Final Project Written Report

Martavin McWilliams

MDS 620

Professor Andrews

September 25, 2025

Introduction

The Dataset I chose to perform an analysis on was the IBM HR Employee Dataset. It is a human resource dataset that details employees, their benefits as well as their habits and decisions. The data included a total of 1470 rows and thirty-five columns. I used this dataset to perform my final project and generate insight. The question I wanted to answer was “What factors are the leading cause to Attrition, employees leaving a company.”

Data Preprocessing and Exploration.

The primary variable that I focused on was Attrition. Attrition is important because it gives companies insight on potentially losing employees. First, I cleaned my data. I ran the info function and describe function to learn more about the dataset. I then cleaned my data by removing null values, dropping potential duplicates, and removing outliers to improve the accuracy in my findings. I found that monthly income was an outlier. I created a boxplot showing monthly income and attrition rate. I applied a filter to remove the outlier. I also cleaned inconsistent data by replacing some column’s category names with cleaner/smaller names. Next, I used the function count plot to create a bar chart for Attrition. The bar chart and boxplot revealed that monthly income and overtime were two factors that majorly affected an employee’s decision to leave a company. I then started setting up my data to run a logistic regression model. By splitting the data into training and testing sets, and also scaling the data, I was able to run the model. The Regression showed me the precision, recall, accuracy, confusion matrix, and F1-score of the model. The model had an accuracy score of 88%. Also, the precision score for employees who stayed (0.89) was better than that of employees who left (0.76). However, the score for both were over 70%, which means it was a decent score for predicting if an employee would stay or leave a company. I chose to use a logistic model because my target variable (attrition) was binary which makes it easier to predict future outcomes. To get more insight, I performed a cluster analysis. The cluster analysis was separated into three groups, cluster 0, cluster 1, and cluster 2. With a silhouette score of only 0.28 percent for the cluster analysis, the logistic regression model proved to perform the best. However, after evaluating the cluster, I again found that employees with low income and high overtime were a high-risk group of leaving a company. Finally, I set up a decision tree to show which variable strongly influenced attrition rate. At the top of the list of the decision tree was overtime as the number one cause of attrition.

Conclusion

After running and reading the dataset, I can predict that employees with low income and high overtime hours are at high risk of leaving a company. My proposal is for organizations to pay each individual what they deserve and cut down the workload to retain top talent.