**Employee Attrition Prediction using IBM HR Analytics Dataset**

Govardhan Narasimha Murthy  
 Department of Computer Science  
 University of Colorado, Boulder  
Boulder CO USA  
Govardhan.Narasimha@colorado.edu

Pradyumna Chippigiri  
 Department of Computer Science  
 University of Colorado, Boulder  
Boulder CO USA  
 prch4538[@colorado.](mailto:email@email.com)edu

Karan Bantia Ram  
 Department of Computer Science  
University of Colorado, Boulder  
Boulder CO USA  
 karan.bantiar@colorado.edu

Ashutosh Naik  
 Department of Computer Science  
 University of Colorado, Boulder  
Boulder CO USA  
 ashutosh.naik@colorado.edu

**ABSTRACT**

This study develops a predictive model for employee attrition, aimed at improving retention in organizations. By using data analytics, the project identifies factors leading to employee turnover, a significant cost concern for companies. The approach includes exploratory data analysis, feature engineering, and predictive modeling to create a reliable model that flags high-risk employees. The goal is to help organizations enhance their retention strategies, thus fostering a better work environment and supporting ongoing success. This research is applicable across various sectors, especially in environments where retaining talent is crucial for competitive advantage.

1**INTRODUCTION AND MOTIVATION**

The Employee Attrition Prediction project aims to leverage data analytics to formulate a robust predictive model for accurately identifying potential employee turnovers within organizations. This project is driven by the importance of retaining valuable talent and reducing the significant costs associated with employee replacements. Additionally, it has the potential to improve employee satisfaction, enhance organizational performance and create a positive work culture. This project is motivated by the importance and opportunity in identifying and using data analytics to make positive changes and positively impact employees and organization as a whole. By tapping into advanced data analytics methodologies, we can extract patterns, relationships, and predictive insights from the dataset, enabling us to devise an efficient employee retention strategy.

This has real-world implications in various sectors, spanning from tech startups to multinational corporations, where retaining talent is pivotal for competitive advantage.

Using techniques like exploratory data analysis, feature engineering, and predictive modeling, we'll delve into the dataset's depths. Our goal is to offer a reliable and interpretable predictive model that accurately flags employees at high attrition risk and in identifying what factors contribute to employees leaving. Ultimately, this initiative seeks to equip organizations with a valuable tool for enhancing their retention strategies, fostering a stable work environment, and ensuring continued growth and success.

2**LITERATURE SURVEY**

N.B. Yahia et. al. [1] propose a people analytics approach to predict employee attrition that shifts from a big data context to a deep data context. This shift emphasizes data quality over quantity. Their approach uses a mixed method to construct a relevant employee attrition model. The attrition prediction approach is based on machine learning, deep learning, and ensemble learning models. The researchers tested this approach on large and medium-sized simulated human resources datasets and a real small-sized dataset, which had a total of 450 responses. The approach achieved high accuracy scores of 0.96, 0.98, and 0.99 for the three datasets, outperforming previous solutions. The authors found that while rewards and payments are often seen as key to retention, "business travel" emerged as a leading motivator for employees. The authors used a wide range of models from Decision Trees, SVMs, to LSTMs, DNNs and Ensemble methods like XGBoost and Random forest.

Al-Darraji et. al. [2] propose a deep neural network approach to the predict the attrition using the IBM Dataset.They address the dataset’s imbalance by creating a synthetic balanced version.The prediction is done using the Deep Neural network model with 7 hidden layers which is trained with various hyper parameters. The model is later validated using the 70-30 train-test and the k-fold validation. This approach outperforms various methods of attrition prediction for this dataset based on the F1-Score, Accuracy, Precision.The accuracy obtained using this method was 94.16%

Francesca Fallucch et. al. [3] implements a Data Mining process on an employee attrition dataset starting with data pre-processing (data cleaning and data exploration) to prepare the dataset for further process. Different kinds of descriptive analysis is run on this dataset to detect key factors which cause employee attrition. The dataset was divided into training (70%) and testing (30%). Different classification machine learning models were used on the training data set and the model was evaluated based on the testing dataset. The different models used were Gaussian Naive Bayes, Logistic Regression classifier, K-nearest neighbors (K-NN) etc. and Gaussian Naïve Bayes was identified as the best classification model based on various performance parameters.

The authors in [4] firstly discuss the reason for attrition in which the major one is work-life balance, Lack of Decision Making Ability and then go through XGBoost which is a boosted tree algorithm using gradient boosting to predict employee attrition. It goes through the advantages of using the approach. Some of it include Tree pruning, Save and Reload which will save the data from previous computation and help in saving time. Also, a high accuracy of more than 90% was obtained by using XGBoost.

3**DESIGN AND IMPLEMENTATION**

3.1**DATASET**

IBM HR Analytics Employee Attrition & Performance is a popular choice for predicting employee attrition. It contains various information about employees, their background, Travelling Distance. Based on these characteristics, the dataset can be used to predict whether the employee will leave the company or not. The dataset contains data of **1470** employees each with **35** attributes. These attributes describe various characteristics such as *Age, Target Class, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Num Companies Worked, Over 18, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Curr Manager.*

The binary target class **‘Attrition’** which is the 2nd column indicates whether an employee leaves the company **‘1 = Yes’** or **‘0 = No’.** This dataset can be used by companies to decide the future of the employees and decide their hike. It can be used to build and test various classification models available and compare the accuracy between them.

3.2 **PLANNED APPROACH**

There are three sections - Data Preprocessing and EDA, Feature Extraction and Model Implementation.

3.2.1 **DATA PREPROCESSING AND EDA**

The dataset was initially examined for missing values, which were found to be absent. Subsequently, skewness within the dataset was checked, revealing a class imbalance, with 1233 rows assigned to the 'No' class and 237 rows to the 'Yes' class. To address this imbalance during exploratory data analysis, it is essential to focus on "percentages" and "relative values" when comparing features between these two classes.

Furthermore, preliminary data visualization was used to gain insight into the data distribution. The analysis of these distributions revealed distinctive characteristics. Specifically, the age distribution displayed a notable right-skew, indicating a greater number of younger employees within the dataset. In contrast, the distribution of daily rate exhibited a relatively uniform profile, devoid of any prominent peaks. Moreover, the distribution for distance from home exhibited right-skewness, suggesting that a considerable proportion of employees reside in proximity to their workplace.

Additionally, the monthly income distribution was right-skewed, indicating that a substantial portion of employees earned lower monthly incomes, with only a few individuals occupying higher income strata. The distribution of total working years also displayed a right-skewed pattern, aligning with the earlier observation of a younger age distribution.

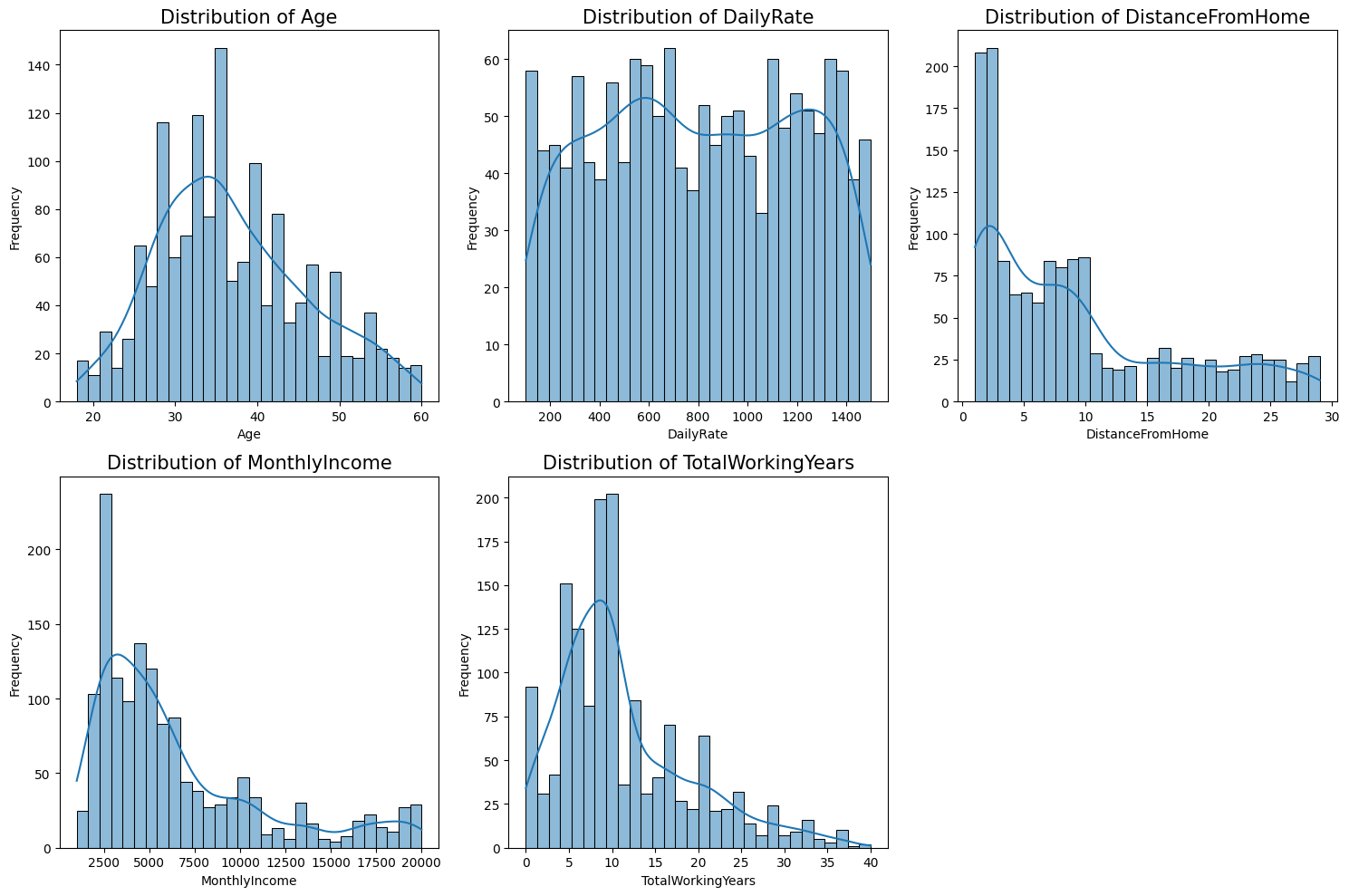


Figure 1: Initial Data Distribution

Following this, a correlation analysis was carried out to examine the relationships between various features. The analysis revealed a substantial positive correlation of 0.68 between ‘TotalWorkingYears’ and ‘Age’, which is expected as employees tend to accumulate more working years as they get older. ‘MonthlyIncome’ exhibited a strong positive correlation of 0.95 with ‘JobLevel’, indicating that as employees progress to higher job levels, their monthly income generally increases. In addition, ‘YearsAtCompany’ displayed positive correlations with both ‘YearsWithCurrManager’ (0.77) and ‘YearsInCurrentRole’ (0.76), suggesting that employees who have been with the company for a longer duration also tend to have spent more time in their current role and working under the same manager.

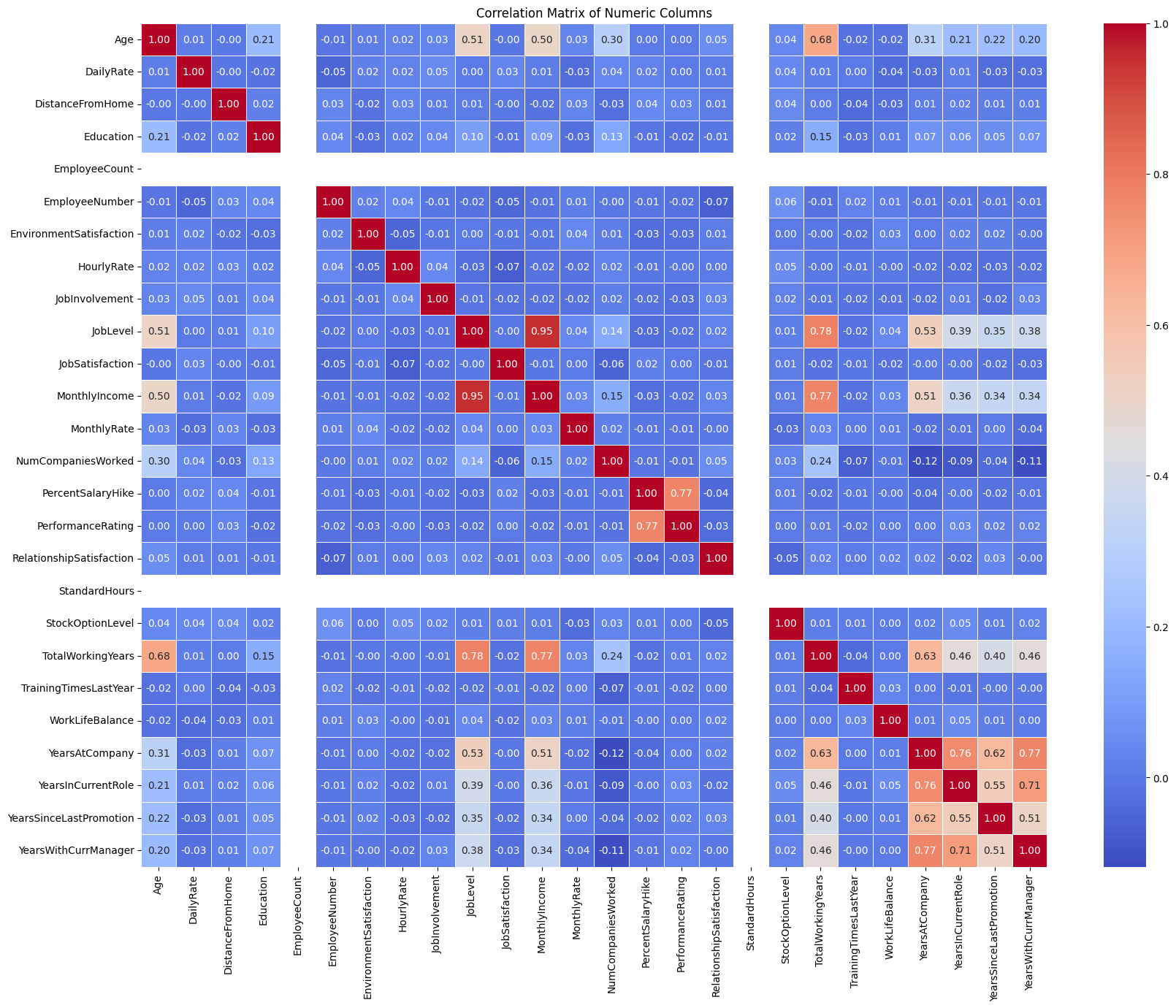


Figure 2: Correlation analysis heatmap

Apart from conducting a correlation analysis for numerical features, the chi-squared statistic was employed to evaluate the relationship between ordinal features. See figure 3. In all cases, the computed p-values were found to be less than 0.05, signifying a statistically significant association between the variables within each pair. For instance, the markedly low p-value observed for the pair (Attrition, JobRole) strongly indicates a substantial and significant relationship between an employee's job role and their decision to depart from the company.

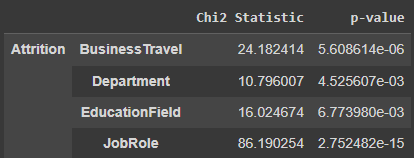


Figure 3: Chi-squared Analysis (ordinal features)

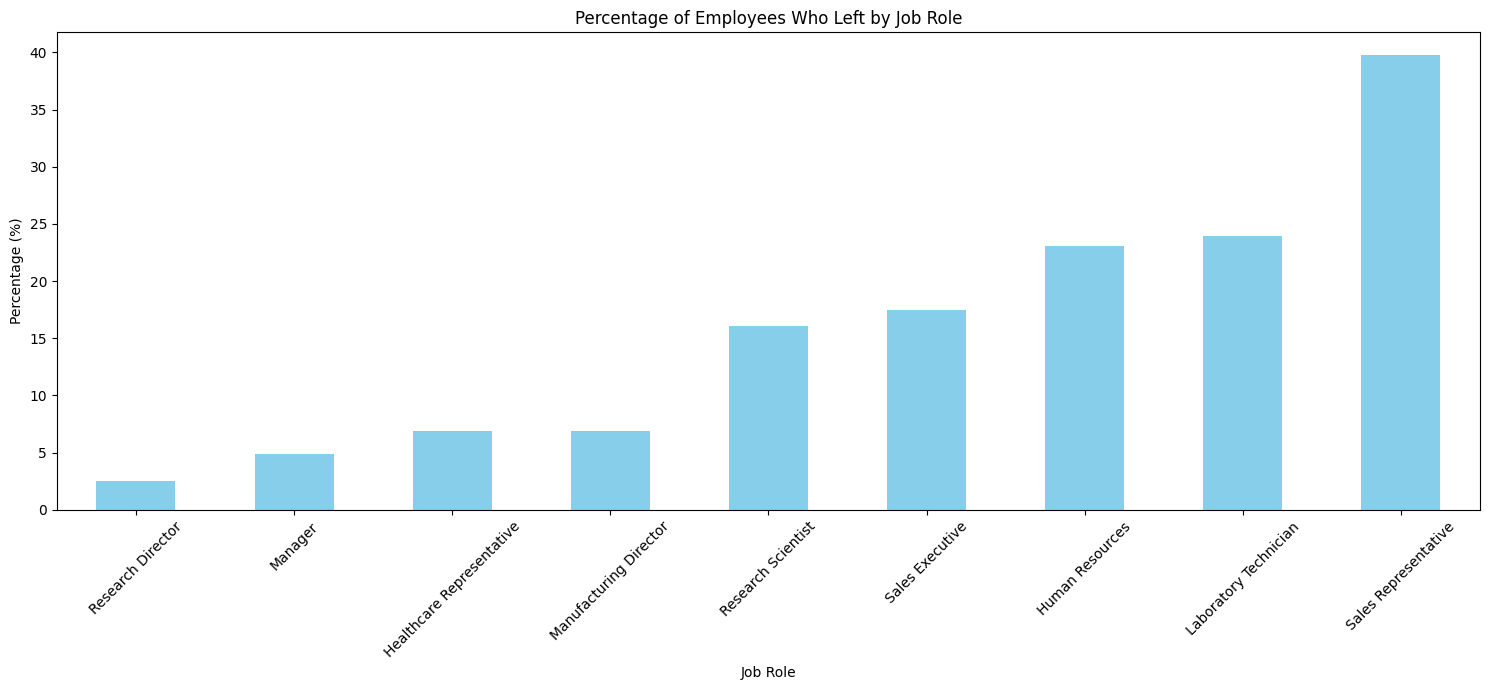


Figure 4: Job role v/s Attrition rate.

Given the substantial correlation unveiled through the chi-squared analysis between job roles and attrition, a visual representation of attrition rates, stratified by job roles, was generated (as depicted in Figure 4). This visualization revealed that Sales Representatives exhibited the highest attrition rates, surpassing 40%. Roles like Research Director and Manager depicted the lowest attrition rates of less than 6%.

Subsequently, the attrition rate was depicted in relation to job travel frequency (as depicted in Figure 5). Notably, employees characterized by frequent business travel exhibited the highest attrition rate, approaching 25%. This observation suggests a connection between frequent business travel and the propensity of employees to depart from the company. In contrast, employees who infrequently engage in business travel displayed a comparatively lower attrition percentage.

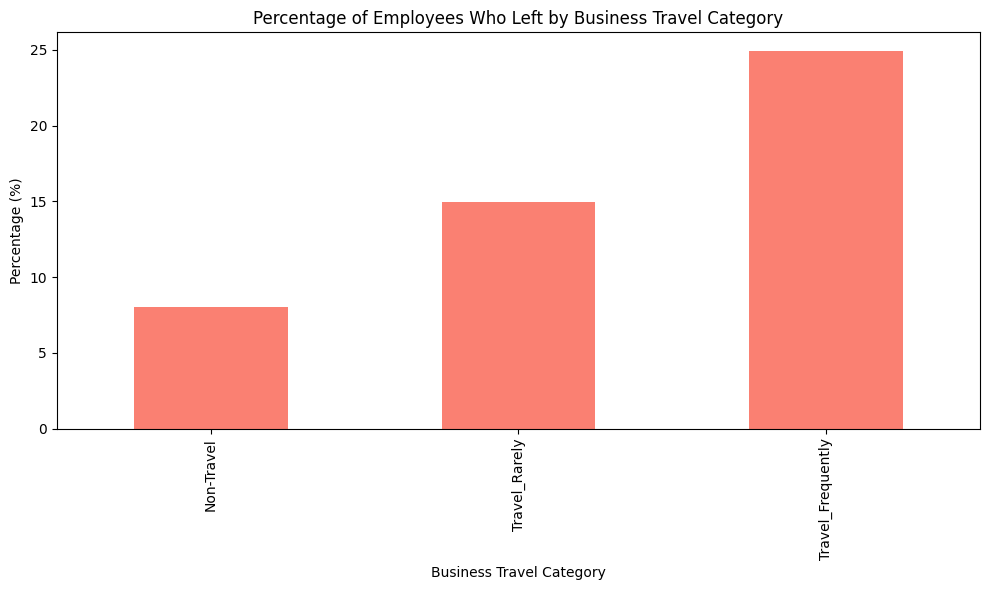


Figure 5: Attrition rates against business travel frequency

Next, we examined the MonthlyIncome distribution categorized by TotalWorkingYears bins for employees who left the company versus those who stayed (as depicted in figure 6). As expected, the median monthly income generally increased with more years of work, reflecting career progression and salary increments over time.

In the early career stages (0-5 years and 6-10 years), employees who left had noticeably lower median monthly incomes compared to those who stayed. In the mid-career stages (11-15 and 16-20 years), the difference in median monthly income between the two groups was minimal or even reversed. For more experienced employees (20+ years), the gap in average monthly income widened again, with those who stayed earning more. These findings suggest that income disparities may have a more significant impact on attrition, especially in the early and later stages of an employee's career.

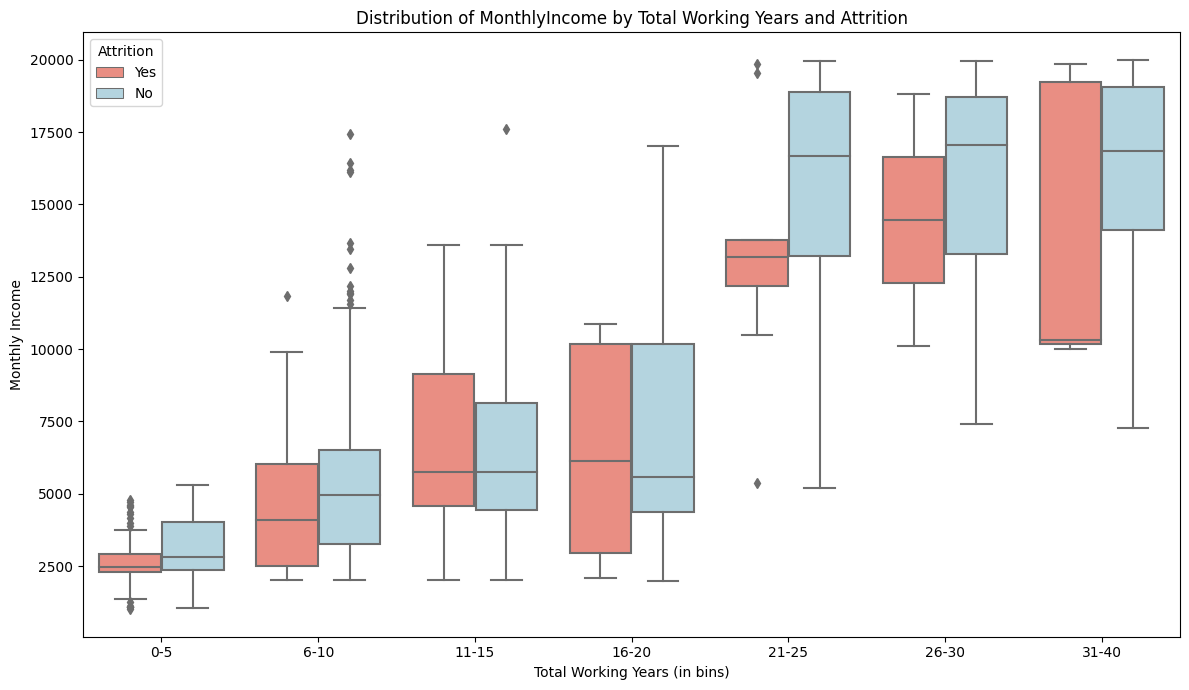


Figure 6: Monthly Income vs Total working years separated by attrition classes.

We additionally computed the relative percentages of attrition with respect to overtime status (see figure 7). Notably, close to 31% of employees who engaged in overtime left the company, while in contrast, only approximately 10% of employees who did not work overtime decided to leave. The disparity in attrition rates between these two groups strongly indicates that working overtime may constitute a substantial factor contributing to employee attrition.

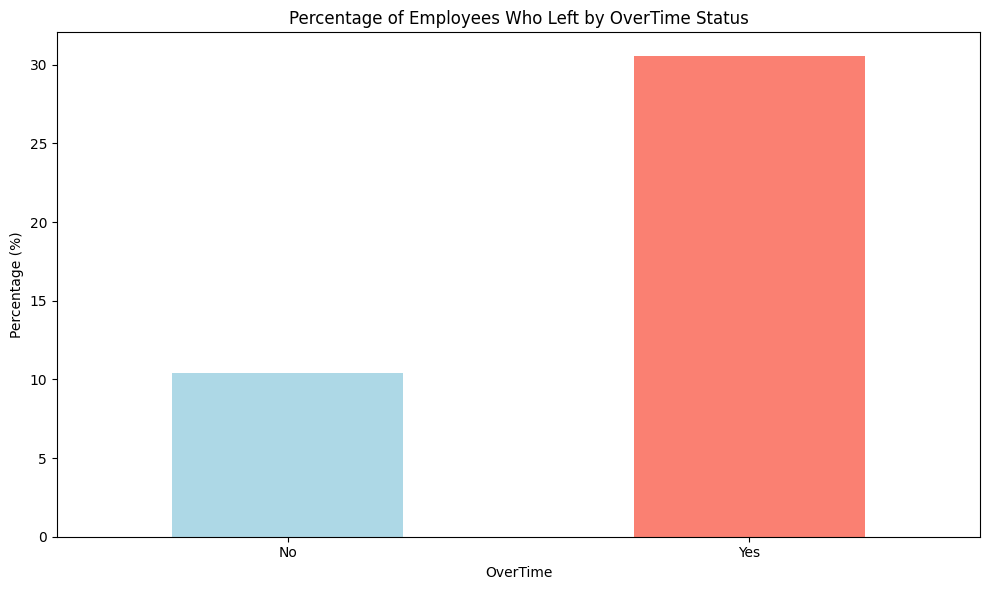


Figure 7: Attrition rate by Over time status.

This observation led to a deeper examination of the attrition rates based on both monthly income bins and overtime status (figure 8). Notably, employees with lower incomes who worked overtime displayed a substantially higher attrition rate compared to their counterparts who did not engage in overtime. As income levels increased, the disparity in attrition rates between the two overtime categories diminished, but it still remained higher for those working overtime. In the case of higher income ranges, the attrition rate was relatively low, even for those working overtime. This discernible pattern indicates that employees with lower incomes who also work overtime are at a heightened risk of departing from the company.

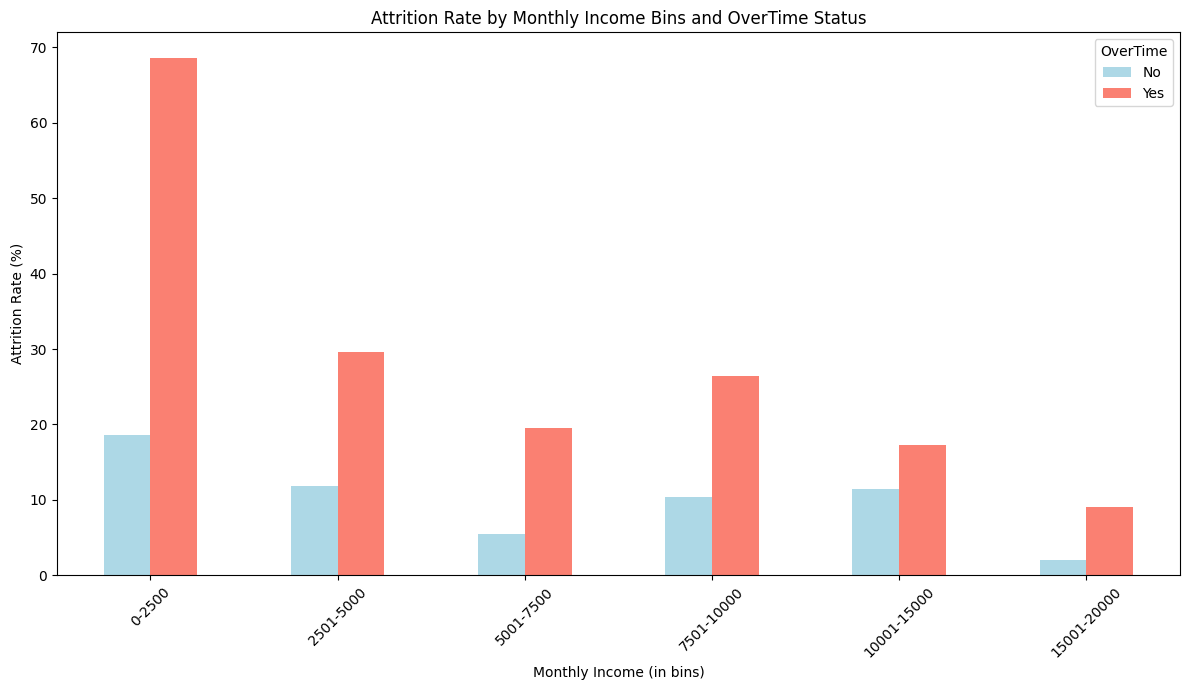


Figure 8: Attrition rate and monthly income bins

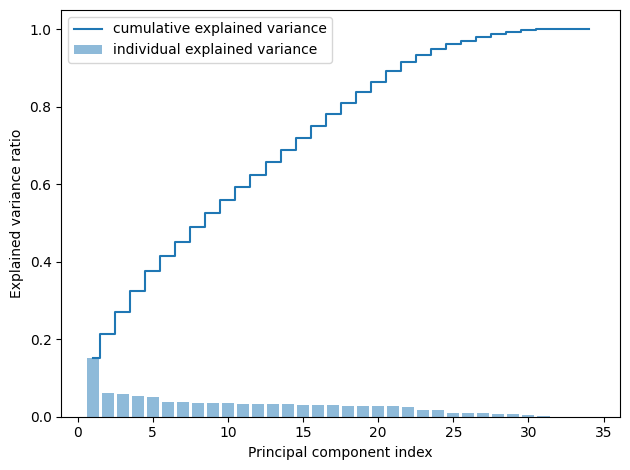
3.2.2 **FEATURE EXTRACTION**

Feature extraction aims to reduce data complexity while preserving essential information. This process converts raw data into a useful, smaller set of features, enhancing model performance by focusing on relevant features and excluding unhelpful or redundant ones, particularly in high-dimensional data.

Principal Component Analysis (PCA) is a key method in this context. PCA works by transforming and standardizing variables into uncorrelated principal components, sorted by their variance. This effectively represents data in a reduced dimension, retaining most of its variability.

By discarding less significant components, PCA streamlines computations, removes duplicative information, and simplifies data analysis and visualization. It is especially valuable for managing and understanding complex, high-dimensional datasets, allowing for data compression without significant loss of important features.

On applying PCA we discovered that the **first 23 components preserve over 90% of the variation in the dataset.**



3.2.3 **MODEL IMPLEMENTATION**

Implementing a predictive model involves several critical steps to ensure its effectiveness in data analysis. Initially, a model is selected based on the nature of the data and the specific task. This is followed by model training, where the model learns and adjusts its parameters using labeled data. Model testing then assesses the model's accuracy on unseen data, ensuring its reliability in practical applications. This structured approach ensures the model's successful application in real-world scenarios, thereby increasing the value and impact of data analysis.

We analyzed various machine learning algorithms like, *Logistic Regression, Artificial Neural Networks, Decision Tree, Naive Bayes, Random Forest, and Gradient Boosting.* Our objective is to conduct a comparative study of these models to identify the optimal model for predicting the Attrition of employees. We also selected a different number of principal components to analyze how the number of components affect the performance of each model on the test data.

The dataset was also imbalanced so we decided to use SMOTE (Synthetic Minority Over-sampling Technique). It plays an important role in addressing imbalances in datasets. When a dataset is unbalanced, with one class significantly outnumbering others, it can lead to biased models that perform poorly on minority classes. SMOTE addresses this by generating synthetic samples for the underrepresented class, balancing the dataset without losing valuable information. By doing so, it ensures that the model learns equally from all classes, improving its ability to generalize and make accurate predictions across diverse scenarios. This technique is crucial for enhancing the fairness and accuracy of predictive models.

In the case of the Artificial Neural Networks, they feature an 8-neuron hidden layer with ReLU activation and a 1-neuron output layer using sigmoid activation.

* **Logistic Regression** a binary classifier is used to determine if an employee is quit or not. The sigmoid function is used to determine the likelihood of attrition based on characteristics and transfer it to a number between 0 and 1 taking advantage of the model's clarity and readability.
* **Artificial Neural Network**s use a 4-neuron hidden layer (ReLU activation) and a 1-neuron output layer (sigmoid activation) for binary classification to predict employee attrition. Each layer's neurons carry out weighted computations using activation functions to capture intricate feature-target correlations.
* **Decision Tree** is a non-linear, interpretable classification technique. It handles categorical variables without requiring considerable preprocessing, capturing intricate relationships to determine employee attrition.
* **Naive Bayes** - Due to the simplicity, efficiency, and good handling of categorical information, Naive Bayes is appropriate for predicting the problem at hand. It assumes feature independence and produces interpretable results, making it helpful in understanding the classification process. Even on this moderate-sized dataset, Naive Bayes can perform well by tuning parameters.
* **Random Forest** uses multiple decision trees to make accurate predictions by randomly selecting data and reducing overfitting. Our dataset is perfect for this approach, capturing intricate relationships between variables and generating dependable predictions through ensemble learning.

4**EVALUATION METRICS**

The performance of our model is evaluated using the following metrics:

1. **Accuracy**: This metric provides a good overview of model performance by showing the ratio of correct predictions to total predictions. Despite its simplicity it might not be a good indicator for imbalance datasets. It can be calculated using the formula:

*Accuracy = (TP + TN) / (TP + TN + FP + FN)*

1. **F-1 Score**: The F1 score gives a complete picture of the model's accuracy. It combines both false positives and false negatives from the confusion matrix. It can be determined using the formula:

F1 Score = *TP / (TP + 0.5 \* (FP + FN))*

1. **ROC AUC**: This statistic assesses the model’s capacity to differentiate at various categorization levels. AUC is an abbreviation for “Area Under the Curve”. It measures the area beneath the ROC curve and aggregates performance across all categorization criteria. The higher the AUC, the better the model is. The X-axis of a ROC curve usually represents the False Positive Rate FPR = FP / (FP + TN) while the Y-axis of the graph represents the True Positive Rate TPR = TP / (TP + FN)

5**RESULTS AND ANALYSIS**

In this section we will discuss our results obtained after performing the model evaluation using different ML models and PCA with varying numbers of components. The following results were obtained.The evaluation metric includes AUC, Accuracy and F1 Score for each model and each value of the component. By comparing the metric for different models and different components we can decide and identify which model best performs for our dataset. Based on the results we can choose the model and the number of the components that yield the best performance. The visual representation of the result table above will help us interpret and analyze the results better.

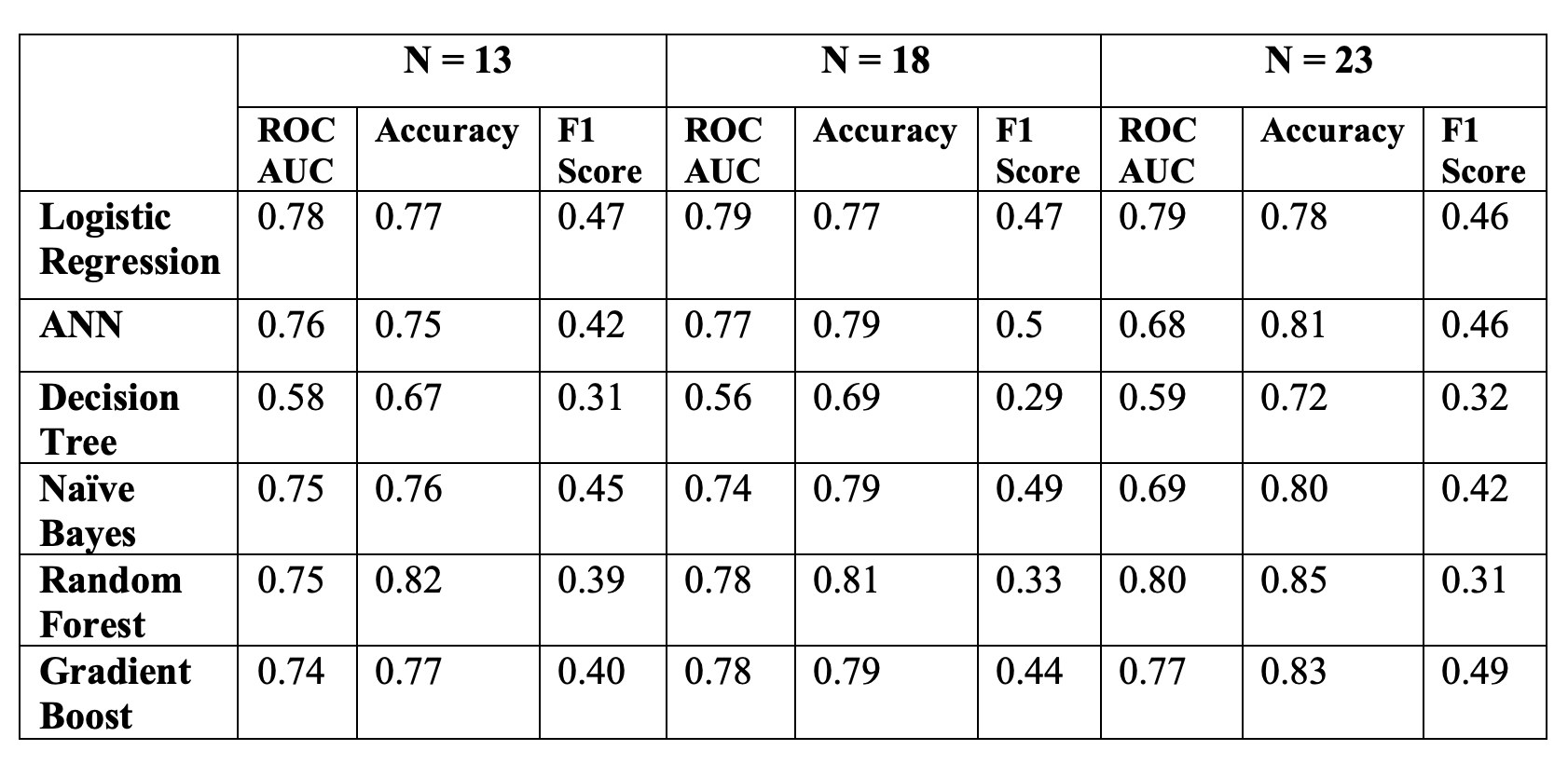
****

Figure 9: Result

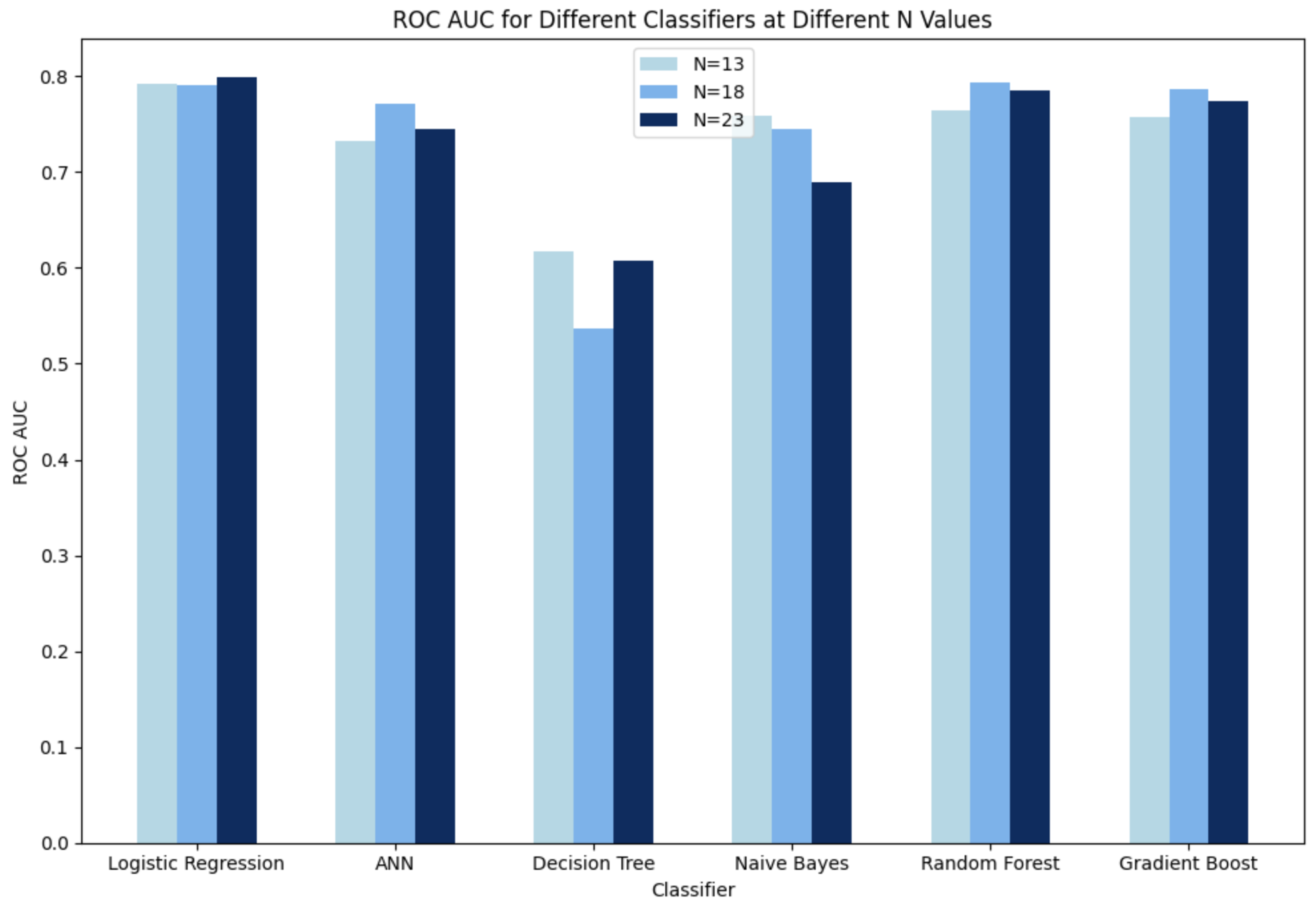
 Figure 10: ROC values for different classifiers at different N values

Figure 10 shows that Logistic regression and Random Forest classifiers show better ROC AUC performance with N. Figure 11 shows that Random Forest achieves highest accuracy and its performance improves with higher N. Figure 12 shows that F1 score varies across classifiers and various N values. Gradient Boost and Naive bayes have highest accuracy for N=23 and N=18 respectively. Logistic regression has a stable performance with N. Gradient Boost algorithm achieves the highest F1 score as the N values increases and in case of Random Forest Classifier the F1 score decreases with N values.So based on the comparison on various models we can choose the model that best suits our requirement.

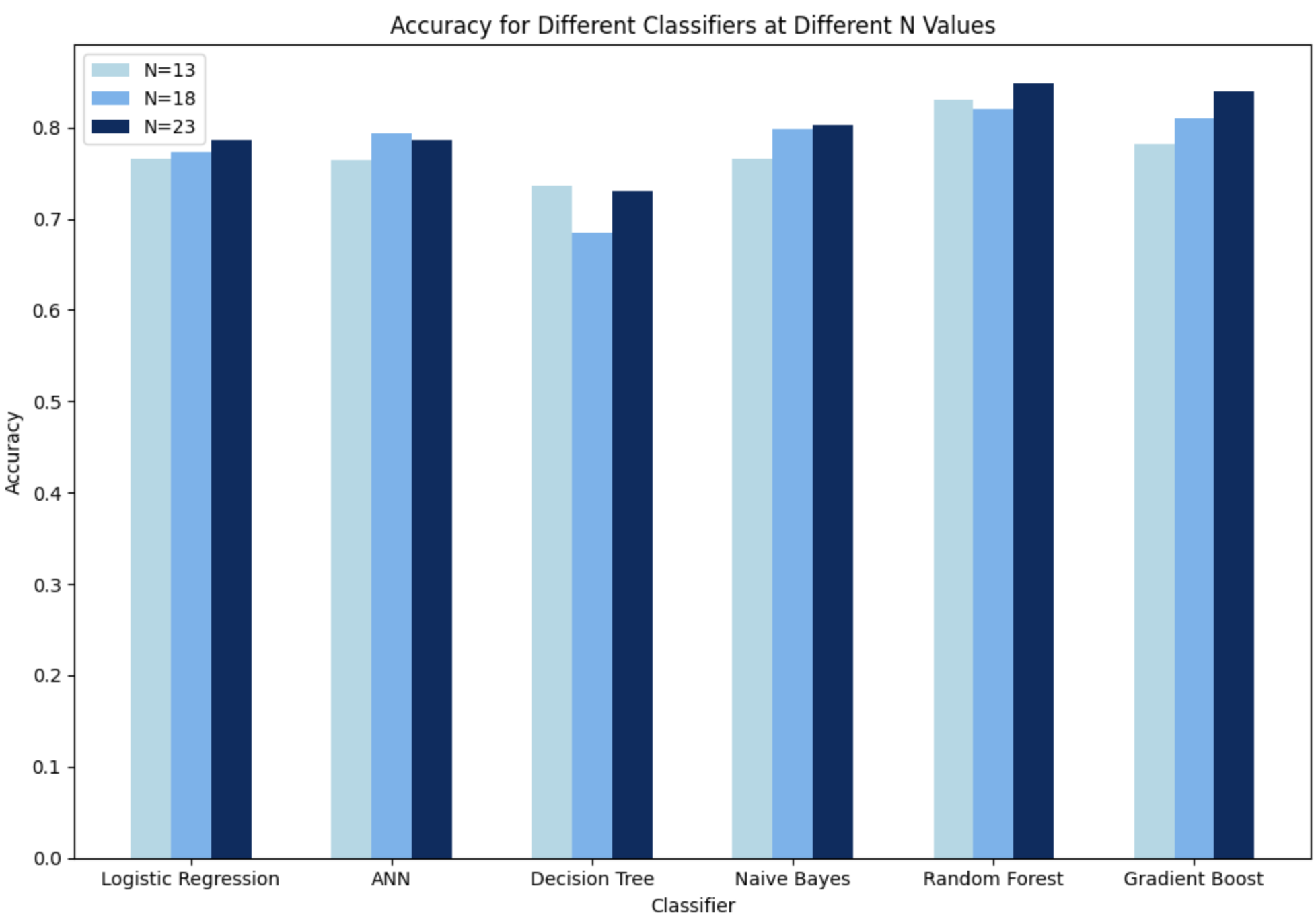
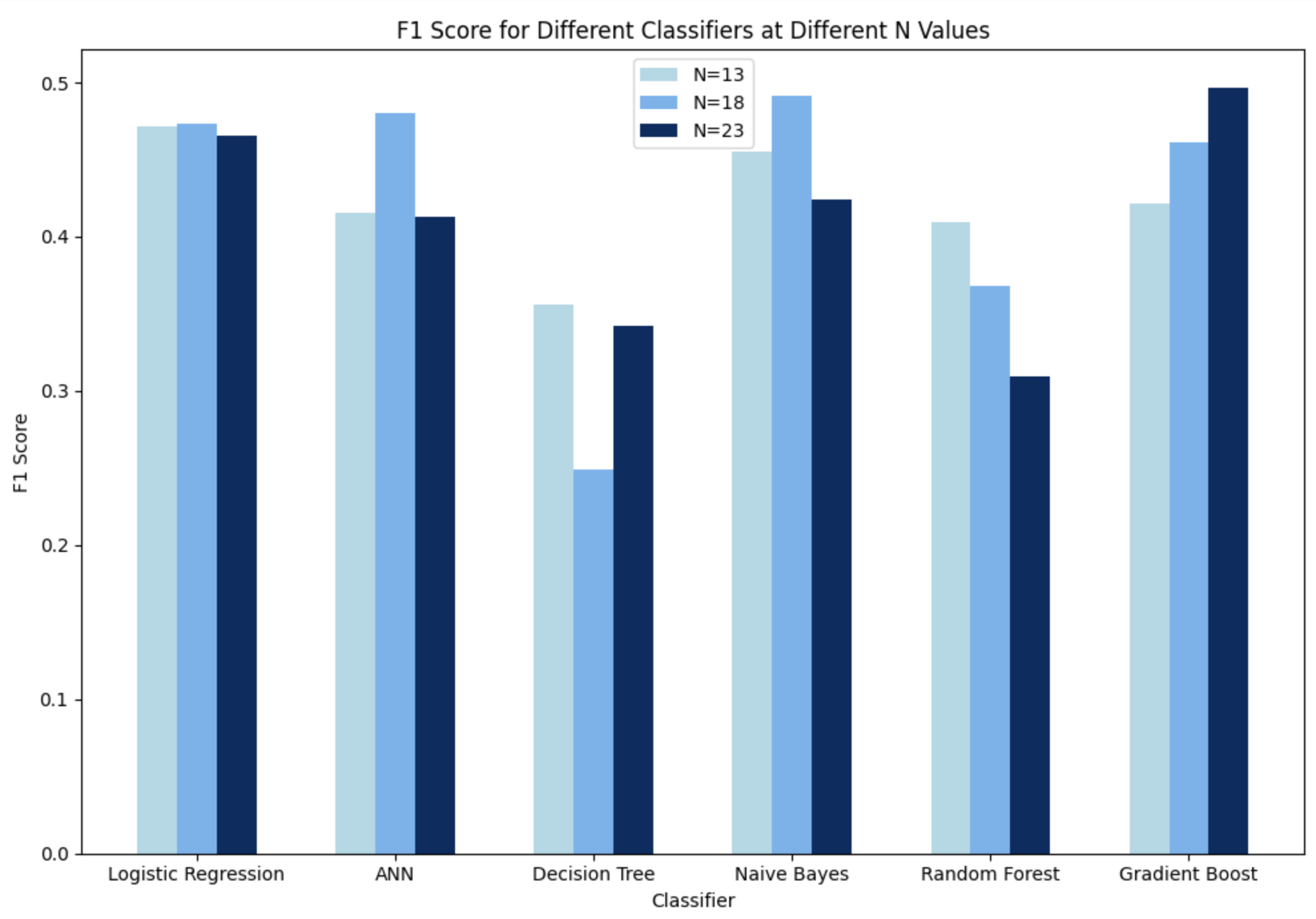


Figure 11: Accuracy for different Classifiers at different N values

Figure 12: F1- score for different classifiers at different N values

6 **FUTURE WORK**

As part of our future plans, we plan to implement two ideas that add to the existing results.

1. We would like to utilize Unstructured data analysis like text data from employee reviews and feedback by leveraging Natural Language Processing to extract sentiment and contextual information. We can also explore relationships between employees using Social Network Analysis to understand how it affects attrition.
2. Implement a User Interface for HR to interpret the outcomes from the model and resolve the factors which are resulting in lower attrition rate.
3. Expanding the analysis to include external factors impacting attrition, such as economic conditions, industry trends, or changes in labor markets, could provide a broader context. Correlating these external factors with internal data might reveal previously unidentified correlations or predictive patterns.

7 **APPENDIX**

7.1**HONOR CODE PLEDGE**

On my honor, as a University of Colorado Boulder student I have neither given nor received unauthorized assistance.

7.2 **INDIVIDUAL CONTRIBUTION**

1. **Ashutosh Naik -** Mainly performed Exploratory Data Analysis, and models like Logistic Regression and Decision Tree were implemented. During EDA many interesting trends were observed which were subject of further analysis. The performance of the models were tested using the evaluation metrics. Also assisted in the Introduction, Abstract and Literature survey in the report and in making the presentation.
2. **Pradyumna Chippigiri -** Worked on data transformation, correlation analysis, and standardization. Performed Naive bayes model prediction for the employee attrition. Helped in creation of the report for Motivation, literature survey and results and conclusion sections and in making of the final presentation.
3. **Karan Bantia Ram -** Worked on SMOTE to reduce dataset imbalance. Performed Random forest and gradient model training and evaluations. Helped in creating the report for evaluation metrics, classifiers definitions and literature survey and in making of the final presentation.
4. **Govardhan Narasimha Murthy -** Worked on Dimensionality Reduction using PCA and analyzed the amount of data preserved after reducing the number of columns. Also worked on Artificial Neural Networks model training and analysis of the results obtained from the model. Assisted in creation of a report for Methodology and the steps included in it and in making the final presentation.

8 **REFERENCES**

[1] N. B. Yahia, J. Hlel and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction," in IEEE Access, vol. 9, pp. 60447-60458, 2021, doi: 10.1109/ACCESS.2021.3074559.

[2] Al-Darraji, S.; Honi, D.G.; Fallucchi, F.; Abdulsada, A.I.; Giuliano, R.; Abdulmalik, H.A. “Employee Attrition Prediction Using Deep Neural Networks”. *Computers* **2021**, *10*, 141. <https://doi.org/10.3390/computers10110141>

[3] Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano and Ernesto William De Luca “Predicting Employee Attrition Using Machine Learning Techniques” , ResearchGate, 2020,

[4] A. K. Mishra, M. K. Pandey, and S. K. Singh, "Predicting Employee Attrition Using XGBoost," 2023 IEEE International Conference on Electrical Computer and Communication Engineering (ICECCE), pp. 1-4, 2023.