reduction, data transformation, discretization, and feature selection are included in the preprocessing steps for the dataset that was utilized in this comparative study. In this study, strategies such as parameter tuning and regularization are utilized for the sake of optimization. These techniques are used to combat overfitting concerns. Following the 88.87% was the highest accuracy achieved by the optimized SVM model when compared with the other classification models that were reviewed, the ANN model came in second, followed by the DT model. The comparative analysis conducted on the three classifiers revealed that the optimized SVM model is the best model for predicting employee turnover.

In 2021 two scholars from the Informatics Institute of Technology, Sri Lanka [20] identified the employee turnover rates in the garment business through their research. According to their study, the average monthly turnover rate in the garment industry from 2000 to 2012 was 6.6%. The majority of employees are leaving due to 1) low pay, which refers to salary discontent as a result of labor cost cuts and the search for low-cost labor, and 2) inadequate benefits. 2) Pay disparity refers to salary volatility among gender wage differences. This term relates to inadequate recognition and reward. 4) The term marriage status change' refers to the departure of the majority of women in the textile industry upon marriage and maternity leave. 5)conflict with relatives 6) Job unhappiness is a result of inadequate training, a lack of managerial assistance, and a negative social image of garment workers as "Juki girls." 7) consider the existence of better work options to be preferable to their current positions. 8) Changing the career outlook of younger employees refers to transferring to a different department or job function. 9) unrealistic goals and a lack of intercommunication. 10. labor scarcity refers to the garment sector machine operators who have departed due to an overwhelming volume of work.

	Preprocessing	ML Algorithms	Evaluation	Recommended
[1]	SMOTE ADASYN Correlation	Random Forest K Nearest Neighbors Logistic Regression Naïve Bayes Decision Trees Gradient Boosting XGBoost LightGBM	ROC	Random Forest
[5]	One Hot Encoding Feature Selection RFECV	Logistic Regression Support Vector Machines Random Forest Decision Trees ADABoost	Accuracy Precision Recall F1 Score	Random Forest
[6]	Ordinal Encoding Univariate ROC Feature Selection Hyperparameter Tuning	Decision Tree Random Forest	Accuracy Recall Precision	Random Forest
[8]	Rescaling Correlation Data Balancing - ADASYN Hyperparameter Tuning	Deep Neural Networks	Accuracy Precision Recall F1 Score	DNN
[12]	SMOTE PCA Fuzzy Interference	CART Naïve Bayes K Nearest Neighbors Random Forest Support Vector Machine XGBoost CatBoost	Accuracy AUC	XGBoost
[16]	Detecting Outliers Parameter Tuning Data Cleaning Data Reduction	Decision Tree Support Vector Machine Artificial Neural Network	Accuracy Error Rate RMSE ROC	ANN
[17]	Feature Selection ADASYN	Support Vector Machine Random Forest K Nearest Neighbors	Accuracy Precision Recall F1 Score	Random Forest
[20]	Data Cleaning Feature Engineering Correlation	Decision Trees Regression Analysis Artificial Neural Network Support Vector Machines Random Forest	Accuracy	Random Forest
[21]	Feature Selection Feature Scaling	Decision Tree AdaBoost Random Forest Logistic Regression Gradient Boost	Accuracy Precision Recall F1 Score AUC	Logistic Regression
[23]	Correlation Feature Selection	Random Forest Logistic Regression K-means Clustering	Accuracy P Value	Random Forest

1.2 Research Gap

According to the literature survey that was completed above, it is clear that the industrial existing employee turnover prediction system is an implementation that is desperately required. We can see that they have covered the following domains so far in the research that was done under the existing employee turnover prediction, which has been carried out up to this point in time.

- IT Industry
- Apparel Industry
- Telecommunication
- Human Resource

The following table compares the characteristics of existing studies to those of the one we designed and implemented.

Features	[1]	[12]	[18]	[20]	[22]	Our Implemented Model
Predict Employee Turnover	Yes	Yes	Yes	Yes	Yes	Yes
Analytical Overview • Correlation • Feature importance • Value distribution	No	No	No	No	No	Yes
Export the Analyzed Data	No	No	No	Yes	No	Yes
Can be accessed at any time, anywhere, with any device over the internet.	No	Yes	No	No	No	Yes
Simplify UI	No	Yes	Yes	Yes	Yes	Yes
Use Existing Algorithms	Yes	Yes	Yes	Yes	Yes	Yes

1.3 Research Problem

The big picture of our research is to automate the organization's processes in human resource management related to labor turnovers. Mainly the system has four research novelties, including existing labor turnover prediction, question chain chatbot with emotion recognition, analyzing chatbot answers to providing summarization with visualization about leaving employees, and Resume Analyze. The system facilitates the HR management to identify leaving employees early, get the real reasons for resigning employees with visualizations, and choose the best candidate among the resumes for the job vacancy. This research will be helpful for organizations to improve their productivity, efficiency, and income. Organizations can reduce their labor turnover by using this system. Employee turnover prediction systems are used in the industry, but not widely. The recent survey we conducted gave us insight into how much people value employee turnover prediction systems and the features of the existing employee turnover prediction systems.

As the individual component is about building an employee turnover prediction system, we first wanted to know whether people have identified this as a problem in the industry. The figure below shows that people also think that employee turnover is a big issue in the industry which should have been solved with 87.1% of Yes responses.

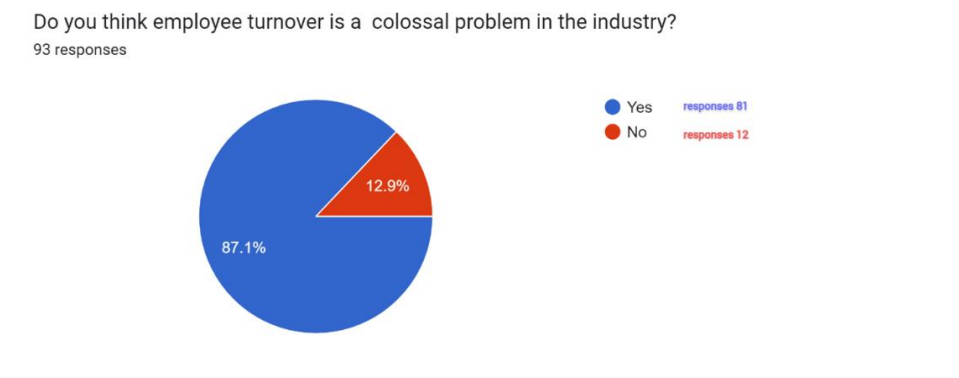


Figure 1 Survey Question 1

Then we asked them what they think about having a turnover prediction system in a company and whether it will be beneficial. And whether their company already has a turnover prediction system in their companies. According to the figure, the

majority of people, 80.6%, think it is beneficial for a company to have such a system.

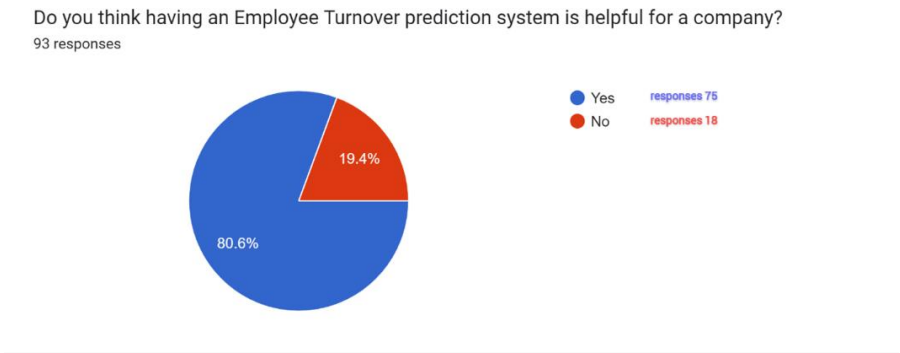


Figure2Survey-Question2

According to the figure, only a few companies with 12.9%, already have a turnover prediction system. So, it is obvious that the employee turnover prediction system is a much-needed implementation for the industry.

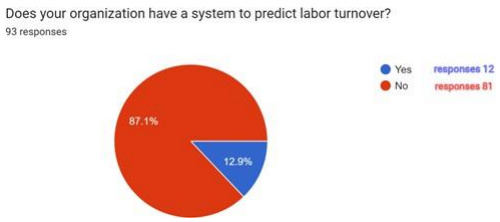


Figure3Survey-Question3

Then we ask the people what features of their system are. According to the figure, we can see that 16.7%of survey participants have the turnover prediction capability systems, and 33.3% have other features along with this capability.

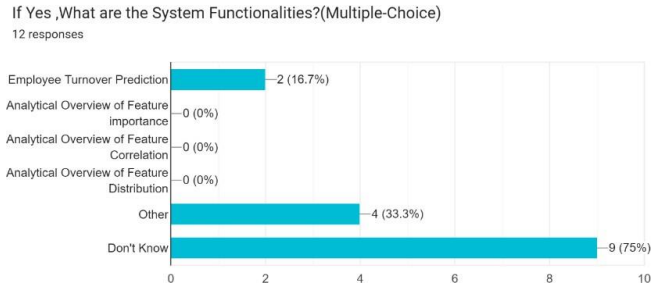


Figure4Survey-Question4

Then, we asked the people what factors they think are most influential for employee turnover, which can be helpful in building the machine learning model later on. According to the responses, 58.1% think it is the huge workload.

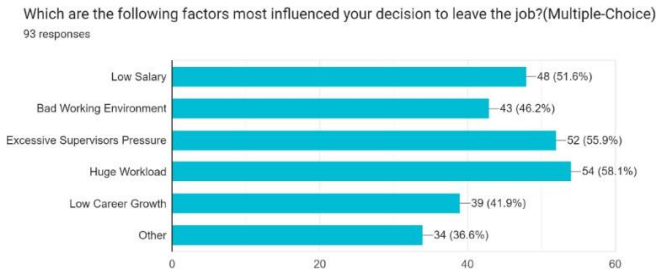


Figure 5 Survey-Question 5

Finally, survey participants are about to leave the company indirectly by asking whether they are happy with their current job, which can be helpful in further analysis.

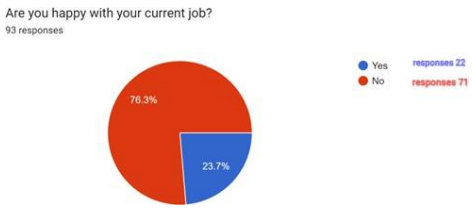


Figure 6 Survey-Question 6

Here is the link to our survey to get some insights.

- https://docs.google.com/forms/d/e/1FAIpQLScm0ca1akqqCbg4frnl3YUxf2RStax6CO9vIrZO8K3PniKv4w/viewform?usp=sf_link

The individual component is to create a labor turnover prediction system that will identify the employees who are about to leave the company and find out the most influential reasons for it. In this component, numerous research questions are intended to be answered. Which are,

- Which are the most suitable preprocessing techniques that can be used for the data?
- How to deal with the imbalanced dataset?
- Are feature selection techniques required for the data?
- How to find the optimal machine-learning algorithm for our data?

- How to increase the accuracy of the machine learning model by hyperparameter tuning?
- How to minimize the execution time of the machine learning model?
- How to optimize the web application to be real-time usable and easy to access?

1.4 Research Objectives

Main Objective

The main objective of my component is to give human resource management an easily understood representation of each employee's turnover rate. More specifically, predict the existing employees who are most likely to leave the company in the near future with more than 70% accuracy. To forecast employee turnover, historical data will be reviewed to gauge each employee's level of satisfaction. This will primarily aid firms in identifying and reducing their potential leavers.

Specific Objectives

- Provide a simple user interface to the user – in this approach, we will implement our system with a user-friendly interface for any non-technical user.
- Provide an analytical overview – This approach mainly focuses on giving more in-depth information about the new that will be inputted into the trained ML model.

2 METHODOLOGY

2.1 Methodology

2.1.1 Requirement Gathering and Analyzing

After several meetings with our supervisor, the existing labor turnover prediction component's criteria were decided. Background research and a literature study were conducted in the domain of existing labor turnover prediction using machine learning techniques in order to identify previous academic implementations of a similar nature. Thanks to the aforementioned essential activities, the research gap between currently applied solutions has been recognized.

The detected research gap and the primary concern that the proposed solution should address constituted the research problem. Prior to deployment, system requirements and analysis were undertaken to guarantee that all requirements were fulfilled. The following is a list of the system requirements for the existing labor turnover prediction component.

- It should be able to read .csv files, which contain the essential data for the prediction.
- It should be able to predict the employees who are most likely to depart the organization.
- It should be able to calculate the feature correlation (shows which attribute has a high or low correlation with another characteristic) for the inputted data.
- It should be able to calculate the feature importance (most important characteristics for Employee Turnover) for the inputted data.
- It should be able to calculate the top 15 feature's value distribution (to Determine the most usual values or value ranges for each characteristic and make decisions appropriately to reduce employee turnover in your company.) for the inputted data.
- It should be able to create downloadable .csv files with the predictions and findings.
- The machine learning model size should be as small as possible
- The latency should be as little as possible.

2.1.2 Choosing the Correct Framework

According to the literature review conducted in a previous step, two possible ways have been found to build the existing labor turnover prediction system. Those can be listed as follows.

- TensorFlow: is a piece of software that manages data sets that are laid out in the form of computational nodes in a graph. When a graph's edges are used to represent multidimensional vectors or matrices, the resulting structures

are called tensors. Tensors can be created when a graph's edges connect its nodes. TensorFlow programs, which utilize a data flow architecture that operates with generalized intermediate outputs of computations, are extremely accessible to large-scale parallel processing applications, neural networks being a typical example. This is because TensorFlow systems have a data flow architecture that operates with generalized intermediate computation outcomes.TensorFlow is the tool that manages all of the complexities that occur behind the scenes. Google developed it. Although it can be used for a wide variety of tasks, its core area of attention is in the training and inference of deep neural networks. This does not mean that it cannot be used for other things.

Neural networks are most successful when a vast amount of data is available for the machine learning model's training. However, we have seen in the above-conducted literature review that neural networks such as ANN and DNN are used in their implementations even though our dataset only has 1,470 records; we will be using ANN in our project to see how well it performs compared to traditional ML algorithms. As we are using deep learning in our project, we will use the TensorFlow framework for neural network training.

- ScikitLearn: The Scikit-learn (Sklearn) library is the most advantageous and trustworthy option when it comes to utilizing Python for machine learning. It offers a standard Python interface for accessing a variety of sophisticated machine learning and statistical modeling techniques, such as classification, regression, clustering, and dimensionality reduction.These tools are used to analyze data and make predictions. The data can be analyzed using these various techniques. This package, which was mostly developed in Python, was constructed with the help of the Python libraries NumPy, SciPy, and Matplotlib. Instead of focusing on loading, editing, and summarizing the data as its core activities, the Scikit-learn library emphasizes modeling the data.

  Classification is part of the supervised learning domain, which is well-known to us, and scikit-learn provides all of the well-known classification methods, such as logistic regression, decision trees, and random forest. As we will be using traditional ML classification algorithms for this project, we have decided to make use of scikit learn for traditional ML algorithms and TensorFlow for neural networks. In addition, when we utilize scikit learn, we are able to employ pre-built ML classifiers within it, which frees us from the necessity of manually programming it from the ground up.

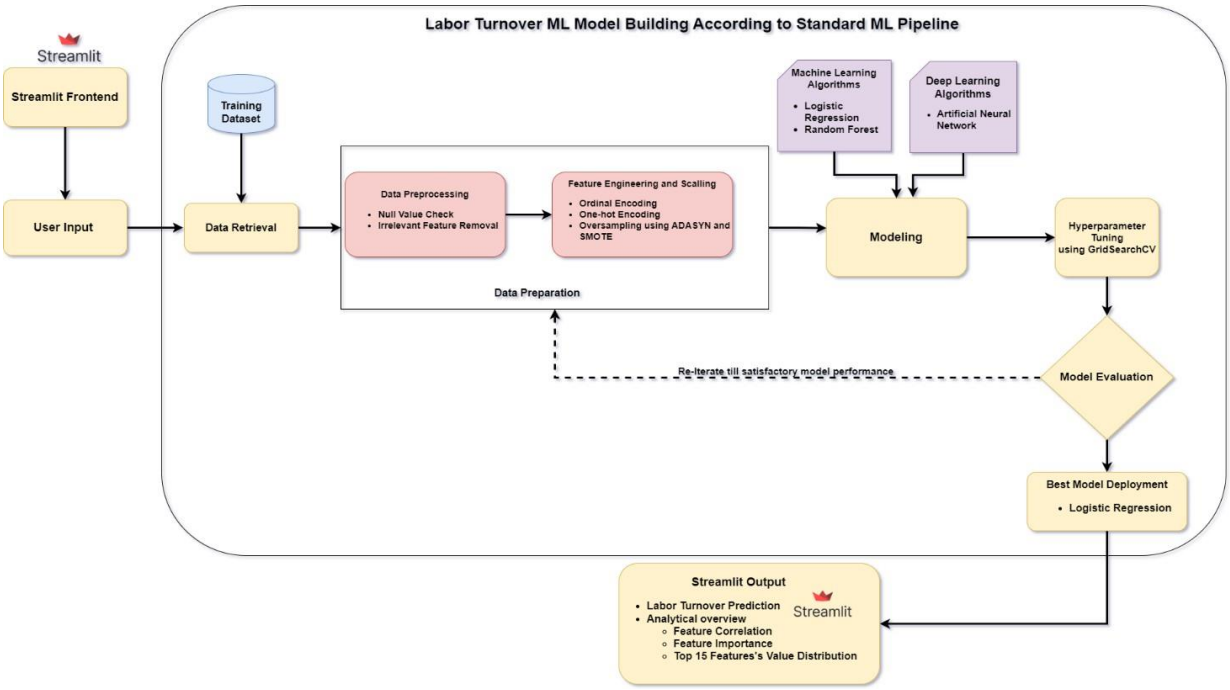
2.1.3 Implementation

This section describes the approach employed to implement the existing labor turnover prediction system in a step-by-step manner. The existing labor turnover prediction component can be subdivided into six major elements for better comprehension according to the standard machine learning pipeline: Data retrieval, Data preparation, Modeling, Model Tuning, Model Evaluation, and Deployment.

2.1.3.1 Design Overview

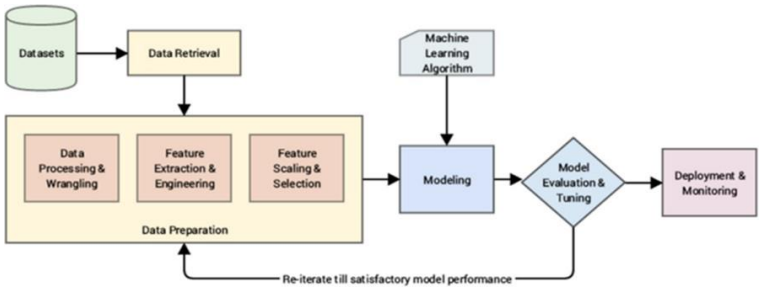
The following figure offers a representation of the design architecture and implementation that is easy to understand. We used Streamlit to deploy the machine learning model for existing labor turnover predictions online, where it is accessible at any time and from any location. The user can input the data that will be used to labor turnover prediction in .csv format. The system shows the inputted data in a table, the uploaded file, its number of entries (rows), and the number of features (columns). Then the labor turnover predictions will be calculated using the pre-trained machine learning model (logistic regression model). Then the system displays the predictions with the people who are most likely to leave the company by their names and IDs. Moreover, the system displays a full report containing all the employees' names, IDs, departments, job roles, and predictions. Finally, the system provides the user with an analytical overview which contains the following.

- Feature correlation: shows which attribute has a high or low association with another feature.
- Feature Importance: demonstrates the most significant factors for Employee Turnover.
- Top 15 feature's value distribution: includes the histograms for each feature, and by exploring them, You may determine the most common values or value ranges for each attribute and make a decision based on this information to reduce employee turnover in your firm..



2.1.3.2 Existing Labor Turnover Prediction

The existing labor turnover prediction section aims to build a machine learning model to predict the employees who are most likely to leave the company and find out the most critical reasons for employee turnover. We will build this machine-learning model according to the standard machine-learning pipeline. A standard machine learning pipeline illustrates the many processes involved in developing a machine learning model. A machine learning pipeline can be separated into six major steps, as illustrated in the figure below. Data retrieval, Data preparation, Modeling, Model tuning, Model evaluation, and deployment are the steps.



1. Data Retrieval

The first step in the process consists of collecting and extracting data. There are many different types of datasets, including structured and unstructured data, the latter of which frequently contains missing or noisy data.

- Import the libraries and tools: Import all the necessary libraries and tools, including those for exploratory data analysis (EDA) and plotting, data preprocessing, modeling, tools for hyperparameter tuning, and model evaluation.

```
# Core. EDA (exploratory data analysis) and plotting libraries
import numpy as np
import pandas as pd
import sys
import sklearn
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
# to use the plots in the notebook
%matplotlib inline

# to avoid warning messages
import warnings
warnings.filterwarnings('ignore')

# Scale Data
from sklearn.preprocessing import StandardScaler

# Models from Scikit-Learn
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier

#Models from tensorflow and Keras
import tensorflow as tf
from tensorflow import keras
from keras.models import Sequential
from keras.layers import Dense

#Oversampling
import imblearn
from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import ADASYN

#Hyperparameter Tuning
from sklearn.model_selection import GridSearchCV
from keras.wrappers.scikit_learn import KerasClassifier

# Model Evaluations
from sklearn.model_selection import train_test_split
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
from sklearn.metrics import roc_curve, roc_auc_score

# Model Save
import pickle
```

- Import the Dataset: This step requires you to import the particular data that is pertinent to the project, which is gathered and stored in the memory by utilizing pandas. The majority of the data is presented in CSV format.

2. Data Preparation

At this point in the process, the data are being prepared and cleaned in preparation

for modeling using a variety of methodologies, including Exploratory Data Analysis (EDA), Data Preprocessing and Wrangling, Feature Extraction and Engineering, Feature Scaling, and Feature Selection. These strategies contribute to reducing the dimension of the dataset, which in turn helps boost both the accuracy and the amount of time spent training. In addition, the dataset is divided into train data and test data, which are respectively employed for training the model and evaluating the performance of the model. In order to improve the quality of the data, pick important information, and make the model as accurate as possible, this stage is necessary while developing a machine learning model.

I. Exploratory Data Analysis (EDA)

By analyzing the dataset, exploratory data analysis helps to discover important facts and features of the data, as well as to understand those facts and characteristics. Even more, can do visualizing for better understanding. According to figure 10, which can be found below, the dataset contains a total of 1470 entries (rows) and 35 columns. The first 26 columns include data in the form of integers, and the next nine columns contain data in the form of objects. This indicates that there are nine categorical features inside this dataset that we will later be required to encode into numerical form.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

Analyzing the statistical information in the dataset allows for the identification of outliers by providing several metrics about the numeric columns in the dataset. These metrics include the mean, standard deviation, median, minimum and

maximum values and etc.

	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.923810	9.135373	18.0	30.00	36.0	43.00	60.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.00	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.00	7.0	14.00	29.0
Education	1470.0	2.912925	1.024165	1.0	2.00	3.0	4.00	5.0
EmployeeCount	1470.0	1.000000	0.000000	1.0	1.00	1.0	1.00	1.0
EmployeeNumber	1470.0	1024.865306	602.024335	1.0	491.25	1020.5	1555.75	2068.0
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.00	3.0	4.00	4.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.00	66.0	83.75	100.0
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.00	3.0	3.00	4.0
JobLevel	1470.0	2.063946	1.106940	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.00	3.0	4.00	4.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.00	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.00	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.00	14.0	18.00	25.0
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.00	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.00	3.0	4.00	4.0
StandardHours	1470.0	80.000000	0.000000	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.00	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.00	3.0	3.00	4.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.00	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.00	3.0	7.00	17.0

A table that displays the correlation coefficients between several sets of variables contained in a dataset is called a correlation matrix. This may indicate as to which independent variables may or may not have an impact on the variable that is the focus of the study.

II. Data Cleaning and Wrangling

This stage is concerned with transforming data into a form that can be used, given that the raw data collection is typically inaccessible to algorithms in their native form. Imputation of missing data, the removal of duplicates, and the identification of outliers are all included in this step.

- Missing Data Imputation: Check whether there are missing values by ".isna().sum()". This will show the sum of the null in each column. According

to figure 12 below, there are no missing values in the dataset.

```
Sum of Missing values in the dataset:

Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

- Irrelevant Feature Removal: At this point, we will eliminate the feature columns from the dataset that are not useful. According to figure 13 below, we can see that all the entries in the "StandardHours," "EmployeeCount," and "Over18" feature has only one constant value, so these features will not be valuable for the prediction. Therefore, we just remove these features to save the dataset's feature space and improve the ML model's accuracy and training time. Moreover, we remove the employee number as it is just an id number of the employee. After performing irrelevant feature removal, the feature columns were reduced from 35 to 31. and the new shape of the dataset is 1470 rows and 31 columns.

```
Value distribution of StandardHours column
80    1470
Name: StandardHours, dtype: int64
Value distribution of EmployeeCount column
1    1470
Name: EmployeeCount, dtype: int64
Value distribution of Over18 column
Y    1470
Name: Over18, dtype: int64
```

III. Train – Test Split

At this point, the data were partitioned into the training and testing datasets. In most cases, we employ a 70-30% split, indicating that we will use 70% of the data for training purposes and only 30% for testing purposes. When partitioning up the data, we make use of the test train split function in scikit-learn. We usually perform the train-test split before feature encoding or selection to avoid data leakage, leading to overfitting or underfitting.

Moreover, in this stage, firstly, we divide the dataset into the X and y sets. X means the predictors, which are all the feature columns except the target variable, and y means the target variable, which is the "Attrition" Column.After the train-test split, the shapes of the X training and testing datasets are as follows.

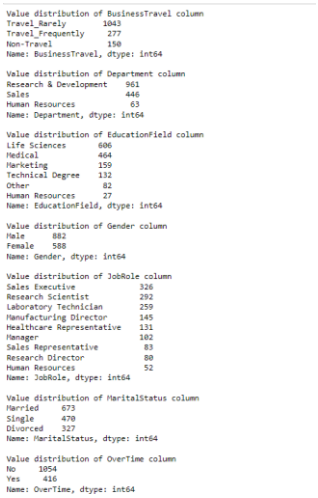
- X_train: 1029 rows and 30 columns.
- X_test: 441 rows and 30 columns.

IV. Data Encoding

At this point, the data is converted into a format that can be used by the machine learning algorithm. Because machine learning algorithms are unable to understand categorical data, it is necessary to first convert the category data into numerical data before feeding it to the algorithms. For this purpose, we can make use of either the pandas dummies or the scikit-learn label encoder. Figure 14 reveals the value distribution of the categorical features.

different categories.

- *EducationField* feature has six different categories
- *JobRole* feature has nine different categories.



This project mainly used two approaches to deal with the above categorical features.

- Ordinal Feature Encoding: ordinal encoding is utilized when the variables in the data are ordinal. Ordinal encoding transforms each label into numeric values, and the encoded data shows the order of the labels. By using ordinal encoding, we can reduce the feature space of the dataset. We used ordinal encoding for the following three features, as described below.
    - Attrition: 1 for Yes and 0 for No (Yes means left the company, and No means staying)
    - OverTime: 1 for Yes and 0 for No (Yes Means overtime, and No means no-overtime)
    - Business Travel: 0 for Non-Travel and 1 for Travel-Rarely, and 2 for Travel-Frequently
- One Hot Encoding: is one approach to transforming data in order to get it ready for an algorithm and obtain a more accurate prediction. Through the use of one-hot encoding, Each category value is transformed into a new categorical column, and then each of those columns is assigned a binary value of either 1 or 0. Each integer value is represented in this

section as a binary vector.We used one-hot encoding for the following five features, as described below.

- Department: There are three different departments in the dataset. This means we have to create three dummy variables for the "Department" column.
- EducationField: There are six different education fields in the dataset. This means we have to create six dummy variables for the "EducatioField" column.
- JobRole: There are nine different job roles in the dataset. This means we have to create nine dummy variables for the "JobRole" column.
- Gender: There are two different genders in the dataset. This means we have to create two dummy variables for the "Gender" column.
- MaritalStatus: There are three different marital statuses in the dataset. This means we have to create three dummy variables for the "MaritalStatus" column.

In order to encode the categorical data for this project, the dummies approach to Pandas was utilized. When the data encoding was completed, the total number of columns changed from 30 to 43. After developing a total of 23 columns for categorical features, we eliminated five columns (one column from each feature) in order to avoid falling into the dummy trap (Dummy Trap: is a case where the independent variables are multilinear). After that, we eliminated all of the category features from the primary dataset before combining the remaining features from the primary dataset with the dummies that were developed and the primary dataset itself.

V. Data Scaling

A vital stage in the data preprocessing process is to normalize the range of the data, which may be accomplished through the use of the data scaling approach. Standardization and Normalization are two examples of the many ways that data can be scaled; both of these procedures can be carried out with the help of the scikit-learn module.

- Standardization: means removing the value that is considered to be the "mean" for each feature and then scaling the remaining value by dividing it by the feature's standard deviation. Following the application of standardization, the standard deviation will equal one, but the means of the features will equal 0. In order to achieve the desired level of standardization, we make use of the sklearn StandardScaler() method.

$$X' = \frac{X - \mu}{\sigma}$$

In formula 2.1 μ is the mean of the feature values, and σ is the standard deviation of the feature values.

VI. Oversampling

When there is a significant imbalance in the class distribution of our training data, we run into an issue referred to as the "imbalanced classification problem." the skew might not be all that bad (it might vary), but we consider imbalanced classification a concern because it can impact how well our machine learning algorithms work. The imbalance may have a significant impact on our Machine Learning system if our algorithm completely neglects the minority class. This is problematic because the class that is labeled a minority is typically the class that receives the most attention from us. Random resampling can primarily be carried out in one of two ways: either oversampling, which involves taking additional samples from a minority group, or under-sampling, which consists in taking fewer samples from a majority group.

According to our dataset, as we only have a small number of entries of 1470, we should obviously use oversampling instead of under-sampling. Mainly there are techniques that can often be used for oversampling.

- SMOTE: Synthetic Minority Oversampling Technique (SMOTE) methodology employs the KNN method by selecting K-nearest neighbors, joining them, and generating synthetic samples in the space. The approach computes the distance between the feature vectors and their nearest neighbors using the feature vectors and their neighbors. The difference is multiplied by a random value between 0 and 1 and added to the feature. SMOTE algorithm is a pioneering algorithm from which numerous subsequent algorithms are derived.[12]

- ADASYN: ADAptive SYNthetic (ADASYN) relies on developing minority data samples adaptively based on their distributions via K nearest neighbor. The technique adjusts the distribution adaptively and made no assumptions about the distribution of the underlying data. Euclidean distance is used for the KNN Algorithm. The major difference between ADASYN and SMOTE is that ADASYN uses a density distribution as a criterion to determine automatically the number of synthetic samples that must be generated for each minority sample by adaptively changing the

weights of the various minority samples to account for skewed distributions. The latter generates the identical amount of synthetic samples for each minority sample [13].

In this project, we use both techniques for oversampling and find the best technique for each machine-learning algorithm.

3. Modeling

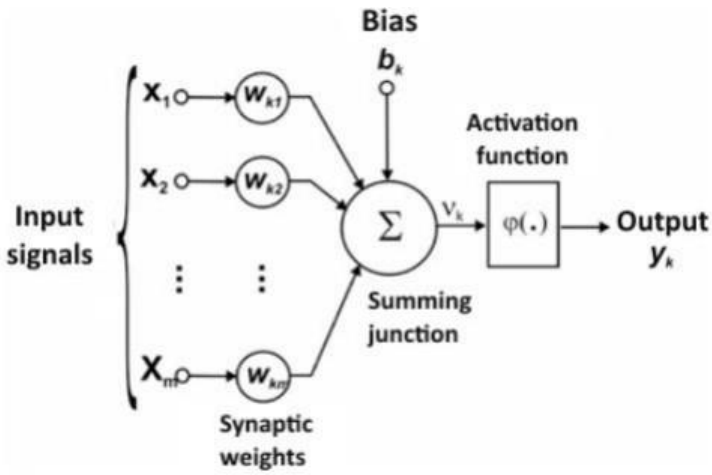
After finishing the literature survey [1, 4, 5, 6, 10, 12, 16], we have found that Logistic Regression, Random Forest, K Nearest Neighbors, Gradient Boost, Decision Trees, and Naïve Bayes are the most recommended machine learning algorithms and Artificial Neural Networks and Deep Neural Networks are the most recommended deep learning algorithms that are used existing labor turnover prediction domain. Among these above-mentioned algorithms, logistic regression, random forest, and artificial neural networks are the most accurate models which will be used in our project to compare how well it performs with our dataset and finally deploy the machine learning model with the most accurate algorithm.

- Logistic regression: is a classification function whose structure is determined by the class, which uses a single multinomial logistic regression model, and which applies a single estimator. Frequently, logistic regression will pinpoint the location of the class boundary. According to a particular technique, it will also claim that the class probabilities depend on the distance from the boundary. This phenomenon approaches the extremes of 0 and 1 more quickly as the size of the data set increases.

- Random Forest: is an ensemble learning method for classification that works by first producing an output that is the classification that is the mode of the classifications produced by the individual trees during the training phase of the method. This mode is determined by creating a large number of decision trees during the training phase of the method.

- Artificial Neural Networks: Artificial Neural Networks are brain-inspired problem-solving systems. The ANN technique differs from digital computing because it emulates the brain structure of sentient, living entities that can learn and experience. ANN can organize their neurons to perform pattern recognition more efficiently than the fastest digital computer. Artificial neurons weigh, and process inputs in one or more buried layers to determine the next layer's output. ANNs use a "learning rule" (typically gradient descent-based backpropagation of mistakes) to adaptively change the hidden layer and output layer neuron weights and biases. The early artificial neural networks appeared to replicate the human brain's great performance in complicated tasks due to their self-adaptive nature. The biological model's neuron is the neural network's

simplest structure, as shown in the figure below.



Perceptron networks are the simplest ANNs for identifying linearly separable patterns. The multilayer perceptron (MLP) architecture modeling involves choosing a layer number and neuron count. These neurons extract and store application knowledge. The non-linear model of one artificial neuron weights input signals, xi, by their synapse weight, wi. If wi is positive, the synapse is excitatory; otherwise, it is inhibitory. The summing junction's output, vk, is the combination of the bias bk (neuron activation threshold) and the input signals (xi), previously weighted by the neuron's synapses (wi). The activation function restricts yk with this outcome.

4. Model Tuning

In the field of machine learning, the problem of finding an ideal collection of hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning. A hyperparameter is a parameter whose value is applied to govern the learning process.In contrast, the values of other parameters, most often node weights, are acquired through the process of learning. In this stage, we will modify the parameters of each machine-learning model and use GridSearchCV to identify the ideal parameters for each model.

- GridSearchCV: is a tool that is included with the scikit-learn package and is used to perform hyperparameter tweaking in order to find the most appropriate values for the parameters of a specific model. This endeavors to test each and every possible combination of hyperparameters and

stores the model's optimal parameter values as a result of its findings.

According to the literature review, which has done in the previous step, [10] shows that there are three possible standard parameters for logistic regression, which are C, penalty, and solver, which are available for hyperparameter tuning.

- The C parameter is a float that governs the severity of the penalty. The lesser the value of c, the lesser the misclassification penalty, and vice versa.
- This parameter is a penalty, also known as regularization, and it reduces the model's overfitting condition by lowering variance. The penalty parameter can accept the values none, l1, l2, and elasticnet.
- The final parameter is the solver, which determines the technique that will be used to enhance the model's performance. It supports five values: lbfgs, loglinear, newton-cg, sag, and saga.

Moreover [6] suggests following hyperparameters for the random forest for better model performance.

Random Forest	Max features	Auto, squre root
	Max depth	10, 20, 30, …, 110
	Min sample split	2,3,4, … 10
	Min sample leaf	2,3,4, … 10
	Bootstrap	True, False
	Criterion	Gini, entropy
	Number classifier	100

The possible parameters that can be tuned in an ANN are as follows.

- Batch Size: The number of subsamples transmitted to the network in a batch before the parameters are updated is known as the batch size. 32 might make a good batch-size starting point. Can also experiment with 32, 64, 128, 256, and so forth.
- Number of Epochs: The number of epochs indicates how often the total training sample is presented to the network during training. Increasing the number of epochs until the validation accuracy decreases while the training accuracy rises. This type of circumstance is known as overfitting.
- Learning Rate: The learning rate determines the rate at which the network parameters are changed. The learning process is slowed by a low learning rate, but it converges gradually. A higher learning rate speeds learning, yet convergence may not occur.
- Momentum: With knowledge of the past steps, momentum helps

determine the direction of the subsequent step. It facilitates the prevention of oscillations. Average momentum values range from 0.5 to 0.9.

- Weight Initialization: Ideal weight initialization procedures would vary according to the activation function utilized on each layer. The vast majority of the time, uniform distribution is used.
- Activation Functions: Activation functions provide nonlinearity into models, enabling deep learning models to learn non-linear prediction boundaries. In general, the most popular function is the rectifier activation function. In the output layer, binary predictions are made using Sigmoid. SoftMax is utilized in the output layer when producing multi-class predictions.
- Number of Hidden Layers: Hidden layers are the layers between the input and output layers. To discover the ideal value for this parameter, we can simply keep adding layers until the test error no longer decreases. A large number of hidden units within a layer can increase accuracy using regularization techniques. A reduced quantity of units could lead to underfitting.

GridSearchCV and KerasTuner will be used to optimize the parameters of our artificial neural network in this project. As we know, hyperparameter tuning is a resource and time-consuming task we only optime few parameters among all the parameters mentioned above. Precisely, we will determine the optimal parameter values for the following artificial neural network parameters.

- Batch Size
- Epochs
- Optimization Algorithm

5. Model Evaluation

The process of evaluating the performance of the build model in accordance with particular criteria is referred to as the model evaluation. It is recommended that you familiarize yourself with the following terms before continuing on with this section. For better understanding, let's assume Attrition Yes means Leaving, and attrition No means Staying.

- True Positive (TP): This value indicates that Leaving person has been correctly classified as Leaving.
- True Negative (TN): this result indicates that the Staying person has been

correctly classified as Staying.

- False Negative (FN): This result demonstrates that an error occurred during the classification process. Where the Leaving person is categorized as Staying, a high value of FN provides a major problem for the company resources as we are unable to detect the people who are about to leave the company.
- False Positive (FP): this value represents incorrect classification decisions in which a Staying person is classified as Leaving. The increase in FP value increases the alert rate, but on the other hand, it is considered to be less harmful than the increasing FN value.

I. Accuracy

Accuracy is defined as the ratio of properly identified samples to the total number of samples.When the dataset is well-balanced, accuracy is an appropriate metric to use. However, in real network contexts, standard samples are significantly more abundant than aberrant samples; as a result, accuracy may not be an appropriate metric to use in these situations.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

II. Precision

Precision (P) refers to the ratio of "true positive" samples to "predicted positive" samples; it indicates how confident the model is in predicting the people who are about to leave the company.

$$P = \frac{TP}{TP + FP}$$

III. Recall

Recall (R) refers to the proportion of "real" positive samples relative to the overall number of "positive" samples.

$$R = \frac{TP}{TP + FN}$$

IV. F1 – Score

The F1-score is the harmonic mean of the recall and precision scores, according to its definition. The application of the following formula combines both accuracy and recall to a single score.

$$F = \frac{2 * P * R}{P + R}$$

V. Confusion Matrix

A confusion matrix is a table that is frequently used to describe the number of correct and wrong predictions made in comparison to the actual findings of the data.

VI. Classification Report

Calculating the accuracy of predictions based on a classification algorithm, the number of accurate predictions, and the number of wrong predictions can be done with the help of the Classification Report. More specifically, True Positives, False Positives, True Negatives, and False Negatives are the four types of results that are utilized in predicting the categorization report's metrics.

VII. ROC Curve

The receiver operating characteristic (ROC) curve is a graphical indication that demonstrates the behavior of a binary classifier system, such as how it is distinct from other thresholds in terms of its ability to discriminate. Tracking the true positive rate in relation to the false positive speed across a range of threshold values results in the formation of a curve. The proportion of accurate predictions relative to the total number of cases investigated is one way to evaluate the model's level of precision

6. Deployment (Web App)

Following the evaluation, we serialize the machine learning algorithm with the highest f1 score, the logistic regression classifier, using pickle. By serializing the machine learning model, training can be omitted when the models are used with new data. The pickled models will be used immediately in the web application to predict new data.

To execute real-time predictions with the developed machine learning model and to boost the model's accessibility, we should deploy the machine learning model as a web application. We primarily have three alternatives for delivering the machine learning model on the web: Flask, Django, and Streamlit. Flask and Django are complicated and resource-hungry, necessitating additional storage space, processing power, and execution time. However, the Streamlit platform appears to be the ideal alternative due to the following benefits.

- Streamlit enables the development of apps for Machine Learning projects with only the fundamental coding skills necessary.
- CSS and HTML expertise are rarely required.
- Additionally, it allows hot-reloading, which enables live-updating the application while making changes and saving the file.
- There is no need to construct a backend, specify unique routes, or process HTTP requests.

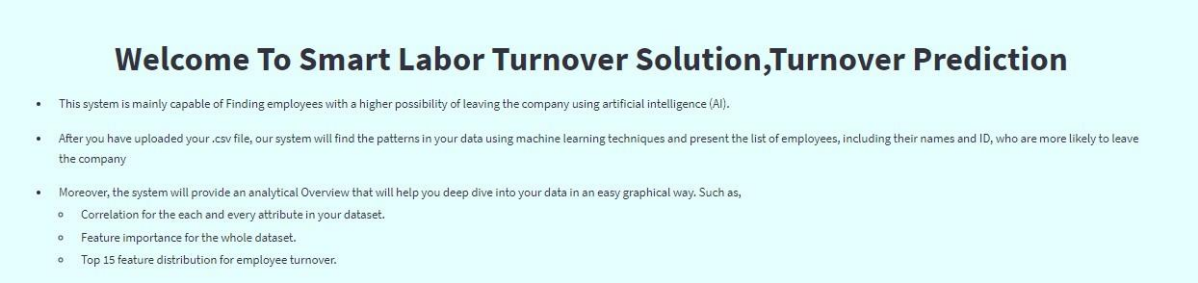
Using machine learning techniques, the website application's primary function is to identify workers who have a greater likelihood of quitting their jobs with the organisation. The following is a list of the web application's technical specs.

- Provide a list of workers, along with their names and identification information, that are more likely to quit their jobs at the organisation.
- The system will give an analytical overview that will assist you in doing a deep dive into your data in a graphical format that is simple to understand. For example,
    - a correlation analysis for every single attribute included in your dataset.
    - Feature importance for the entire dataset.
    - Top 15 feature distribution influencing employee turnover.

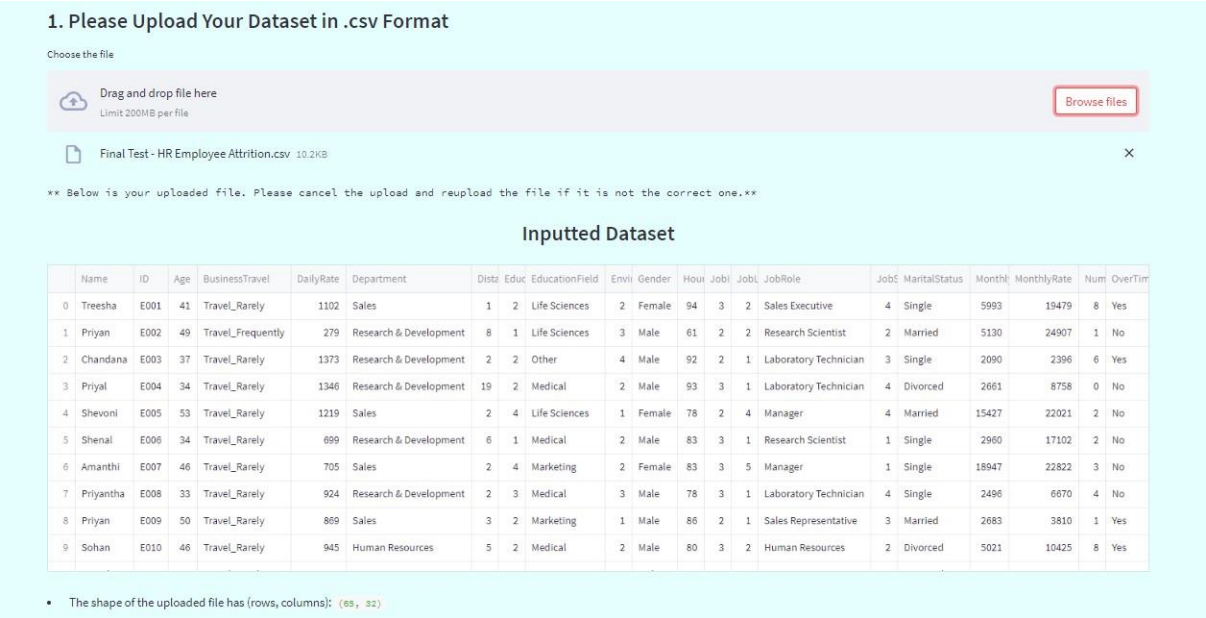
The Web app mainly consists of four sections as follows.

1. Home Section: provides an overview of the system and its features. In addition, it gives the user with instructions on how to utilize the web

application.



2. User Input Section: enables the user to upload data in .csv format. The system will display the uploaded file name; if it is incorrect, the user can reupload the file. Moreover, after the successful upload, the system will return the dataset in a table view and display the dataset's shape (number of rows and columns).



	Name	ID	Age	BusinessTravel	DailyRate	Department	Dista	Educ	EducationField	Envi	Gender	Hou	Jobl	JobL	JobRole	JobS	MaritalStatus	Monthl	MonthlyRate	Num	OverTim
0	Treesha	E001	41	Travel_Rarely	1102	Sales	1	2	Life Sciences	2	Female	94	3	2	Sales Executive	4	Single	5993	19479	8	Yes
1	Priyan	E002	49	Travel_Frequently	279	Research & Development	8	1	Life Sciences	3	Male	61	2	2	Research Scientist	2	Married	5130	24907	1	No
2	Chandana	E003	37	Travel_Rarely	1373	Research & Development	2	2	Other	4	Male	92	2	1	Laboratory Technician	3	Single	2090	2396	6	Yes
3	Priyal	E004	34	Travel_Rarely	1346	Research & Development	19	2	Medical	2	Male	93	3	1	Laboratory Technician	4	Divorced	2661	8758	0	No
4	Shevoni	E005	53	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	Female	78	2	4	Manager	4	Married	15427	22021	2	No
5	Shenal	E006	34	Travel_Rarely	699	Research & Development	6	1	Medical	2	Male	83	3	1	Research Scientist	1	Single	2960	17102	2	No
6	Amanthi	E007	46	Travel_Rarely	705	Sales	2	4	Marketing	2	Female	83	3	5	Manager	1	Single	18947	22822	3	No
7	Priyantha	E008	33	Travel_Rarely	924	Research & Development	2	3	Medical	3	Male	78	3	1	Laboratory Technician	4	Single	2496	6670	4	No
8	Priyan	E009	50	Travel_Rarely	869	Sales	3	2	Marketing	1	Male	86	2	1	Sales Representative	3	Married	2683	3810	1	Yes
9	Sohan	E010	46	Travel_Rarely	945	Human Resources	5	2	Medical	2	Male	80	3	2	Human Resources	2	Divorced	5021	10425	8	Yes

1. Labor Turnover Prediction Section: displays the employee's name and id who are most likely to leave the company. Moreover, displays the total number of employees who are about to turnover. Then after that displays the full report of the company's employee turnover, which contains employee name, id, department, job role, and attrition prediction. Additionally, the user is able to

download the data as a .csv file

2. Here is Your Prediction

Below are the people who are most likely to leave the company.

	Name	ID
21	Dilki	E022
23	Dilsha	E024
24	Gehan	E025
26	Namal	E027
28	Kalhara	E029
29	Tinuri	E030
30	Kasun	E031
31	Pooja	E032
34	Dihan	E035
41	Asoka	E042

- According to the prediction, Totally 20 people can leave the company.

Here is the Full report of your company's employee Turnover.

- In the Prediction column, 1 means the employee will leave, and 0 means the employee will stay.

	Name	ID	Department	JobRole	Pred
0	Treesha	E001	Sales	Sales Executive	1
1	Priyan	E002	Research & Development	Research Scientist	0
2	Chandana	E003	Research & Development	Laboratory Technician	1
3	Priyal	E004	Research & Development	Laboratory Technician	0
4	Shevoni	E005	Sales	Manager	0
5	Shenal	E006	Research & Development	Research Scientist	0
6	Amanthi	E007	Sales	Manager	0
7	Priyantha	E008	Research & Development	Laboratory Technician	0
8	Priyan	E009	Sales	Sales Representative	0
9	Sohan	E010	Human Resources	Human Resources	0

- To download the full report, please click the download button below.

Click here to download the full report

- To download the original file with predictions, please click the download button below.

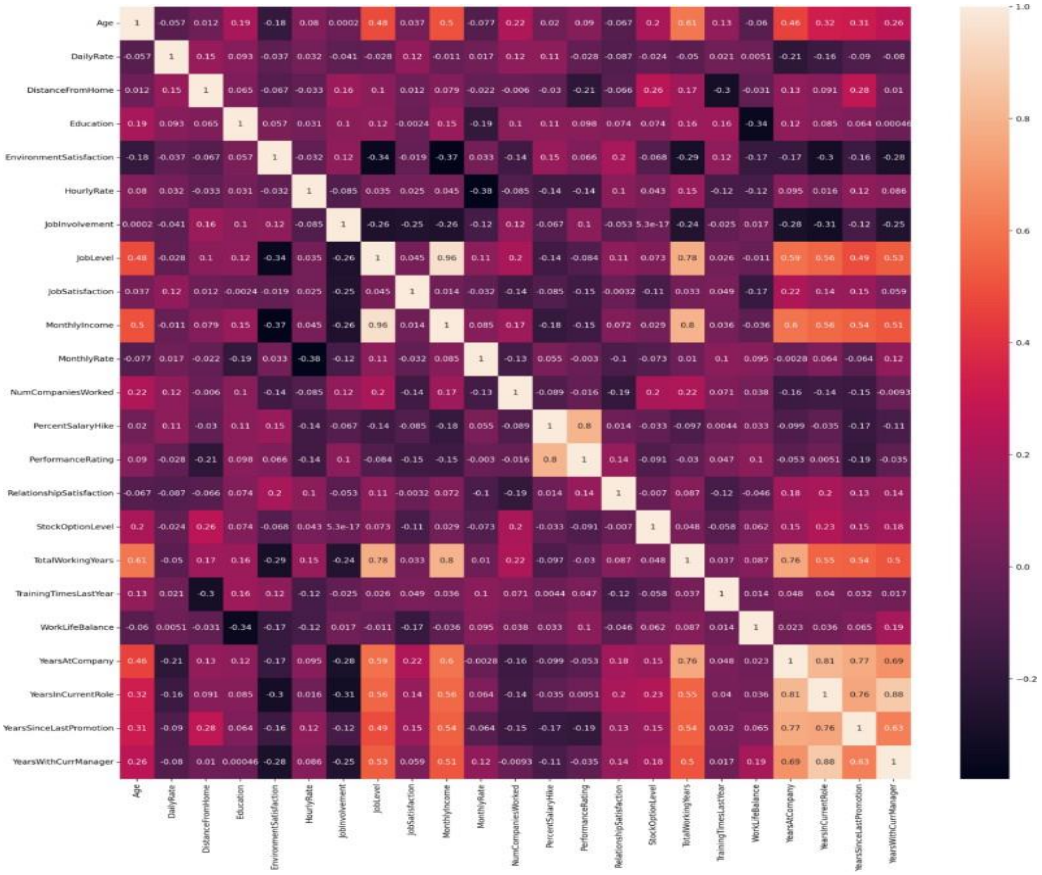
Click here to download the original file with predictions

2. Analytical Overview Section: this section mainly consist of three parts as follows.
    - Feature Correlation: shows which attribute is having a high or low correlation in respect to another attribute.

3. The Analytical Overview

3.1 Feature Correlation.

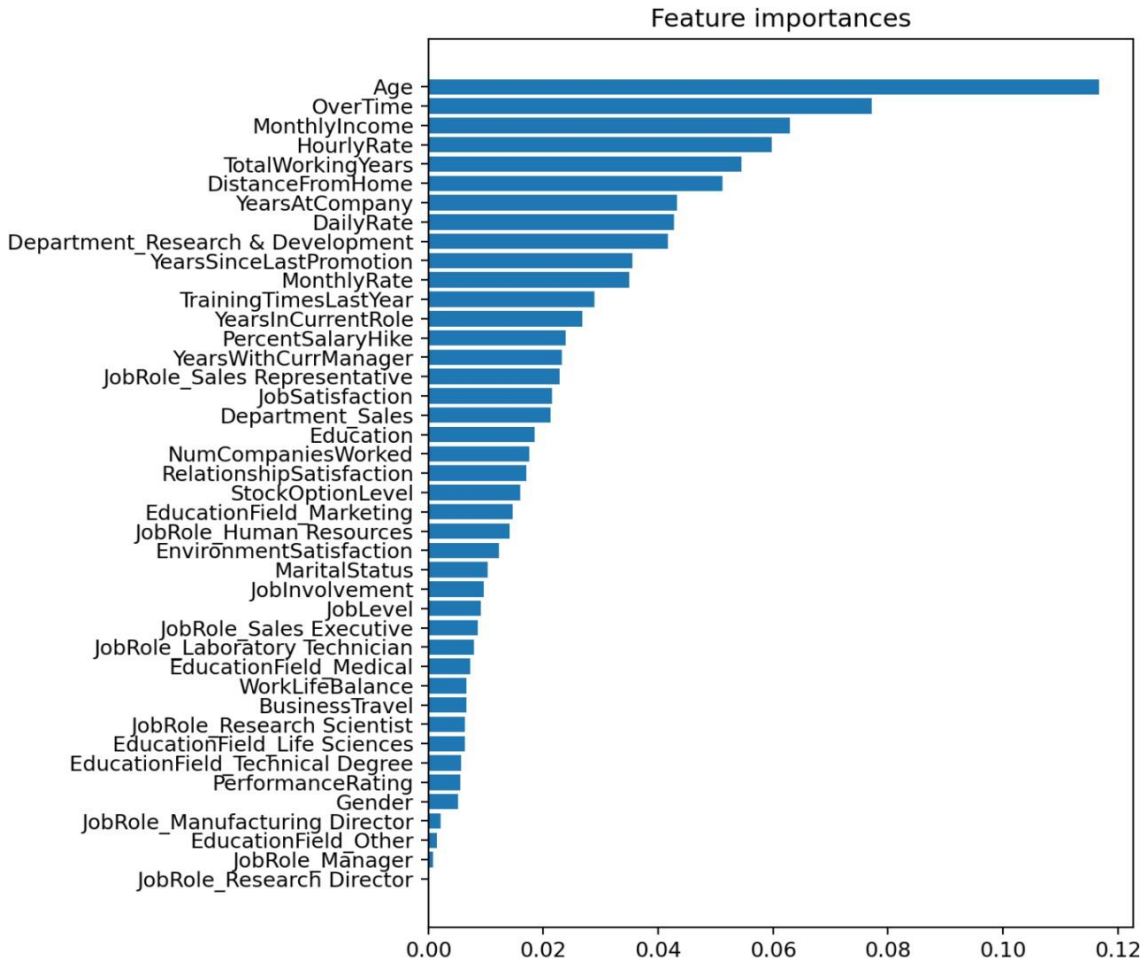
- The correlation matrix below, shows which arrtribute is having a high or low correlation in respect to another attribute



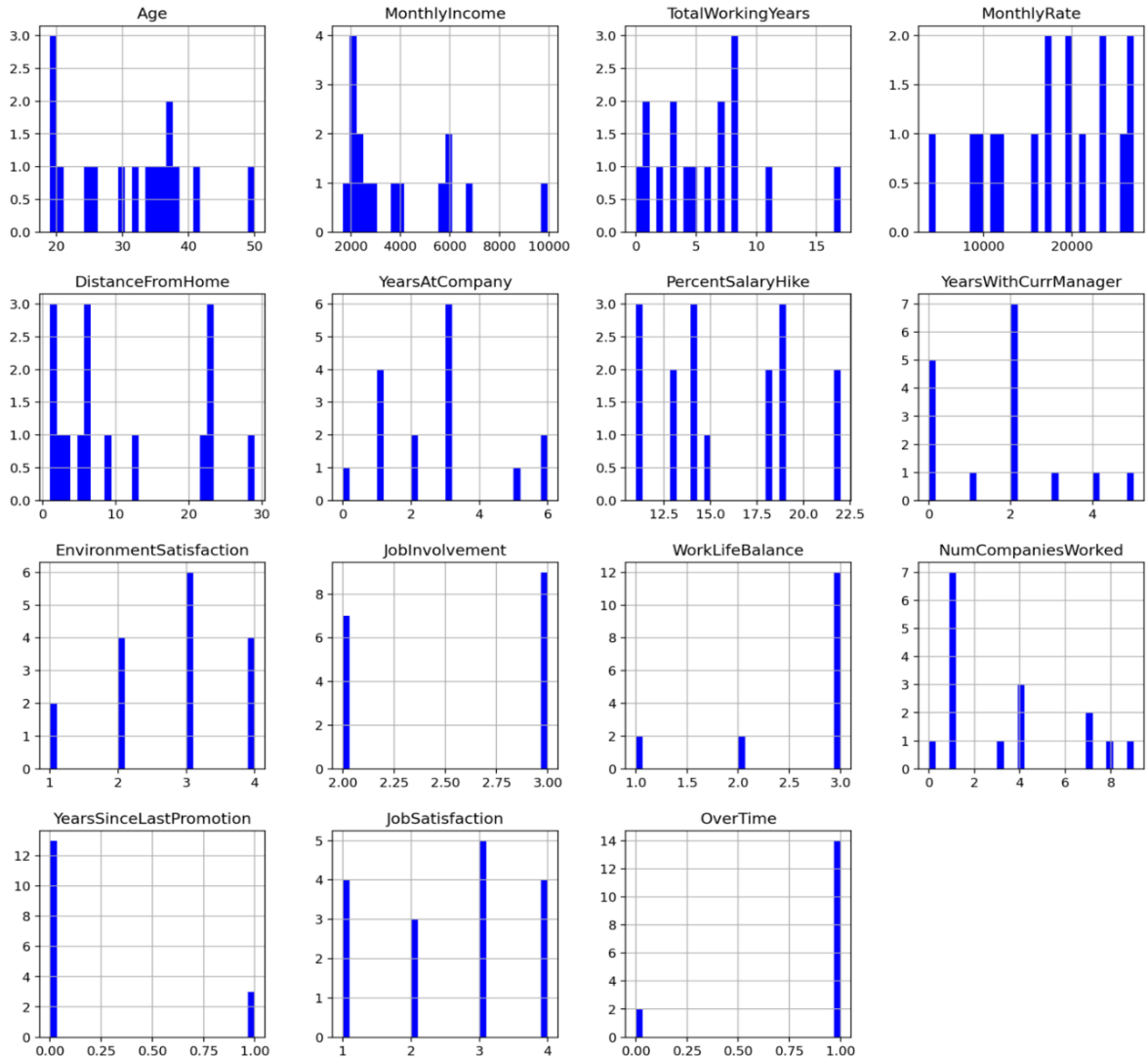
- Feature Importance: shows the most critical attributes for your companies' Employee Turnover. The features are listed here in descending order.

3.2 Feature Importance.

- Feature Importance shows the most critical attributes for your companies' Employee Turnover. The features are listed here in descending order.



- Top 15 Feature's Values Distribution: by exploring the histograms below, you can find out the most common values or value ranges for each attribute and make decisions accordingly to decrease your organization's employee turnover rate.



2.1.4 Testing

The name "Smart Labor Turnover Solution System" will be given after completing the project. Our team then learns how to test website functionality. Our project's final product test plans are these. This step of the procedure tests a web application's functioning. Source code testing verifies software functioning.

Functional testing includes:

- Locating data sources and input points.

- Testing execution.
- Accurate function recognition is necessary because program functioning depends on function integration.
- Results must be analyzed.

"Smart Labor Turnover Solution System " Web app testing is more important than ever due to predicted concerns. However, web application testing is complex. It depends on browser compatibility, application performance, user experience, user acceptability, security, and more. Companies require professional testers to test the website across all platforms, browsers, and devices. Web application testers must follow best practices to get accurate and reliable test results without wasting time.

Non-functional testing allows a few users to utilize this web app. HRs and certain others qualify. After receiving their feedback, we modified the system and got their opinion. Performance and usability are assessed during non-functional testing. Examine the system to ensure it meets the criteria. It fixes everything functional testing doesn't. Non-functional training checklists included performance, checklist, and documentation testing.

2.1.5 Dataset, Tools, and Libraries

This section describes the dataset, tools, and libraries used in the "Existing Labor Turnover Prediction" system.

1. Existing Labor Turnover Prediction Dataset

The dataset that is used is the 'IBM HR Analytics Employee Attrition & Performance' dataset, which can download from the following URL.

- https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

This dataset has a total number of 1470 entries (rows) and 35 features (Columns), as shown in the figure below.

	0	1	2	3	4	5	6	7
Age	41	49	37	33	27	32	59	30
Attrition	Yes	No	Yes	No	No	No	No	No
BusinessTravel	Travel_Rarely	Travel_Frequently	Travel_Rarely	Travel_Frequently	Travel_Rarely	Travel_Frequently	Travel_Rarely	Travel_Rarely
DailyRate	1102	279	1373	1392	591	1005	1324	1358
Department	Sales	Research & Development	Research & Development	Research & Development	Research & Development	Research & Development	Research & Development	Research & Development
DistanceFromHome	1	8	2	3	2	2	3	24
Education	2	1	2	4	1	2	3	1
EducationField	Life Sciences	Life Sciences	Other	Life Sciences	Medical	Life Sciences	Medical	Life Sciences
EmployeeCount	1	1	1	1	1	1	1	1
EmployeeNumber	1	2	4	5	7	8	10	11
EnvironmentSatisfaction	2	3	4	4	1	4	3	4
Gender	Female	Male	Male	Female	Male	Male	Female	Male
HourlyRate	94	61	92	56	40	79	81	67
JobInvolvement	3	2	2	3	3	3	4	3
JobLevel	2	2	1	1	1	1	1	1
JobRole	Sales Executive	Research Scientist	Laboratory Technician	Research Scientist	Laboratory Technician	Laboratory Technician	Laboratory Technician	Laboratory Technician
JobSatisfaction	4	2	3	3	2	4	1	3
MaritalStatus	Single	Married	Single	Married	Married	Single	Married	Divorced
MonthlyIncome	5993	5130	2090	2909	3468	3068	2670	2693
MonthlyRate	19479	24907	2396	23159	16632	11864	9964	13335
NumCompaniesWorked	8	1	6	1	9	0	4	1
Over18	Y	Y	Y	Y	Y	Y	Y	Y
OverTime	Yes	No	Yes	Yes	No	No	Yes	No
PercentSalaryHike	11	23	15	11	12	13	20	22
PerformanceRating	3	4	3	3	3	3	4	4
RelationshipSatisfaction	1	4	2	3	4	3	1	2
StandardHours	80	80	80	80	80	80	80	80
StockOptionLevel	0	1	0	0	1	0	3	1
TotalWorkingYears	8	10	7	8	6	8	12	1
TrainingTimesLastYear	0	3	3	3	3	2	3	2
WorkLifeBalance	1	3	3	3	3	2	2	3
YearsAtCompany	6	10	0	8	2	7	1	1
YearsInCurrentRole	4	7	0	7	2	7	0	0
YearsSinceLastPromotion	0	1	0	3	2	3	0	0
YearsWithCurrManager	5	7	0	0	2	6	0	0

2. Tools

Anaconda: is an R and Python distribution for Machine Learning and Data Science

applications. Anaconda acts as a hub for the libraries required for data science and machine learning, as any library may be installed via the anaconda command line.

Jupyter Notebook: is an environment that allows users to play with code in the browser without having to return the complete code each time. This enables users to experiment on only a small portion of the entire code, which is incredibly versatile.

Visual Studio Code: simplifies debugging, task execution, and version control. It provides just the tools needed for a fast code-build-debug cycle, leaving sophisticated procedures to full-featured IDEs like Visual Studio IDE.

Streamlit : is an open-source application framework developed in Python. It facilitates the creation of web applications for data science and machine learning. It works with important Python libraries such as scikit-learn, Keras,, NumPy, pandas, and Matplotlib, PyTorch, SymPy(latex) and many more.

Tool	Version
Anaconda Navigator	2.1.1
Jupyter Notebook	6.4.5
Visual Studio Code	1.70.1
Streamlit	1.11.0

1. Libraries

Pandas: is an open-source Python library for data manipulation and analysis.Pandas is simple to use and includes a comprehensive range of tools for working with data frames. One of the key functions of pandas is to modify data so that it may be used with machine learning methods.

Numpy: is an abbreviation for numerical Python, which is the foundation of all numerical and scientific computations. The key reason for using Numpy is that it is quick because it is written in C language.

Matplotlib: is a Python charting package that may be used to visualize data by plotting various graphs. This library can generate plots, histograms, bar charts, error charts, and scatter plots.

Seaborn: is a data representation or visualization package based on the Python matplotlib library, which provides a high-level interface for appealing drawings and instructive statistical visuals.

Scikit-Learn: sometimes known as sklearn, is an open-source Python machine learning package. This library provides Python implementations for a variety of tasks,

including data preprocessing and the implementation of several algorithms for classification, regression, and clustering.

TensorFlow: TensorFlow is an open-source machine learning library that trains and infers deep neural networks. Google uses it as a symbolic math library for research and production. In 2015, Google Brain distributed TensorFlow under Apache License 2.0 for internal usage.

Pickle: serialises Python object structures. It turns a Python object into a byte stream to save it to a file/database, keep programme state between sessions, or send data over a network. Unpickling the byte stream restores the object hierarchy.

Imblearn: is a Python module that provides various re-sampling techniques often employed in datasets with high between-class imbalance. Imblearn approaches are methods for collecting data with an equal proportion of classes. This method of data collecting would facilitate the generalization of the prediction model. It is compatible with scikit-learn and incorporated into scikit-learn-contrib projects.

Library	Version
Pandas	1.3.4
NumPy	1.20.3
Matplotlib	3.4.3
Seaborn	0.11.2
Scikit Learn	1.0.2
TensorFlow	2.5.0
Pickle	4.0
Imblearn	0.7.0

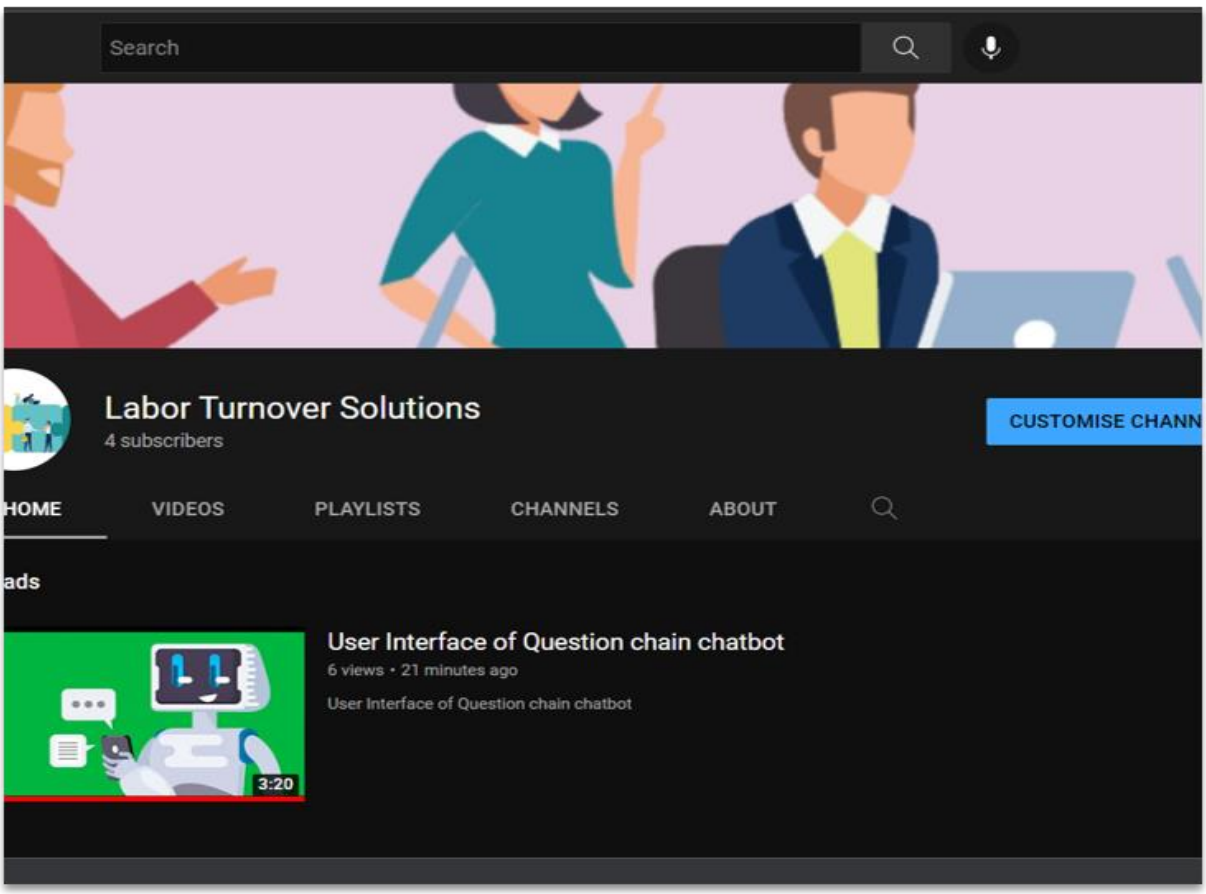
2.2 Commercialization Aspect of the Product

The existing labor turnover prediction component that was developed offers a number of benefits in its own right. This prediction system can be used to identify the employees who are most likely to leave cooperation and shows the most critical data, which results in a turnover. Since we have trained the machine learning model for several departments, such as research & development, sales, and human resources, this machine learning model can extend to other domains that need to predict employee turnover. The system allows the user to download the complete inputted dataset with the made prediction in it. So later, we can sell this dataset on the open market who wants to build a turnover system with more data. Moreover, the machine learning model can sell to cooperations for further development.

Our overall product goal of this research is to automate the organization's processes in Human resource management related to labor turnovers. Mainly the system has four research novelties, including existing labor turnover prediction,

question chain chatbot with emotion recognition, analyzing chatbot answers to providing summarization with visualization about leaving employees, and Resume Analyze. The system facilitates the HR management to identify leaving employees early, get the real reasons for resigning employees with visualizations, and choose the best candidate among the resumes for the job vacancy, which will be helpful for organizations to improve their productivity, efficiency, and income. Organizations can reduce their labor turnover by using this system. So mainly focus on selling our overall product to small and medium size companies, business groups, and enterprises operating at the industrial level. The budget for the overall product can be summarized as follows.

We have created a youtube channel and facebook group for the marketing our product as figure 27. We hope to post more videos, Advertisements and posters in the social media platforms.



Resources	Price(RS)
Domain and Web Hosting	10000.00
Traveling Cost	8000.00
Internet Usage	7500.00
Stationery	3000.00
Documentation	5000.00
Other expenses	4000.00
Total	37500.00

My individual component mainly focuses on selling the product to the HR departments of small and medium size industrial levels, companies, and business groups. The individual component's budget is as follows.

Resources	Price (RS)
Machine Learning Workshop	3000.00
Internet usage	5000.00
Traveling cost	3000.00
Stationery	1000.00
Other expenses	2500.00
Total	14500.00

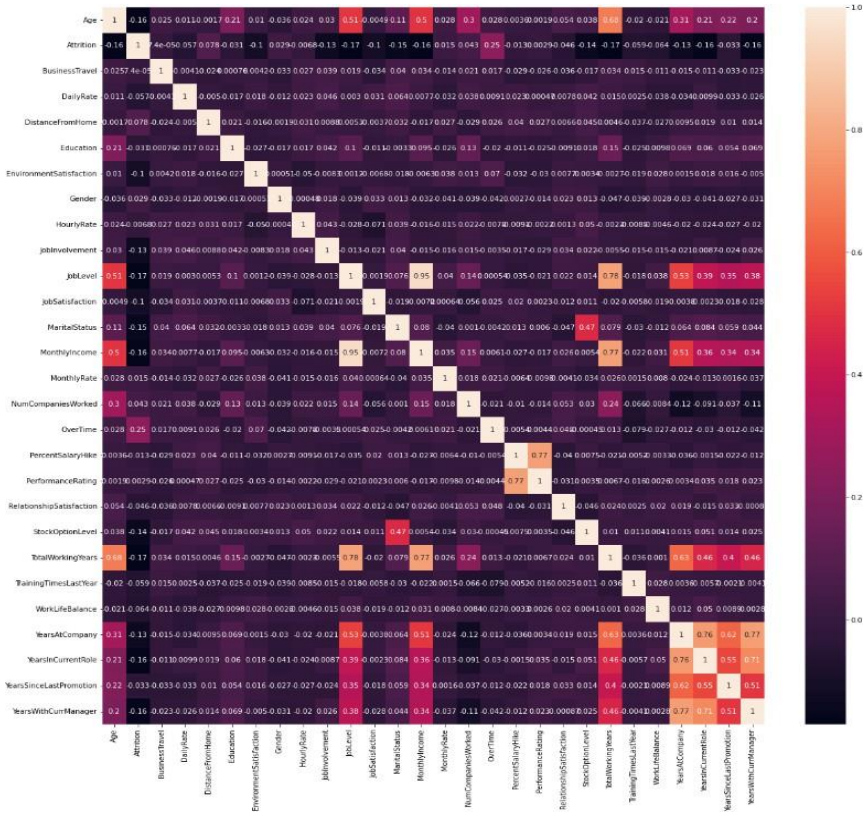
3 RESULTS & DISCUSSIONS

3.1 Results

1. Data Analysis and Findings

According to the correlation matrix (Figure 24 below), we can see that the following features are highly correlated.

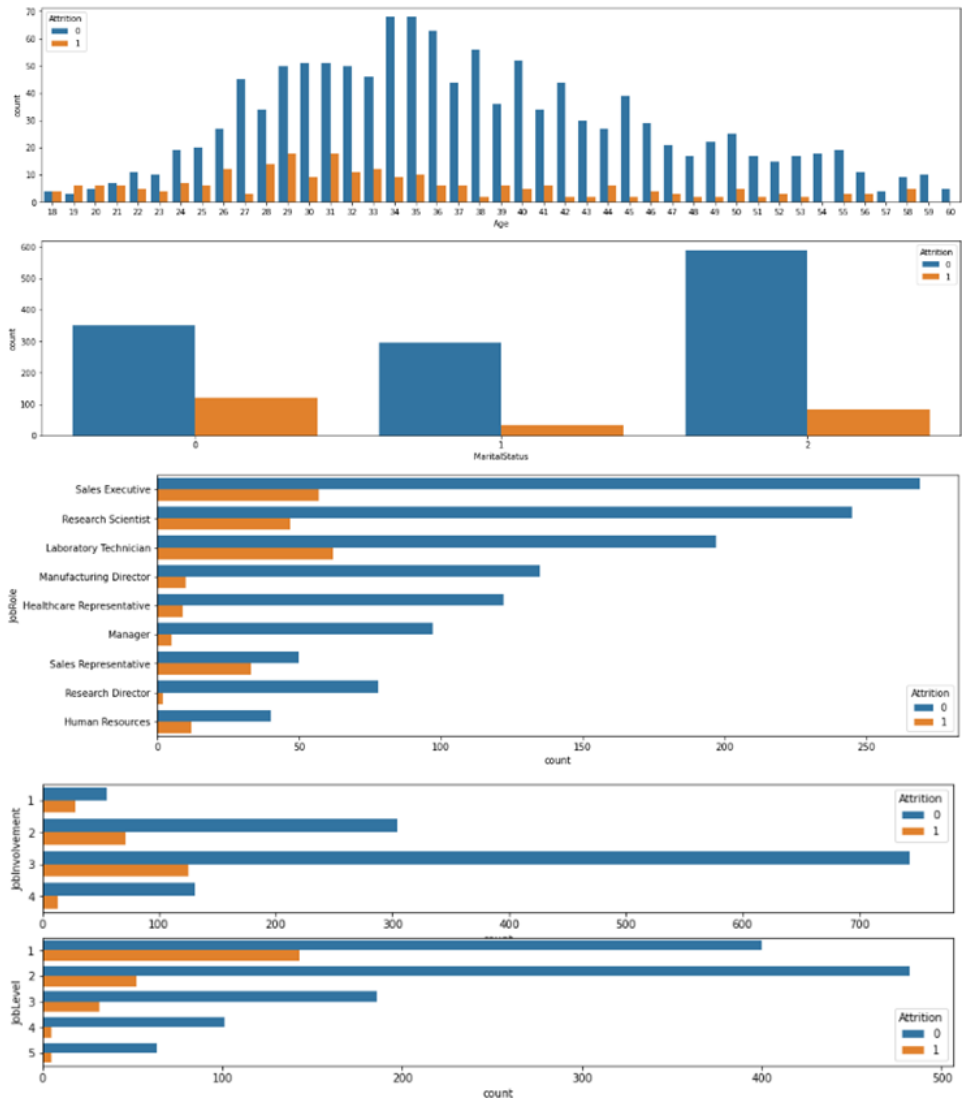
- JobLevel and MonthlyIncome
- JobLevel and TotalWorkingYears
- MonthlyIncome and TotalWorkingYears
- PrecentSalaryHike and PerformanceRating
- YearsAtCompany and YearsInCurrentRole
- YearsAtCompany and YearsWithCurentManager
- YearsInCurrentRole and YearsWithCurrentManager



Moreover, when we compare the value distribution of "Age," "MartialStaues," "JobRole," "JobLevel," and "JobInvolvement" features with the target variable "Attrition," we can conclude the findings as follows. (As shown in figure 25 below)

- People over 58Y rarely leave, and more employees tend to leave between 18Y and 21Y.
- Single employees tend to leave compared to married and divorced (0: single, 1: married, 2: divorced)

- Sales Representatives tend to leave compared to any other job.
- Less involved employees tend to leave the company. (1:'Low', 2:'Medium', 3:'High', 4:'Very High')
- Less experienced (low job level) tend to leave the company.



2. Model Performance Comparison of Balanced and Imbalanced Datasets

The original dataset consists of a total of 1029 records for training, and it was highly imbalanced, with only 166 records (16.13%) for class 1 and 863 records (83.67%) for class 0. So we can see that if we use this dataset without balancing, it will be highly accurate in detecting class 0 and poorly accurate in detecting class 1. As we only

have 1029 records in the dataset, we can see that undersampling will not be a good choice, and we used two oversampling techniques, which are SMOTE and ADASYN, to work with this imbalanced dataset. We compare the two oversampling methods with different algorithms and choose the best technique for each ML algorithm.

After performing SMOTE oversampling method, we successfully generated 697 new synthetic samples for class 1 to balance the two classes. After the SMOTE technique, we got 863 records (50%) for both classes, and our total training dataset has increased to 1726 records. Then we applied ADASYN oversampling method for class 1 and increased to 839 records (49.29%) by making synthetic records. Moreover, after the ADASYN technique, the training dataset has increased to 1702 records.

Original Training Dataset			
Class	Count	Precentage	Total
0 - Stay	863	83.67%	1029
1 - Leave	166	16.13%	
SMOTE			
0 - Stay	863	50%	1726
1 - Leave	863	50%	
ADASYN			
0 - Stay	863	49.29%	1702
1 - Leave	839	50.71%	

In this project, we used 2 ML algorithms and 1 DL algorithm, which are random forest, logistic regression, and artificial neural networks. As we deal with an imbalanced dataset, we use the AUC score as our primary evaluation metric. The following table shows how well each model performed with original and oversampled datasets. All the models performed better in the oversampled datasets than in the original dataset. Moreover, we can conclude that LR and RF performed at their best with SMOTE dataset, and ANN performed best with the ADASYN dataset.

Original Dataset					
Classifier	Accuracy	Precision	Recall	F1 Score	AUC
LR	0.89	0.88	0.69	0.74	0.69
RF	0.86	0.88	0.56	0.57	0.56
ANN	0.87	0.77	0.69	0.72	0.69
SMOTE					
LR	0.80	0.67	0.72	0.68	0.72
RF	0.84	0.70	0.66	0.68	0.66
ANN	0.78	0.64	0.69	0.66	0.69
ADASYN					
LR	0.79	0.66	0.71	0.68	0.71
RF	0.83	0.67	0.61	0.63	0.61

ANN	0.79	0.65	0.70	0.67	0.70

3. Model Performance Comparison After Hyperparameter Tuning

Hyperparameter tuning was done after balancing the dataset using SMOTE and ADSYN oversampling techniques. GridSearchCV was used for hyperparameter tuning. According to [6] and [10], we tested 20000 different parameter value combinations for LR, 13440 for RF, and 3000 for ANN, as shown in the table below. Even though many parameters and values can be tuned for each ML model, we only use the following, as hyperparameter tuning is a highly time and resource-consuming task.

As shown in the table below, the following are the optimal parameter values for each parameter of the ML models. Moreover, with hyperparameter tuning, we used cross-validation to ensure the model performance and reduce overfitting. We used 10-fold cross-validation for LR and 5-fold cross-validation for RF and ANN. After hyperparameter tuning LR and ANN models AUC score increased from 0.72 and 0.70 to 0.74 and 0.71, respectively. However, the RF model's AUC score decreased from 0.66 to 0.64. the reason for that can be that we haven't tuned enough parameter values for RF, or the initial RF model was overfitted, and after the cross-validation was performed, the model AUC score decreased.

Logistic Regression			
Parameters	Grid Values	Initial Values	Best Values
C	np.logspace(-4,4,200)	1	117.585
penalty	l1, l2	l2	l2
solver	lbfgs, liblinear, sag, saga, newton-cg	lbfgs	lbfgs
AUC Score		0.72	0.74
Total fits	20000		
Tuning time	3min 21s		
Random Forest			
Parameters	Grid Values	Initial Values	Best Values
n_estimators	10, 50, 100, 200, 250, 400, 500	100	250
max_depth	None, 3, 5	None	None
min_sample_split	2, 3, 4, 5	2	2
min_sample leaf	2, 3, 4, 5	1	2
max_features	auto, sqrt	sqrt	auto
Bootstrap	True, False	True	False
AUC Score		0.66	0.64
Total Fits	13440		
Tuning Time	30min 47s		
Artificial Neural Network			
Parameters	Grid Values	Initial Values	Best Values
Epochs	100, 200, 300, 400, 500, 600, 700	100	100
Batch size	64, 128, 256	50	64

Optimizer	SGD, Adadeta, RMSprop, Adagrad, Adam	Adam	SGD
AUC Score		0.70	0.71
Total Fits	3000		
Tuning Time	9min 47s		

4. Summary of Models

The random forest model has the lowest AUC score of 0.64 with SMOTE oversampled dataset and n_estimators = 250, max_depth = None, min_sample_split = 2, min_sample leaf = 2, max_features = “auto”, and Bootstrap = False as the best parameter values.

The artificial neural network model has an AUC score of 0.71 with ADASYN oversampled dataset and epochs = 100, Batch size = 64, and Optimizer = "SGD" as the best parameter values. The ANN is the second best-performing model. Moreover, in the ANN model, we have two hidden layers, and we used "relu" as the activation function in the hidden layers. The first hidden layer has 120 neurons and the second hidden layer has 60 neurons in it.

The logistic regression model has the highest AUC score of 0.74 with SMOTE oversampled dataset, and C = 117.585, penalty = "l2", and solver = "lbfgs" were selected as the optimal parameter values for the model.

Logistic Regression					
	Accuracy	Precision	Recall	F1 Score	AUC
Initial	0.89	0.88	0.69	0.74	0.69
SMOTE	0.80	0.67	0.72	0.68	0.72
ADASYN	0.79	0.66	0.71	0.68	0.71
Hyperparameter Tuned	0.81	0.68	0.74	0.70	0.74
Random Forest					
Initial	0.86	0.88	0.56	0.57	0.56
SMOTE	0.84	0.70	0.66	0.68	0.66
ADASYN	0.83	0.67	0.61	0.63	0.61
Hyperparameter Tuned	0.83	0.68	0.64	0.66	0.64
Artificial Neural Network					
Initial	0.87	0.77	0.69	0.72	0.69
SMOTE	0.78	0.64	0.69	0.66	0.69
ADSYN	0.79	0.65	0.70	0.67	0.70
Hyperparameter Tuned	0.75	0.63	0.70	0.65	0.71

In comparing all three models, the LR model is chosen as the final model for the

existing labor turnover prediction, with the highest performance and AUC scores among the other two models.

5. Feature Importance

The Feature Importance approach gives a score to input features depending on how helpful those traits are when it comes to predicting the variable being looked for. The logistic regression model, which has been chosen as the best model, contains a total of 43 features, and the contribution that each feature makes to the overall accuracy of the model can be seen in the figure below. We can see that monthly income, age, daily rate, total working years, and overtime are the most influential reasons for employee attrition.

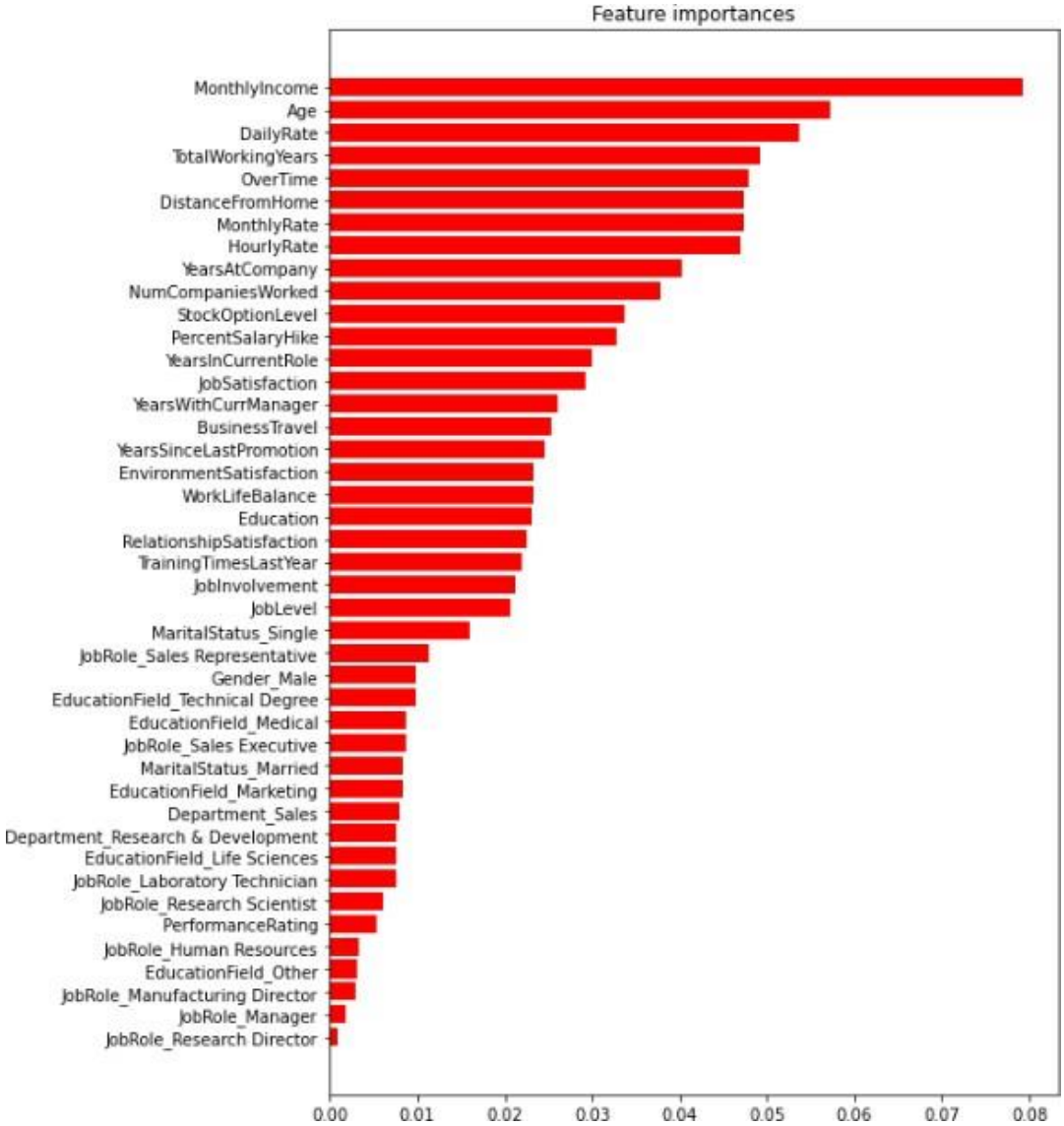


Figure 26 Feature importance

3.2 Research Findings

The primary focus of the research is how to use machine learning and deep learning techniques to solve the problem of existing labor turnover by predicting the employees who would most likely quit the job in the future, finding the most influential reasons for the attrition, and minimizing it.

We found that when it comes to employee attrition prediction, simple Ml classifiers, specifically logistic regression, perform well than the more complex ensemble methods, specifically random forest or neural networks, specifically artificial neural networks with two hidden layers with 120 and 60 neurons, respectively.

Moreover, we balanced the imbalanced dataset and found that all the ML algorithms work well with the balanced dataset. We mainly used SMOTE and ADASYM oversampling techniques and figured that RF and LR algorithms are working well with the SMOTE approach and ANN works well with the ADASYN approach.

We also found that even though hyperparameter tuning is a very time and resource-consuming task. It can help optimize the best parameters for ml algorithms and increase the model performance. Moreover, cross-validation can be used to overcome the overfitting problem while building the ML Models.

Finally, we have found that monthly income, age, daily rate, total working years, and overtime are the most compelling reasons for employee attrition by feature importance.

3.3 Discussion

The IBM dataset includes individuals from various backgrounds, at various career stages, with various degrees of salary and performance. It is evident that statistical analysis or neural network will most likely perform better than an ensemble or tree-based method, taking into account the numerous themes and categories naturally present in the data. We can see that LR is the best-performing model, ANN is the second highest-performing model, and RF is the least-performing model. Moreover, oversampling helped to increase the model accuracy, and different oversampling methods performed differently with other ML models. Hyperparameter tuning and cross-validation helped not only to make the ML models more accurate but also to reduce overfitting.

3.4 Summary of Student's Contribution

Member	Component	Tasks
T.J.A Udayana (IT19199672)	Existing Labor Turnover Prediction	• Created the User Interface using Streamlit Framework. • Deployed the web application in the streamlit cloud. • Collected the Actual data from the employees. • Integrated the web application components. • Implemented the machine learning model using the LR,RF and ANN. • Balanced the Training dataset using Smote and ADAsyn. • Improve the model accuracy and AUC using the hyperparameter optimization. • Tested the System with Non-IT users.

4 CONCLUSION

This paper presented machine learning and deep learning techniques to predict employee turnover. The focus is on using different algorithms and combinations of several data preprocessing techniques, oversampling techniques, and hyperparameter tuning to utilize the effectiveness of the employee attrition prediction, as Employee Attrition is one of the biggest Business Problems. In the data preprocessing, we have performed duplicate and null feature removal, ordinal encoding, and irrelevant feature removal. Two oversampling techniques have been used to work with the imbalanced dataset: SMOTE and ADASYN. For Modelling, two machine learning algorithms: Logistic Regression, Random Forest, and one deep learning algorithm: Artificial Neural Network, are performed on the dataset. Finally chose logistic regression model with a 0.74 AUC score was the best model due to its performance which was evaluated accordingly to evaluation matrices.

REFERENCES

[1] G. Marvin, M. Jackson, and M. G. R. Alam, "A machine learning approach for employee retention prediction," in 2021 IEEE Region 10 Symposium (TENSYMP), 2021.

[2] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, "An improved machine learning-based employees attrition prediction framework with emphasis on feature selection," Mathematics, vol. 9, no. 11, p. 1226, 2021.

[3] H. Zhang, L. Xu, X. Cheng, K. Chao, and X. Zhao, "Analysis and prediction of employee turnover characteristics based on machine learning," in 2018 18th International Symposium on Communications and Information Technologies (ISCIT), 2018.

[4] K. Bhuva, and K. Srivastava, "Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition," in IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 2018.

[5] S. Yadav, A. Jain, and D. Singh, "Early prediction of employee attrition using data mining techniques," in 2018 IEEE 8th International Advance Computing Conference (IACC), 2018

[6] A. Nurhindarto, E. W. Andriansyah, F. Alzami, P. Purwanto, M. A. Soeleman, and D. P. Prabowo, "Employee Attrition and Performance Prediction using Univariate ROC feature selection and Random Forest," Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control, 2021.

[7] M. Subhashini and R. Gopinath, "Employee Attrition Prediction in Industry using Machine Learning Techniques," 2021.

[8] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee attrition prediction using deep neural networks," Computers, vol. 10, no. 11, p. 141, 2021.

[9] R. Yedida, R. Reddy, R. Vahi, R. Jana, A. GV, and D. Kulkarni, "Employee attrition prediction," arXiv preprint arXiv:1806.10480, 2018.

[10] R. Maharjan, "Employee Churn Prediction using Logistic Regression and Support Vector Machine," 2021.

[11] T. Juvitayapun, "Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods," in 2021 13th International Conference on Knowledge and Smart Technology (KST), 2021.

[12] M. K. Sharma, D. Singh, M. Tyagi, A. Saini, N. Dhiman, and R. Garg, "Employee Retention And Attrition Analysis: A Novel Approach On Attrition Prediction Using Fuzzy Inference And Ensemble Machine Learning," Webology (ISSN: 1735-188X), 19(2), 2022.

[13] R. A. Danquah, "Handling Imbalanced data: A case study for binary class problems," arXiv [stat.ML], 2020.

[14] T. P. Salunkhe, "Improving employee retention by predicting employee attrition using machine learning techniques," (Doctoral dissertation, Dublin Business School), 2018.

[15] J. Judrups, R. Cinks, I. Birzniece, and I. Andersone, "Machine learning based solution for predicting voluntary employee turnover in organization," in 20th International Scientific Conference Engineering for Rural Development Proceedings, 2021.

[16] N. Mansor, N. S. Sani, and M. Aliff, "Machine Learning for Predicting Employee Attrition," International Journal of Advanced Computer Science and Applications, 12(11), 2021.

[17] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," in 2018 International Conference on Innovations in Information Technology (IIT), 2018

[18] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," Appl. Sci. (Basel), vol. 12, no. 13, p. 6424, 2022

[19] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, 2020.

[20] D. S. Rodrigo and G. S. Ratnayake, "Employee Turnover Prediction System: With Special Reference to Apparel Industry in Sri Lanka," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-9, doi: 10.1109/I2CT51068.2021.9418108.

[21] A. Qutub, the King Abdulaziz University, Information Systems Department, Faculty of Computing and Information Technology Jeddah, Saudi Arabia, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, "Prediction of employee

attrition using machine learning and ensemble methods," Int. J. Mach. Learn. Comput., vol. 11, no. 2, pp. 110–114, 2021.

[22] R. Punnoose and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," Int. j. adv. res. artif. intell., vol. 5, no. 9, 2016.

[23] S. Yang and M. T. Islam, "IBM Employee Attrition Analysis," arXiv [cs.CY], 2020.

[24] S. N. Khera and Divya, "Predictive modelling of employee turnover in Indian IT industry using machine learning techniques," Vis. J. Bus. Perspect., vol. 23, no. 1, pp. 12–21, 2019.

[25] A. Frye, C. Boomhower, M. Smith, L. Vitovsky, and S. Fabricant, "Employee Attrition: What Makes an Employee Quit?," SMU Data Science Review, 1(1), 2018.

[26] K. K. Mohbey, "Employee's attrition prediction using machine learning approaches," in Machine Learning and Deep Learning in Real-Time Applications, IGI Global, pp. 121–128, 2020.

[27] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," Appl. Soft Comput., vol. 24, pp. 994–1012, 2014.

[28] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in Advances in Intelligent Systems and Computing, Cham: Springer International Publishing, pp. 737–758, 2019.

[29] J. D. Novaković, A Veljović, S. S. Ilić, Z. Papić, and T. Milica, "Evaluation of classification models in machine learning," Theory and Applications of Mathematics & Computer Science, 7(1), 39-46, 2017.

[30] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015.

R. Jain and A. Nayyar, "Predicting employee attrition using XGBoost machine learning approach," in 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), 2018.