PREDICTION OF EMPLOYEE ATTRITION USING BAYES RISK POST-PRUNING IN DECISION TREE

WIN PA PA MAY PHYO AUNG

M.C.Sc. **JANUARY 2023**

PREDICTION OF EMPLOYEE ATTRITION USING BAYES RISK POST-PRUNING IN DECION TREE

BY

WIN PA PA MAY PHYO AUNG

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

UNIVERSITY OF COMPUTER STUDIES, YANGON

JANUARY 2023

ACKNOWLWDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my sincere thanks to **Dr. Mie Mie Khin**, Rector, University of Computer Studies, Yangon, for her kind permission to submit this thesis.

My heartfelt thanks and respect go to **Dr. Nang Soe Soe Aung**, Principal and Pro-rector of Computer University (Lashio), for her invaluable and administrative support.

I would like to express my deeply thanks to **Dr. Si Si Mar Win,** Processor and **Dr. Tin Zar Thaw**, Professor, Course coordinator of Master's (CS), Faculty of Computer Science, University of Computer Studies, Yangon, for their painstaking suggestion and encouragement throughout the development of the thesis.

My sincere thanks and regards go to my supervisor, **Dr. Nilar Aye,** Professor, Head of Faculty of Information Science, University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

My sincere thanks and regards go to **Dr. Khaing Khaing Wai**, Professor, Head of Department of Information Technology Support and Maintenance, University of Computer Studies, Yangon, for her kind management throughout the completion of this thesis.

Additionally, I would like to express my deep appreciation to **Daw Toe Toe**, Associate Professor, Head of Faculty of Information Science, Computer University (Lashio), who gave me the opportunity to develop this thesis for her general guidance during the period of study.

I also wish to express my deepest gratitude to **Dr. Zin Mar Kyu**, Associate Professor, Faculty of Information Science, Computer University (Lashio), for her kind direction, inspiration, and practical advice throughout the completion of this thesis as well as for her numerous supportive recommendations and good ideas.

I would also like to thank the thesis board examiners for their insightful remarks and edits, which helped me produce a quality product. I would like to thank **Daw Mya**

Hnin Mon, Associate Professor, Department of English at the University of Computer Studies in Yangon, for all of her help and for proofreading my thesis from a linguistic standpoint.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation. I especially thank to my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

I would like to thank to everybody who has helped me in any ways to make my thesis beautiful. I really appreciate their help very much. So, let me say that I'm very grateful to everyone with all my heart.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…………………………. ……………………………………………………

Date Win Pa Pa May Phyo Aung

ABSTRACT

Employee attrition is the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death, or retirement. Attrition is widely understood to be one of the major problems affecting organizations today. Losing employees has many direct and indirect impacts across a company. It occurs employee attrition when an employee leaves and is not replaced at all or for a significant amount of time, resulting in a reduction of the workforce. In this system, Decision Tree (ID3) classifier is used to analyze the causes of employee attrition. And then Bayes Risk Post-Pruning (PBMR) technique is applied to reduce the condition of overfitting on decision tree. The proposed system performance is evaluated various evaluation standards such as precision, sensitivity and F1 score values based on IBM Human Resource Analytic Employee Attrition and Performance dataset from Kaggle site. The proposed system compares the accuracy between before post-pruning and after Bayes Risk post pruning was applied. The proposed approach findings help organizations overcome employee attrition by improving the factors that cause attrition. This system is implemented by using python programming language with Google Collab Drive.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF EQUATIONS

Page

CHAPTER 1
INTRODUCTION

The ability of a machine to mimic intelligent human behavior is known as artificial intelligence. One subfield of artificial intelligence is machine learning. Data—numbers, photographs, or text is the foundation of machine learning. Examples of data include bank transactions, pictures of individuals or even specific bakery goods, repair records, time series data from sensors, or sales reports. The information is collected and made ready to be served as training data, or the material on which a machine learning model will be trained. The application works best when there is more data. Machine learning (ML) has been developed and effectively applied to a wide number of real-world areas, making it one of the fastest-growing disciplines of study. This study presents a comparative analysis of prediction employee attrition before pruning tree and after pruning tree.

Employee attrition in a firm is the term used to describe the loss of employees due to conventional ways such as retirement and resignation, elderly clients dying, or layoffs resulting from a change in the organization's target demographics. The high incidence of personnel attrition inside a company is a critical issue as it has a significant impact on them. With priceless tacit knowledge, which is frequently the source of a company's competitive advantage, employees leave a company [10]. The cost of recruiting and training new employees, as well as any resulting downtime for the business, are paid by the employer. On the other hand, increased retention reduces the costs of acquiring and training new employees as well as enables the progressive influx of more seasoned workers into the workforce. Nowadays, businesses are very interested in knowing the reasons behind staff attrition in order to reduce employee turnover. As a result, an organization must focus on anticipating employee turnover and identifying the primary reasons of attrition in order to enhance its human resource strategy [8].

If an employee is fired by their employer for any reason, there could be a number of reasons why they leave, such as a lower salary, a lack of job satisfaction, personal reasons, or environmental concerns. On the other hand, voluntary attrition is sometimes referred to as the employee who leaves an employer. If the person is talented, the organization loses out from this type of attrition. Everyone wants a bigger pay and job security in the current environment. Because of this, workers quit their jobs right away if they have better opportunities elsewhere.

Machine learning methods are becoming a significant part of predicting employee attrition in the modern field of computer science. Based on the employee's past performance data, such as age, experience, education, previous promotion, and so on, these approaches offer forecasts. The HR department is aware of employee attrition in advance based on the forecast results. As a backup plan for the worker who intends to quit in the near future, the HR department has already begun recruiting new workers.

The IBM Human Resource Analytic Employee Attrition and Performance dataset use in this study is a freely accessible dataset from the Kaggle Dataset Repository. Employee satisfaction, salary, seniority, and demographic information are the four main components of the dataset. The dataset includes a number of attributes that affect the predicted variable "Attrition," which indicates whether an employee left the organization or not based on 1,470 cases and 35 attributes. Attrition is the identified class, and there are 237 instances of "Yes" and 1233 instances of "No" in it. The proposed system compares employee attrition predictions before and after decision tree pruning in order to determine whether method is more accurate and efficient at predicting employee attrition.

1.1 Objectives of the Thesis

The main objectives of the thesis are

- To analyze employee attrition using Decision Tree classifier
- To comprehend Bayes Risk Post-Pruning Algorithm
- To reduce recruitments, hiring and training costs
- To control the growing employee attrition rate
- To maintain qualified and strong Human Resource processes

1.2 Related Works

The numerous studies that have been offered in the field of employee attrition prediction are analyzed in this section, along with their benefits and drawbacks, in detail.

The performance of three different classifiers—the artificial neural network classifier and the decision tree classifier—were compared [5]. The authors of this study employed the Employee Attrition and Performance dataset from IBM Human Resource

Analytic. For optimization, they also used techniques like regularization and parameter adjustment. The objective of this study is to forecast employee attrition using machine learning classification models.

The post-pruning method discussed in this study [6] takes into account a number of evaluation criteria, including attribute choice, accuracy, tree complexity, time spent pruning the tree, precision/recall scores, TP/FN rates, and area under the ROC. The Zoo, Iris, Diabetes, Labor, and Blogger datasets were among the five used by this system. This study demonstrated that the suggested approach generated classification accuracy that was superior to Reduced-error Pruning (REP) and Minimum-error Pruning (MEP). The results of the experiments demonstrated that, across all test datasets, the suggested technique generated classification accuracy that was superior to REP and MEP.

In order to predict employee attrition, the four cutting-edge machine learning techniques Extra Trees Classifier (ETC), Support vector machine (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC) were used in this work [3]. In order to identify the causes causing employee attrition, the Employee Exploratory Data Analysis (EEDA) was performed. The primary elements that contribute to employee attrition, according to their analysis, are monthly income, hourly wage, job level, and age. Results of the study were based on the IBM HR employee attrition dataset. Compared to other machine learning algorithms, the Extra Trees Classifier (ETC) that proposed more accurate.

This study compares the effectiveness of Naive Bayes, SVM, decision trees, random forests, and logistic regression as machine learning techniques. The result has been supplied that it will assist us in determining the conduct of employees. It can be taken into account the following time. In comparison to other machine learning techniques, experimental data show that the logistic regression strategy can achieve up to 86% accuracy.

1.3 Scope of Thesis

There are five chapters in this thesis. The thesis's first chapter lays out its introduction, objectives, and overall structure.

The background theory and literature review for data mining and matching learning methodologies are presented in Chapter 2.

The suggested system's design is covered in Chapter 3. Additionally included in this chapter are the suggested system's approach, detailed description, and system overview.

Chapter 4 goes into considerable detail about the system architecture, implementation, experimental findings, and system discussions.

Chapter 5 of this thesis covers the conclusion, as well as the benefits, drawbacks, and prospective future developments of this system.

CHAPTER 2
BACKGROUND THEORY

This chapter presents the background theory that the research project is tied to. All businesses should make an effort to predict staff turnover since doing it. So, it can help them better understand their customers and forecast future sales. Additionally, it can help the business pinpoint and enhance its human resource weaknesses. The personnel data in many firms has been the subject of a great deal of research, but there is still more to be discovered. Results for data from different industries are different.

2.1 Data Mining

The process of "mining" knowledge from vast amounts of data is known as data mining. The phrase itself is misleading. Just be aware that the phrase used to describe the extraction of gold from rocks or sand is gold mining, not rock or sand mining. Data mining should thus have been given the regrettably lengthy moniker "knowledge mining from data," which is more fitting. The emphasis on mining from enormous volumes of data cannot be adequately conveyed by the term "knowledge mining," which is shorter. Nevertheless, mining is a colorful term that describes the process of extracting a few priceless gems from a large quantity of raw material. So, an erroneous term that includes both "data" and "mining" gained popularity.

Data mining's insightful data analysis has improved business decision-making. These studies' supporting data mining approaches have two main objectives: either characterizing the target dataset or predicting results using machine learning algorithms. Data is organized and filtered using these strategies to highlight the most important information, such as fraud detection, user patterns, bottlenecks, and even security breaches.

The four basic processes of data mining are typically goal-setting, data collection and preparation, data mining algorithm application, and outcome evaluation.

1. Set the company objectives first. This is sometimes the most challenging step in the data mining process, and many businesses neglect to give it enough attention. The definition of the business issue by data scientists and business stakeholders is essential since it guides the development of the data queries and project parameters.

2. Data preparation: Once the problem's breadth is known, it is simpler for data scientists to determine which collection of data will be useful in supplying the information needed to address the business's specific inquiries. Data cleaning, which eliminates noise like duplicates, missing numbers, and outliers, will be done once they have gathered the pertinent data.
3. Model development and pattern mining: Data scientists can look into any intriguing data linkages, such as correlations, association rules, or sequential patterns, depending on the sort of study they are performing.
4. Results evaluation and knowledge application: After the data has been compiled, the outcomes need to be assessed and explained. Results should be valid, original, applicable, and comprehensible when they are finalized.

2.2 Machine Learning

Machine learning is a branch of computer science and artificial intelligence (AI) that focuses on simulating human learning by using data and algorithms to improve the system's accuracy over time. Machine learning is frequently used to solve complex issues or complete jobs that demand a lot of data. It is a great answer for more complex data and delivers faster, more accurate findings. It helps a business identify profitable opportunities or any unforeseen hazards [16].

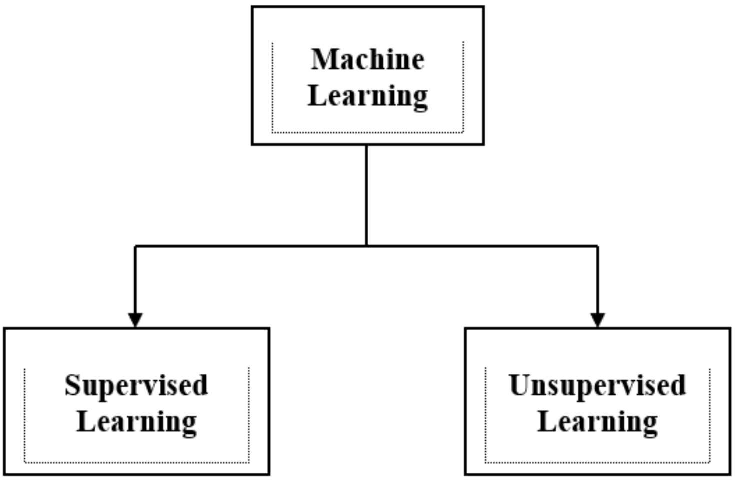


Figure 2.1 Type of Machine Learning

2.2.1 Supervised Learning and Unsupervised Learning

The method of classifying data involves two steps. A model describing a predetermined collection of data classes or concepts is constructed in the first stage.

The model is built by looking at database tuples with attributes. According to the class label property, one of the attributes, each tuple is deemed to belong to a preset class. Data tuples are sometimes referred to as samples, examples, or objects when used in the classification context.

The training data set consists of all of the data tuples that were examined to create the model. The individual tuples that comprise the training set are chosen at random from the sample population and are known as training samples. This step is also known as supervised learning since the class label for each training sample is provided (i.e., the model's learning is "supervised" because it is aware of the class to which each training sample belongs). Contrast this with unsupervised learning (also known as clustering), where the number or set of classes to be taught cannot be specified in advance and each training sample's class label is unknown.

Classification rules, decision trees, or mathematical formulas are all examples of representations for the learned model. For instance, classification rules can be developed to categorize clients as having either excellent or fair credit ratings given a database of customer credit information. Future data samples can be categorized using the rules, and they can also help us understand the contents of the database.

The model is applied for classification in the following phase. The model's prediction accuracy is estimated first. A test set of class-labeled samples is used in the straightforward holdout method. These samples are chosen at random, and they are separate from the training samples. A model's accuracy on a given test set is measured by the proportion of test set samples that the model correctly classifies. The learnt model's class prediction for each test sample is contrasted with the sample's actual class label.

Future data tuples or objects for which the class label is unknown can be classified using the model if the accuracy is deemed acceptable. These data are also referred to as "unknown" or "previously unseen" data in machine learning. Prediction can be understood as the creation and use of a model to evaluate the kind of unlabeled sample or to evaluate the value or ranges of a particular attribute that a sample is likely to possess.

According to this perspective, the two main types of prediction are classification and regression. Regression is used in classification to predict class labels, and prediction is used in regression techniques to predict continuous values. This point of view is widely held in data mining. There are many uses for classification and prediction, including selective marketing, medical diagnosis, and credit approval.

2.3 Data Classification Process

The process of data classification involves two parts. They are classification and learning.

1. **Learning**: A classification algorithm examines training data. Credit rating serves as the class label attribute in this instance, while the classification rules serve as a representation of the learnt model or classifier.
2. **Categorization:** Test data are used to gauge the degree to which the classification rules are accurate. The rules can be used to categorize fresh data tuples if the accuracy is deemed acceptable [2].

2.4 Classification Methods

Types of Classification to generate models representing significant data classes or to forecast future data trends, two types of data analysis are classification and prediction. Prediction simulates continuous-valued functions, whereas classification models' categorical labels (classes).

1. Logistic Regression

To predict a binary outcome—that is, whether something happens or not—logistic regression is used. This can be expressed as Yes/No, Pass/Fail, Alive/Dead, etc.

The binary outcome, which falls into one of two groups, is determined by analyzing independent variables. The dependent variable is usually categorical, although the independent variables might be either category or quantitative. Written as follows:

$$P(Y=0|X) \text{ or } P(Y=1) \qquad \textbf{2.1}$$

Given independent variable X, it determines the likelihood that dependent variable Y will occur.

Using this, one can assess whether a word has a positive or negative connotation (0, 1, or on a scale between). As an alternative, it can be used to assign a probability between 0 and 1 to each item in a picture, such as a tree, flower, or blade of grass.

2. **Naive Bayes**

A probabilistic machine learning model called a Naive Bayes classifier is utilized for classification tasks. The Bayes theorem serves as the foundation of the classifier.

Bayes Theorem:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad \textbf{2.2}$$

The likelihood of A happening given that B has already happened, according to the Bayes theorem. Here, A is the hypothesis and B is the supporting evidence. Here, it is assumed that the predictors and features are independent. That is, the presence of one feature does not change the behavior of another. The term "naive" is a result.

3. **K-Nearest Neighbors**

K-Nearest Neighbor is one of the most basic supervised learning-based machine learning algorithms. The K-NN algorithm places the new instance in the category that resembles the current categories the most, presuming that the new case and the previous cases are comparable. After storing all the previous data, a new data point is categorized using the K-NN algorithm based on similarity. This indicates that new data can be reliably and quickly categorized using the K-NN approach.

The K-NN technique can be used for regression even though classification problems are where it is most typically applied. K-NN makes no assumptions about the underlying data because it is a non-parametric approach. As a result of saving the training dataset rather than instantly learning from it, the method is also referred to as a lazy learner. Instead, it performs an action while classifying data by using the dataset. The KNN approach simply stores the data during the training phase and categorizes fresh data into a category that is very similar to the training data.

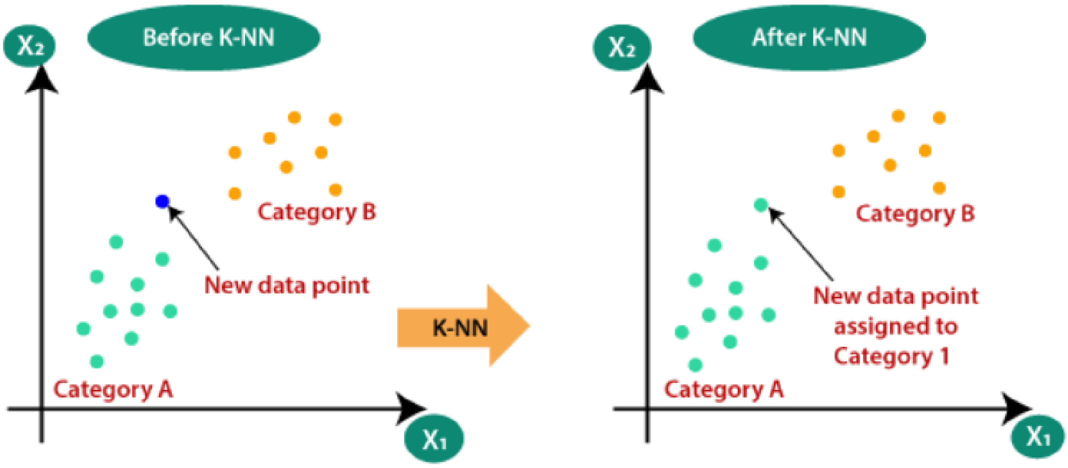


Figure 2.2 K-Nearest Neighbor (KNN)

4. Decision Tree

A decision tree is a supervised learning method that specializes at classification issues because it can rank classes precisely. Like a flow chart, it divides data points into two comparable categories at a time, starting with the "tree trunk" and moving through the "branches," "leaves," and finally the "leaves," where the categories become more finitely similar. Through the creation of categories within categories, this enables organic classification with minimal human oversight.

5. Support Vector Machines

Classification or regression issues can be solved using the "Support Vector Machine" (SVM) method of supervised machine learning. However, the most common application of it is in classification-related concerns. Each data point is represented by a point in n-dimensional space (n is the number of features you have) when using the SVM algorithm, with the value of each feature being the value of a certain coordinate. The components of a support vector are the coordinates of each individual observation. The SVM classifier performs the best in terms of separating the two classes (hyper-plane/line). The objective of the support vector machine algorithm is to find an N-dimensional hyperplane (where N is the number of features) that distinctly classifies the input points.

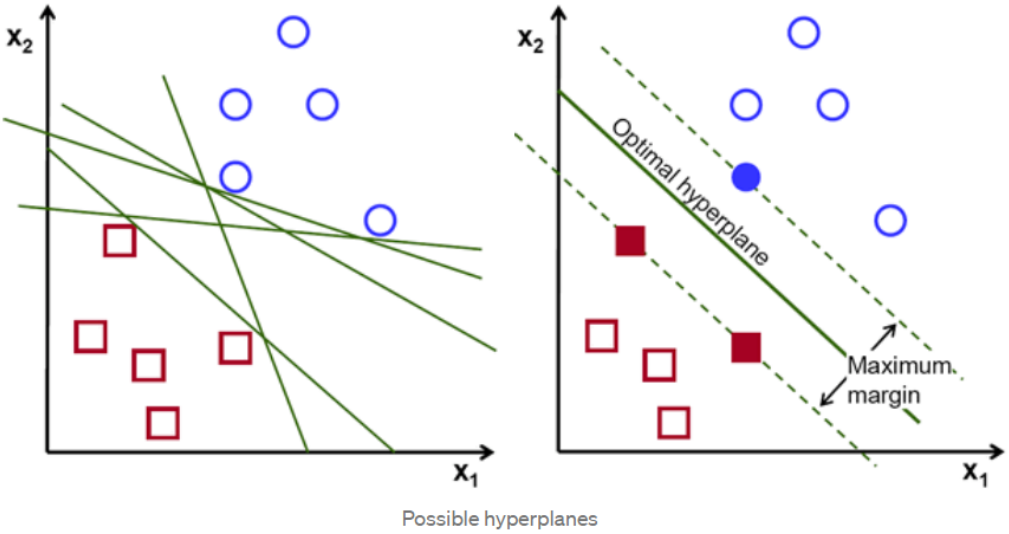


Figure 2.3 Support Vector Machine

6. Random Forest

Random forest and other supervised machine learning algorithms are frequently used in classification and regression problems. It builds decision trees from different samples, using their average for categorization and majority vote for regression. One of the most important features of the Random Forest Algorithm is its capacity to handle data sets containing both continuous variables, as in regression, and categorical variables, as in classification. It produces better results when it comes to classification problems.

The random forest algorithm is a development of the decision tree, in which a huge number of decision trees are first created using training data, and then new data is fitted inside one of the trees to form a "random forest." It merely averages your data and connects it to the closest tree on the data scale. Because they avoid the problem of the decision tree overly "forcing" data points into a category, random forest models are helpful.

2.5 Decision Tree

In decision theory, methods for assessing issues incorporating risk, uncertainty, and probabilities are collectively referred to as "decision analysis." This approach uses probability analysis to aid in choosing the appropriate corrective measures. When choosing a course of action that will ultimately have unpredictable repercussions, managers can find this procedure to be helpful. Decision trees, or alternatively the decision tree analysis, is one particular technique for decision analysis that aids management in structuring the decision problem by mapping out all practical alternative managerial actions contingent on the potential states of nature (i.e., chance events) in a hierarchical manner. They are especially useful for analyzing sequential investment choices where uncertainty is resolved at specific, discrete moments in time.

They offer a highly effective framework within which the decision-maker may clearly outline the issue, challenge each choice, and research the potential repercussions of each option. In other words, they make it feasible to properly analyze the potential effects of a choice. They help to create a balanced picture of the risks and benefits associated with each potential course of action by offering a framework to measure the worth of outcomes and the odds of obtaining them. In general, decision trees assist in making the best choices based on the facts at hand and best assumptions.

Technically speaking, a decision tree is a graphic depiction of a decision problem that is meant to aid management in making better judgments. This is a graph of decisions and their potential outcomes, which can include resource costs and hazards while developing a strategy to achieve a goal. Each node depicts either a decision made by management or a probabilistic result. To decide who gets chosen, a decision tree might be employed. It has two different types of nodes: choice nodes, which signify that management has control over the course of action and are represented by squares, and outcome nodes, which indicate that the decision maker has no influence over and are represented by circles. The outcomes of the various options are listed at the conclusion of each branch of the decision tree.

The decision-maker would pick the course of action that maximizes the projected Net Present Value after accounting for risk. The best initial decision must be identified by going backward from the end of the tree to the beginning. The management must calculate the anticipated risk-adjusted discount for each step by going backward. NPV a choice is made by averaging all of the NPV values obtained at the preceding stage with their associated probability of occurrence. Four steps are commonly involved in a tree analysis: the payoffs connected to a certain path along the tree serve as the problem's end nodes in a tree-like structure, Giving events represented on the tree subjective probabilities choosing a course of action and allocating rewards for consequences of analyses.

A decision tree is a tool for supporting decisions that employs a graph or model that resembles a tree to represent options and potential outcomes, such as chance event outcomes, resource costs, and utility costs. In operations research, decision trees are frequently used, particularly in decision analysis to find the approach most likely to succeed. Decision trees can also be used as a descriptive method for computing conditional probabilities.

2.5.1 Decision Tree (ID3)

The first of Ross Quinlan's three Decision Tree implementations, ID3 or (Iterative Dichotomize) was created. It begins with a set of objects and a description of their properties and generates a decision tree for the supplied data in a top-down manner. Materials and Information To divide the object set, one property is checked at each node of the tree using the principles of information gain maximization and entropy minimization. This procedure is repeated until the set in a particular sub-tree is homogeneous (i.e. it contains objects belonging to the same category). Utilizing a greedy search is the ID3 algorithm. It chooses a test based on the information gain criterion and never considers any other options.

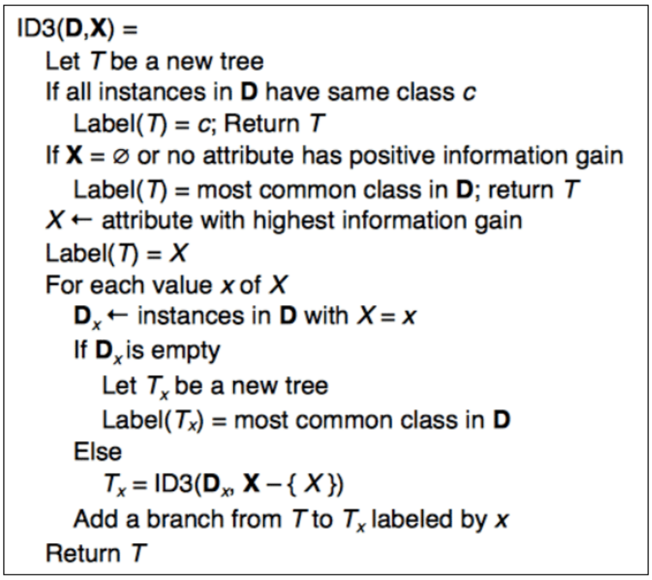
In the decision tree method, the appropriate property for each node of a created decision tree is often determined using the information gain methodology. As a result, we can choose the characteristic of the current node that has the biggest information gain (entropy reduction at the maximum level). This will result in the least amount of data being required to classify the training sample subset produced via later partitioning. Therefore, the degree of mixture of various kinds for all generated sample subsets will be minimized to a minimum when this property is used to divide the sample set included in the present node. Therefore, using an information theory technique will successfully lower the necessary number of object classification divisions.

The information gain for each and every attribute is calculated for the goal of creating a decision tree, and the attribute with the highest information gain is chosen as the root node. To show the remaining possible values, arcs are employed. The next step is to decide if each of the possible outcome scenarios belongs to the same class. Instances of the same class are denoted by a single name class; otherwise, splitting attributes are used to classify the instances.

The Concept Learning System (CLS) algorithm, which is a recursive top-down divide-and-conquer algorithm, is the foundation of ID3. Information theory is used by the ID3 family of decision tree induction algorithms to choose the attribute shared by a group of instances to split the data on next. In this manner, attributes are selected repeatedly until a comprehensive decision tree that categorizes each input is obtained.

Some of the initial occurrences can be incorrectly classified if the data is noisy. In the case of noisy data, it might be possible to prune the decision tree to lower classification errors. This learning algorithm learns rather quickly, and the decision tree classification system it produces learns fairly quickly as well.

The ID3 algorithm can handle continuous attributes by discretizing them or by examining their values directly to identify the appropriate split point by applying a threshold to the attribute values.



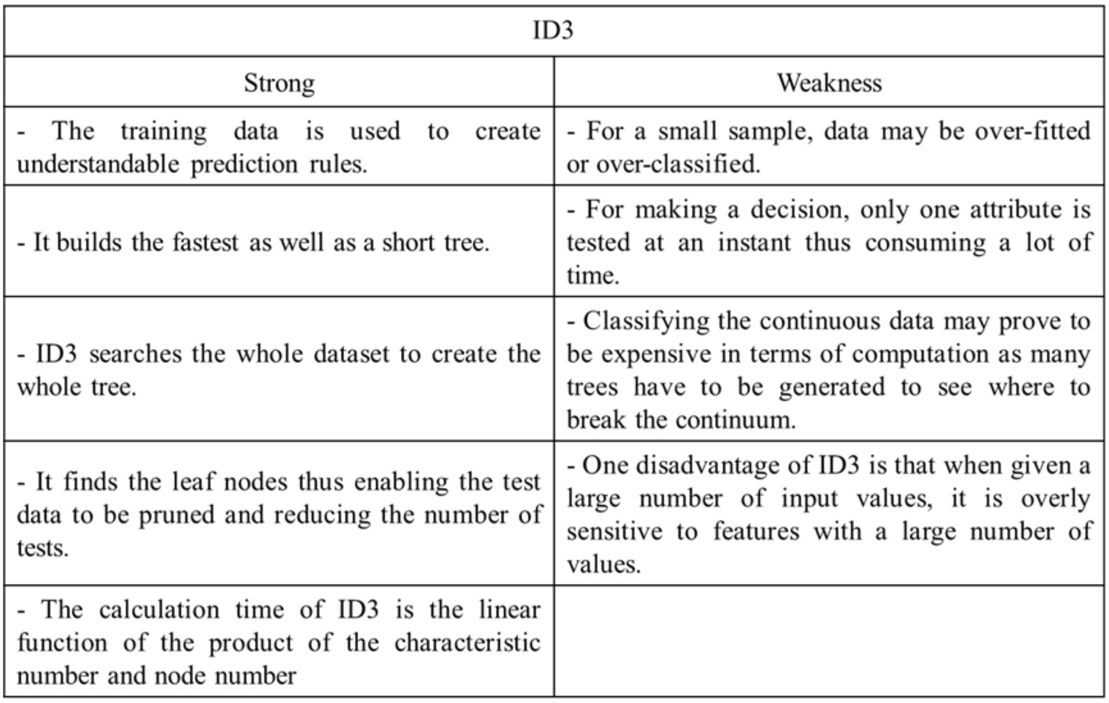
```
ID3(D,X) =
  Let T be a new tree
  If all instances in D have same class c
    Label(T) = c; Return T
  If X = ∅ or no attribute has positive information gain
    Label(T) = most common class in D; return T
  X ← attribute with highest information gain
  Label(T) = X
  For each value x of X
    D_x ← instances in D with X = x
    If D_x is empty
      Let T_x be a new tree
      Label(T_x) = most common class in D
    Else
      T_x = ID3(D_x, X – { X })
    Add a branch from T to T_x labeled by x
  Return T
```

Figure 2.4 Pseudocode of ID3 algorithm

Strong and Weakness of Decision Tree (ID3) algorithm

The decision tree (ID3) algorithm's strengths and weaknesses are displayed in the table below.

Table 2. 1 Strong and Weakness of ID3



ID3	
Strong	Weakness
- The training data is used to create understandable prediction rules.	- For a small sample, data may be over-fitted or over-classified.
- It builds the fastest as well as a short tree.	- For making a decision, only one attribute is tested at an instant thus consuming a lot of time.
- ID3 searches the whole dataset to create the whole tree.	- Classifying the continuous data may prove to be expensive in terms of computation as many trees have to be generated to see where to break the continuum.
- It finds the leaf nodes thus enabling the test data to be pruned and reducing the number of tests.	- One disadvantage of ID3 is that when given a large number of input values, it is overly sensitive to features with a large number of values.
- The calculation time of ID3 is the linear function of the product of the characteristic number and node number	

2.6 Decision Tree Pruning

Pruning is a data compression technique used in machine learning and search algorithms that reduces the size of decision trees by removing elements of the tree that are redundant and useless for categorizing occurrences. Pruning reduces the complexity of the final classifier, increasing projected accuracy by lowering overfitting. One of the problems that arises in a decision tree algorithm is the appropriate tree size. The risk of an extremely large tree is that it will overfit the training set and perform poorly on fresh data. A small tree could be unable to provide crucial structural information about the sample space.

Because it is impossible to tell whether adding just one more node can considerably reduce error, it is challenging to establish when a tree method should end. This problem is known as the horizon effect. A common strategy is to grow the tree until each node only contains a few instances, at which point nodes that do not provide any new information are deleted using pruning. A learning tree's size should be reduced through pruning while maintaining its cross-validation set-measured predictive accuracy. Different tree trimming algorithms use different measurements to maximize performance.

Pruning procedures fall into two categories (pre- and post-pruning).

1. **Pre-pruning Tree:** By substituting a stop () criterion (such as maximum Tree depth or information gain (Attr)>minGain) in the induction algorithm, Pre-pruning approaches preclude a full induction of the training set. Pre-pruning techniques are said to be more effective since they do not cause a full set; rather, the trees are left tiny from the beginning. The horizon effect is a typical issue with Pre-pruning techniques. This is to be viewed as the stop () criterion's unwanted early termination of the induction.
2. **Post-pruning Tree:** Post-pruning is the technique used most frequently to simplify trees (or just pruning). To make things easier, leaves are employed in this layout in place of nodes and subtrees. Pruning can significantly reduce the size of invisible entities in addition to helping classify them more properly. The accuracy of the assignment on the train set could deteriorate even while the classification qualities of the tree are generally categorized more accurately.

2.7 Post-Pruning Tree

This method is applied following the creation of a decision tree. When a decision tree's depth is expected to be extremely high and the model overfits, this strategy is employed. Another name for it is reverse pruning. When a decision tree has expanded indefinitely, this method is used.

2.7.1 Bayes Minimum Risk

The Bayes Risk decision rule, which employs Bayes to reduce expectations of loss or risk, is crucial because costs associated with its implementation correspond to misclassification errors.

The cost of classifying the data with attribute x into class Ci, where the correct class is C j, where j=1, 2,..., m, and the risk is equal to that cost. Equation: When x is placed in class Ci, conditional risk is defined (1).

$$\boldsymbol{l^i(x) = \sum_{j=1}^{2} \lambda_{j,i} \, \mathrm{Pr}(C_j|x)}$$

2.3

where; $\boldsymbol{\lambda_{j,i}}$ = cost of classifying the data into class $\boldsymbol{C_j}$,

$\boldsymbol{C_j}$ = true class

$\mathbf{Pr}\ (\boldsymbol{C_j}|\boldsymbol{x})$= probability of a subject with attribute $\boldsymbol{x}$ predicted in class $\boldsymbol{C_i}$

Pr$(Cj\ |\boldsymbol{x})$ is calculated using Bayes’ Theorem given in equation (2.4).

$$\mathbf{Pr}(\boldsymbol{Cj}\ |\boldsymbol{x}) = \frac{\mathbf{Pr(x \cap Cj)}}{\mathbf{Pr(x)}} = \frac{\mathbf{Pr(x|Cj).Pr(Cj)}}{\mathbf{Pr(x)}} = \frac{\mathbf{Pr(x|Cj).Pr(Cj)}}{\sum_{l=1}^{m}\mathbf{Pr(x|C}_l\mathbf{).Pr(C}_l\mathbf{)}}$$

2.4

One special case of risk matrix is zero-one-loss which has the same cost when misclassifying (classifying a subject with the i class as a j class or vice versa) as in equation (2.5).

$$\boldsymbol{\lambda_{j,i}} = \begin{cases} \mathbf{1}, \boldsymbol{if\ i \neq j} \\ \mathbf{0},\ \boldsymbol{if\ i = j} \end{cases}$$

2.5

A post-pruning method is run by assessing the risk of each subtree using Bayes Risk, starting at the bottom (leaf node) and working its way up (root node). Equation shows the loss risk associated with each parent node t based on a zero-one basis (2.6).

$$\boldsymbol{R_t^i}(\boldsymbol{x}) = \sum_{\boldsymbol{j=1,j\neq i}}^{\mathbf{2}} \boldsymbol{\lambda_{j,i}}\,\mathbf{Pr}(\boldsymbol{C_j}|\boldsymbol{x})$$

2.6

where: R_t^i $(\boldsymbol{x})$ = risk associated with node t when classifying subject with attribute $\boldsymbol{x}$ into class $\boldsymbol{C_i}$

Pr $(C_j\ |\boldsymbol{x})$ = probability of a subject with attribute $\boldsymbol{x}$ predicted in class C_j

The risk associated with the leaf node of its parent node t is shown in equation (2.7).

$$\boldsymbol{R_l} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{tl}} \boldsymbol{R_l^i}(\boldsymbol{x})$$

2.7

where: R_l^i $(\boldsymbol{x})$ = risk associated with leaf node l when classifying subject with attribute $\boldsymbol{x}$ into class $\boldsymbol{C_i}$

tl = total leaf nodes in the subtree

2.8 Data Pre-processing

Unprocessed data must be transformed into a format that a machine learning model can use. This is known as data preparation. This is the first and most important stage in developing a machine learning model. The data is not always cleaned and ready when working on a machine learning project. Furthermore, data must always be organized and cleaned up before being used in any task.

Since real-world data frequently includes noise, missing values, and can be in an unattractive format, it is challenging to directly create machine learning models utilizing this type of data. Data preprocessing, which is necessary to clean the data and get it ready for the model, improves the accuracy and efficacy of a machine learning model. Finding missing values, label encoding, dividing a dataset into training and test sets, and feature selection are examples of data preparation procedures.

2.8.1 Handling Missing Data

The next phase of data preparation involves dealing with missing data in the datasets. If the dataset has some missing data, the machine learning model could have a very difficult time. The dataset consequently has missing values that need to be dealt with. The following are the two main methods for handling missing data:

1. By removing the specific row: The first approach is frequently used to handle null data. By doing this, the specific empty row or column is simply removed. However, since removing data could result in information loss and incorrect output, this tactic is worthless. By removing the specific row: The first approach is often used to handle null data. In this manner, the particular row or column with no data is removed. However, since removing data could result in information loss and incorrect output, this tactic is worthless.

2. By finding the mean: By doing this, find the mean of every column or row that contains a missing value and replace it with that value. For parameters like age, salary, year, and others that have a numerical component, this strategy works effectively.

2.8.2 Splitting dataset into training and test set

During the preprocessing phase of machine learning, the dataset is divided into a training set and test set. One of the most crucial data pretreatment techniques since it can increase how well a machine learning model works.

2.8.2.1 Training Data

The training dataset is used to create or fit the machine learning model and is the size-wise largest subset of the initial dataset. Before the ML algorithms can learn how to make predictions for the given task, training data is necessary. Whether supervised learning or unsupervised learning techniques are used, the training data varies. Since the inputs are not labeled with the appropriate outputs in unsupervised learning, the training data contains unlabeled data points.

Models must draw patterns from the available training datasets in order to create predictions. For supervised learning, in contrast, labels are included in the training data to aid in the model's training and prediction. The kind of training data that are provided to the model determines its precision and propensity for prediction in a significant way. It indicates that the model will function more effectively the better the training set of data is. More than or equal to 60% of the total data in a typical ML project is made up of training data.

2.8.2.2 Testing Data

It is time to test the model on the test dataset after it has been trained on the training dataset. This dataset assesses the model's performance and offers confidence in its ability to generalize effectively to new or untested datasets. The test dataset includes a different subset of the original data than the training dataset. Given that it has some comparable traits and a comparable class probability distribution, it acts as a benchmark once model training is over.

A clean dataset known as test data contains information for each kind of event the model might encounter in the actual world. A typical ML project's test dataset comprises 20–25% of the total original data. The model's testing and training correctness, or how accurate it is when used with various datasets, can now be verified and compared. If the model's accuracy on training data is better than its accuracy on testing data, it is said to have overfitted. The first dataset must include the test results, and it must be sufficiently large to allow for precise prediction.

2.9 Exploratory Data Analysis

Exploratory data analysis is a crucial component of a data analyst's or scientist's everyday routine (EDA). It enables a comprehensive analysis of the dataset, the formulation or rejection of hypotheses, and the creation of prediction models with a strong foundation. The relationship between the various factors and how they can impact a corporation are described and understood using a variety of statistical tools and data manipulation approaches.

Any project involving data analysis or data science should begin with exploratory data analysis, or EDA. Exploratory data analysis is a crucial technique for conducting preliminary research on data in order to identify patterns, identify anomalies, test hypotheses, and triple-check presumptions using summary statistics and graphical representations.

Exploratory data analysis (EDA) is the process of analyzing a dataset to search for patterns and irregularities (outliers) and formulating hypotheses based on comprehending the dataset. To make the numerical data in the dataset easier to grasp, EDA involves producing summary statistics for each number and producing several graphical representations.

2.10 Employee Attrition

Employee attrition in a business is the term used to describe the loss of staff through usual means such as retirement and resignation, clients dying of old age, or layoffs brought on by a change in the organization's target demographics because it has a significant impact on an organization, a high rate of staff attrition is severed issue. Employees leave a company with invaluable tacit knowledge that is frequently the source of the competitive advantage of the organization [10].

Employee turnover places a financial burden on the organization in the form of hiring and training expenses as well as disruption to operations. On the other hand, greater retention reduces the expense of employing new employees and putting them through training, as well as enables the progressive influx of more seasoned workers into the workforce. Today's organizations have a strong commercial interest in comprehending the reasons behind staff attrition in order to reduce employee turnover. Consequently, in order to enhance its human resource strategy, a business must focus on anticipating employee turnover and identifying the primary causes of attrition [11].

Employee attrition can occur for a number of reasons, such as a lower wage, a lack of job satisfaction, personal factors, or environmental concerns if an employee is fired for whatever reason. It's referred to as involuntary attrition (Kaur & Vijay, 2016). Conversely, voluntary attrition is referred to as the person who leaves their job with them by their side. If the employee is talented, this type of attrition is a loss for the organization. Everyone in the current situation wants a bigger pay and job security. As a result, when presented with a superior opportunity elsewhere, employees quit their positions right away.

In the modern era of computer science, machine learning techniques are crucial for predicting staff attrition. Based on previous data about the employee, such as age, experience, education, last promotion, and so on, these approaches offer forecasts. The HR department has prior knowledge about employee turnover based on the prediction results. The HR department has also planned on hiring replacement workers for the worker who intends to leave in the near future.

CHAPTER 3
DESIGN OF THE PROPOSED SYSTEM

In this chapter, the design of the proposed system and the methodology will be explained in detail. The main goal of this thesis is to analyze and compare the performance of Decision Tree before pruning and after pruning Techniques with IBM Human Resource Analytic Employee Attrition and Performance dataset from Kaggle site. The experiments for the study are carried out using the Python programming language.

3.1 Overview of the Proposed System

Maintaining a single employee costs a company five to 10 times more than hiring a new one. Predictive algorithms can precisely identify potential employees in the near future in order to provide a retention solution. A unique prediction model based on Data Mining (DM) and Machine Learning techniques is presented by this proposed system.

Data preparation, EDA analysis, feature selection and engineering, model building, and model evaluation are the six phases that make up the proposed model. The used data set has 35 attributes and 1470 instances. 1029 examples are utilized as the testing set, and 441 instances are used to train the model.

The system flow of the proposed system is depicted in Figure 3.1. The system uses 70% of dataset as training data and 30% of dataset as testing data. The main goal is to assess how well two data mining methods—decision tree and Bayes minimum risk—perform in predicting employee attrition. Getting the data from the IBM Human Resource Analytic Employee Attrition and Performance dataset is the first and most important stage. The dataset typically includes a wide variety of errors and noisy data. The pre-processing phase has been completed, and the data has been cleansed to make them usable. The output of the data pre-processing is noise-free data that may be used for further processing stages.

3.1.1 Software Requirement

Jupyter Notebook: The Jupyter Notebook App is a server-client program that enables web browser-based editing and running of notebook papers. The Jupyter Notebook App can be used offline, without an internet connection, on a local desktop, as detailed in this paper, or it can be deployed on a remote server and accessed online. The Jupyter Notebook App features a "Dashboard" (Notebook Dashboard), a "control panel" that displays local files and enables opening and closing of notebook papers as well as displaying, editing, and executing notebook documents.

Jupyter Notebook makes it simple to display your intended audience the complete project's process by allowing users to collect all components of a data project in one location. Users can build data visualizations and other parts of a project using the web-based application, which they can then share with others on the platform.

Python Programming Language: The programming language Python is high-level and versatile. Code readability is a priority in its design philosophy, which uses substantial indentation. Both Python's types and trash collection are dynamic. Structured, object-oriented, and functional programming are just a few of the programming paradigms it supports. Python is frequently used for creating websites and applications, automating repetitive tasks, and analyzing and displaying data. Python has been used by many non-programmers, including accountants and scientists, for a variety of routine activities including managing finances since it is very simple to learn.

Tkinter GUI: Tkinter is the name of the Python binding for the Tk GUI toolkit. It is the official Python interface to the Tk GUI toolkit and acts as the de facto default GUI for Python. Tkinter is preinstalled on all Linux, Windows, and macOS installs of Python. The word Tkinter comes from the Tkinter interface. Python provides a variety of choices for GUI development (Graphical User Interface). Tkinter is the approach used the most frequently among all GUI approaches. It is a typical Python interface for the Python-supplied Tk GUI toolkit. The fastest and simplest approach to construct GUI apps is with Python and Tkinter. It's simple to build a GUI using Tkinter.

3.1.2 System Flow Diagram

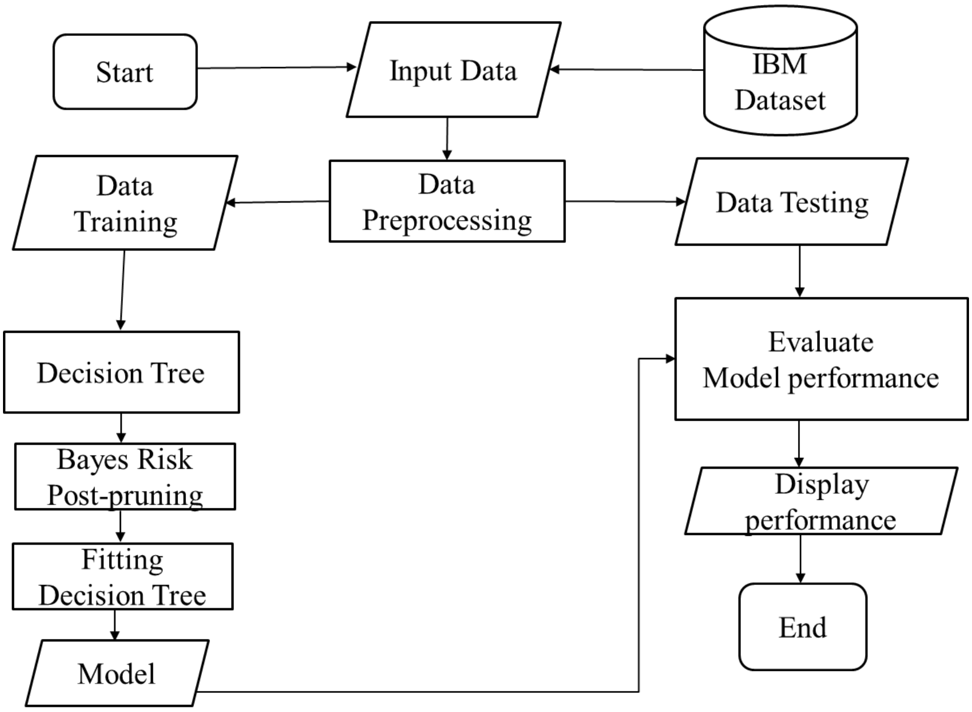


Figure 3.1 Flow Chart of the Proposed System

Set input data from IBM Human Resource Analytic Employee Attrition and Performance dataset. The dataset is available from Kaggle Dataset Repository. The next step is preprocessing part. The input data are divided into two parts as Data Training and Data Testing. ID3 Algorithm is applied to build decision tree on the training data. After getting the decision tree, some branches of the decision tree may contain noise or outliers. So, the system uses Bayes Risk Post-Pruning method to remove unnecessary branches or nodes. Then the proposed system gets fitting decision tree. Testing dataset is used for evaluating model. After generating the prediction model, the system will evaluate the model performance. Finally, the system will display the model performance.

3.2 Dataset Information

The information about the attribute of an employee is provided details in the following Table 3.1.

Table 3.1 Description of Attributes

No.	Feature Name	Type of Data	Data Description
1	Age	Continuous	The age of individual employee
2	Attrition	Categorical	Employee leaving the company (Yes, No)
3	Business Travel	Categorical	Business travel frequency (No Travel, Travel Frequently, Travel Rarely)
4	DailyRate	Continuous	Salary Level
5	Department	Nominal	Employee department (HR, R&D, Sales)
6	DistanceFromHome	Continuous	The distance from work to home
7	Education	Categorical	Level of education attained (1 = ‘Below Collage’, 2 = ‘College’, 3 = ‘Bachelor’, 4 = ‘Master’, 5 = ‘Doctor’)
8	EducationField	Nominal	Field of education (HR, Life Sciences, Marketing, Medical Sciences, Others, Technical)
9	EmployeeCount	Continuous	Count of instance
10	EmployeeNumber	Continuous	Employee ID
11	EnvironmentSatisfaction	Categorical	Employee satisfaction with the environment (1 = ‘Low’, 2 = ‘Medium’, 3 = ‘High’, 4 = ‘Very High’)
12	Gender	Categorical	Female, Male
13	HourlyRate	Continuous	Hourly Salary
14	JobInvolvement	Categorical	Job Involvement (1 = ‘Low’, 2 = ‘Medium’, 3 = ‘High’, 4 = ‘Very High’)
15	Job Level	Categorical	Level of Job (1 to 5)
16	JobRole	Categorical	(1=Hc Rep, 2=Hr, 3=Lab Technician, 4=Manager, 5=ManaginDirector, 6=ResearchDirector, 7=Research Scientist,8=SalesExecutive, 9=Sales Representative)
17	JobSatisfaction	Categorical	Satisfaction with the job (1 = ‘Low’, 2 = ‘Medium’, 3 = ‘High’, 4 = ‘Very High’)
18	MaritalStatus	Categorical	(1=Divorced,2=Married, 3=Single)

19	MonthlyIncome	Continuous	Monthly Salary
20	MonthlyRate	Continuous	Monthly Rate
21	NumCompaniesWorked	Continuous	No. of Companies Worked At
22	Over18	Categorical	(1=Yes, 2=No)
23	OverTime	Categorical	(1=Yes, 2=No)
24	PercentSalaryHike	Continuous	Percentage Increase in salary
25	PerformanceRating	Categorical	Performance Rating
26	RelationshipSatisfaction	Categorical	Relations Satisfaction (1 = ‘Low’, 2 = ‘Medium’, 3 = ‘High’, 4 = ‘Very High’)
27	StandardHours	Continuous	Standard Hours
28	StockOptionLevel	Categorical	Stock Options
29	TotalWorkingYears	Continuous	Total Years Worked
30	TrainingTimesLastYear	Continuous	Hours Spent Training
31	WorkLifeBalance	Categorical	Time Spent Between Work and Outside (1 ‘Bad’, 2 ‘Good’, 3 ‘Better’, 4 ‘Best’)
32	YearsAtCompany	Continuous	Total Number of Years at the company
33	YearsInCurrentRole	Continuous	Years in Current Role
34	YearsSinceLastPromotion	Continuous	Last Promotion
35	YearWithCurrManger	Continuous	Years Spent with Current Manager

3.3 Collection of Data

The IBM Human Resource Analytic Employee Attrition and Performance dataset is used in the model to compare the performance of two machine learning algorithms: Decision Tree (ID3) and Bayes Minimum Risk pruning. Employee satisfaction, salary, seniority, and demographic information are the four main components of the dataset. 35 attributes and 1470 cases are included in the dataset. With 237 instances of "Yes" and 1233 cases of "No," the identified class is designated as "Attrition." The dataset is divided into two portions, referred to as the training dataset 70% and testing dataset 30% respectively.

3.4 EDA Analysis

The experiment design is followed by the EDA Analysis in the proposed system. Python programming is used to carry out the experiments of the proposed system. Data analysis utilizing visual methods is called exploratory data analysis (EDA). With the use of statistical summaries and graphical representations, it is used to identify trends, patterns, or to verify assumptions. The EDA analysis is processed in this proposed system, and each stage is discussed in more detail below.

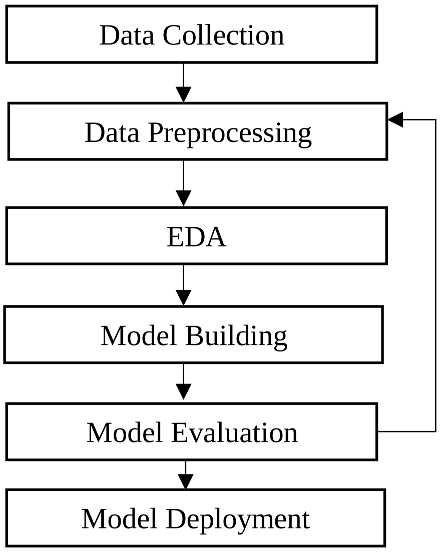


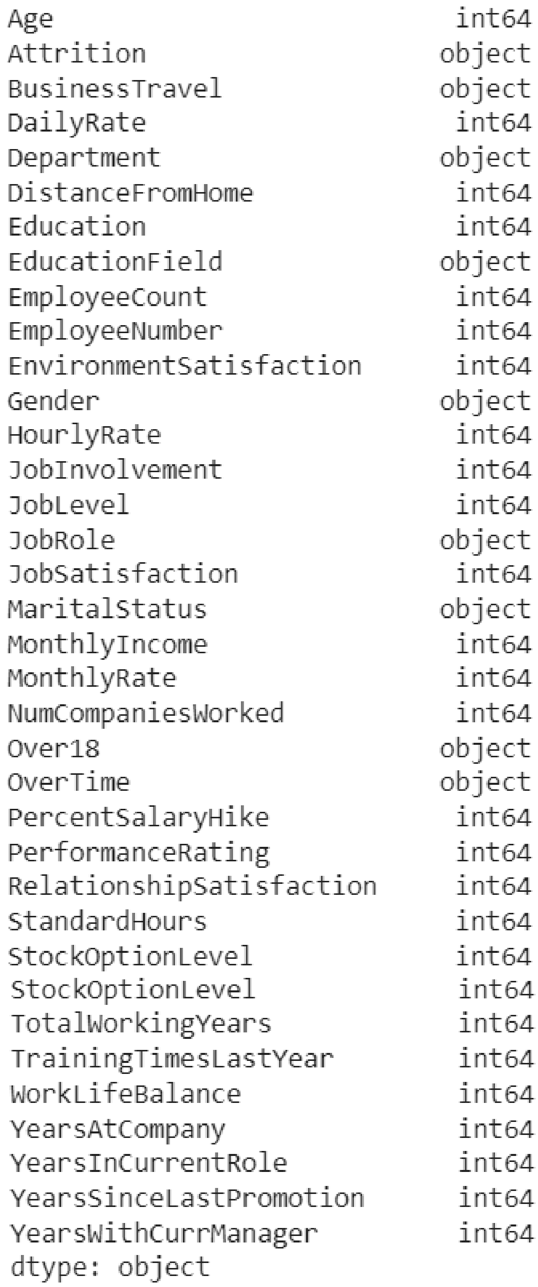
Figure 3.2 EDA Analysis

3.5 Pre-processing and Exploratory Data Analysis

In machine learning, data preprocessing is used to transform the dataset's raw data into clean data. Data preprocessing entails filling in missing data with median values, deleting unnecessary aspects from the dataset (such as Customer ID), and identifying the key features needed to train a high-performing model. In the proposed system, data cleaning, data transformation, and data normalization are performed in this stage.

After that, the exploratory data analysis (EDA) step is continued to process. Exploratory Data Analysis (EDA) is a technique for analyzing the dataset in order to draw conclusions from its key properties. The dataset includes a predictive feature (Attrition or No Attrition), numerical features, and categorical features. It is really difficult to glance at a large spreadsheet and identify key details about the data. Therefore, EDA is crucial in certain situations. The purpose of conducting EDA is to make sense of the data and help to be better understand each feature before feeding them to machine learning models, thereby, making the modelling more efficient.

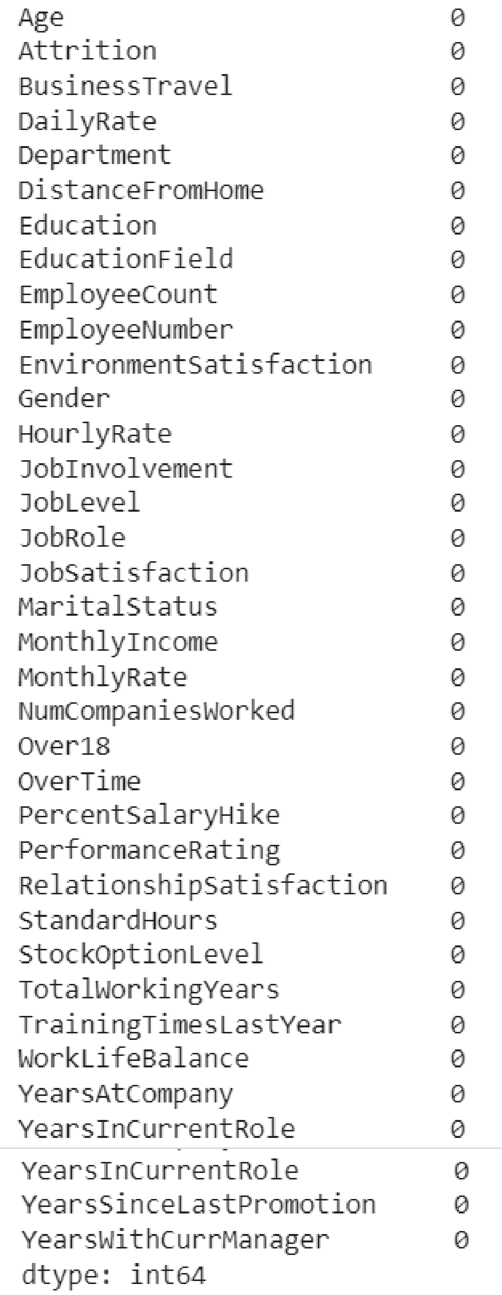
At the beginning of EDA, “pandas.DataFrame.info” method is used to get information about the data. This method outputs a brief summary of the data frame that includes the names of the columns and their data types, the number of non-null values, and the memory consumption.



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

Figure 3.3 Summary of the Data Frame

The data set has 35 columns and 1470 observations, as seen above. The data collection does not appear to include any null values, yet the column. There is no missing values in this 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
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

Figure 3.4 Checking Missing Values

These observations are eliminated from the dataset because they seemed to be conflicting. Then EDA Analysis is carried out to get the domain knowledge and to understand the data in employee attrition dataset.

3.5.1 Data Visualization

Data visualization is the process of displaying data using popular images like infographics, charts, and even animations. These informational visual representations convey intricate data relationships and data-driven insights in a way that is simple to comprehend. The proposed system includes visualization techniques, such as: tables, pie charts and bar charts, heat maps and tree maps.

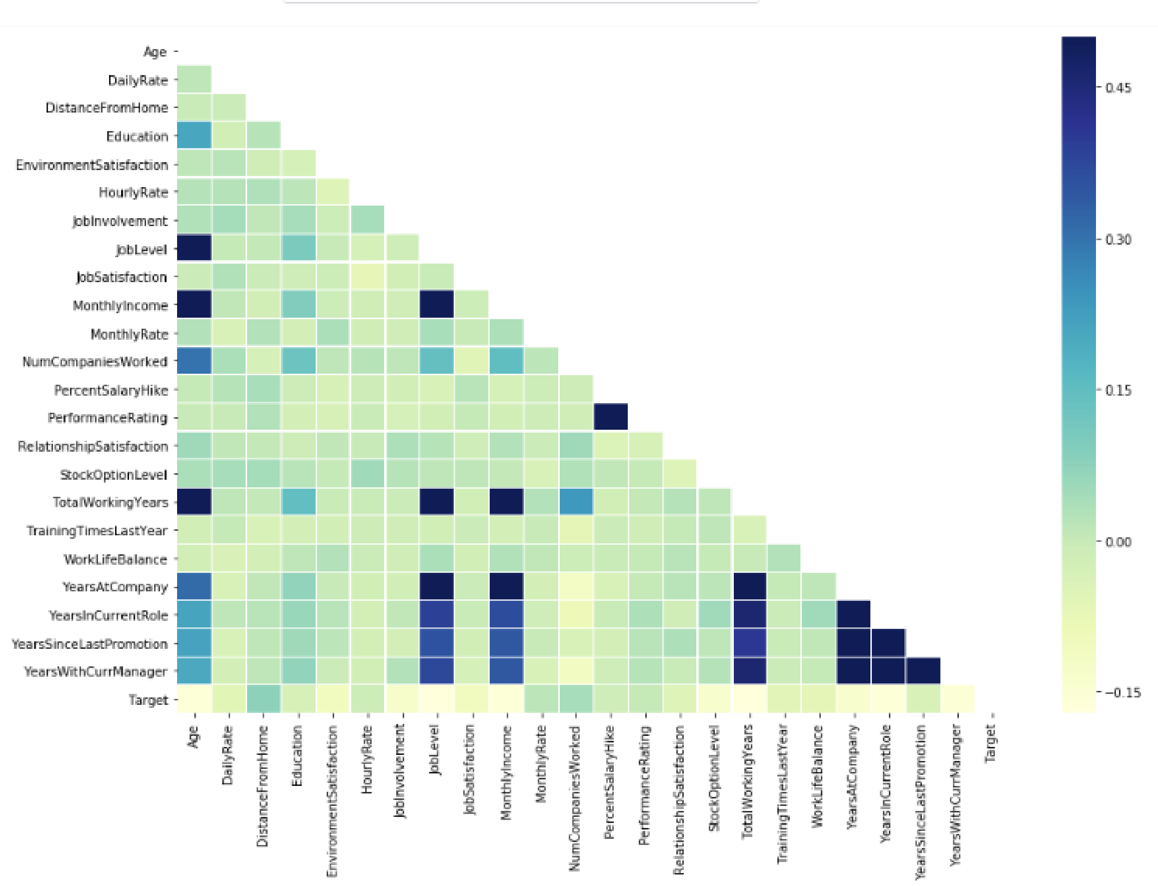


Figure 3.5 Heatmap Table of the Proposed System

3.5.2 Data cleaning and reduction

The dataset has 35 properties, which makes it high dimensional. It is best to eliminate any extraneous characteristics that do not advance the study's goals. The characteristics "EmployeeCount," "StandardHours," and "Over18" can be removed based on the data quality report and data visualization since their cardinality/distinction is "1," which means that they have the same values across the data. Aside from that, "EmployeeNumber" can be removed from the dataset because it is determined to be unhelpful for the modeling and prediction process.

3.5.3 Data Normalization

The process of extracting features from the data and converting them into a format appropriate for the machine learning model is known as "data normalization". Before training the model, all categorical features in the dataset are converted into numerical labels because the majority of machine learning techniques demand them. The discretization process and changing the attribute type from numerical to nominal were both carried out as part of the data cleaning and reduction process. Based on the findings above, four (4) features were eliminated, leaving the remaining 30 attributes.

3.6 Splitting the data in training and testing sets

Splitting the data into two groups, known as the training and testing sets, is the first stage in creating a model. The machine learning algorithm creates the model using the training set. The test set is used to assess the performance of the model and contains samples that are not included in the learning process. To ensure an impartial assessment, it is crucial to evaluate the model's quality using unobserved data.

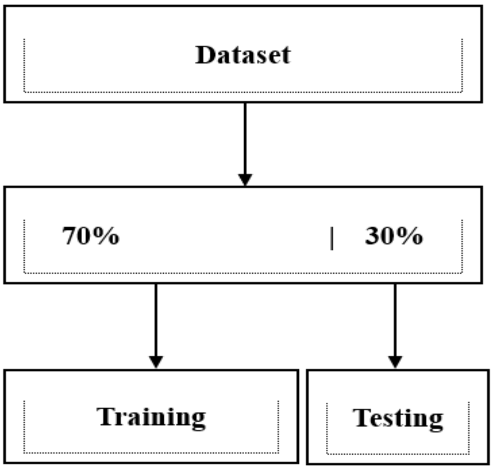


Figure 3.6 Data Splitting

3.7 Methodology of the Proposed System

In this stage, the preprocessed data is utilized to create the machine learning model to forecast employee attrition. For building accurate and comprehensible attrition prediction models, the methods used in this proposed system are Decision Tree (ID3) and Bayes Minimum Risk Pruning. The goal of these method applied in this system is to evaluate and contrast the effectiveness of Decision Tree and Bayes Risk Post-Pruning

Techniques in Predicting Employee Prediction dataset. This proposed system uses a dataset as input and builds prediction model using it. The input dataset for a classification task is typically split into train and test datasets. With the training dataset is used to develop the prediction model and the test dataset is used to evaluate the model's performance using evaluation metrics.

CHAPTER 4
IMPLEMENTATION OF THE PROPOSED SYSTEM

This chapter represents a step-by-step comparison of employee attrition prediction system. The effectiveness of two models is assessed by comparison in two different contexts. The supervised machine learning models are evaluated for their prediction abilities in this chapter. The accuracy, precision, recall, and F1-score of each model are evaluated in these comparisons. The analysis results of these comparisons are shown by figures, and these comparisons indicate that how the performance vary.

4.1 Performance Evaluation

Evaluation of attrition prediction model performance to ascertain the model's generalizability is a crucial step in the attrition prediction process. Some of the common evaluation metrics utilized by various academics for employee attrition prediction evaluation include accuracy [15], precision and recall, and F-Measure.

In this suggested system, the performance of the models is assessed using a variety of metrics. After each model's confusion matrix has been constructed, accuracy, precision, recall, and specificity are calculated [23]. The number of samples that are true positive (TP), true negative (TN), false positive (FP), and false negative samples is calculated (FN). A typical method for evaluating a classification model's efficacy is the confusion matrix. The actual and projected classifications of a classifier are displayed using a confusion matrix.

True positives (TP): the number of employees who actually left and the classification model has identified them correctly as attrition.

True negatives (TN): the number of employees who do not actually left and the classification model has identified them correctly as no attrition.

False positives (FP): the number of employees who do not left but the classification model incorrectly determined them as attrition.

False negatives (FN): the number of employees who left but the classification model incorrectly determined them as no attrition.

Table 4.1 A confusion matrix for a binary classifier

	Predicted Classes		
Actual Classes	Class=Yes /+/ Attrition	Class=Yes /+/ Attrition TP (true positive)	Class=No/-/No-Attrition FN (false negative)
	Class=No/-/No-Attrition	FP (false positive)	TN (true negative)

Accuracy: the portion of the total number of correctly predicted cases, is calculated as follows:

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

4.1

Precision: the fraction of predicted attrition that do leave, is calculated as follow:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}}$$

4.2

Recall: the fraction of real attrition which are correctly determined as leave is calculated as follow:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP +FN}}$$

4.3

F-Measure: can be considered a weighted average of precision and recall, with the best value being 1 and the poorest being 0. Precision and memory make equal relative contributions to the F1 score. It is calculated as follow:

$$\textbf{F-measure} = \frac{\textbf{2 * Precision * Recall}}{\textbf{Precision + Recall}}$$

4.4

4.2 Experimental Setup

This chapter's objective is to represent implementation of the proposed system, design, and performance assessment. High-value employees who are expected to leave soon can be kept by using the employee attrition prediction system. Python is the programming language used to implement this proposed system.

4.3 Implementation of the System

The “Welcome” page of the proposed system is shown in Figure 4.1. Firstly, “Welcome” button is clicked.

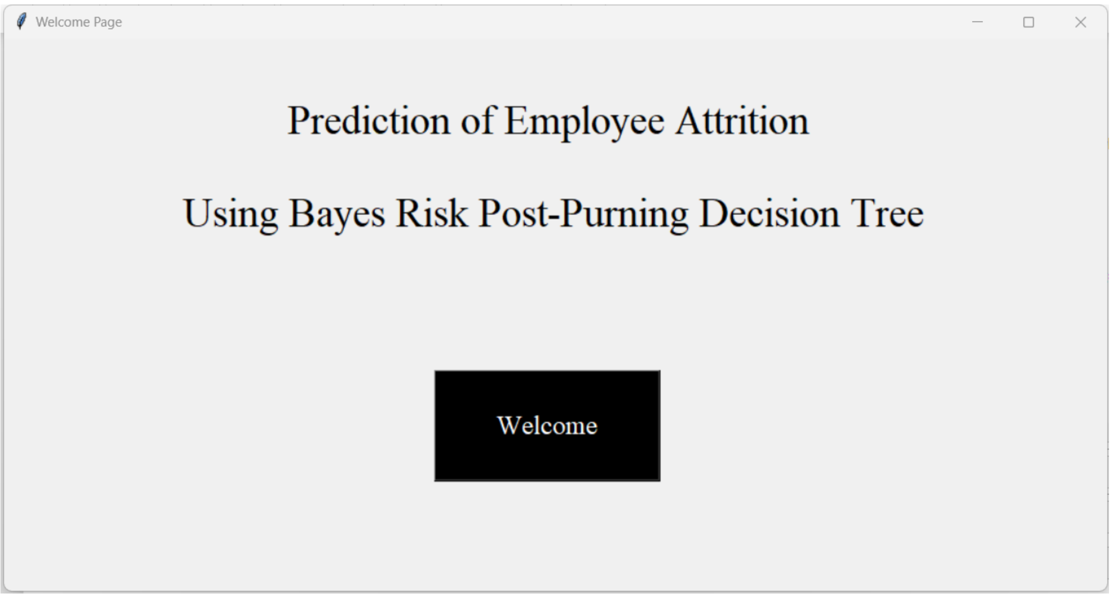

Welcome Page
Prediction of Employee Attrition
Using Bayes Risk Post-Purning Decision Tree
Welcome

Figure 4.1 Welcome Page of the Proposed System

When the "Welcome" button is clicked, “Home” page of this system, as depicted in Figure 4.2, is displayed. This page includes “Accuracy Results” button of the proposed system, “Upload Data Set”, “Decision Tree Before Pruning”, and “Decision Tree After Pruning” buttons.

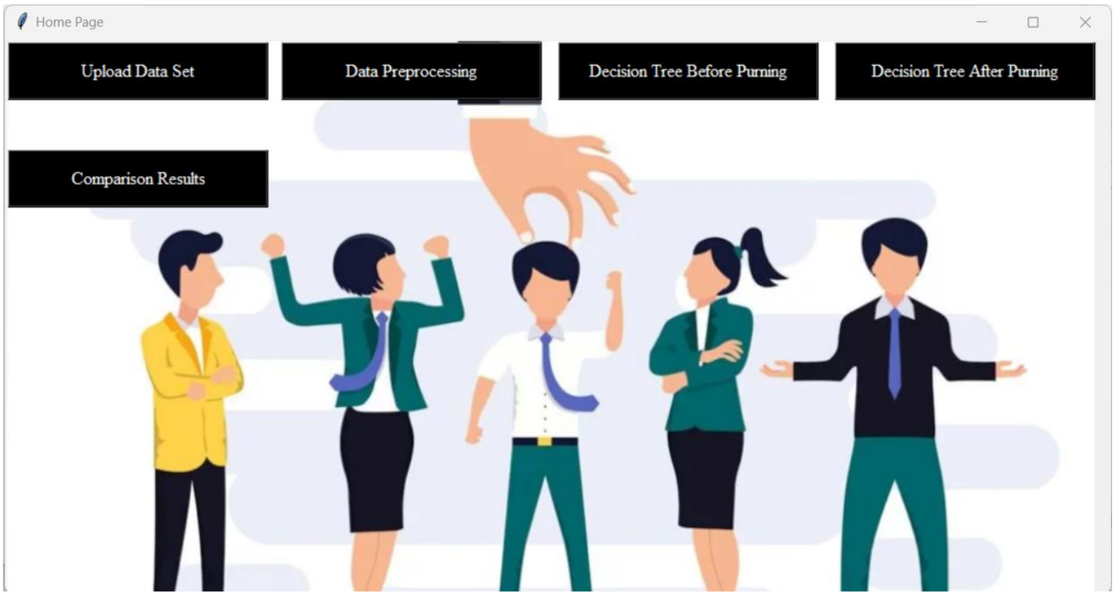
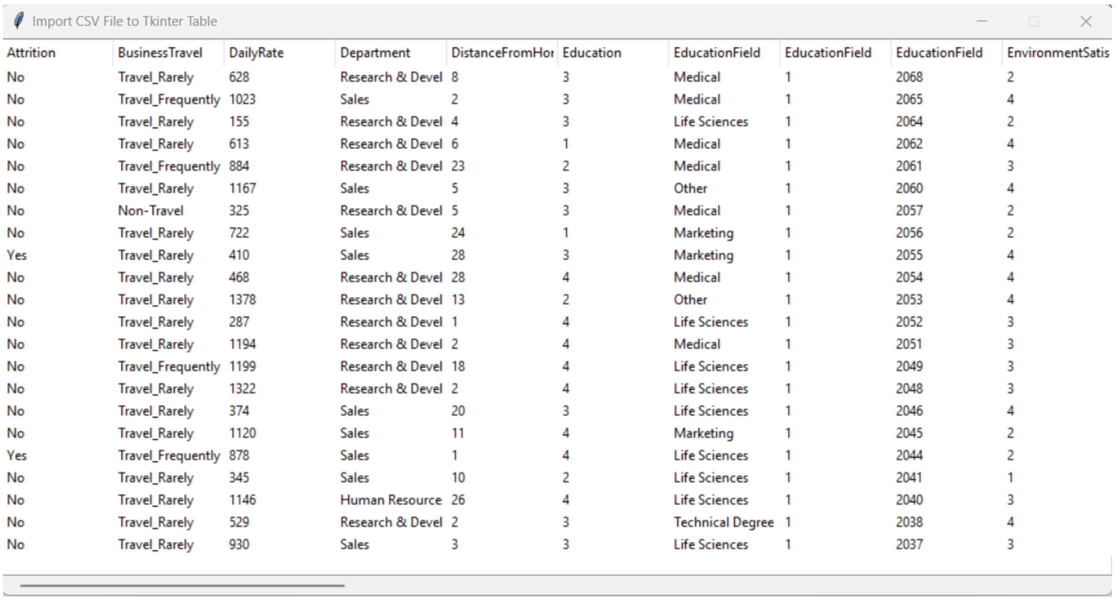


Figure 4.2 Home Page of the Proposed System

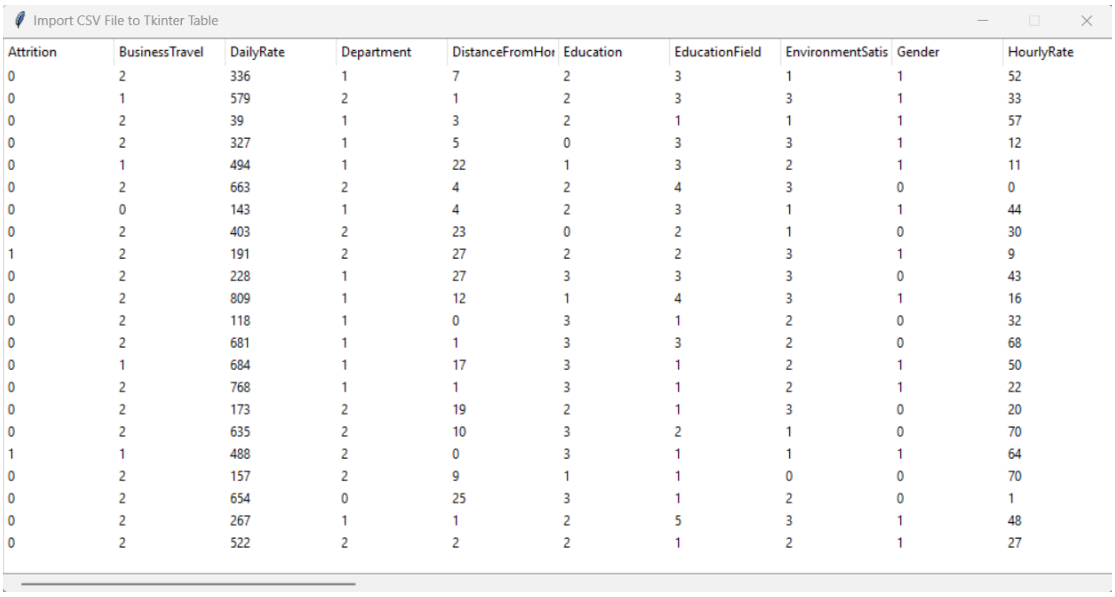
The selected feature dataset that is used in this proposed system can be viewed by clicking the "Upload Data Set" button. The result is shown in Figure 4.3.



Import CSV File to Tkinter Table

Attrition	BusinessTravel	DailyRate	Department	DistanceFromHor	Education	EducationField	EducationField	EducationField	EnvironmentSatis
No	Travel_Rarely	628	Research & Devel	8	3	Medical	1	2068	2
No	Travel_Frequently	1023	Sales	2	3	Medical	1	2065	4
No	Travel_Rarely	155	Research & Devel	4	3	Life Sciences	1	2064	2
No	Travel_Rarely	613	Research & Devel	6	1	Medical	1	2062	4
No	Travel_Frequently	884	Research & Devel	23	2	Medical	1	2061	3
No	Travel_Rarely	1167	Sales	5	3	Other	1	2060	4
No	Non-Travel	325	Research & Devel	5	3	Medical	1	2057	2
No	Travel_Rarely	722	Sales	24	1	Marketing	1	2056	2
Yes	Travel_Rarely	410	Sales	28	3	Marketing	1	2055	4
No	Travel_Rarely	468	Research & Devel	28	4	Medical	1	2054	4
No	Travel_Rarely	1378	Research & Devel	13	2	Other	1	2053	4
No	Travel_Rarely	287	Research & Devel	1	4	Life Sciences	1	2052	3
No	Travel_Rarely	1194	Research & Devel	2	4	Medical	1	2051	3
No	Travel_Frequently	1199	Research & Devel	18	4	Life Sciences	1	2049	3
No	Travel_Rarely	1322	Research & Devel	2	4	Life Sciences	1	2048	3
No	Travel_Rarely	374	Sales	20	3	Life Sciences	1	2046	4
No	Travel_Rarely	1120	Sales	11	4	Marketing	1	2045	2
Yes	Travel_Frequently	878	Sales	1	4	Life Sciences	1	2044	2
No	Travel_Rarely	345	Sales	10	2	Life Sciences	1	2041	1
No	Travel_Rarely	1146	Human Resource	26	4	Life Sciences	1	2040	3
No	Travel_Rarely	529	Research & Devel	2	3	Technical Degree	1	2038	4
No	Travel_Rarely	930	Sales	3	3	Life Sciences	1	2037	3

Figure 4.3 Dataset of Proposed System

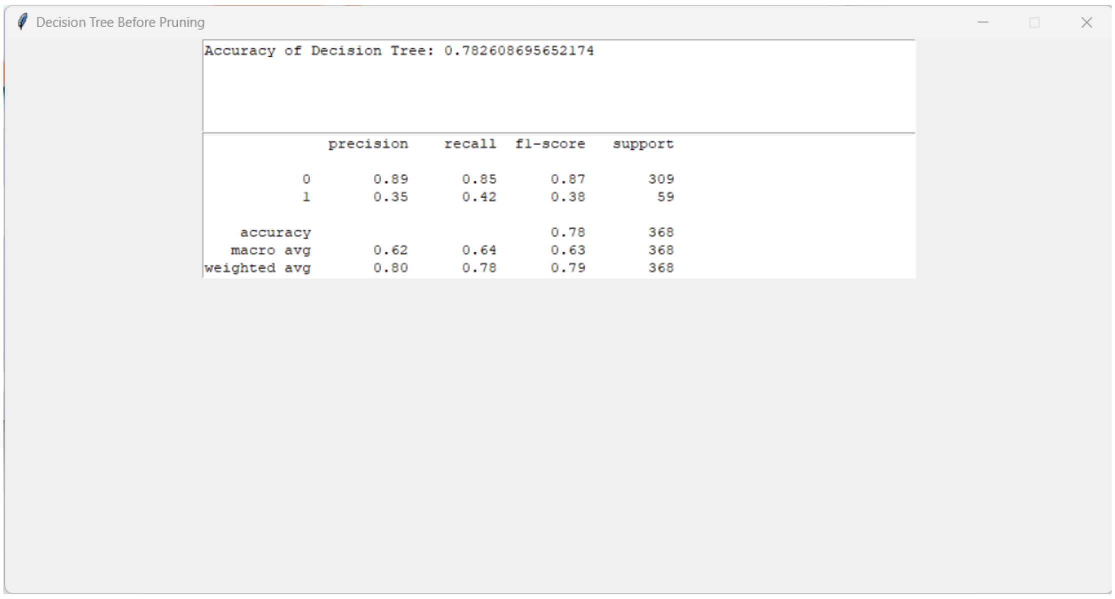


Import CSV File to Tkinter Table

Attrition	BusinessTravel	DailyRate	Department	DistanceFromHor	Education	EducationField	EnvironmentSatis	Gender	HourlyRate
0	2	336	1	7	2	3	1	1	52
0	1	579	2	1	2	3	3	1	33
0	2	39	1	3	2	1	1	1	57
0	2	327	1	5	0	3	3	1	12
0	1	494	1	22	1	3	2	1	11
0	2	663	2	4	2	4	3	0	0
0	0	143	1	4	2	3	1	1	44
0	2	403	2	23	0	2	1	0	30
1	2	191	2	27	2	2	3	1	9
0	2	228	1	27	3	3	3	0	43
0	2	809	1	12	1	4	3	1	16
0	2	118	1	0	3	1	2	0	32
0	2	681	1	1	3	3	2	0	68
0	1	684	1	17	3	1	2	1	50
0	2	768	1	1	3	1	2	1	22
0	2	173	2	19	2	1	3	0	20
0	2	635	2	10	3	2	1	0	70
1	1	488	2	0	3	1	1	1	64
0	2	157	2	9	1	1	0	0	70
0	2	654	0	25	3	1	2	0	1
0	2	267	1	1	2	5	3	1	48
0	2	522	2	2	2	1	2	1	27

Figure 4. 4 Dataset of Proposed System After Pre-processing

If "Decision Tree before pruning" button is clicked, result of this model is shown in the following Figure 4.5.

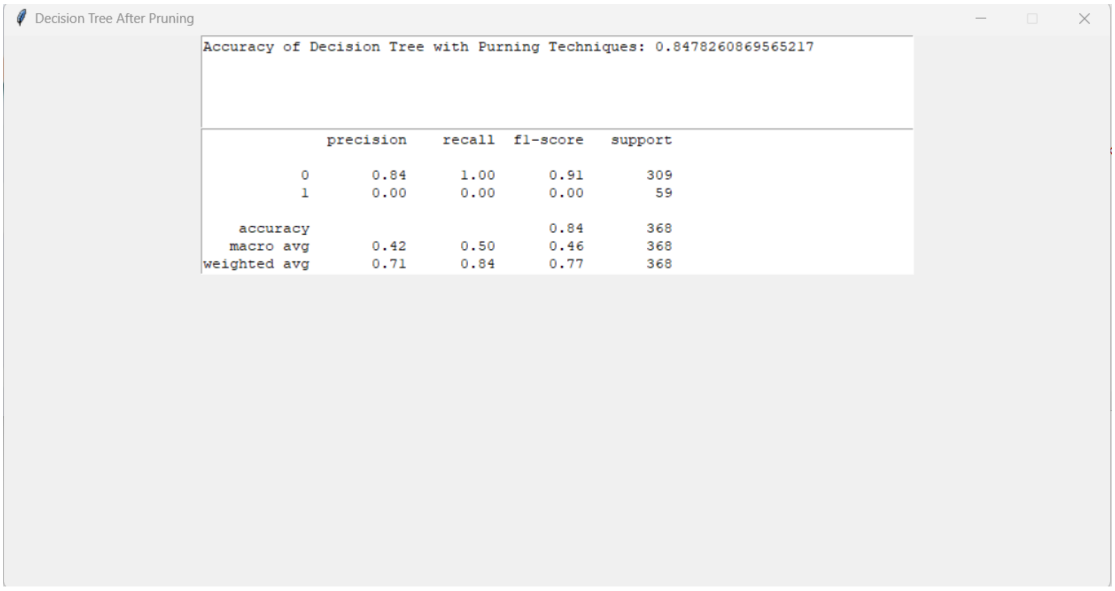


```
Accuracy of Decision Tree: 0.782608695652174
```

```
              precision    recall  f1-score   support

           0       0.89      0.85      0.87       309
           1       0.35      0.42      0.38        59

    accuracy                           0.78       368
   macro avg       0.62      0.64      0.63       368
weighted avg       0.80      0.78      0.79       368
```

Figure 4.5 The Result of Decision Tree before Pruning

If “Decision Tree After Pruning” button is clicked, result of this model is displayed in the following figure 4.5.



```
Accuracy of Decision Tree with Purning Techniques: 0.8478260869565217
```

```
              precision    recall  f1-score   support

           0       0.84      1.00      0.91       309
           1       0.00      0.00      0.00        59

    accuracy                           0.84       368
   macro avg       0.42      0.50      0.46       368
weighted avg       0.71      0.84      0.77       368
```

Figure 4.6 The Result of Decision Tree after Pruning

4.4 Experimental Results

The proposed system is constructed utilizing the ID3 classification method. After obtaining the decision tree, some of its branches can contain noise or outliers. In order to eliminate unneeded branches or nodes, the system employs the Bayes Risk

Post-Pruning approach. Decision Tree's accuracy after using Bayes Risk Post-Pruning was 84%; the accuracy before was 78%.

Between the two experiments—without pruning and after pruning system—this one has the highest accuracy.

4.4.1 Experimental Result

The following results are obtained from the initial dataset (without pruning). The results of Decision Tree (ID3) are shown in the figure 4.1 .

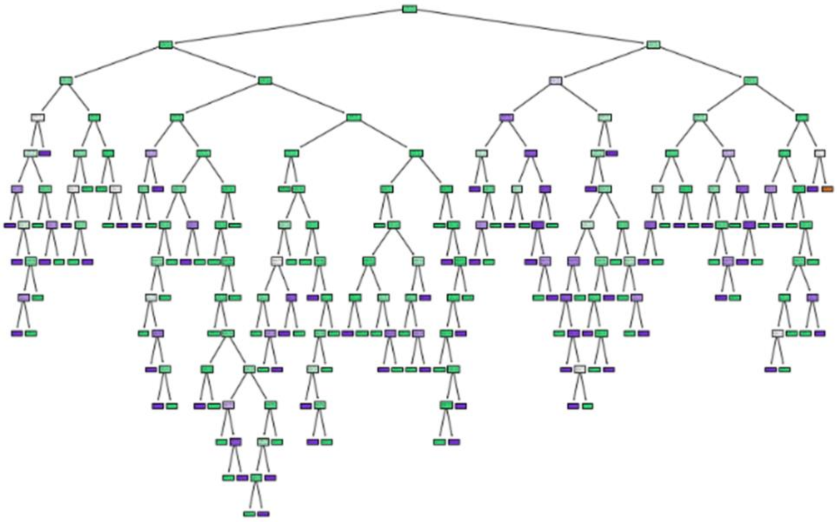


Figure 4.7 Decision Tree without Post-Pruning

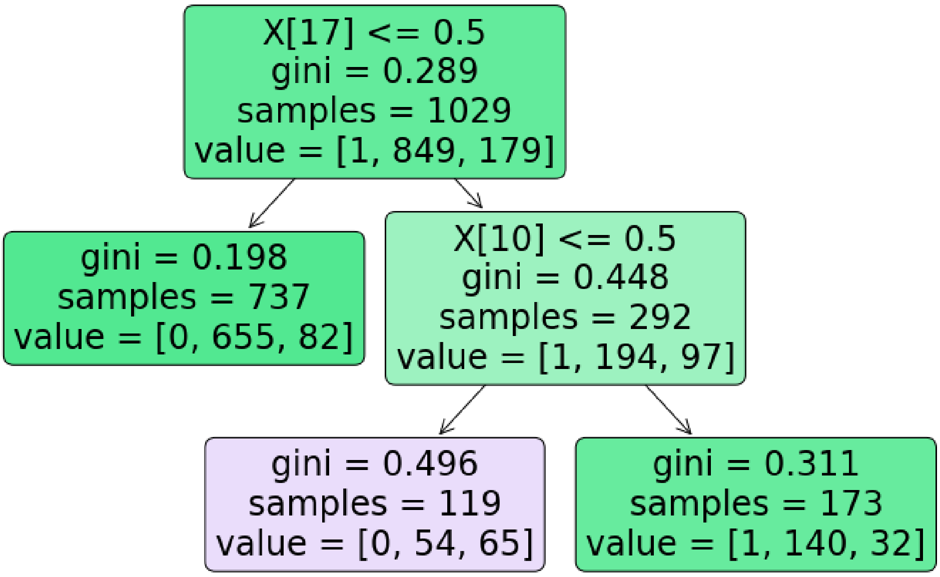
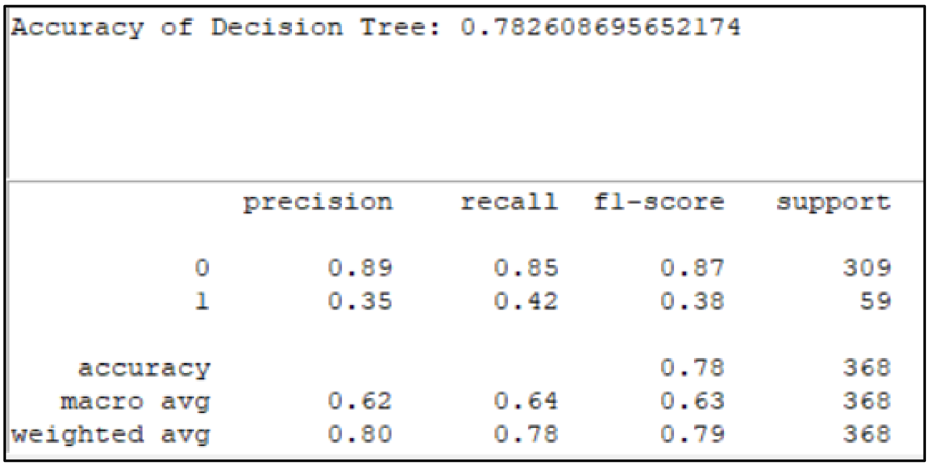
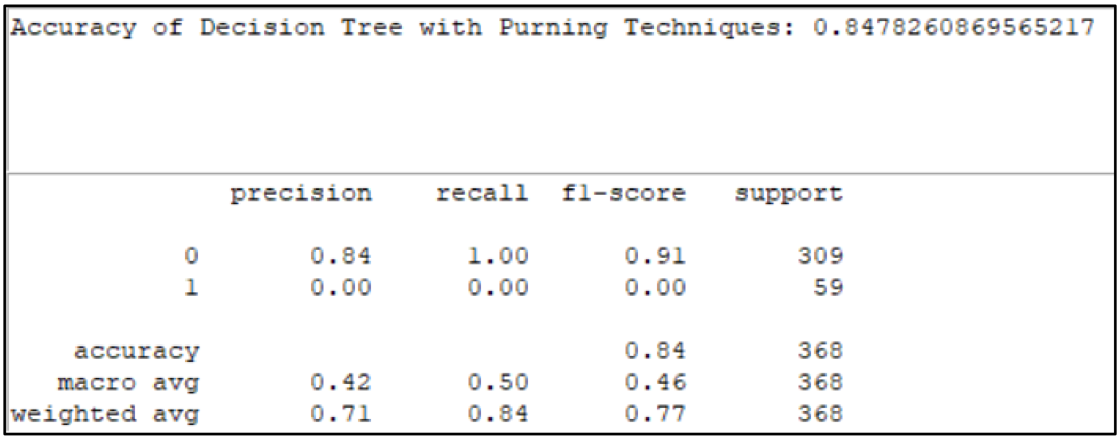


Figure 4.8 Decision Tree after Post-Pruning



```
Accuracy of Decision Tree: 0.782608695652174

              precision    recall  f1-score   support

           0       0.89      0.85      0.87       309
           1       0.35      0.42      0.38        59

    accuracy                           0.78       368
   macro avg       0.62      0.64      0.63       368
weighted avg       0.80      0.78      0.79       368
```

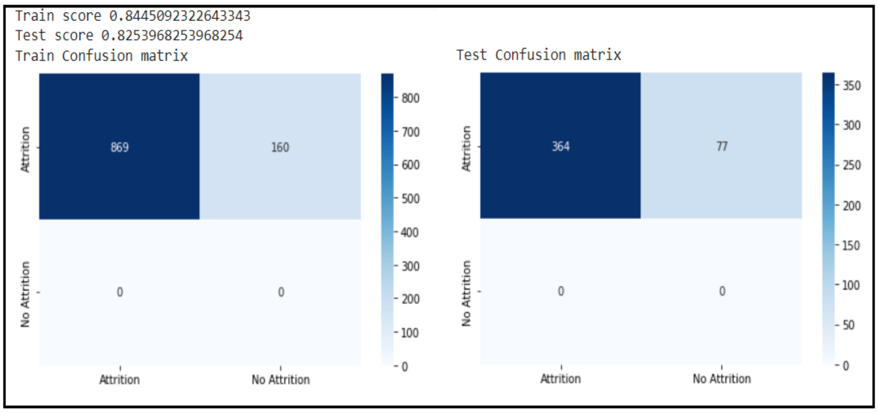
Figure 4.9 Decision Tree (ID3) Result of Proposed System



```
Accuracy of Decision Tree with Purning Techniques: 0.8478260869565217

              precision    recall  f1-score   support

           0       0.84      1.00      0.91       309
           1       0.00      0.00      0.00        59

    accuracy                           0.84       368
   macro avg       0.42      0.50      0.46       368
weighted avg       0.71      0.84      0.77       368
```

Figure 4.10 Decision Tree Result of Proposed System After Pruning

The training data in Decision Tree, which revealed that 869 and 0 are, respectively, true positives and false positives, or correctly detected cases, predicted the likelihood of a client departing. Of the total 1029 occurrences, these cases make up 869, or 84% of the total. Testing data Model properly identified 369 and 0 data points, respectively, and achieved 82% accuracy.

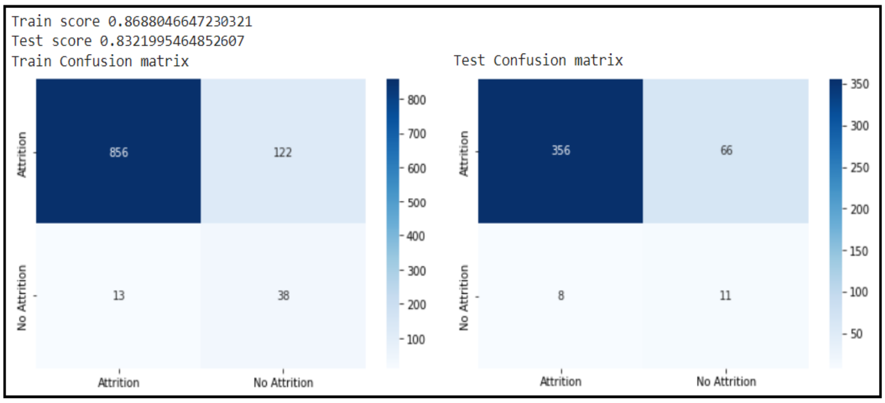


```
Train score 0.8445092322643343
Test score 0.8253968253968254
Train Confusion matrix
```

Test Confusion matrix

Figure 4.11 Confusion Matrix of Proposed System

The training data in after pruning Tree, which revealed that 856 and 38 are, respectively, true positives and false positives, or correctly detected cases, predicted the likelihood of a client departing. Of the total 1029 occurrences, these cases make up 894, or 86% of the total. Testing data Model properly identified 356 and 11 data points, respectively, and achieved 83% accuracy.



```
Train score 0.8688046647230321
Test score 0.8321995464852607
Train Confusion matrix
```

Test Confusion matrix

Figure 4.12 Confusion Matrix of After Pruning Decision Tree of Proposed System

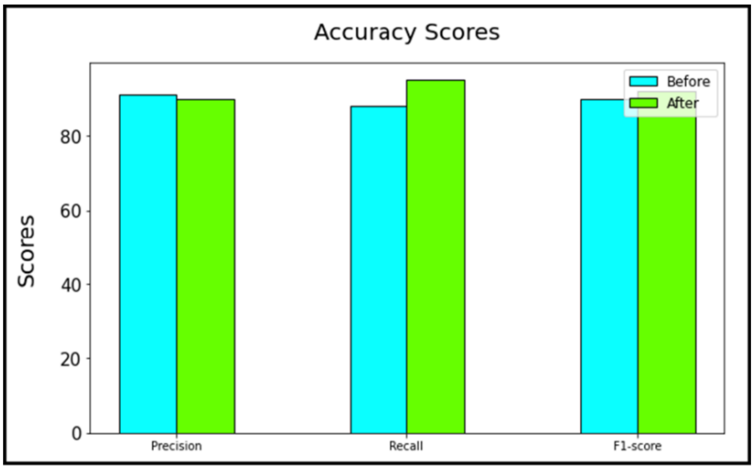
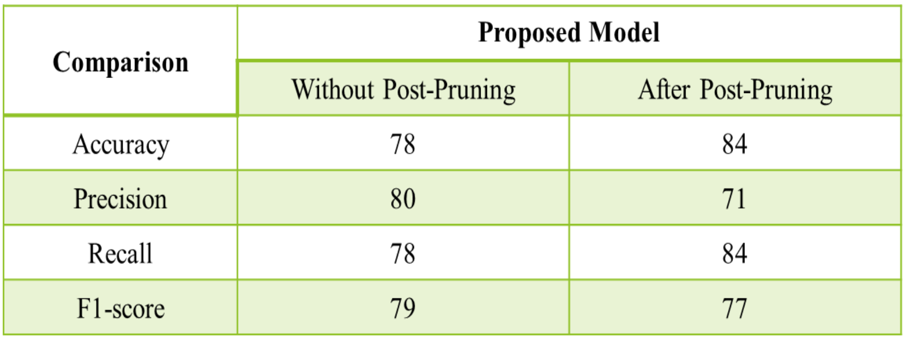


Figure 4.13 Performance Evaluation of Proposed System

4.5 Model Comparison

The original model, the initial model with feature selection, and the proposed models with multicollinearity do not differ significantly in terms of how well they predict the outcome overall. However, due to the large difference in the number of variables, the simpler model is leaner, faster, and less resource-intensive. In the following Figure 4.6, the performances of the original model and the final model are compared.



Comparison	Proposed Model	
	Without Post-Pruning	After Post-Pruning
Accuracy	78	84
Precision	80	71
Recall	78	84
F1-score	79	77

Figure 4.14 Models Comparison of proposed system

CHAPTER 5
CONCLUSION

This chapter summarizes the whole work of this thesis. It also presents the conclusion of employee attrition prediction based on employee information using IBM dataset. Furthermore, future work of the thesis is also outlined.

Any business must expect to experience attrition. Customers or workers can leave an organization due to attrition. The decision tree classification technique is used in this study's comparison to predict employee attrition. The IBM Human Resource Analytic Employee Attrition and Performance dataset serves as the basis for the suggested model's operation. For the purpose of overcoming overfitting issues, this study uses Bayes Risk Post-Pruning to condense a decision tree. When Bayes Risk Post-Pruning is used, the system compares the model's performance to that of the testing dataset. When the post-pruning procedure is not used, the proposed system had accuracy scores of 78%. The proposed approach acquires a score of 84% accuracy after applying Bayes Risk post-pruning. The system's precision, recall, and f1-score accuracy scores were 80% before the pruning stage. The system then scores with 84% accuracy following the pruning stage. Compared to the decision tree without post-pruning, Bayes Risk Post-Pruning enhances the decision tree model's capacity to forecast fresh data. The greater accuracy, precision, and recall numbers when Bayes Risk Post-Pruning is used as evidence of this. The system will improve the dataset feature space in the following study to produce more accurate findings utilizing various machine learning methods.

5.1 Limitation and Future Extension

The system will be improved the dataset feature space in the following study to produce more accurate findings utilizing various machine learning methods. With the suggested approach, it is hoped that by examining additional model combinations and ensembles using a bigger and more realistic dataset, a greater accuracy rate and a lower generalization error can be attained. Deep learning techniques will be used by this system to forecast staff attrition. Additionally, by utilizing deep learning techniques, the dataset feature space will be expanded to produce findings that are more accurate.

PUBLICATION

[1] Win Pa Pa May Phyo Aung, Nilar Aye, “Prediction of Employee Attrition using Bayes Risk Post-Pruning in Decision Tree”, University of Computer Studies, Yangon, Myanmar, 2022.

REFERENCES

[1] Aggarwal, S.; Singh, M.; Chauhan, S.; Sharma, M.; Jain, D. Employee Attrition Prediction Using Machine Learning Comparative Study. *Smart Innov. Syst. Technol.* **2022**, *265*, 453–466.

[2] Ahmed Mohamed Ahmed, Ahmet Rizaner, Ali Hakan Ulusoy, "A Novel Decision tree classification based on Post-Pruning with Bayes Minimum Risk", April 4,2018, PLoS ONE 13(4): e0194168

[3] Ali Raza, Kashif Munir, Mubarak Almutairi, Faizan Younas, and Mian Muhammad Sadiq Fareed, "Predicting Employee Attrition Using Machine Learning Approaches", Appl. Sci. 2022, Published: 24 June 2022

[4] Baldomero-Naranjo, M.; Martínez-Merino, L.I.; Rodríguez-Chía, A.M. A robust SVM-based approach with feature selection and outliers detection for classification problems. *Expert Syst. Appl.* **2021**, *178*, 115017.

[5] Bhartiya, N.; Jannu, S.; Shukla, P.; Chapaneri, R. Employee Attrition Prediction Using Classification Models. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 29–31 March 2019.

[6] Devina Christianti, Sarini Abdullah, Siti Nurrohmah, "Bayes Risk Post-Pruning in Decision Tree to overcome overfitting problem on Customer churn classification", Conference paper. January 2020, DOI: 10.4108/eai.2-8-2019.2290487, ICSA 2019, August 02-03, Bogor, Indonesia

[7] Employee Retention Statistics That Will Surprise You. 2022. Available online: **https://www.apollotechnical.com/employee-retention-statistics/** (accessed on 6 May 2022).

[8] F.M. Javed Mehedi Shamart, Sovon Chakraborty, Md. Masum Billah, Protiva Das, "A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm", 5th International Conference on Trends in Electronics and Informatics (ICOEI 2021) Tirunelveli, India, 3-5, June 2021, DOI: 10.1109/ICOEI51242.2021.9452898

[9] Ganthi, L.S.; Nallapaneni, Y.; Perumalsamy, D.; Mahalingam, K. Employee Attrition Prediction Using Machine Learning Algorithms. *Lect. Notes Netw. Syst.* **2022**, *288*, 577–596.

[10] Habous, A.; Nfaoui, E.H.; Oubenaalla, Y. Predicting Employee Attrition using Supervised Learning Classification Models. In Proceedings of the 2021 Fifth International Conference on Intelligent Computing in Data Sciences (ICDS), Fez, Morocco, 20–22 October 2021.

[11] Himani Sharma and Sunil Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (IJSR), April 2016.

[12] HR-Employee-Attrition-Dataset by Aaizemberg| Data.World. Available online: **https://data.world/aaizemberg/hr-employee-attrition** (accessed on 6 May 2022).

[13] Joseph, R.; Udupa, S.; Jangale, S.; Kotkar, K.; Pawar, P. Employee attrition using machine learning and depression analysis. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; pp. 1000–1005.

[14] Kaya, İ.E.; Korkmaz, O. Machine Learning Approach for Predicting Employee Attrition and Factors Leading to Attrition. *Cukurova Univ. J. Fac. Eng.* **2021**, *36*, 913–928.

[15] Krishna Kumar Mohbey, "Employee Attrition prediction using Machine Learning Approaches", 06 January 2020, DOI: 10.4018/978-1-7998-3095-5.ch005

[16] Lai, H.; Hossin, M.A.; Li, J.; Wang, R.; Hosain, M.S. Examining the Relationship between COVID-19 Related Job Stress and Employees'

[17] Maswadi, K.; Ghani, N.A.; Hamid, S.; Rasheed, M.B. Human activity classification using Decision Tree and Naïve Bayes classifiers. *Multimed. Tools Appl.* **2021**, *80*, 21709–21726.

[18] Mate, Y.; Potdar, A.; Priya, R.L. Ensemble Methods with Bidirectional Feature Elimination for Prediction and Analysis of Employee Attrition Rate During COVID-19 Pandemic. *Lect. Notes Electr. Eng.* **2022**, *806*, 89–101

[19] Moderating Role of Perceived Organizational Support: Evidence from SMEs in China. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3719.

[20] Najafi-Zangeneh, S.; Shams-Gharneh, N.; Arjomandi-Nezhad, A.; Zolfani, S.H. An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. *Mathematics* **2021**, *9*, 1226.

[21] Nesserullah, “The Pros and Cons of pruning in classification”, Proceedings of Academics era 32nd International Conference, London, United Kingdom, 18th -19th October 2018

[22] Norsuhada Mansor, Nor Samsiah Sani and Mohd Aliff, “Machine Learning for predicting Employee Attrition”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 11, 2021

[23] O.O Adeyemo & T.O Adeyemo and D. Ogunbiyi, “Comparative Study of ID3/C4.5 Decision Tree and Multilayer Perception Algorithms for the Prediction of Typhoid Fever”, African Journal of Computing & ICT, Vol 8 No.1, March 2015.

[24] Peng, B. Statistical analysis of employee retention. In Proceedings of the International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), Nanjing, China, 26–28 November 2021; Volume 12163, pp. 7–15.

[25] Pratt, M.; Boudhane, M.; Cakula, S. Employee attrition estimation using random forest algorithm. *Balt. J. Mod. Comput.* **2021**, *9*, 49–66.

[26] Qutub, A.; Al-Mehmadi, A.; Al-Hssan, M.; Aljohani, R.; Alghamdi, H.S. Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. *Int. J. Mach. Learn. Comput.* **2021**, *11*, 110–114.

[27] Rahul Yedida, Rahul Reddy, Rakshit Vahi, Rahul J, Abhilash, Deepti Kulkarni, “Employee Attrition Prediction”

[28] Raza, Ali, Kashif Munir, Mubarak Almutairi, Faizan Younas, and Mian Muhammad Sadiq Fareed. 2022. "Predicting Employee Attrition Using Machine Learning Approaches" *Applied Sciences* 12, no. 13: 6424.

[29] Sadana, P.; Munnuru, D. Machine Learning Model to Predict Work Force Attrition. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Pune, India, 2–4 April 2021.

[30] Shobhanam, K.; Sumati, S. HR Analytics: Employee Attrition Analysis using Random Forest. *Int. J. Perform. Eng.* **2022**, *18*, 275.