

**CISC7107 DATA MINING AND DECISION SUPPORT SYSTEMS**

**Final Project**

**IBM HR EMPLOYEE ATTRITION ANALYSIS**

By: **JUSTICE ABBAN**

**MC353070**

**Introduction**

Employee attrition at IBM refers to the voluntary departure of employees from the company. It poses challenges such as increased costs and loss of knowledge. Factors influencing attrition include personal reasons, career opportunities, work-life balance, job satisfaction, compensation, and organizational culture. IBM addresses attrition through retention initiatives and data analytics to identify at-risk employees. The goal is to create a supportive work environment and reduce turnover rates.

**Objectives and Goal**

Using a mix of DM tools to Identify key attrition factors by analyzing various features and attributes in the dataset.

The goal is to build a predictive model that can forecast attrition risk for individual employees within the organization and also find the relationships between the factors that lead to attrition.

**Software’s used**

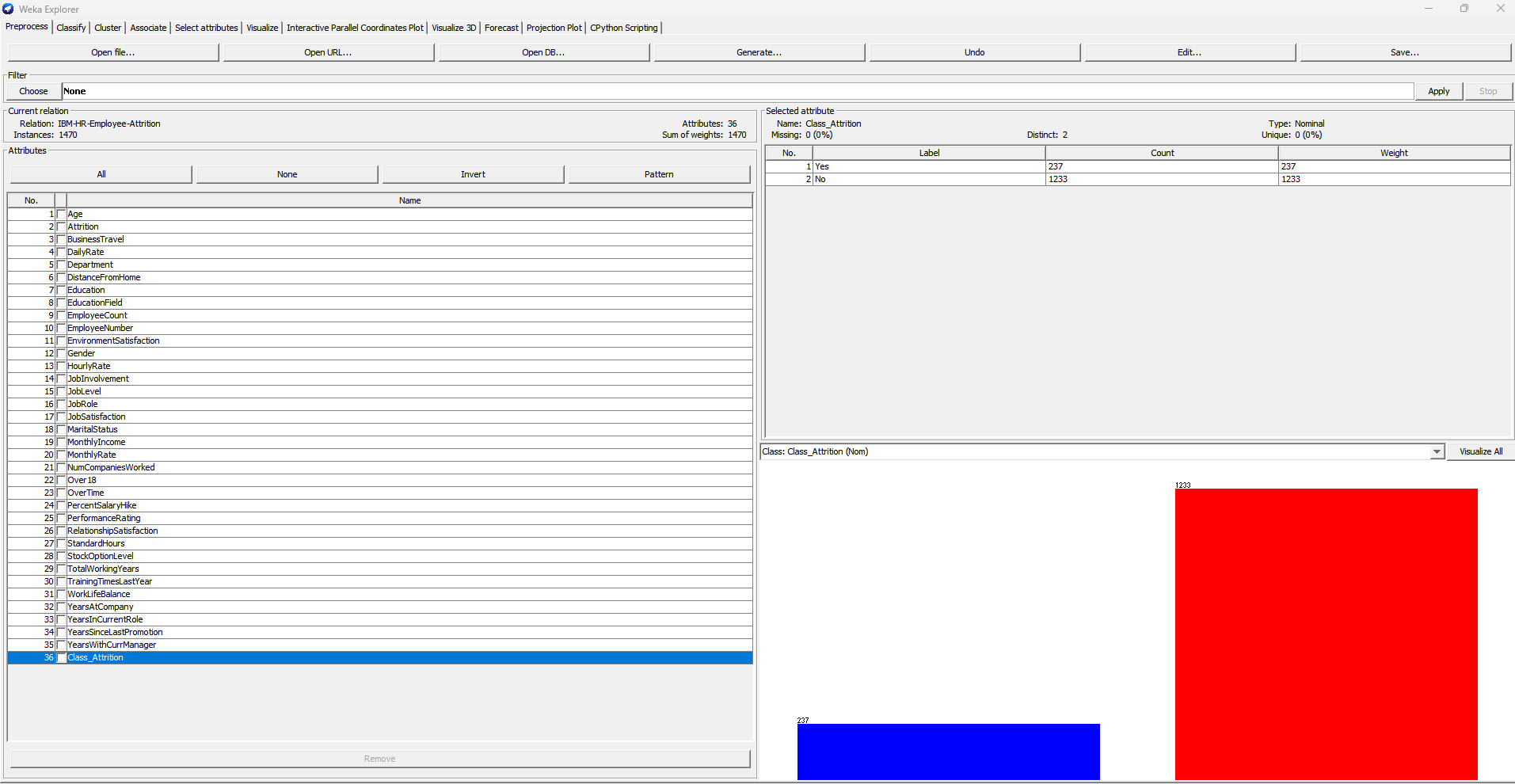
I used Weka for my Classification model, Text Mining, Feature Selection (Information Gain AE). I also used Orange for my generating decision Tree, Info Gain and also for Correlation.

**Dataset and Employee Review**

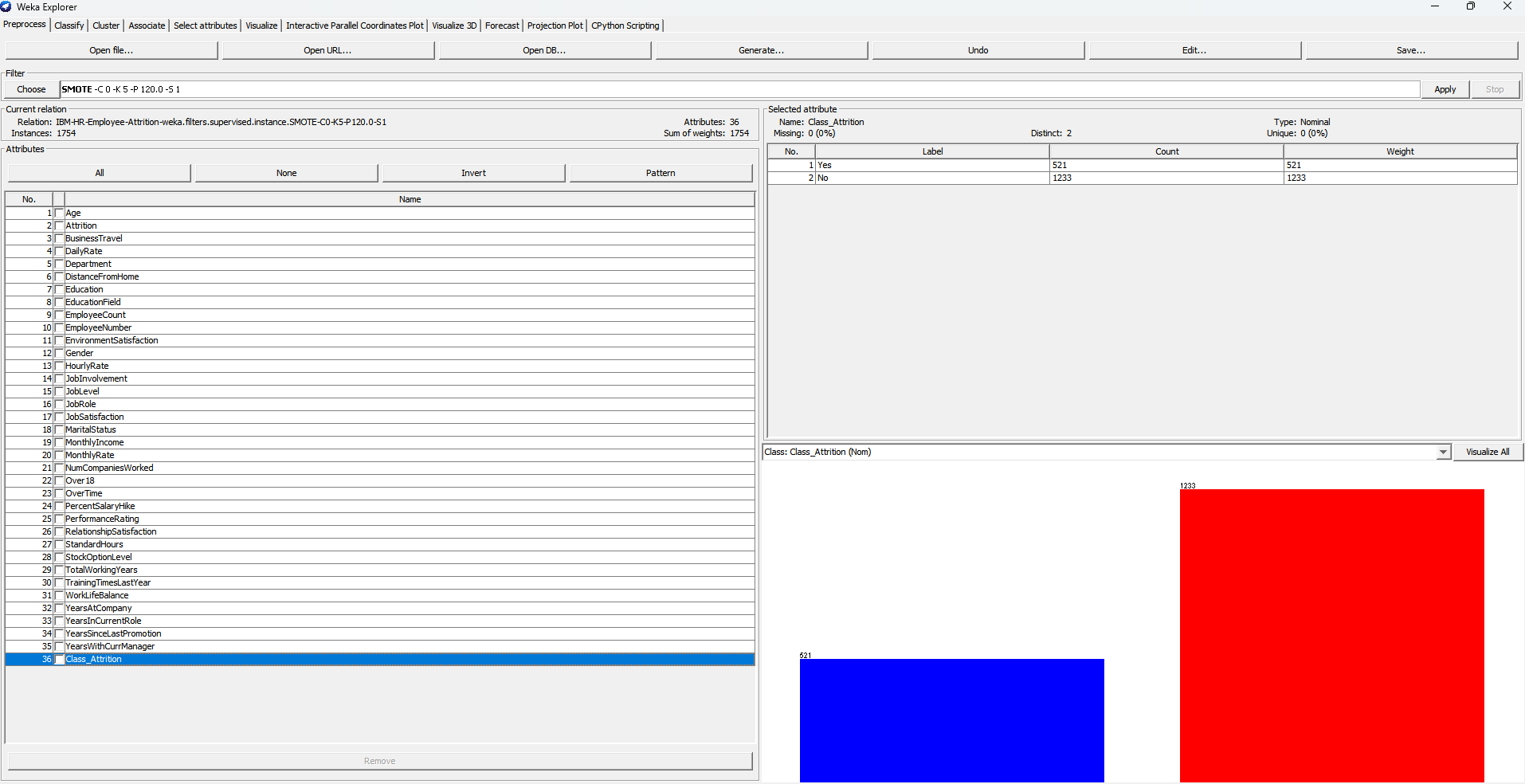
<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

[IBM Reviews: What Is It Like to Work At IBM? | Glassdoor](https://www.glassdoor.com/Reviews/IBM-Reviews-E354_P4.htm?filter.iso3Language=eng)

**Data Preprocessing**

**Figure 1. Imbalance Dataset**

**Figure 2. Balanced Dataset**



In this phase I used the SMOTE function to balance my Dataset for enhance me get an accurate model. The No class had fewer instances compared to the Yes so we had to do re balance the dataset.

**Figure 2. Extreme Values**

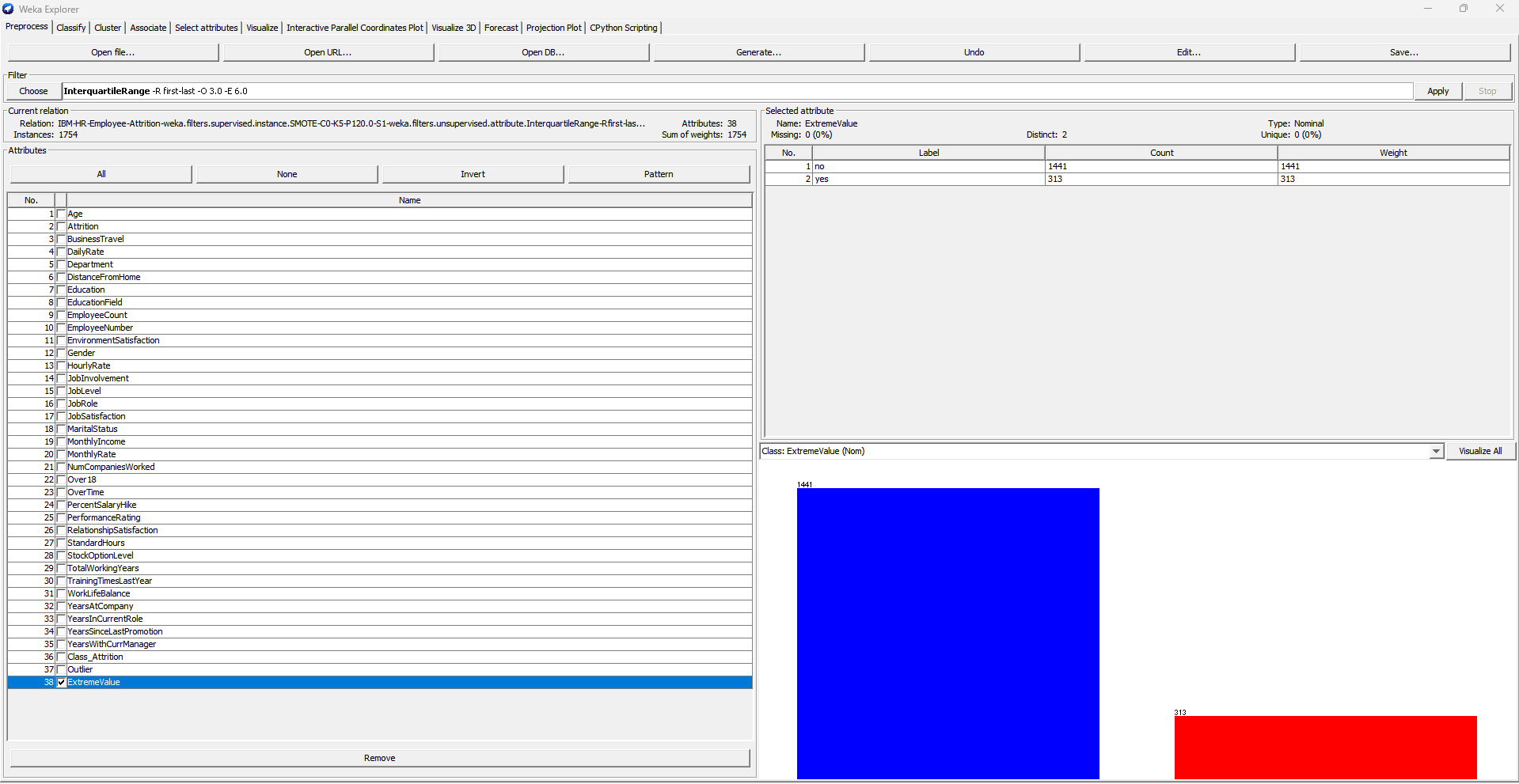
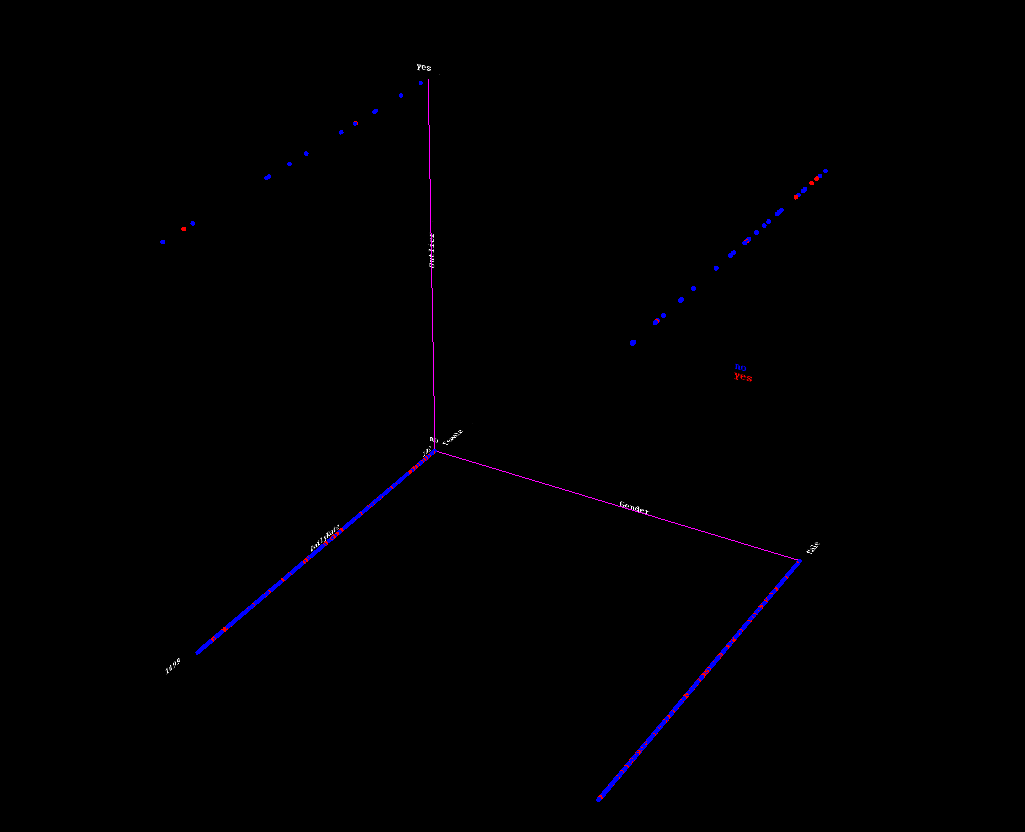


Figure 2 shows the Outliners and Extreme Values in my Dataset. Here I used the InterQuartileRange function to detect these outliners.

**Figure 3. Visualizing Outliners and Extreme Values in my Dataset**

Figure 3 below shows the detected outliners visualized in 3D. We visualized this because we wanted to really see where the outliners and extreme values were actually are before removing them and see if the changes will apply. On our X-axis was defined by the Gender, Y-axis Daily Rate, color- outliners and z-axis Extreme Values.



**Figure 4. Outliner Removed**

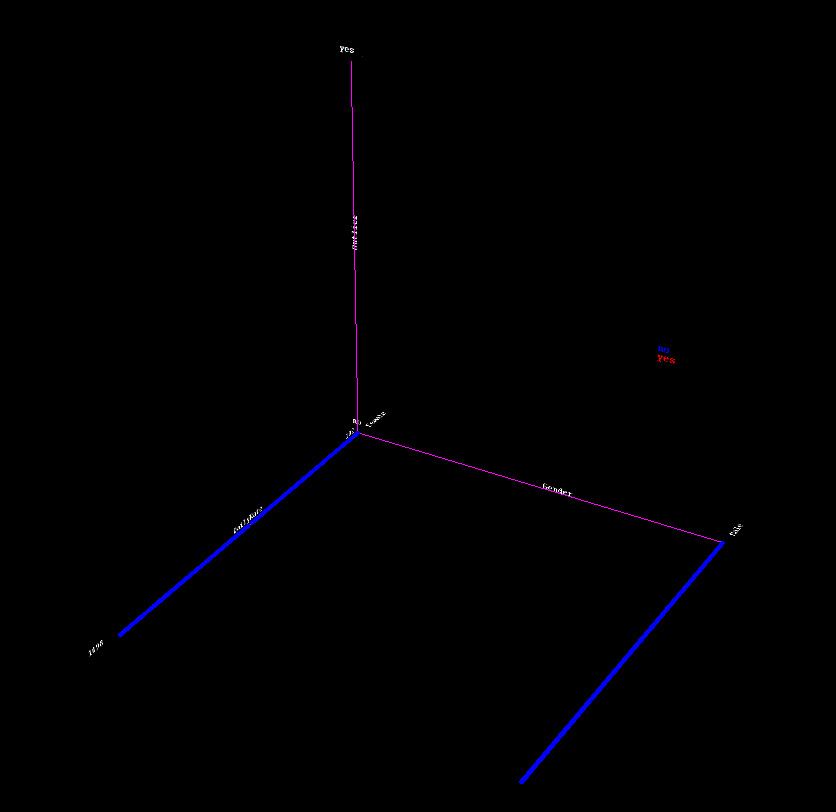


Figure 4 above the visuals of the removed outliners and Extreme Values. In this figure we can see that all the outliners in figure 3 above has been removed successfully.

**Building my Classification Model**

**Figure 5. Result of J48 on Use Training Set**

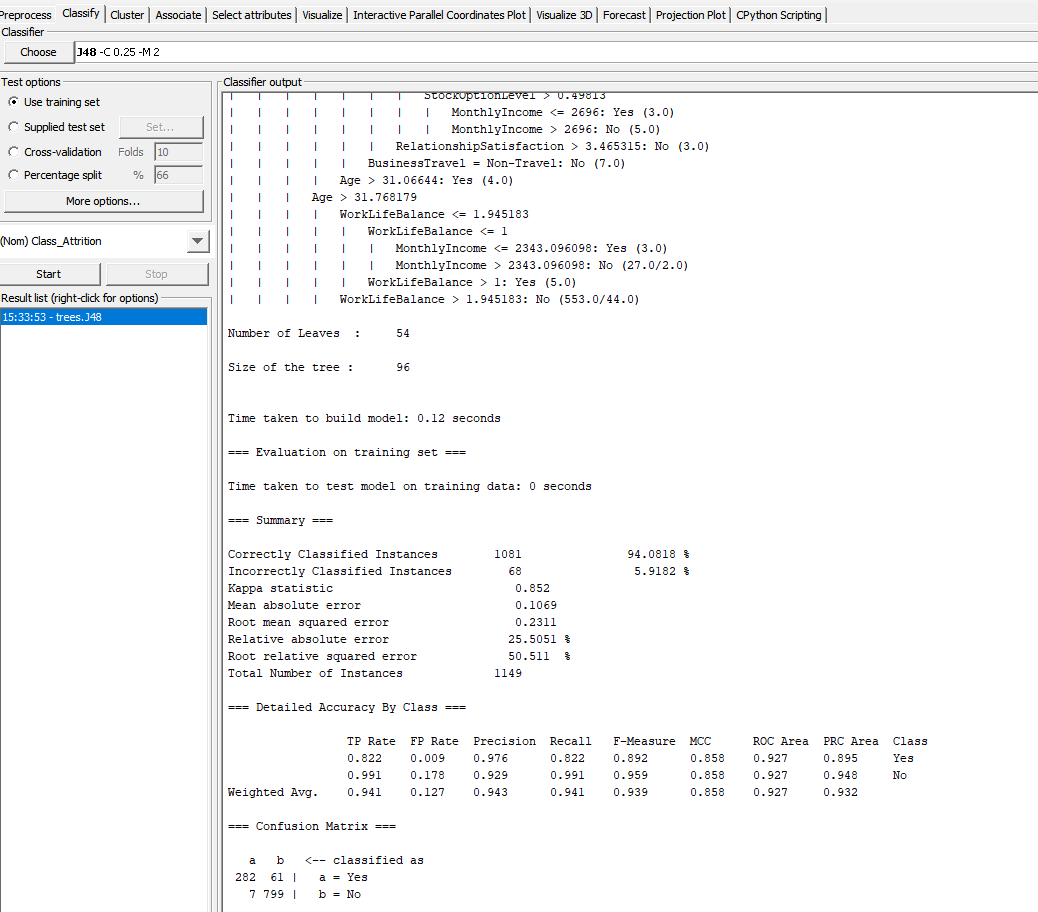


Figure 5 above shows the result of our classification model on the training set. Here we can see that the accuracy was 94.0818%, Kappa statistic – 0.852. We can also see from the Confusion Matrix that, the number of instances classified as Yes were 282 and misclassified was 61. Also, the number of instances classified as No were 799 and misclassified was 7.

**Figure 6. Result of J48 on Supplied Test Set**

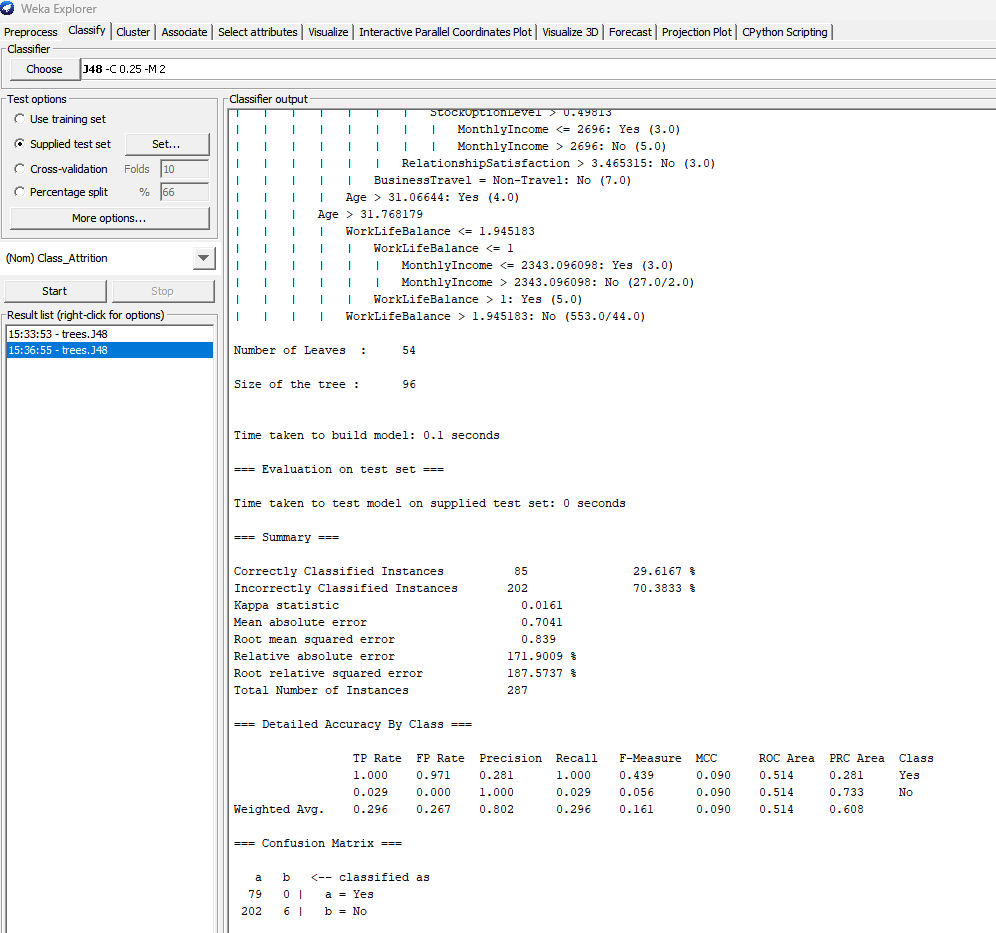


Figure 6 above shows the result of our classification model on the Supplied Test set. Here we can see that the accuracy was 29.6167%, Kappa statistic – 0.0161. We can also see from the Confusion Matrix that, number of instances classified as Yes were 79 and misclassified was 0. Also, the number of instances classified as No were 6 and misclassified was 202.

**Figure 7. Result of J48 on 10 Folds Cross-Validation**

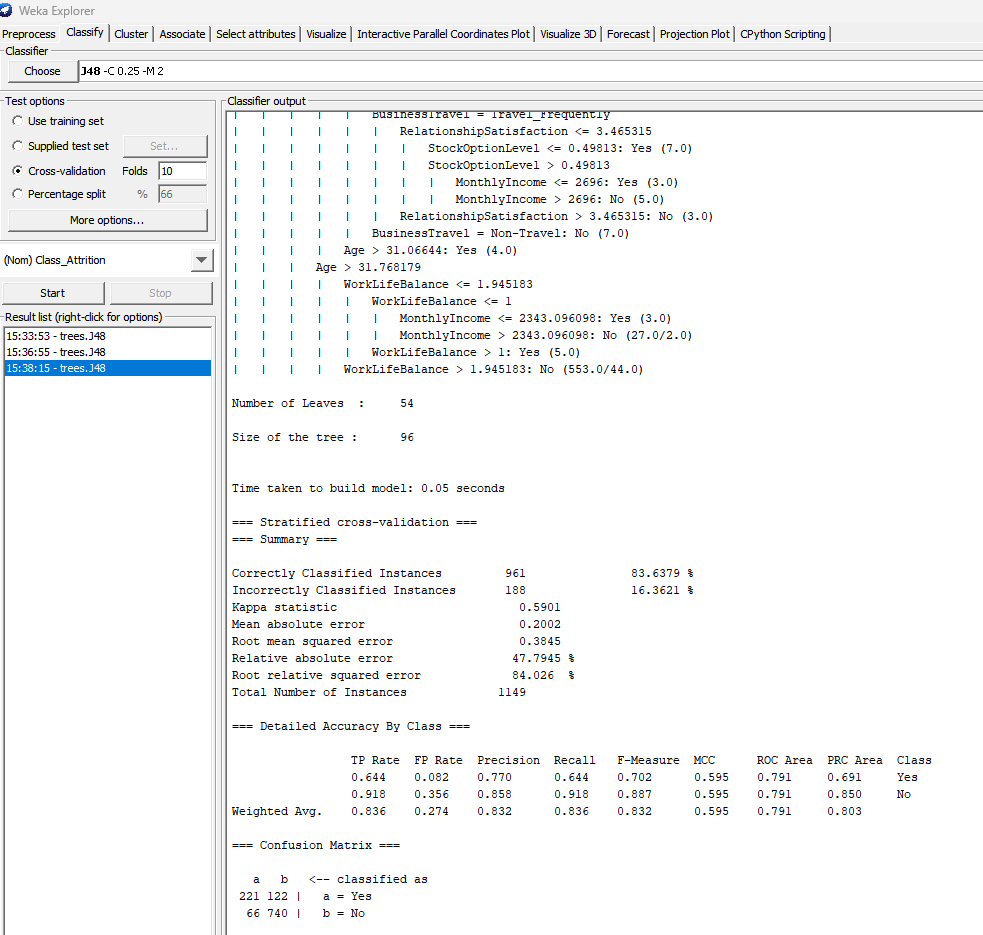


Figure 7 above shows the result of our classification model on the 10-Fold Cross Validation. Here we can see that the accuracy was 83.6379%, Kappa statistic – 0.5901. Looking at the Confusion Matrix, we can see that the number of instances classified as Yes were 221 and misclassified was 122. Also, the number of instances classified as No were 740 and misclassified was 66.

**Figure 8. Result of J48 on Percentage Split**

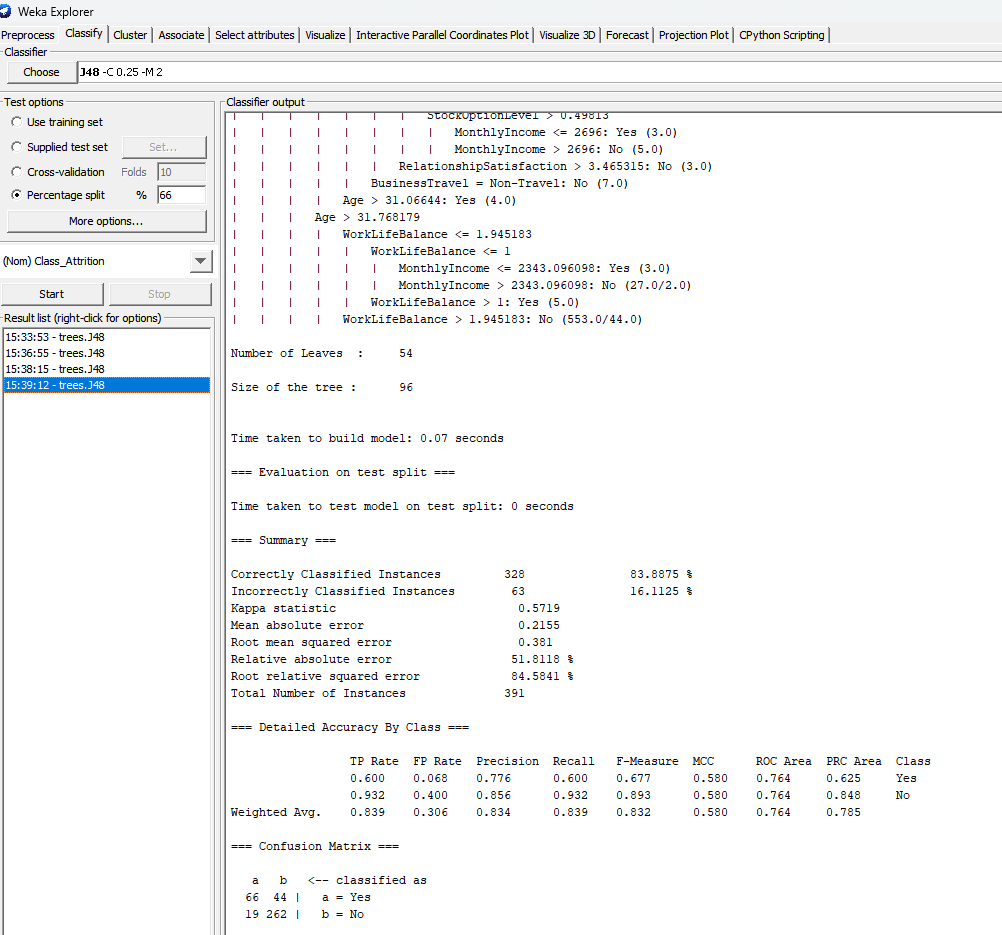


Figure 8 above shows the result of our classification model on using the percentage split. Here we can see that the accuracy was 83.8875%, Kappa statistic – 0.5719. Looking at the Confusion Matrix, we can see that the number of instances classified as Yes were 66 and misclassified was 44. Also, the number of instances classified as No were 262 and misclassified was 19.

**Figure 9. Final Classification Result**

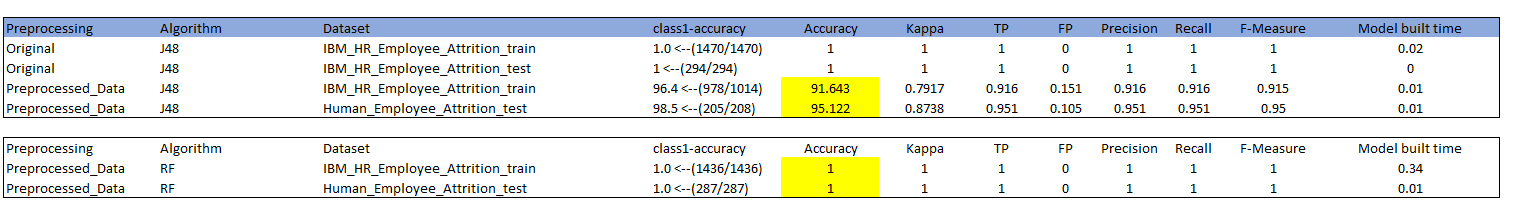


Figure 9 shows the Final Result of building our classification model on the raw dataset and also after preprocessing. In this Table we can see that preprocessing our data and running the test, the results indicated in yellow changed because the initial dataset was imbalanced and therefore we had to use the SMOTE function to balance it to have a non-bias result.

**Using the Attributed Selected Classifier**

**Figure 10. Result on Use Training Set**

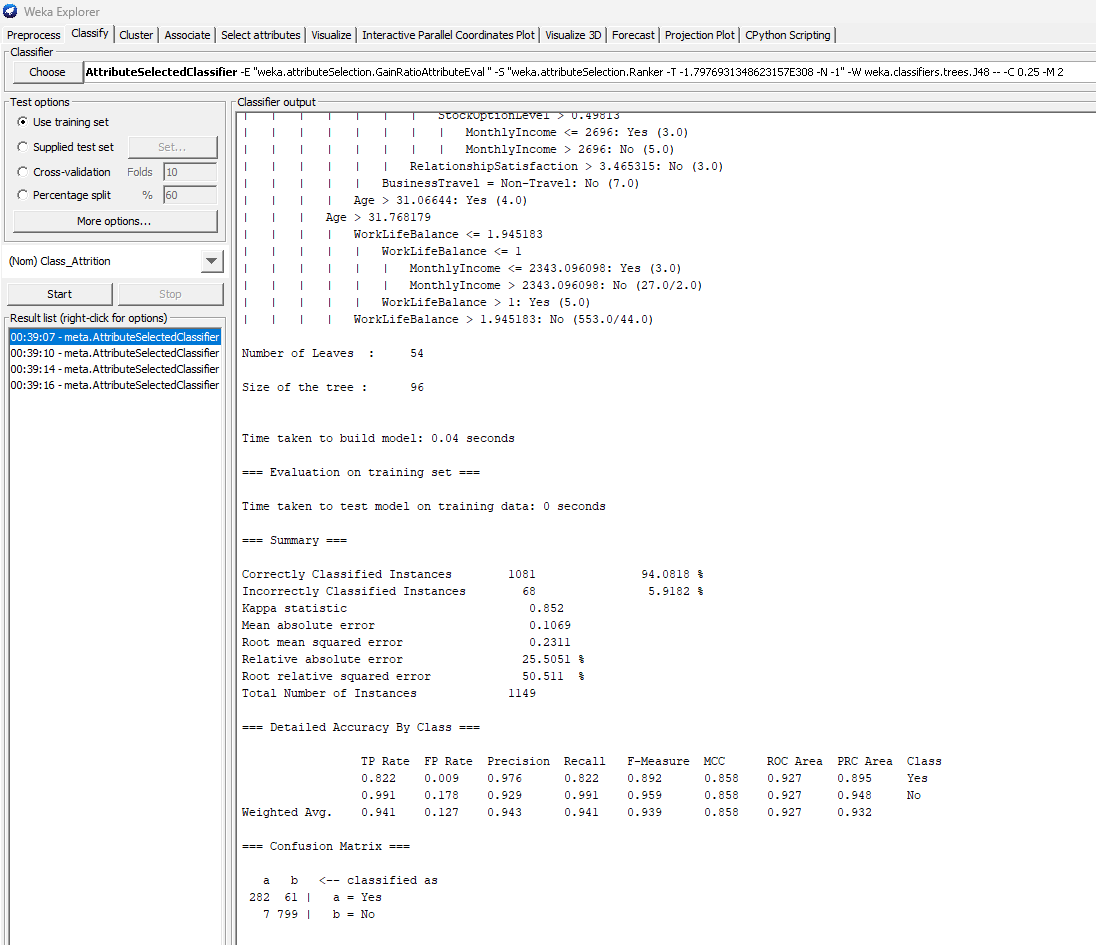


Figure 10 above shows the result of our classification model using the Attribute Selected Classifier and on the Use Training Set option. Here we can see that the accuracy was 94.0818%, Kappa statistic – 0.852. Looking at the Confusion Matrix, we can see that the number of instances classified as Yes were 282 and misclassified was 61. Also, the number of instances classified as No were 7 and misclassified was 799. We can see that our model didn’t really improve and therefore had to perform and explore some more test options.

**Figure 11. Result on Supplied Test Set**

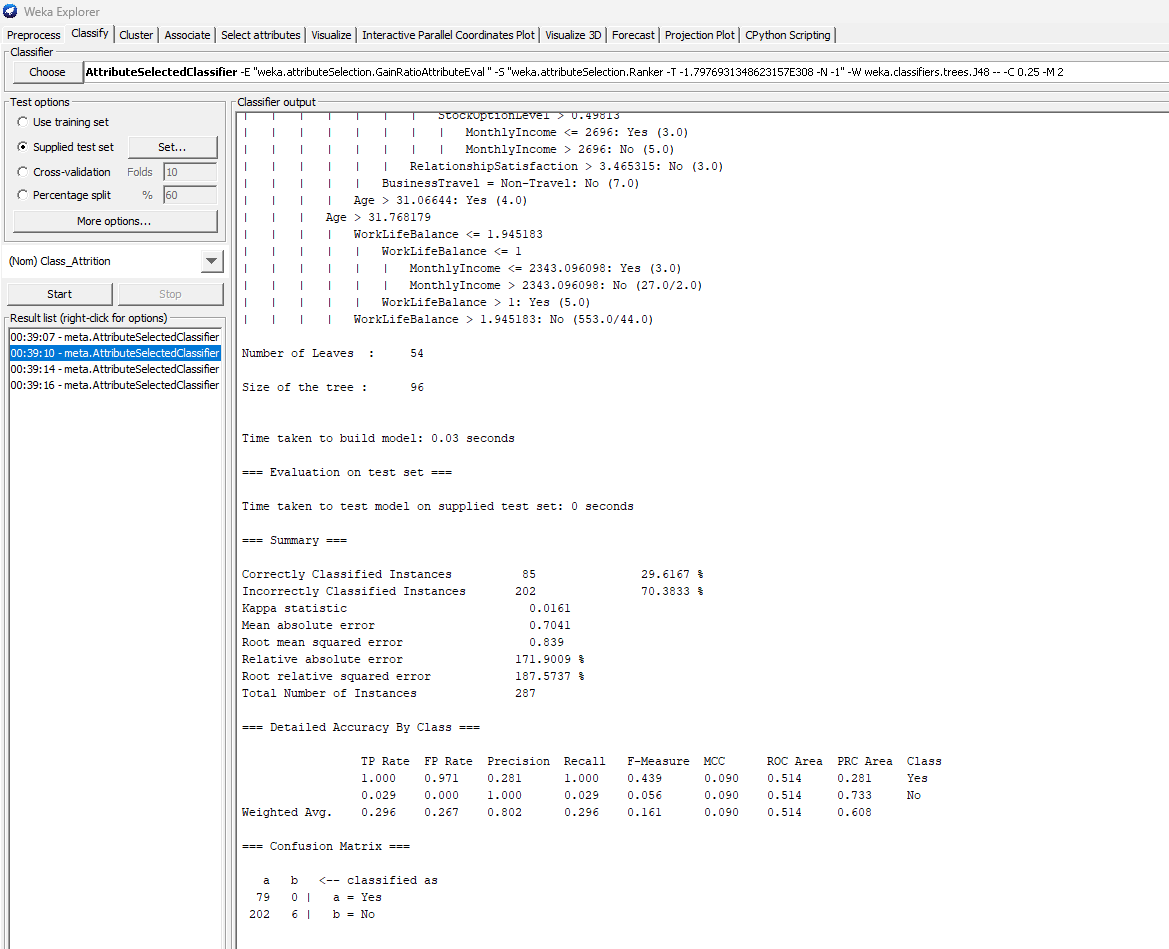


Figure 11 above shows the result of our classification model using the Attribute Selected Classifier and on the Supplied Test Set option. Here we can see that the accuracy was 29.6167%, Kappa statistic – 0.0161. Looking at the Confusion Matrix, we can see that the number of instances classified as Yes were 79 and misclassified was 0. Also, the number of instances classified as No were 6 and misclassified was 202. We can see that our model has a problem with the classification of No class and therefore had to explore some more test options.

**Figure 12. Result on 10 Folds Cross Validation**

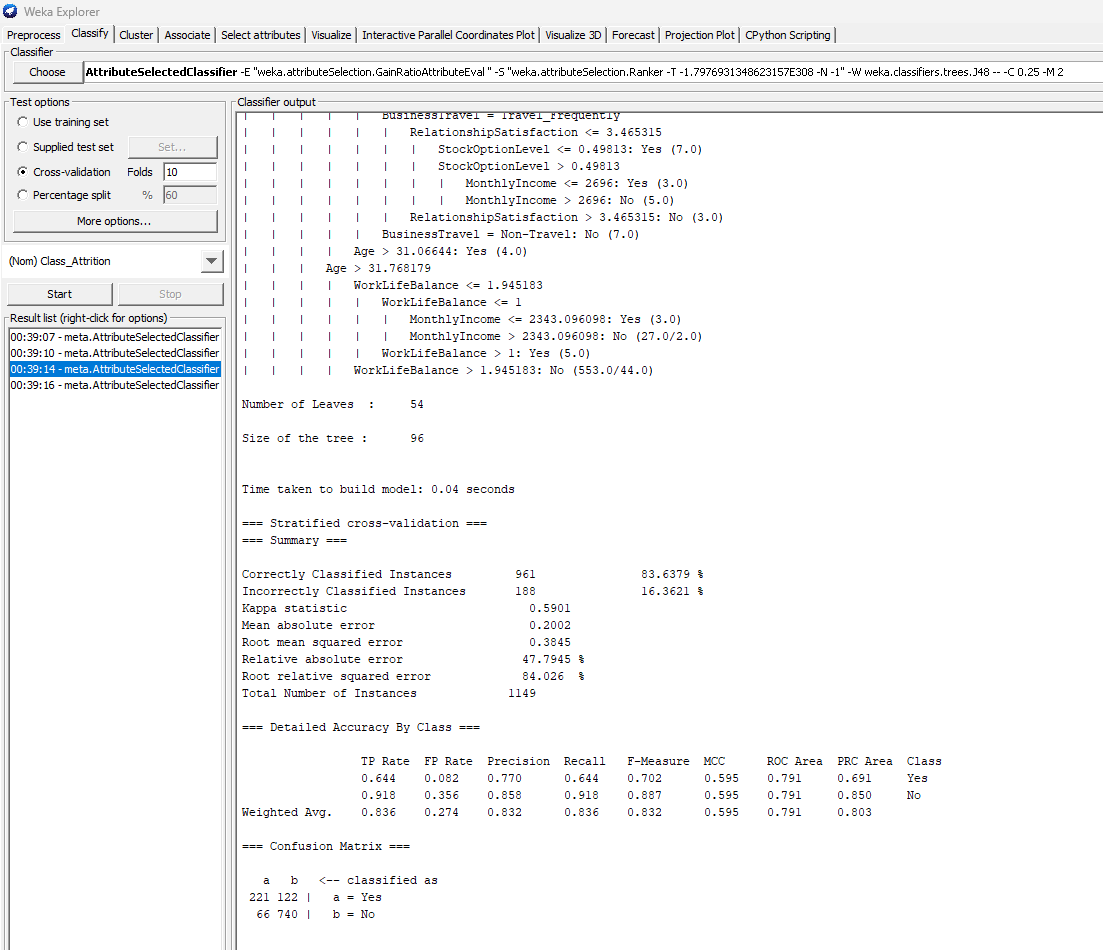


Figure 12 above shows the result of our classification model using the Attribute Selected Classifier and on the 10 Folds CV Test option. Here we can see that the accuracy was 83.6379%, Kappa statistic – 0.5901. Looking at the Confusion Matrix, we can see that the number of instances classified as Yes were 221 and misclassified was 122. Also, the number of instances classified as No were 66 and misclassified was 740.

**Figure 13. Result on Percentage Split**

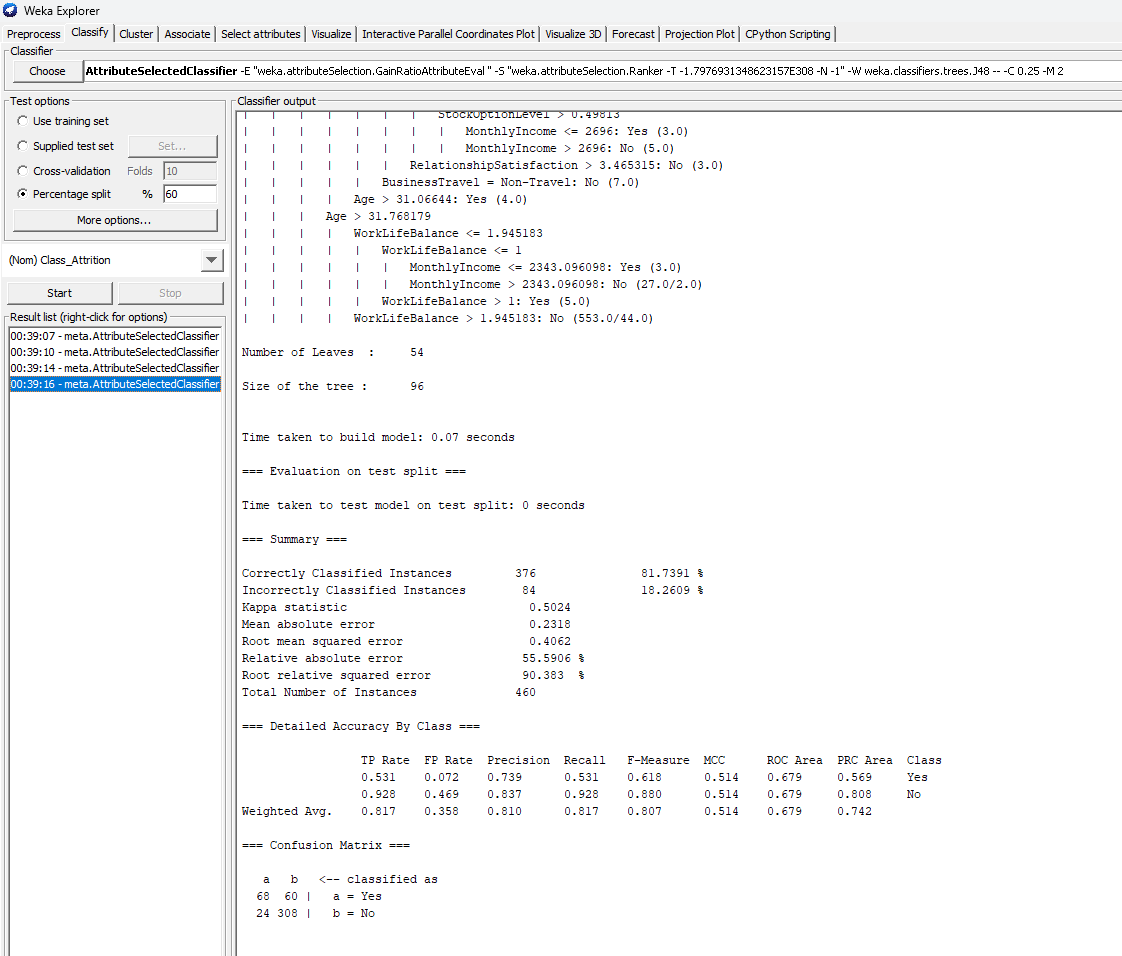


Figure 13 above shows the result of our classification model using the Attribute Selected Classifier and on the Percentage Split Test option. Here we can see that the accuracy was 81.7391%, Kappa statistic – 0.5024. Looking at the Confusion Matrix, we can see that the number of instances classified as Yes were 68 and misclassified was 60. Also, the number of instances classified as No were 308 and misclassified was 24. He we can see that our performance decreased comparing to the original J48 classification model.

**Figure 14. Final Results on all the Classified Model using the AttributeSelectedClassifier**

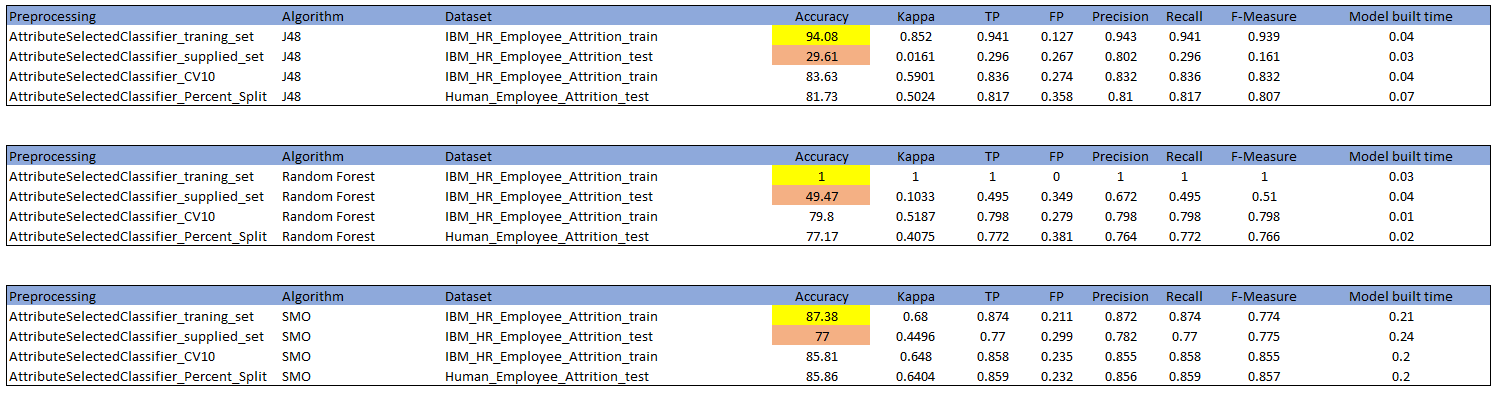
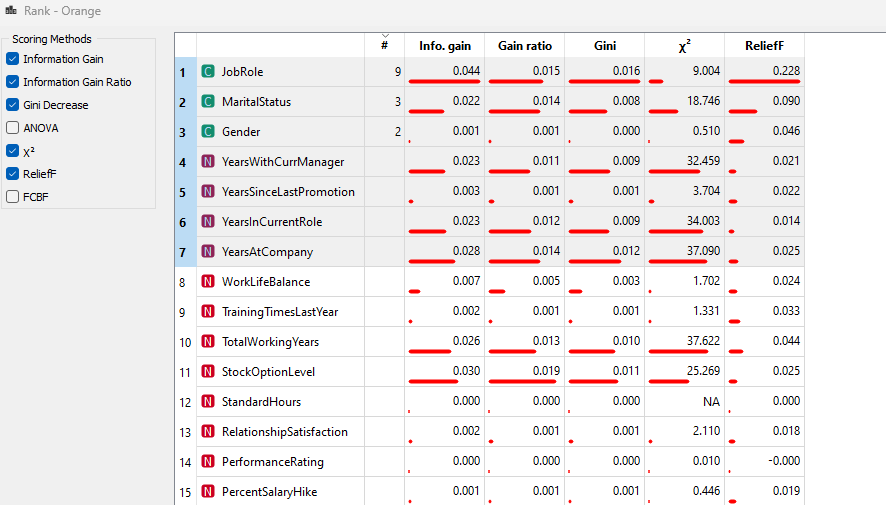


Figure 14 shows the Final Result of our classification model using the Attribute Selected Classifier on the preprocessed dataset. In this Table we can see that every algorithm has its best test options that best suit the dataset. In the J48 algorithm, we can see that test on the training test set had the best accuracy of 94.08%, Random Forest training on the test set also had the best accuracy of 1% and also on the SMO, and classification on the training set had the best accuracy of 87.38%. But one thing we can observe here is that when it comes to classification using the supplied test set, only SMO algorithm performed much well comparing to the other algorithms.

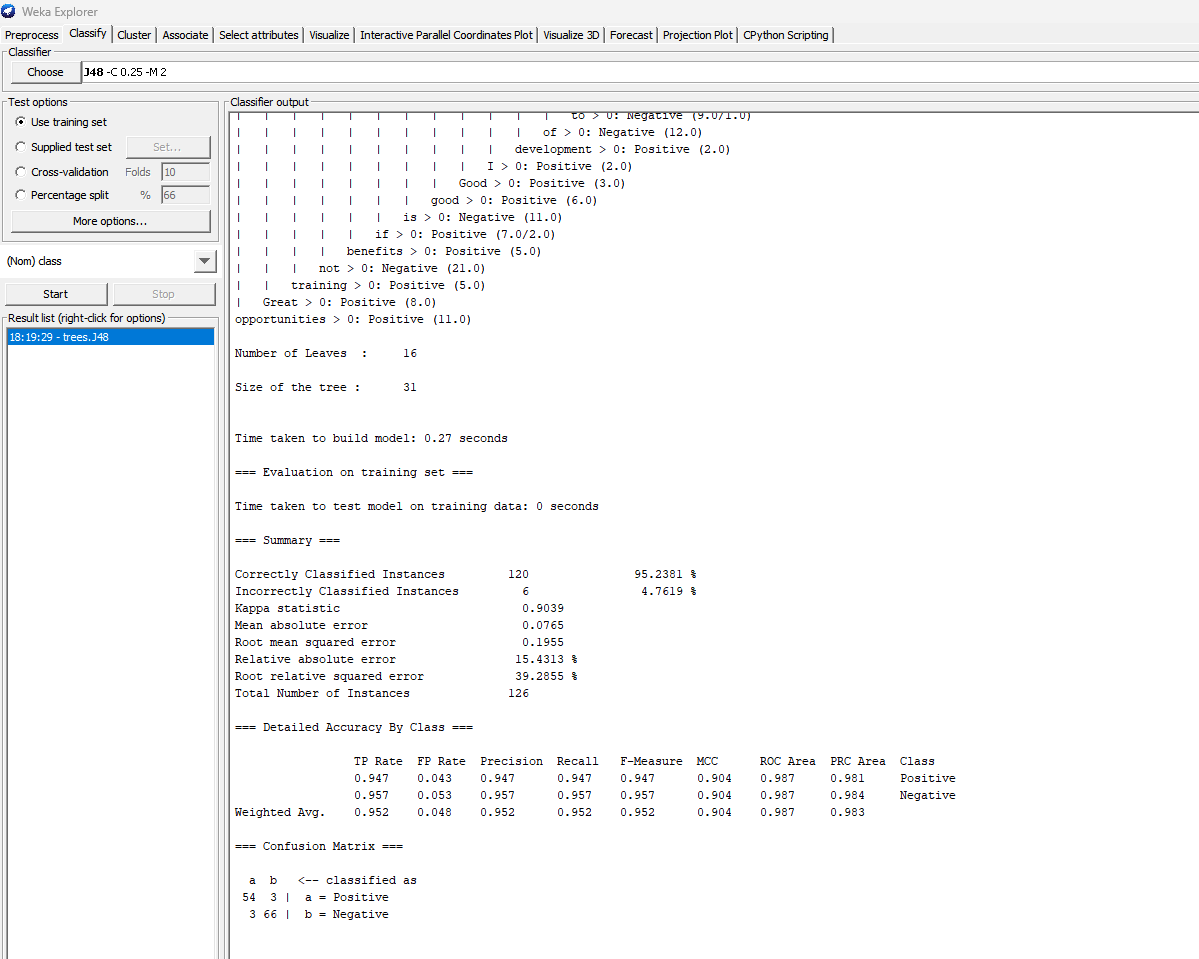
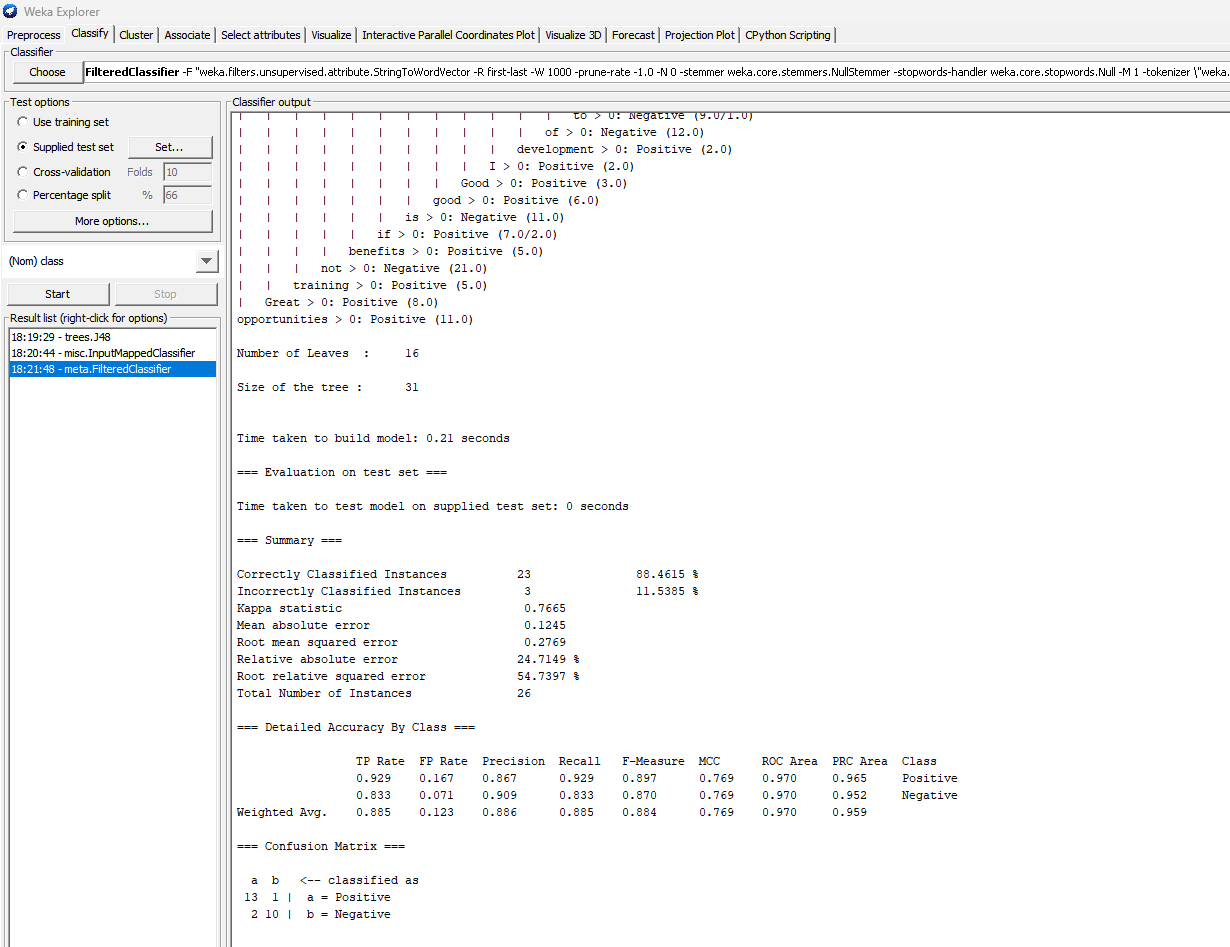
**Figure 15. Info Gain Using Orange**



In figure 15 above, we leverage on the Info Gain function in Orange software to find significant features that has much impact in our attributes. Here we can also observe that JobRole has the highest number in our figure. This figure shows us the best to 7 attributes in our features.

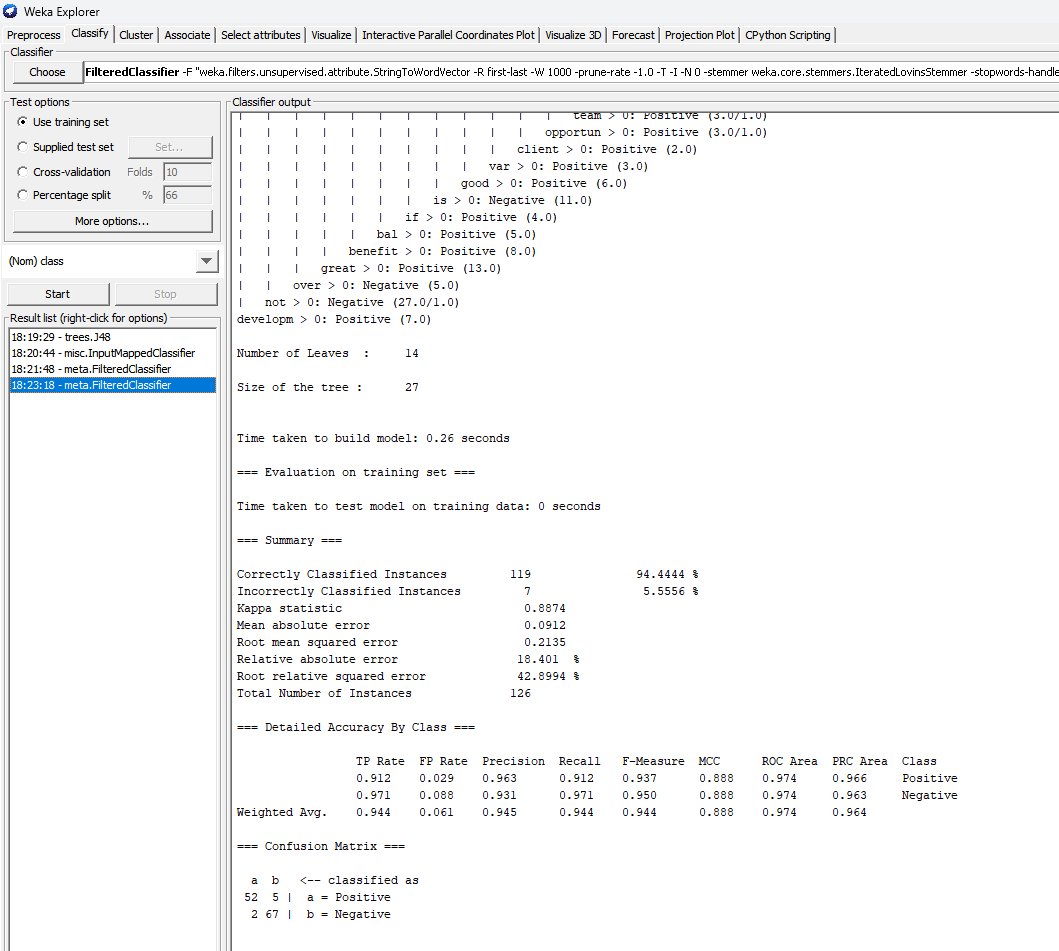
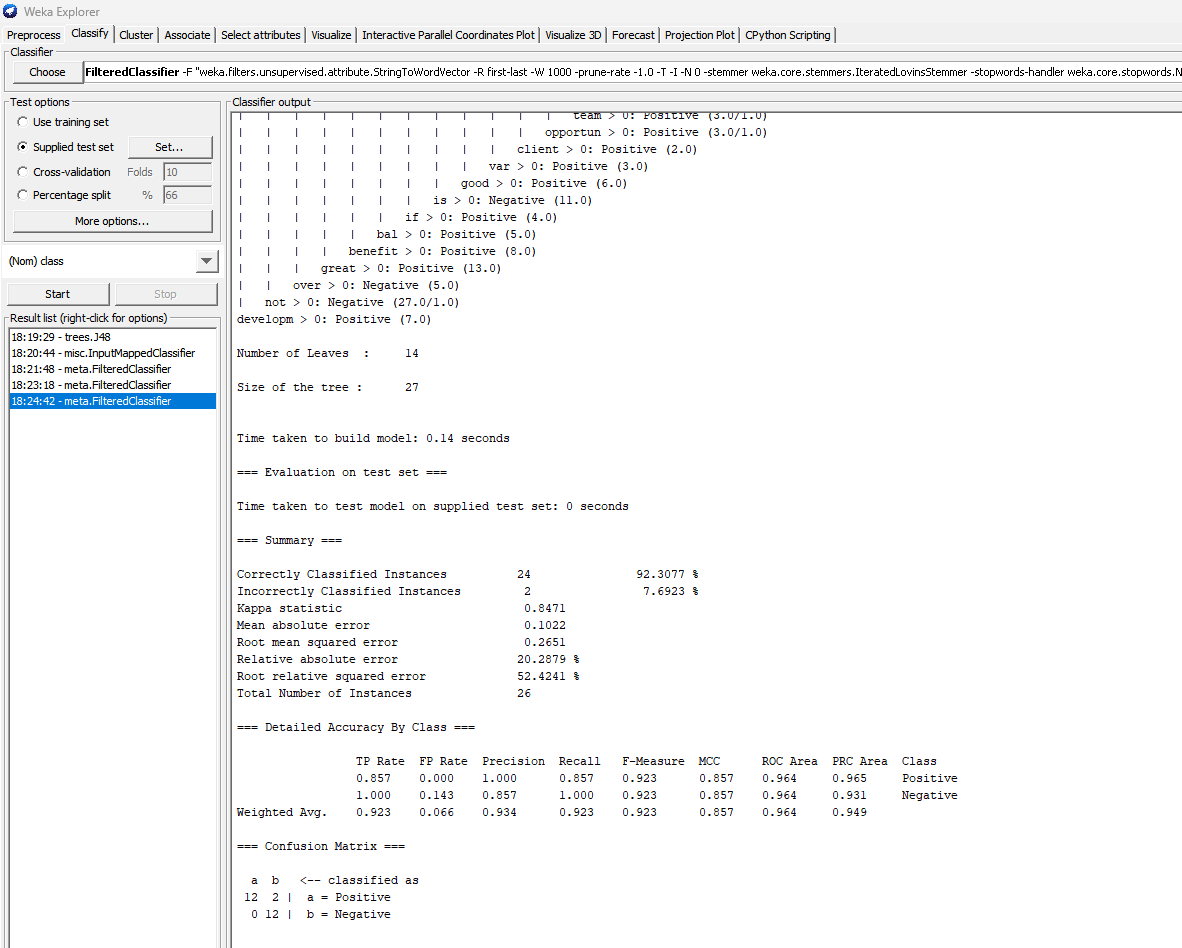
**Text Mining and Visual Mining**

Figure 16. Original Dataset on J48 Use Training Set and Supplied Test Set



The first figure shows the original training test performance using J48. In this experiment the accuracy was 95.2381%. The second figure shows themodel performance for the supplied test set which was 88.4615%.

**Figure 17. Optimized Dataset using FilteredClassifier on Use training set & Supplied Test**



The first figure shows the optimized result of training test performance using the FilteredClassifier Attribute. In this experiment the accuracy decreased to 94.4444 which is 10% of the original experiment 95.2381%. The second figure shows the accuracy after optimization using the FilteredClassifier on the Supplied Test set. In this experiment, the accuracy increased from 88.4615% to 92.3077% which was a great improvement.

**Figure 18. Decision Tree of J48 Original Dataset on Training Set**

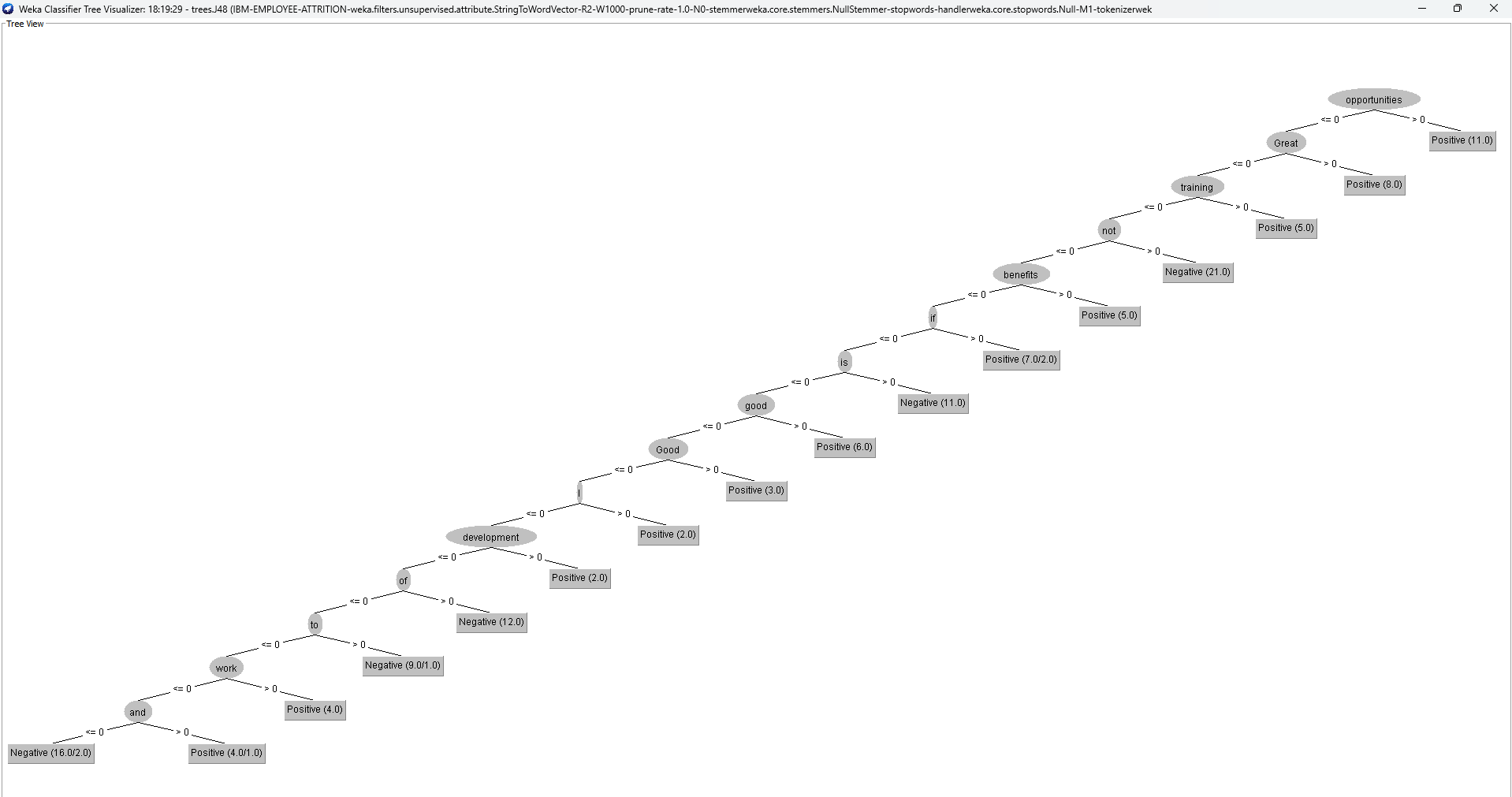


Figure 18 is a visualized decision tree of our original model. Here we can see the number of nodes are 14 with opportunities as the most relevant term. Each node in the decision tree corresponds to a term, and the decision rules at each node determine the path to follow based on the presence or absence of that feature in the document.

**Figure 19. Decision Tree of J48 Optimized Dataset on Training Set**

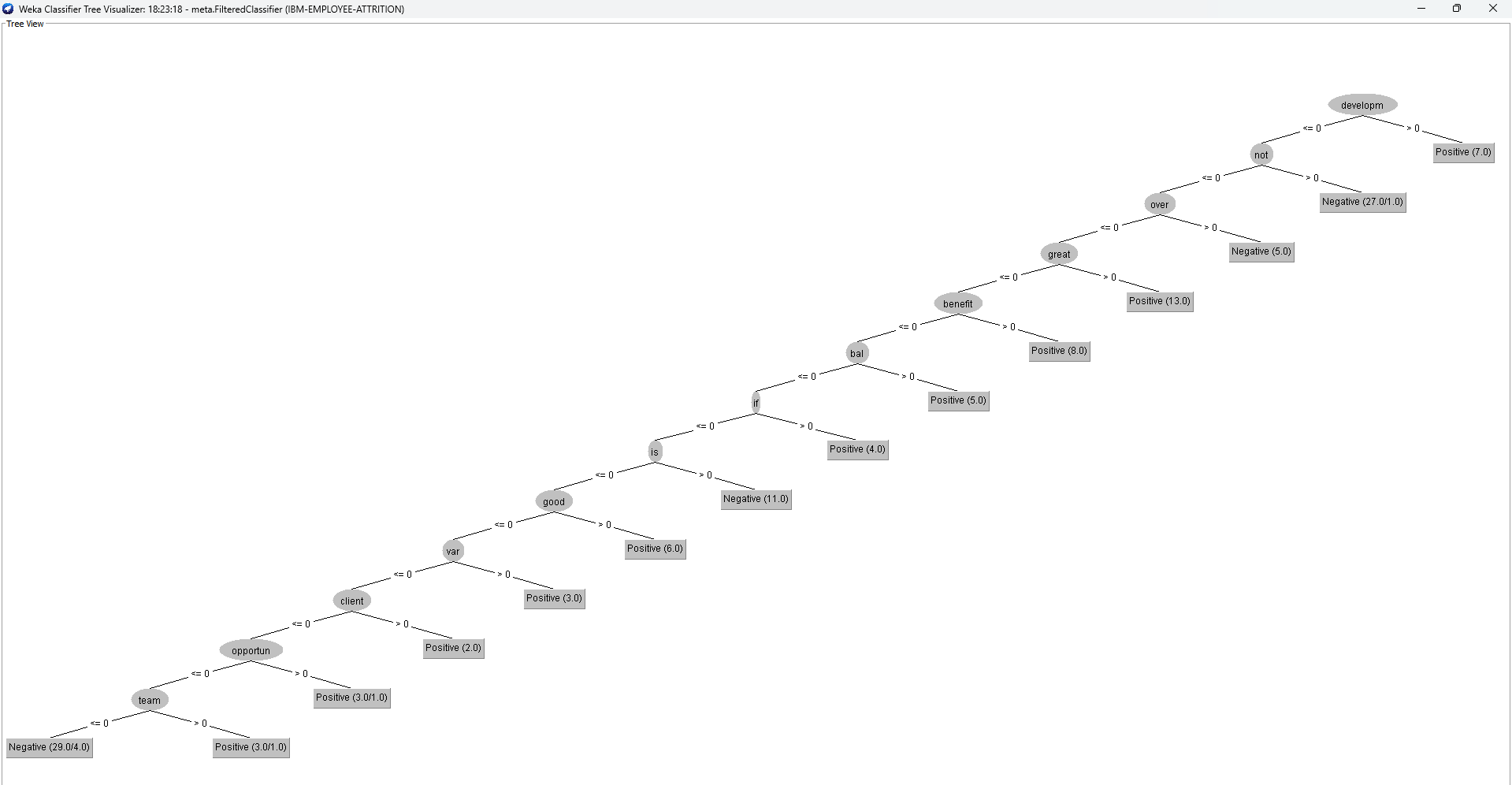


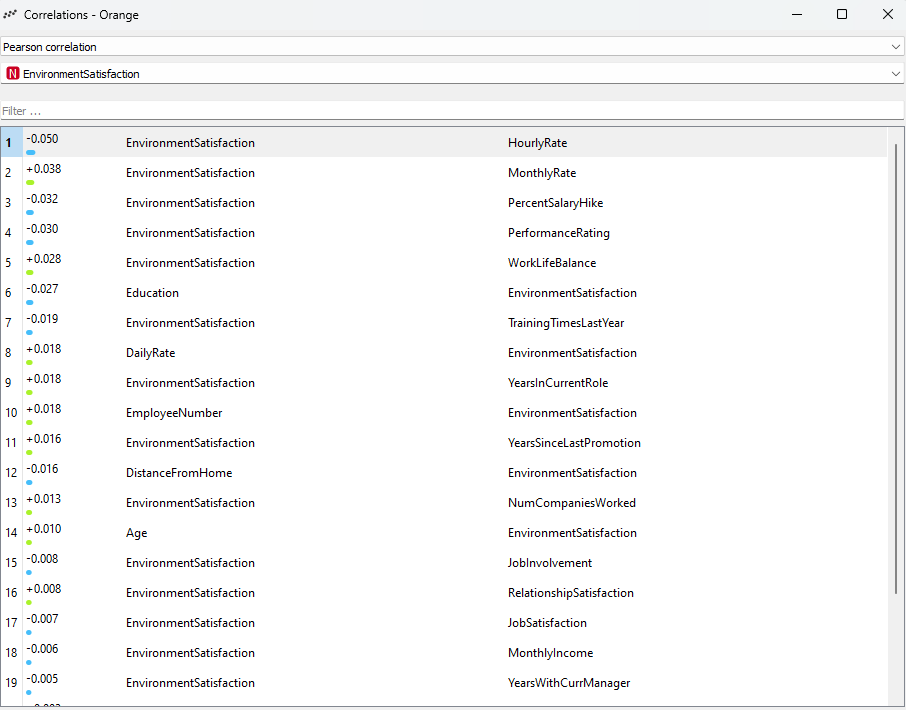
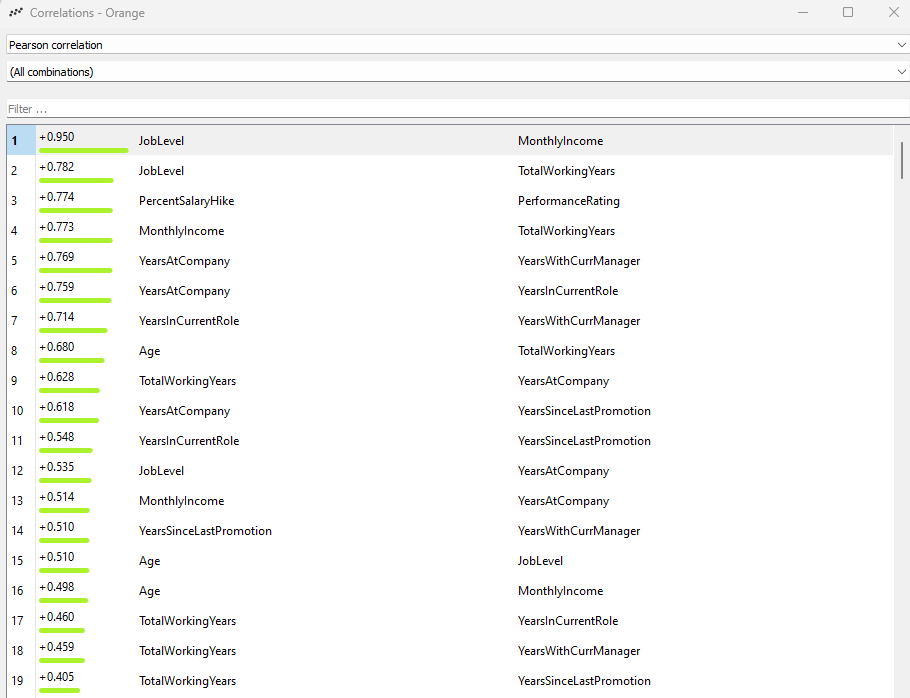
Figure 19 above shows the visualized decision tree of our optimized model. Here we can see the number of nodes reduced 13 with development as the most important term in our text. We can conclude that after the optimization our decision tree reduced by one node.

**Figure 20. Visual Mining**



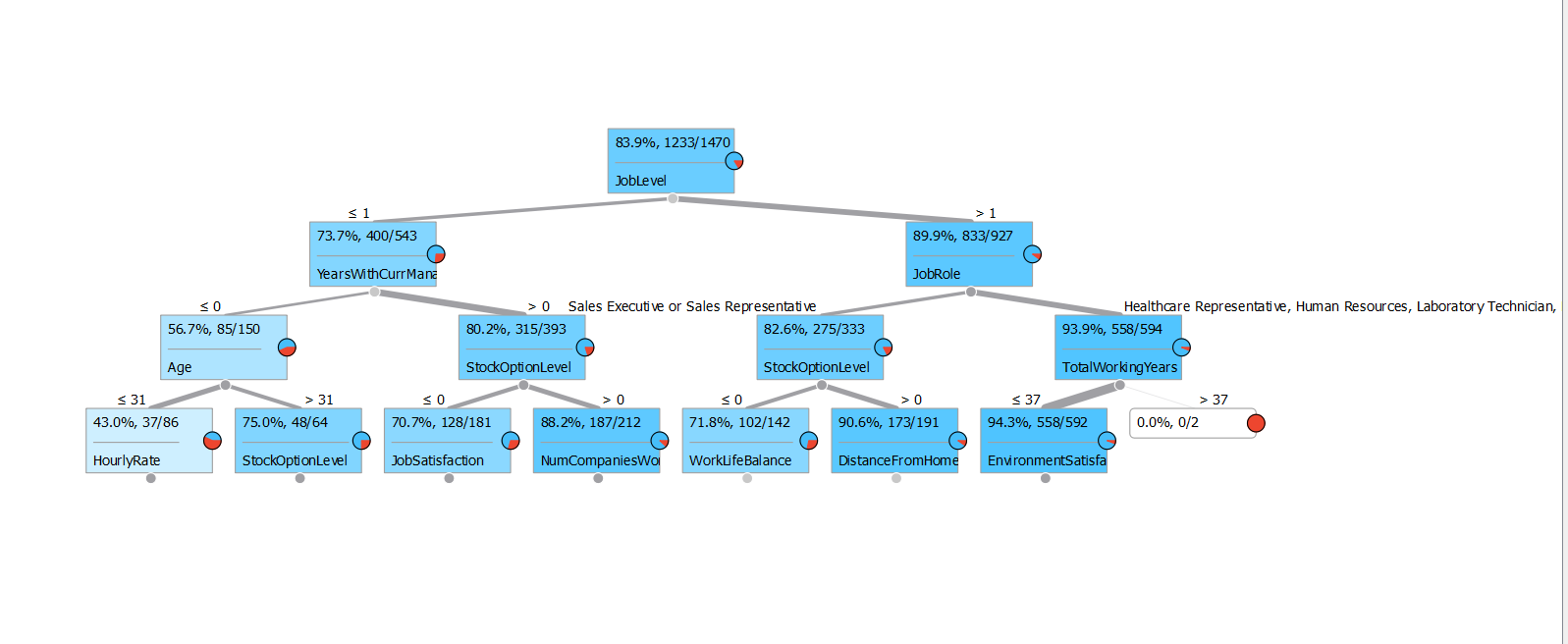
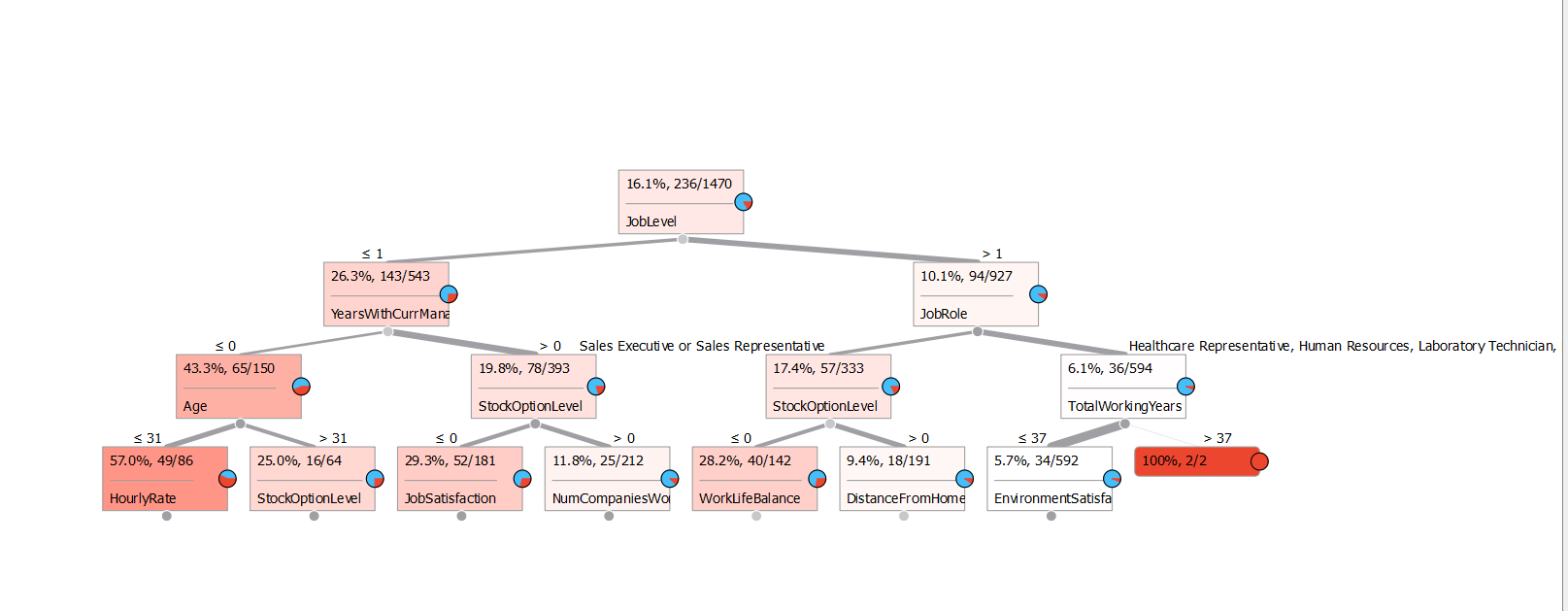
In figure 20, we used Visual Mining techniques to helps us identifying patterns, the positive and negative comments by using word cloud to visually represent, analyze and quickly recognize patterns that may require further investigation, insights and hypotheses leading to better understanding and decision-making.

**Figure 21. Correlation**



In figure 21, we can see the correlation between two different features. This assisted us in the feature selection process by identifying the relationship between our input features and the target variable (class label). It also helped us determine which features had a stronger association with the target variable and are more likely to contribute to the building of the classification model.

**Figure 22. Decision Tree**



In this decision tree we can see that JobLevel is the best predictor to classify the Attrition Rate(**NO class**). Followed by YearsWithCurrManager and JobRole. This Decision Tree helps us know the important features within our dataset. Each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or class label.

**Conclusion**

According to my classification model results, we can know that our finding are in line with people’s behavior in the real world. I utilized Random Forest, J48 and SMO classifier to build my model. Random Forest had the best result of **100%** accuracy and prediction, followed by J48 with **94%** accuracy. I used Text Mining techniques to again find the relationships between the features to fish out some key attributes that can possibly lead to iteration in IBM. I noticed that **MontlyIncome, JobInvolvement, EnvironmentSatisfaction, Relationship Satisfaction, YearsWithCurManager and YearsIncurrent role plays a vital role when it comes to employee attrition.** Then I found Age, high job level, high job satisfaction, high monthly income, number of years worked with the company and also years worked with current manger, these kinds of people are not likely to leave. However, different people have various intention, we need to do further and detailed analysis to find relationships by using correlation and attribute gain. In my correlation analysis, **I found out that JobLevel and MonthlyIncome, JobLevel and TotalWorking Years, Years at Company and Years With Company, Age and JobLevel, Age and MonthlyIncome were highly correlated. This means that people are likely to resign if they don’t fall into these categories or find themselves lower than the category.** Finally, there are other interesting findings in my analysis in terms of number of companies worked, people who worked in 2 - 4 companies are less likely to leave. To evaluate the model performance, we trained and tested the dataset to predict the employee attrition, split it into two parts (80% for training, 20% for testing), and recorded the test set’s accuracy. Random Forest and J48 accuracy were 1.0 and 0.94, respectively, which meant J48 fitted better and was more suitable for prediction in our dataset.

**Recommendation**

Since attrition has been a serious problem for IBM, data analysts would like to estimate whether their employees are leaving IBM to avoid further increases in the attrition rate and keep the employees in their positions. With the data collected recording the employees’ basic information, classification models can be built to estimate their attrition status of them, and by looking into the results, analysts will be able to figure out the reasons why some of the employees have left and provide better working conditions to prevent others from resignation. Also I want to make some suggestions to the company through this analysis, hoping that they will care more about their employees and improve their job satisfaction. The company should allow employees to have enough time to rest and spend time with their families. There is a general belief that employees who take regular breaks are more productive.