EMPLOYEE ATTRITION PREDICTION

AVINASH SANKAR-A20525471
JAICHIDDHARTH R. KARTHIGEYAN-A20527281

ABSTRACT

- Organizations are very concerned about employee attrition because it leads to the loss of important talent, increased expenses, and diminished productivity. Organizations may take proactive steps to keep their best performers and lower attrition rates by using employee attrition prediction. In this project, we created prediction models for employee attrition using logistic regression, decision trees, random forests, and support vector machines. Using a variety of performance criteria, we assessed the models' performance and determined which model was most effective at predicting employee attrition

DATA COLLECTION

Finding the Dataset: The first step in our investigation was to locate the right dataset. On Kaggle, we looked for employee attrition-related datasets, and there we discovered the IBM HR Analytics Employee Attrition & Performance dataset.

Dataset Downloading: We located the dataset and downloaded it from the Kaggle website. The dataset was offered in CSV format.

Link of the dataset:” https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset”

DATA PRE-PROCESSING

Any project involving data analysis must start with data preprocessing. For this research, we underwent several preparation procedures to get the dataset ready for analysis.

Removing Unnecessary Variables: We started by getting rid of any variables that weren't important for our analysis. As a result, the dataset's complexity was decreased, and the classification models' performance was enhanced.

Identifying Missing Values: We looked for any missing values in the dataset.

Looking for Outliers: We looked for any outliers in the dataset. We had to recognize the outliers and deal with them correctly because they could affect the analysis. To find any outliers in the dataset, we employed summary statistics and box plots.

DATA PRE-PROCESSING

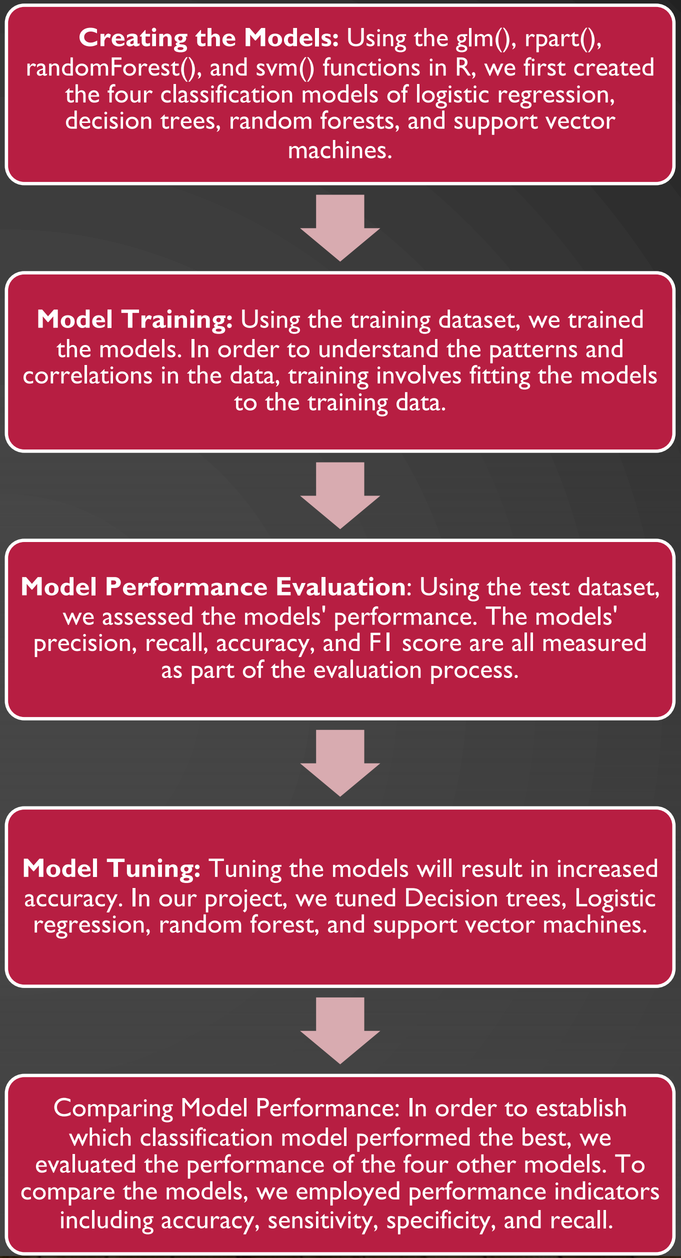
Handling Outliers: After locating the outliers, we took the proper action. For numerical variables, we made the decision to substitute the median for the outliers. As a result, the influence of the outliers on the analysis was not observed as there are no outliers.

Scaling Numerical Variables: In order to make sure that the numerical variables were on the same scale, we scaled them. Scaling is a typical machine learning approach that makes sure that each variable is handled equally throughout the investigation

Converting Categorical Variables into Factors: We changed categorical variables into factors to improve the analysis of the data by the classification models.

Splitting the Dataset into Training and Test Sets: Lastly, we created training and test sets from the dataset.

BUILDING MODEL



Creating the Models: Using the glm(), rpart(), randomForest(), and svm() functions in R, we first created the four classification models of logistic regression, decision trees, random forests, and support vector machines.

Model Training: Using the training dataset, we trained the models. In order to understand the patterns and correlations in the data, training involves fitting the models to the training data.

Model Performance Evaluation: Using the test dataset, we assessed the models' performance. The models' precision, recall, accuracy, and F1 score are all measured as part of the evaluation process.

Model Tuning: Tuning the models will result in increased accuracy. In our project, we tuned Decision trees, Logistic regression, random forest, and support vector machines.

Comparing Model Performance: In order to establish which classification model performed the best, we evaluated the performance of the four other models. To compare the models, we employed performance indicators including accuracy, sensitivity, specificity, and recall.

- **Model Selection:** Finally, we chose the best classification model based on how well it performed. As the top model, we selected the one with the highest accuracy, sensitivity, specificity, and recall.

MODEL EVALUATION

- A crucial phase of every data analytics project is model assessment. In this study, we used a variety of performance indicators to assess the effectiveness of the four classification models: logistic regression, decision trees, random forests, and support vector machines.
- **Performance metrics:** Performance indicators including accuracy, sensitivity, specificity, and recall offer a quantitative assessment of the model's effectiveness.

LOGISTIC REGRESSION:

- The logistic regression model accurately identifies 86.59% of the employees as either likely to go or likely to stay, according to its accuracy score of 86.59%.
- The algorithm successfully predicts 32% of the employees are likely to quit, according to the recall score of 0.32. The model's overall performance is the highest among all, as indicated by the F1-score of 0.43

DECISION TREE

- Compared to the logistic regression model, the decision tree model's accuracy is 21%, which is less accurate.
- The model's recall score is 0.8, which is higher than logistic regression(0.32), showing that the model's predictions are less accurate than those of the logistic regression model.
- The precision score for the model is 0.22. The model performs moderately all around, although not better than the logistic regression model.

RANDOM FOREST:

- The accuracy of the random forest model is 16%, which is higher than that of the logistic regression model.
- The recall score is 1, which is comparable to the logistic regression model, and the precision score is 0.16.
- The model's overall performance is low, according to the F1-score of 0.277, which is somewhat lower than that of the logistic regression model.
- It exhibits the lowest performance.

SUPPORT VECTOR MACHINES

- The accuracy of the SVM model is 16 %, which is comparable to the accuracy of the logistic regression model.
- Its recall score is lower than the logistic regression model, at 1, and its accuracy score is lower than the logistic regression model, at 0.16.
- The model's overall performance is considered to be good by the F1-score of 0.277, which is similar to the logistic regression model. It exhibits the lowest performance as in logistic regression.

CONFUSION MATRIX:

- Using the confusion matrix, we computed several performance measures for each model.

LOGISTIC REGRESSION:

- The confusion matrix shows that the logistic regression model correctly predicted 358 employees who are likely to stay and 23 employees who are likely to leave.
- However, it also misclassified 11 employees who are likely to leave as likely to stay, and 48 employees who are likely to stay as likely to leave.

	Predicted no	Predicted yes
Actual no	358	11
Actual yes	48	23

DECISION TREE:

- The decision tree model's confusion matrix reveals that it correctly predicted 54 employees who were likely to go and 20 employees who were likely to stay.
- However, it incorrectly identified 349 people as likely to quit as well as 17 employees were likely to stay on.

	Predicted no	Predicted yes
Actual no	20	349
Actual yes	17	54

RANDOM FOREST:

- The random forest model correctly predicted 0 employees who are likely to stay and 66 employees who are likely to leave the company, according to the confusion matrix for the model.
- However, it incorrectly classified 5 people as likely to stay and 369 employees who are likely to go as likely to stay.

	Predicted no	Predicted yes
Actual no	0	369
Actual yes	5	66

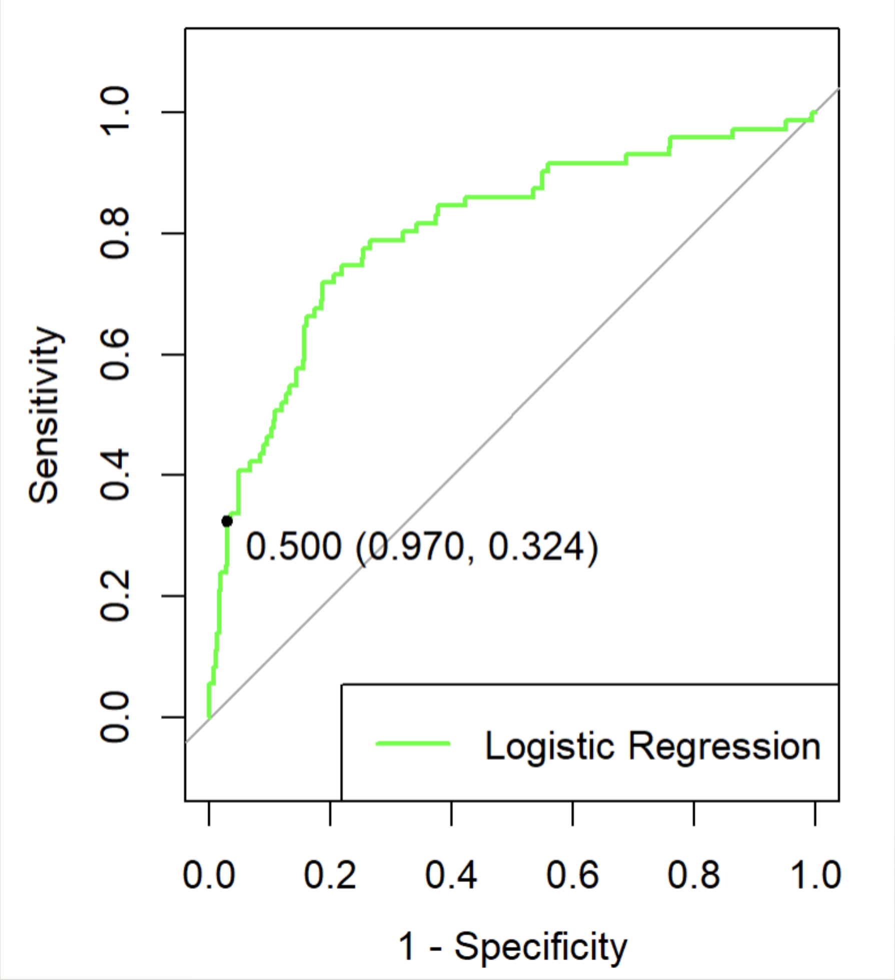
SUPPORT VECTOR MACHINES

- The SVM model correctly predicted 0 employees who are likely to stay and 71 employees who are likely to leave the company, according to the confusion matrix for the model.
- However, it incorrectly classified 0 people as likely to stay and 369 employees who are likely to go as likely to stay.

	Predicted no	Predicted yes
Actual no	0	369
Actual yes	0	71

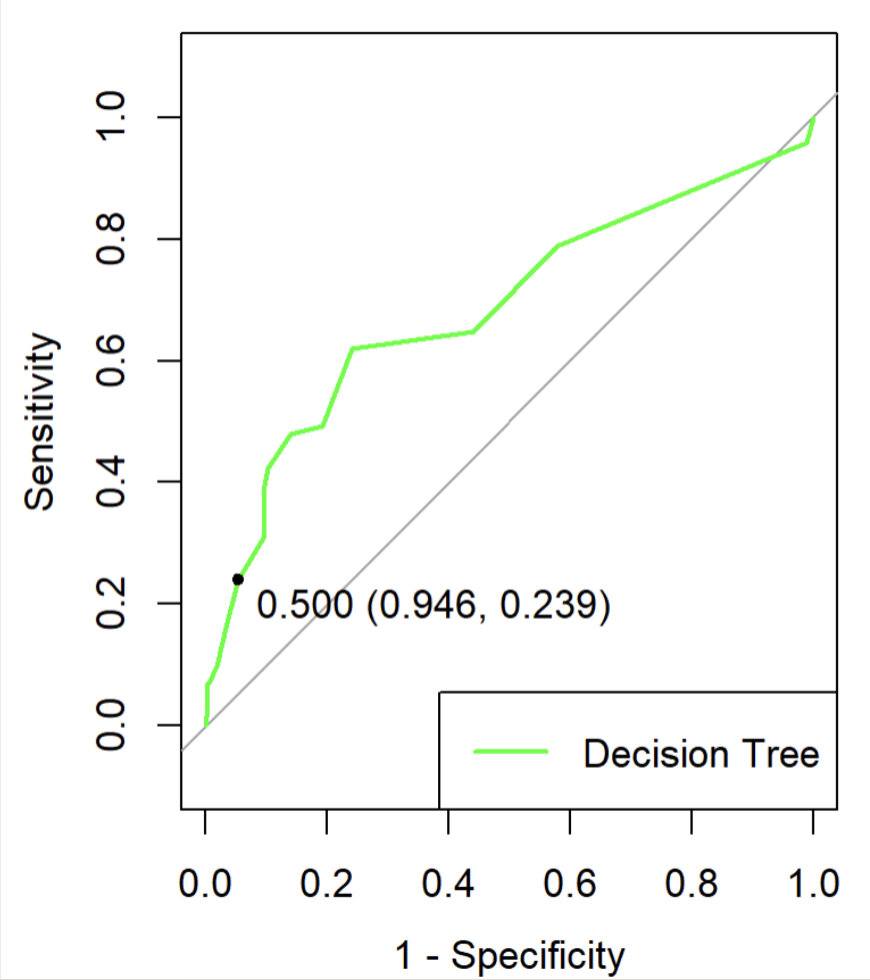
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE:

- For each model, the ROC curve was also executed. An illustration of the trade-off between true positives and false positives is the ROC curve. It is useful to see how the model performs at various classification thresholds.



LOGISTIC REGRESSION:

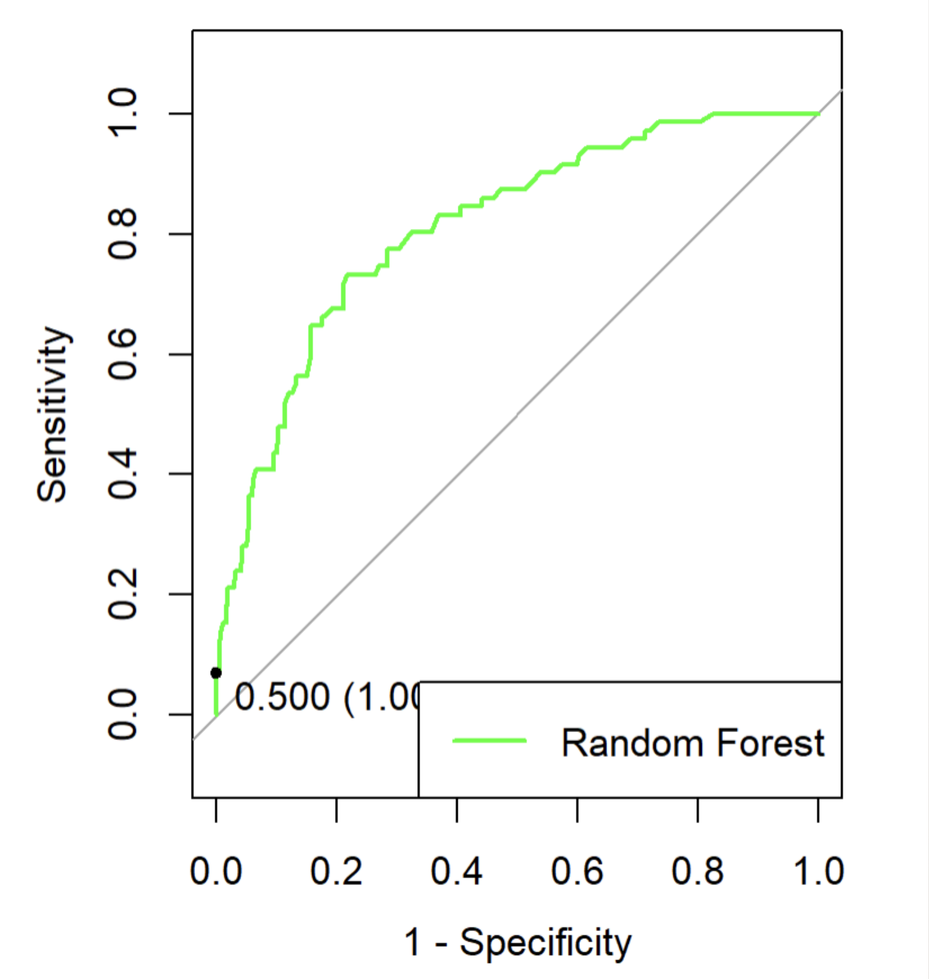
- The logistic regression model provides high performance, as seen by the ROC curve closer to the upper left corner compared to the other models. This model's AUC, which is 0.80, is high among all the models.



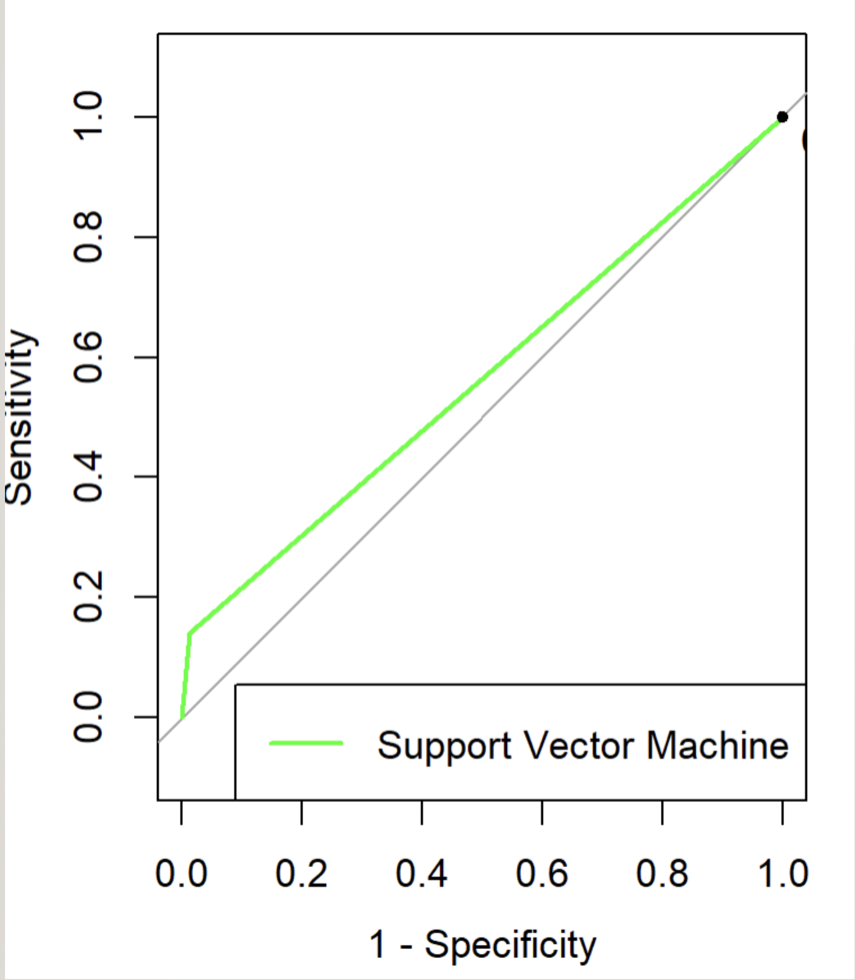
DECISION TREE:

- The decision tree model's ROC curve is not as steep as the logistic regression model's, but gave a decent performance. The AUC of the decision tree model is for this one, which is 0.69.

RANDOM FOREST:



- The random forest model's ROC curve is superior to all of the models. This model's AUC is 0.81, which is the highest among all models.



SUPPORT VECTOR MACHINES

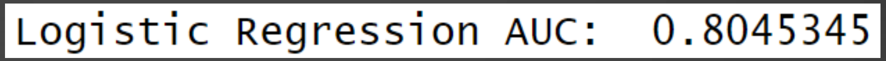
- The SVM model's ROC curve is inferior to all other models and exhibits poor performance among all. It has the lowest AUC (0.56) which is the lowest of any other comparatively.

AREA UNDER THE CURVE (AUC):

- For each model, we determined the area under the ROC curve (AUC). The model's total performance is summed up by a single value called the AUC. Its value is between 0 and 1, where 1 denotes flawless classification and 0.5 denotes random classification.

LOGISTIC REGRESSION

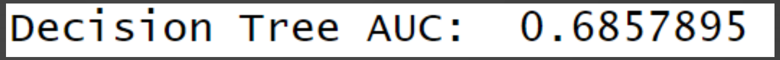
- The logistic regression model has the best AUC of all the models at 0.8. This shows that the model performs better and has a good capacity to distinguish between positive and negative classes.



```
Logistic Regression AUC:  0.8045345
```

DECISION TREE

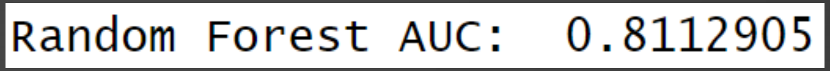
- The decision tree model has the lowest AUC of all the models at 0.69. This shows that the model performs poorly compared to but higher than the support vector machine and could require more optimization



```
Decision Tree AUC:  0.6857895
```

RANDOM FOREST

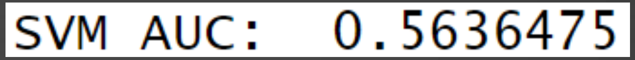
- The random forest model's AUC is 0.81, which is the highest AUCs of all other models. This shows that the model can discriminate between positive and negative classes well, and better than the logistic regression model.



```
Random Forest AUC:  0.8112905
```

SUPPORT VECTOR MACHINES

- The AUC for the SVM model is 0.56, which is the lowest. This suggests that compared to other models, the model may not well discriminate between positive and negative classes.



```
SVM AUC:  0.5636475
```

BEST MODEL

- The logistic Regression model seems to be the best model for predicting employee attrition based on the outcomes of the various evaluation metrics. The reasons are as follows:
- ROC and AUC: The model's ability to accurately detect true positives and true negatives while reducing false positives and false negatives is measured by the ROC curve and the AUC. The logistic regression model performs well in differentiating between positive and negative classes, as seen by its greatest AUC value of 0.80.
- Confusion Matrix: By displaying the number of true positives, true negatives, false positives, and false negatives, the confusion matrix offers a summary of the model's performance. The confusion matrix for the logistic regression model has a high proportion of true positives and true negatives and a low proportion of false positives and false negatives.
- Overall, the logistic regression model outperforms other models across all assessment measures. The logistic regression model is the most accurate model for predicting employee attrition, it may be said.

CONCLUSION

- As a result, it can be said that the logistic regression model consistently beat the others and performed well across all assessment measures, making it the best model for forecasting staff attrition. For businesses trying to forecast and stop staff loss, the insights gleaned from this initiative might be helpful.