

**ASSIGNMENT COVER SHEET**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PROGRAMME | : | Masters in business Analytics | | |
| SUBJECT CODE AND TITLE | : | BAA5123 AI and Decision Making | | |
| ASSIGNMENT TITLE | : | Predicting Employee Turnover with Random Forest Classification, Logistic Regression and SVM | | |
|  |  |  | | |
| LECTURER | : | Dr Tang Tiong Yew | ASSIGNMENT DUE DATE: | 9th of April 2024 |

STUDENT’S DECLARATION

1. I hereby declare that this assignment is based on my own work except where acknowledgement of sources is made.
2. I also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University/College or other institutions.

[ Submit “Turn-it-in” report (please tick √): Yes \_\_\_\_\_ No \_\_\_\_\_]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NO. | NAME | STUDENT ID NO. | SIGNATURE | DATE |
| 1. | Chia Poh Yang | 23138555 | CHIA | 9th of April 2024 |
| 2. | Joshua Wong Jun Wen | 16104812 | WONG | 9th of April 2024 |
| 3. | Harresh Ragunathan | 19076090 | HARRESH | 9th of April 2024 |
| 4. | Abdul Hakim Bin Kamalur Rahman | 24015257 | HAKIM | 9th of April 2024 |
| 5. |  |  |  |  |
| 6. |  |  |  |  |

E-mail Address / Addresses (according to the order of names above):

|  |  |
| --- | --- |
| 1. 23138555@imail.sunway.edu.my | 4. 24015257@imail.sunway.edu.my |
| 2. 16104812@imail.sunway.edu.my | 5. |
| 3. 19076090@imail.sunway.edu.my | 6. |

APPROVAL FOR LATE SUBMISSION OF ASSIGNMENT (If applicable)

IF extension is granted, what is the revised due date? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of Lecturer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |
| --- |
| Marker’s Comments: |

Marks and / or Grade Awarded: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ADDENDUM**

**USE OF ARTIFICAL INTELLIGENCE (A.I.) DECLARATION**

Students are allowed to use AI to support completion of assessments. However, students are reminded to do so ethically and transparently. This is so that (a) submissions can be fairly and accurately marked; and (b) feedback can be provided on the content that reflects student ability, in order to help with future submissions. Students are also reminded that in accordance with the University’s Academic Malpractice Policy, Item 4.11.2, “*… the representation of work: written, visual, practical or otherwise, of any other person, including another student or* ***anonymous web-based material*** *[emphasis added], or any institution, as the candidate’s own*” is considered malpractice.

**Declaration**

[ ] I / We used the following A.I. tools to produce content in this submission:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Purpose** | **Prompts** | **Sections where AI output was used / Outcome(s) in the submission** |
| *ChatGPT* | * *Generating necessary coding that would help the study.* * *Restructuring paragraphs* | * *To suggest better coding for visualizations and ideas for analytics purposes.* * *Rewrite the paragraphs written for better structure and wording.* | *Does not include any specific parts but adapt accordingly when its necessary.* |
| *Grammarly* | *Correcting grammar and spelling, improving sentence structure* | *N/A* | *Grammarly suggestions were used for all sections of the essay* |
|  |  |  |  |
|  |  |  |  |

*Note: Add additional rows if necessary.*

**OR**

[ ] I / We did not use any A.I. tools to produce any of the content in this submission.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NO. | NAME | STUDENT ID NO. | SIGNATURE | DATE |
| 1. |  |  |  |  |
| 2. |  |  |  |  |
| 3. |  |  |  |  |
| 4. |  |  |  |  |
| 5. |  |  |  |  |
| 6. |  |  |  |  |

E-mail Address / Addresses (according to the order of names above):

|  |  |
| --- | --- |
| 1. | 4. |
| 2. | 5. |
| 3. | 6. |

**Table of Contents**

[**Introduction** 5](#_Toc163559500)

[**Business Problems** 6](#_Toc163559501)

[**Solution Methodology Lifecycle** 7](#_Toc163559502)

[Data Collection 7](#_Toc163559503)

[Exploratory Data Analysis 7](#_Toc163559504)

[Handling Outliers 12](#_Toc163559505)

[**Solution Design** 14](#_Toc163559506)

[Recursive Feature Elimination (RFE) 14](#_Toc163559507)

[Implemented Models 15](#_Toc163559508)

[Choosing the Evaluation Metrics 16](#_Toc163559509)

[**Solution implementation** 19](#_Toc163559510)

[Model Building 19](#_Toc163559511)

[Splitting Dataset into training and testing 19](#_Toc163559512)

[Model training. 20](#_Toc163559513)

[Model Evaluation Metrics 21](#_Toc163559514)

[K fold Cross-Validation 21](#_Toc163559515)

[Precision & Recall 22](#_Toc163559516)

[Confusion Matrix Heatmap 22](#_Toc163559517)

[ROC Receiver Operating Characteristics 23](#_Toc163559518)

[Synthetic Minority Oversampling Technique (SMOTE) 24](#_Toc163559519)

[Feature Importance analysis 25](#_Toc163559520)

[**Solution Output** 26](#_Toc163559521)

[**Recommendations** 42](#_Toc163559522)

[**Future Work Recommendation** 43](#_Toc163559523)

[**Conclusion** 44](#_Toc163559524)

[**References** 45](#_Toc163559525)

# **Introduction**

Employee turnover is defined as the rate at which employees leave and are replaced. This poses a great threat to any company as it has the potential to affect the productivity, morale, and profits of any company. This issue makes it crucial for managers and HR departments to understand the cause of this issue to maximize the profit of the company (Allen et al.,2010).

In recent years, employee turnover has gained more traction and attention due to the financial fluctuation of companies. The dynamics of the business may be impacted by the recruiting and training of new employees (Griffeth et al., 2000). Any business that experiences significant employee turnover may find it challenging to hold onto its best employees and preserve its good reputation (Tett & Meyer, 1994).

The purpose of this study is to investigate and comprehend the causes of employee turnover to find mitigating techniques that will reduce it. The aim is to guarantee that businesses may establish a more favorable workplace where employees feel more valued and inspired to remain. In addition, this research aims to assist businesses in optimizing profit margins and staff management (Shaw et al., 2005).

The goal of this research study is to understand and comprehend the cause of employee turnover by discovering mitigation strategies to reduce overall employee turnover. The end goal is to ensure that businesses can create a more conducive working environment where employees feel valued and inspired in their respective workplaces. Furthermore, this research paper is aimed at assisting companies in retaining their current employees while also maximizing their profit. (Shaw et al., 2005)

Deploying various machine learning tools makes it easier to understand the underlying patterns and trends around employee turnover. By utilizing machine learning algorithms can discover important key factors that affect employee turnover such as low work satisfaction and lack of career development. Evaluating and understanding exploratory data of employees such as their demographics, job scope, and performance ratings, managers can partner with these algorithms to create retention strategies such as career advancement and better work-life to better sustain employee longevity and maximize the company’s profit.

# **Business Problems**

The main business problem in this research study is to understand what factors can reduce the costs of recruitment as hiring costs are a burden to companies. Failing to find the right solution can potentially cause the company to spend more on training time, thus causing a decrease in time to gain profit due to the lack of productivity. Moreover, this can also cause companies to lose talent which can disrupt the company’s morale and reputation.

According to research conducted by Tyler Jadah (2023), the individual mentioned that the turnover rate in the US is nearly 50% annually including both voluntary and involuntary turnovers. In 2021, 47.4 million employees in the US left their jobs voluntarily and the numbers increased to over 50 million in 2022. In a literature review by Ongori (2007), he mentioned that employee turnovers are heavily influenced by job-related stress, lack of commitment to the organization, lack of growth opportunities, and role stressors like ambiguity of job expectation in a job-related sense. He also mentioned organizational factors that contribute to the turnover of employees such as organizational instability, high levels of inefficiency in the organization, lack of employee empowerment, and a toxic working environment that contributes to the turnover of employees. By understanding and analyzing the trends that contribute to employee turnover, companies can find suitable solutions to reduce employee turnover and maximize the company’s growth and profit.

**Research Objectives**

1. The research objective is to identify which of the machine learning algorithms could accurately predict the event of employee turnover.
2. To identify which of the variables plays a significant role in affecting employee turnover.
3. To discover the appropriate strategies to countermeasure the likelihood of turnover based on the findings of the analysis.

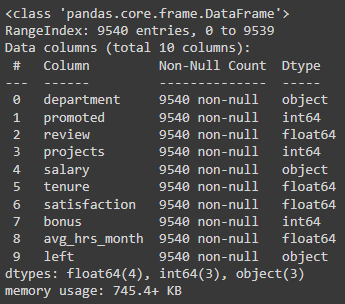
# **Solution Methodology Lifecycle**

## **Data Collection**

This research begins by utilizing a dataset regarding employee turnover from Kaggle. This dataset was retrieved from a large US company, gathered by the Human Resources (HR) department, with a sample size of approximately 10,000 employees who left the company from 2016 to 2020. The HR department retrieved information from exit interviews, performance reviews, and employee records. This information includes the employee’s department, whether they have been promoted within the past 24 months, their composite score, how many projects they are involved in, their salary (low, medium, or high), their tenure, their satisfaction, where they have received a bonus in the past 24 months, their average hours worked in a month, and whether they left.

## **Exploratory Data Analysis**

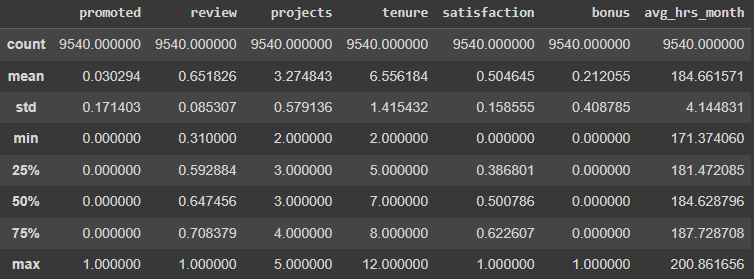
With this dataset, the study may be conducted by performing the Exploratory Data Analysis (EDA). During the EDA process, the main objective is to summarize and visualize the data to gauge a better understanding before running the model. This is done by checking for null values, understanding the nature of the dataset through visualizations, and handling outliers. By using the code “**df.info()**”, the variable and type of data are identified for each column. Furthermore, the code checks if there are any null values in each column, which as shown, there are no null values in each column.



*Diagram 1.0: Checking null values.*

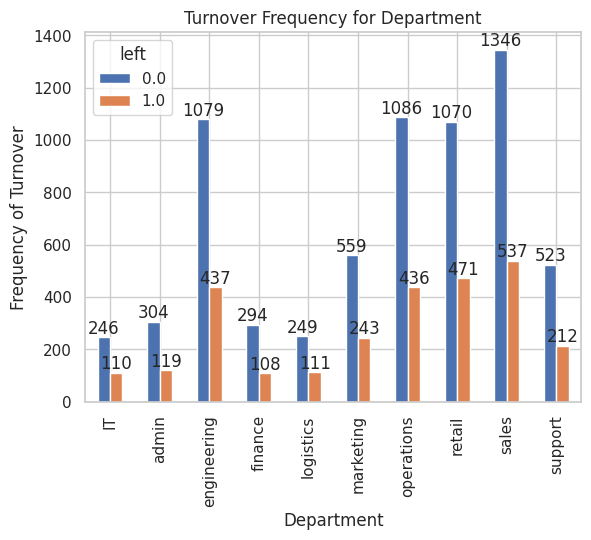
As seen above, the type of data is identified for each variable and there are no null values found.

To understand the nature of the dataset, the descriptive statistics need to be identified. This was done by using the code “df. describe()”, which outputs the following.



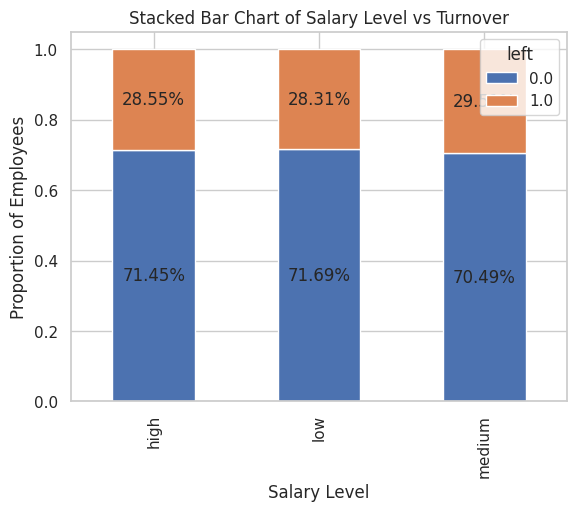
*Diagram 2.0: Descriptive Statistics*

As seen above, the output includes each variable’s number of non-blank values, mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, and maximum. To gain a further understanding of the dataset, the dataset may also be visualized in the form of charts, specifically bar charts, and histograms.



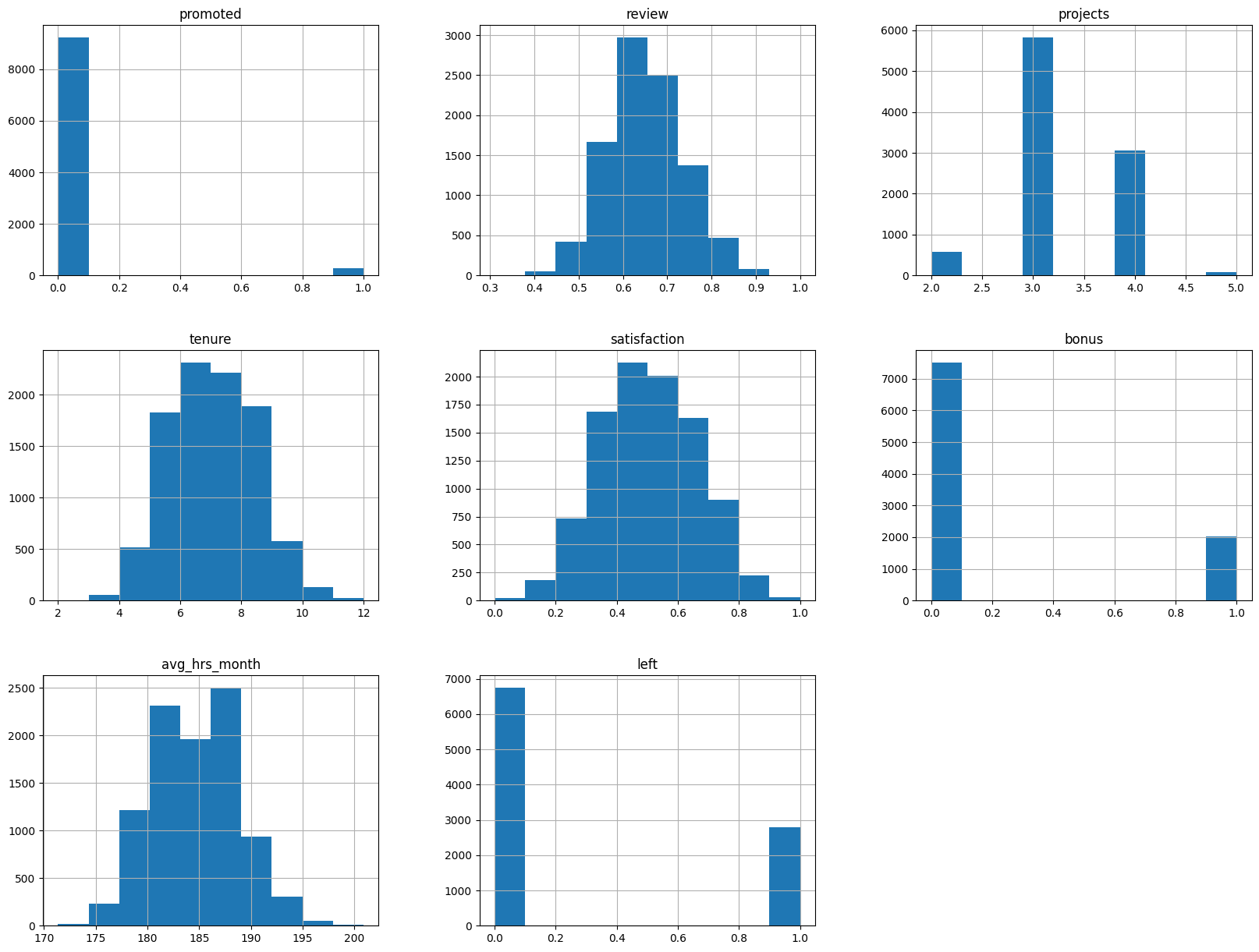
*Diagram 3: Department vs Turnover Histogram*

The following findings of the categorical variable “department” against “left”/ known as “Turnover”. The usage of projecting the histogram provides the idea of which department had the highest turnover rate. The sales department may have faced the highest turnover rate as compared to the others, whereas the lowest turnover rate falls under the finance department.



*Diagram 4.0: Salary vs Turnover Barchart*

The following stacked Bar chart shows the categorical variable of “salary” vs “left” known as turnover. The findings discovered that the percentile of employees leaving are employees who receive medium-ranged income as compared to low and high income.

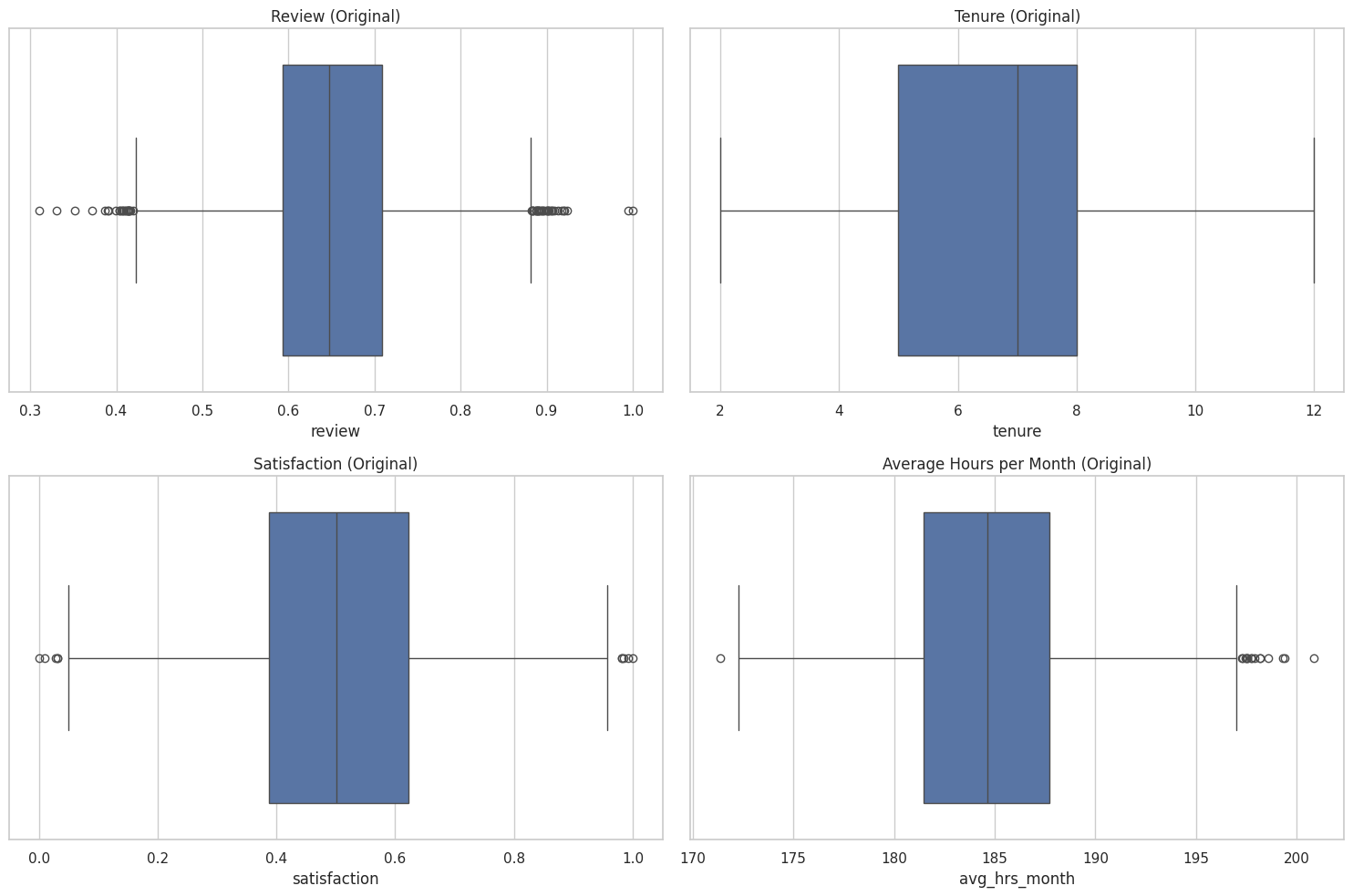


*Diagram 5.0: Histogram of variables in the study*

As seen from the bar charts above, it may be observed that, in general, the number of employees leaving does not exceed the number of employees who remained. Furthermore, by observing the histogram, the overall scope of the data within each variable can be comprehended further. The histogram above also shows a comprehensive discovery where most of the employees did not receive many benefits in the workplace, which could be attributed to the fact that promoted employees are lesser than the employees who were not promoted, and bonuses that are received by employees stand at the lowest as compared to employees that did not receive bonuses.

## **Handling Outliers**

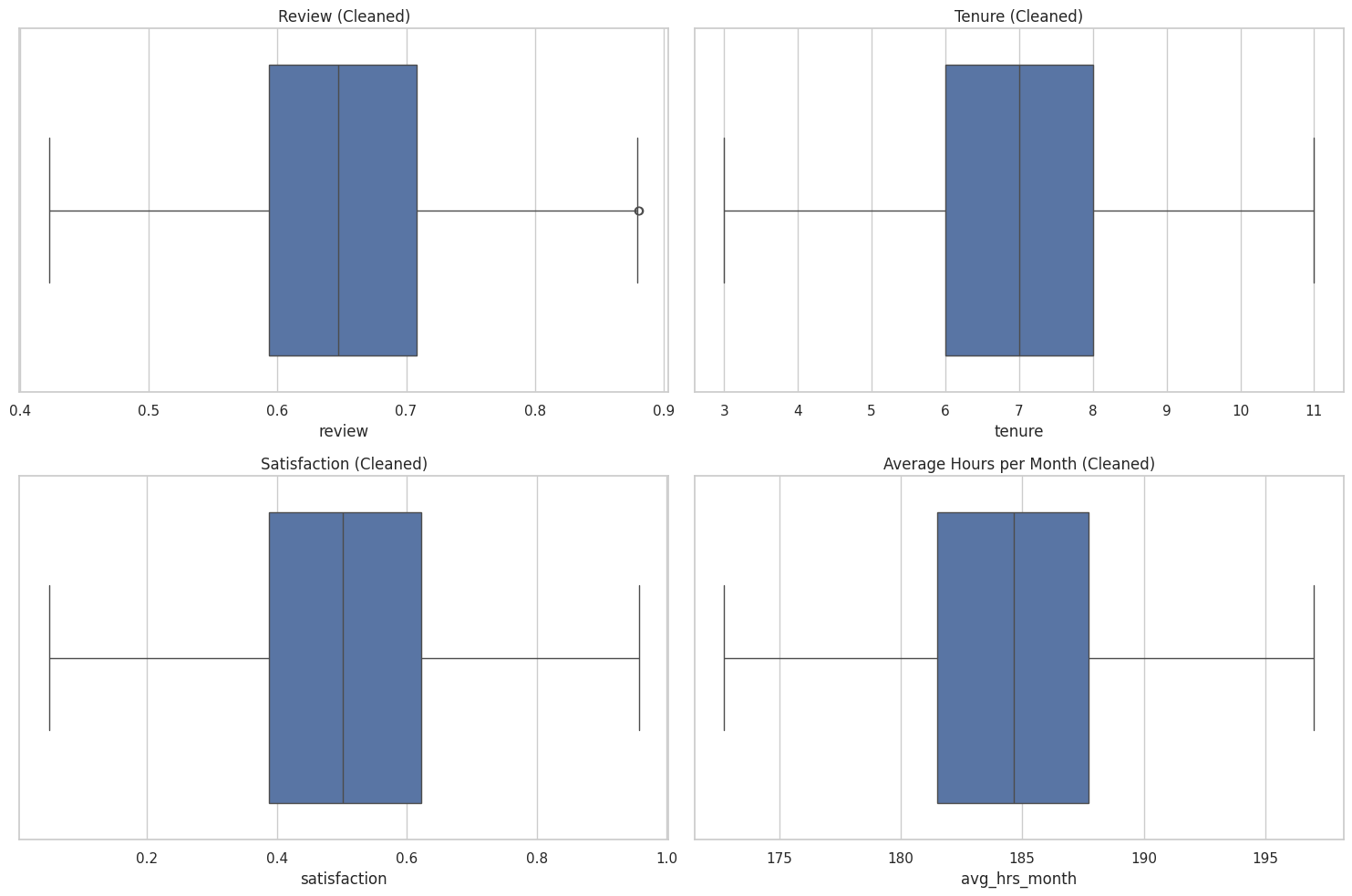
Lastly, to ensure the reliability of the data, the outliers were identified and withdrawn. The outliers need to be withdrawn to prevent any hindrance that may affect the model and these outliers were identified with the use of box plots. The variables that were represented by box plots were the “Review”, “Tenure”, “Satisfaction”, and “Average hours per month” as these are numerical values that vary.



*Diagram 6.0: Boxplots of numeric values*

As seen above, the variables are represented by box plots, and it may be observed that there are indeed outliers that may affect the model. Hence, these outliers need to be eliminated to prevent hindrance of the results of the model. To remove these outliers, the Interquartile Range (IQR) method was applied. The IQR method measures the spread of the median of data, 50% of data. The first quartile is subtracted by the third quartle whereby the first quartile represents the median of half of the data below whereby the third quartile represents the median of data above. The IQR measurement is justified from the range values projected, it represents most of the data. It is a technique that is prone to sensitivity towards outliers aside from the mean and standard deviation. The following method is followed by the quantile() function where it calculates the Q1 and Q3 and filters out the values that are not in the range of Q1- 1.5\*IQR and Q3 + 1.5\*1QR using the function of ‘**df\_clean = remove\_outliers\_iqr(df\_clean, column)’.** Lastly create a cleaned boxplot using the following function

**‘sns.boxplot(x=df\_clean['review'], ax=axes[0, 0])’**



*Diagram 7.0: Cleaned Boxplots of numeric values*

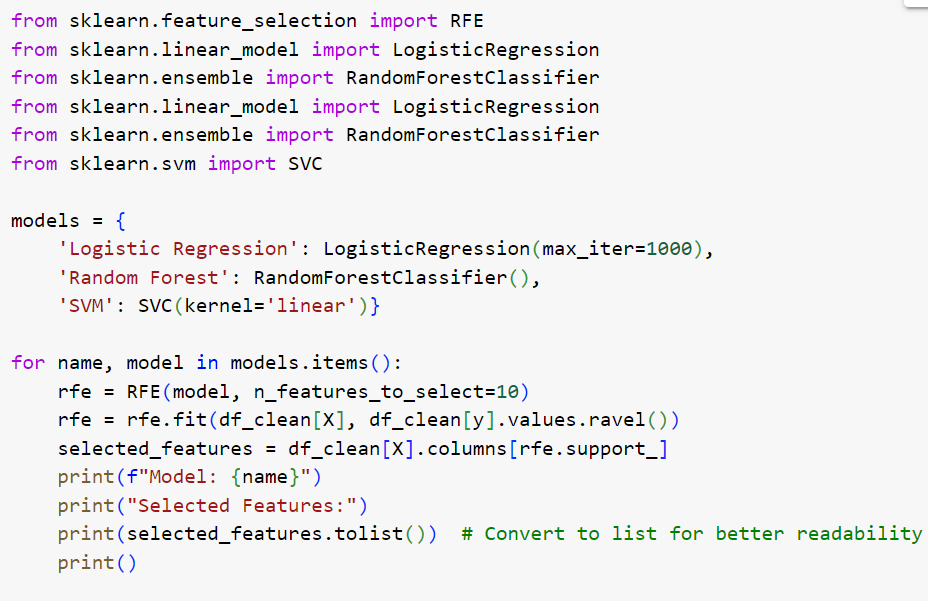
The boxplot above represents the four variables after the outliers have been eliminated, which changed the dimensions of the dataset from (9540, 10) to (9457, 10).

# **Solution Design**

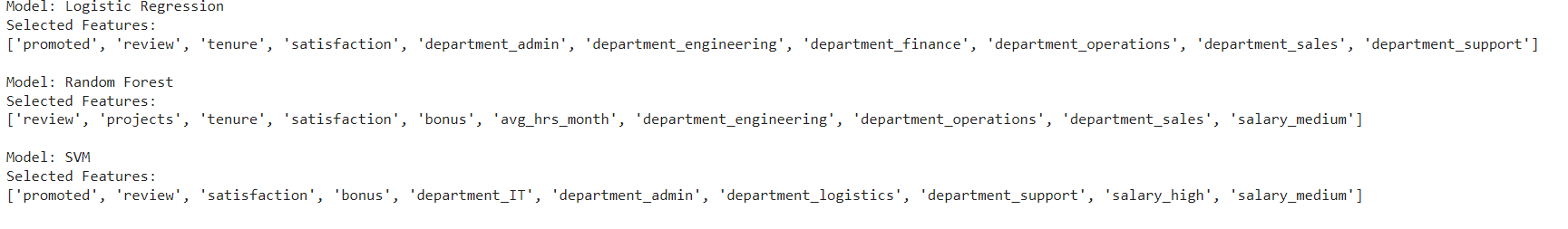
## **Recursive Feature Elimination (RFE)**

After the data preprocessing and collection, as a preliminary step before choosing and running the models, the significant variables need to be selected. This was done using the Recursive Feature Elimination (RFE), which determines the significant variables for each model. It works by recursively removing features and helps in building a model on the remaining features until the desired number of features is reached.

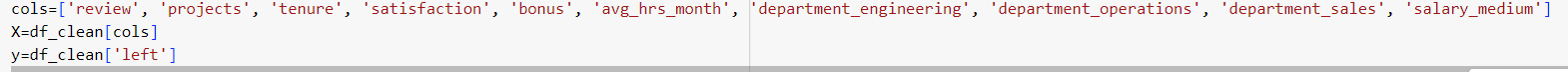
The diagram below shows the coding that was used in this business study: -



The results after the coding were run are as follows: -



From the results, the significant variables are shown for all three chosen models, which are Random Forest Classifier, Logistic Regression, and Support Vector Machine (SVM). Through experimentation, it was found that the variables for the Random Forest Classifier were the most accurate. This was discovered by testing each variable set, from each model, using the random forest model. It was found that each test produced desirable accuracies (more than 80%). Thus, the variables for the Random Forest Classifier were chosen for this study.



## **Implemented Models**

**Random Forest Classifier**

As mentioned in the previous section, the models chosen were Random Forest Classifier, Logistic Regression, and Support Vector Machine (SVM). Random forest is considered one of the most popular and effective machine learning methods that can perform classification and regression tasks (Chen et al., 2022). This model creates many decision trees, also known as a forest, and generally, the larger the number of trees in the forest, the stronger and more accurate the prediction is. Random forest may be useful in model prediction because it is capable of handling missing values and maintaining accuracy if a large amount of data is missing. Furthermore, it is capable of handling larger datasets with larger dimensions (Do et al., 2018). The advantages of using random forests are prominent in many studies. For example, in a study titled “*Using HR Analytics to Support Managerial Decisions: A Case Study*”, the researchers used random forest, along with many other methods, to predict a model. During their evaluation metrics process, they found that random forest performed the best (Liu et al., 2020).

**Logistic Regression**

Logistic Regression is a classification and regression method that uses linear discriminants, and the output revolves around the likelihood that an input point corresponds to a specific class (Komarek et al., 2004). The model illustrates a linear border, and the input space is split into two groups depending on their probabilities. Logistic regression is typically used when the general objective is to classify data elements, in this case, whether the employee will leave or not. Hence, the target variable in logistic regression is binary (1 or 0) (Mehrolia et al., 2020). Logistic regression is another common method in prediction models, as used by a study titled “*HR Analytics: Early Prediction of Employee Attrition using KPCA and Adaptive K-means based Logistic Regression*”, whereby the researchers used logistic regression to predict a model and they found that it was the best-performing model when conducting their evaluation metrics (Pratibha & Hegde, 2022).

**Support Vector Machine SVM**

Support Vector Machine (SVM) is a classification and regression method that can solve nonlinear problems by converting them to quadratic programming (Tian et al., 2012). It is also able to “reduce overfitting by selecting the maximal margin hyperplane in the feature space”. Furthermore, SVM is capable of handling datasets that have large dimensions, making it suitable for datasets with many variables. SVM is commonly used in case studies such as “*Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine*”, where the researchers used SVM for their prediction model and found that during the evaluation metrics, SVM performed the best (Ren et al., 2019).

## **Choosing the Evaluation Metrics**

To determine each model’s accuracy and overall performance, a form of quantitative evaluation needs to take place, which is also known as Evaluation Metrics. These evaluation metrics are ultimately used to determine which out of the three models used performed the best. The evaluation metrics that are used in this study are the K-fold cross-validation, precision and recall, and receiver operating characteristic (ROC).

**K-fold cross-validation**

K-fold cross-validation is an evaluation metric that involves splitting the dataset into k subsets and, repeatedly, different k subsets are used as a validation set, while the rest of the data (k-1 subset) is used as the training set (Anguita et al., n.d.). These are known as iterations where the k-1 subsets are different depending on when the k-subset is being used. Once the model is trained using the k-1 subsets the k subset is used to test the model, computing the evaluation for that single iteration (Wong & Yeh, 2020). After all, iterations are complete, each iteration’s evaluation metric is average, forming the overall evaluation metric of the model.

**Precision recall**

(PR) is an evaluation metric where the area under the PR curve is measured to determine the model’s performance. In the PR curve, precision measures the purity or accuracy of the results, while recall (also known as sensitivity) measures the completeness of the results, and they have inverse relationships (Buckland & Gey, 1994). After the PR curve has been illustrated, the area under the curve is calculated, a larger area generally indicating that the model is performing well.

**Receiver Operating Characteristics ROC**

Lastly, Receiver Operating Characteristic (ROC) is an evaluation metric that essentially is a graph that illustrates detection vs false alarm (true positive rate vs false positive rate) (Candy et al., 2013). The ROC curve can provide many individual metrics such as sensitivity, specificity, and cost-benefit analysis and analyze the area under the curve. The perfect ROC curve is achieved when the true positive rate is equal to 1 while the false positive rate is equal to 0. This implies that a well-performing model will have its practical ROC curve’s vertex as close as possible to the top left corner. Furthermore, the area under the ROC curve may be computed, ranging from 0 to 1, to determine its discrimination ability.

**Technique Implemented**

**Synthetic Minority Oversampling Technique (SMOTE)**

Justifying the nature of the datasets that were collected, the number of employees remaining in the industry could have been observed to exceed the number of employees that left. In the event of model building, this could be classified as an imbalanced dataset, a high concentration of employees remaining in the dataset would lead the model to overtrain the class of ‘0.0’/ ‘remain’ during the training phase. Contrary to the ‘leave’, leaving classes will be undertrained. Thus, leading the model to tend to predict employees staying even if they could be at risk of leaving.

The Synthetic Minority Oversampling Technique **SMOTE** is known as the statistical technique that generates synthetic minority samples (Chawla et al., 2002). It is often used to perform the prediction of abnormalities such as the study of detecting breast cancers, network distributions, or rather predicting the distribution of species (Blagus & Lusa, 2013). The implementation of this algorithm in the current study would countermeasure the overfitting problem, where SMOTE allows the models to focus on decision boundaries between classes, and improve model performance, whereby the accuracies of prediction would be significantly better and in favor of preventing the possibilities of values biases (Waqar et al., 2021).

# **Solution implementation**

The following research studies shall be further conducted through GoogleColab under the usage of Python’s coding. The breakdowns on the following implementation are conducted thoroughly through the sequence of model building, performance evaluation, and results output. Instead of explaining each of the code snippets individually for Random Forest classification, Logistic regression, and Support Vector Machine, the following models shall be simplified into one understandable term. Whereby the usage of code remains the same for all the models, but the only difference lies under the variable's name such as ‘rf. predict’, ‘logreg. predict’, or ‘svm. Predict’. This approach shall reduce the length of explanation and reduce the complexity for individuals to understand the codings.

## **Model Building**

### **Splitting Dataset into training and testing**

The following steps are known as the most crucial steps while building a machine learning model. Datasets are segregated mainly into 2 sets, training and testing. Training sets are used for the model to learn and familiarize with the algorithm to predict an outcome, whereas testing sets are used to evaluate the performance through unseen data and examine the accuracies of the trained model.

The dataset splits into 20% for testing under the code of **‘test\_size=0.2’** and the rest 80% shall be used for training purposes. This translates into the following code of importing the **‘train\_test\_split’** function from the **‘sklearn.model\_selection’** to split the dataset into training and testing sets. Aside from that, the coding further creates functions where it generates 4 different arguments mainly **‘X\_train, x\_test, y\_train, and y\_test’** while setting reproducible results with a **‘random\_state’** value of 42 to ensure the accuracy for the results remained unchanged.

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

### **Model training.**

The following steps emphasized the model training among the 3 classifiers, random forest classification, logistic regression, and support vector machine. The accuracies are being tested out among the best-performing models that were being examined. Following are the codings that were implemented with the usage of importing **‘RandomForestClassifier’, ‘LogisticRegression’, and ‘SVC’ (Support Vector Machine)** from **sklearn. ensemble, sklearn.linear\_model, and sklearn.svm** simultaneously. The accuracies of the trained model can be captured through the function of ‘print’.

from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(random\_state=42)

rf.fit(X\_train, y\_train)

from sklearn.metrics import accuracy\_score

print('Random Forest Accuracy: {:.3f}'.format(accuracy\_score(y\_test, rf.predict(X\_test))))

from sklearn.linear\_model import LogisticRegression

from sklearn import metrics

logreg = LogisticRegression(max\_iter=1000,random\_state=42)

logreg.fit(X\_train, y\_train)

print('Logistic regression accuracy: {:.3f}'.format(accuracy\_score(y\_test, logreg.predict(X\_test))))

from sklearn.svm import SVC

svm = SVC(random\_state=42)

svm.fit(X\_train, y\_train)

print('Support vector machine accuracy: {:.3f}'.format(accuracy\_score(y\_test, svm.predict(X\_test))))

## 

## **Model Evaluation Metrics**

The following section further provides insights into the performance and effectiveness of the chosen machine-learning model. As the study introduces 3 different classifiers, this could further compare all 3 machine learning models through the evaluation metrics analysis such as Precision and recall, K fold Cross-validation, and ROC (Receiver Operating Characteristic).

### **K fold Cross-Validation**

Following K-fold cross-validation is performed under 3 different classifiers in achieving the average accuracies of the models. The following code

‘ **kfold=model\_selection.KFold(n\_splits=10,shuffle=True )**’

is made to justify their will be a total of 10 splits that will be shuffled before splitting. The approach of including **‘(max\_iter=10000, solver = ‘lbfgs’)’** under the logistic regression model is to ensure that the solver may have enough iterations to converge into the solution without prompting a warning.

from sklearn import model\_selection

from sklearn.model\_selection import cross\_val\_score

kfold = model\_selection.KFold(n\_splits=10, shuffle=True)

modelCV1 = RandomForestClassifier(random\_state=42)

modelCV2 = LogisticRegression(max\_iter=10000, solver='lbfgs',random\_state=42)

modelCV3 = SVC(random\_state=42)

scoring = 'accuracy'

results1 = model\_selection.cross\_val\_score(modelCV1, X\_train, y\_train, cv=kfold, scoring=scoring)

results2 = model\_selection.cross\_val\_score(modelCV2, X\_train, y\_train, cv=kfold, scoring=scoring)

results3 = model\_selection.cross\_val\_score(modelCV3, X\_train, y\_train, cv=kfold, scoring=scoring)

print("Random Forest 10-fold cross validation average accuracy: %.3f" % (results1.mean()))

print("Logistic Regression 10-fold cross validation average accuracy: %.3f" % (results2.mean()))

print("SVC 10-fold cross validation average accuracy: %.3f" % (results3.mean()))

### **Precision & Recall**

Precision & Recall is further used as one of the other evaluation metrics. The following evaluation metrics project’s f1-score, precision score, and recall score. **‘Classification\_report’** is being imported from sklearn. metrics, followed by projecting the results.

from sklearn.metrics import classification\_report

print(classification\_report(y\_test, rf.predict(X\_test)))

### **Confusion Matrix Heatmap**

Aside from that, the confusion matrix heatmap has also been utilized to justify the usefulness of random forest classification as the ideal model for the current study.

**‘y\_pred = rf. predict(X\_test)**’. **‘Rf’** is known as a Random Forest classifier and it is used to predict the independent variable (X\_test) test sets, similar to the other approach for Logistic regression known as **‘logreg. predict’** and Support Vector Machine known as **‘svm. predict’.** The **‘confusion\_matrix’** function is further utilized in comparing the values of **‘y\_pred’ and ‘y\_test’** to generate a confusion matrix table. The **‘sns. heatmap’** generates heatmap visualizations for the confusion matrix through the imported library, seaborn. **‘annot=True’** is further used to generate numerical values on the heatmap. **‘ .2f**’ meant by numerical values generated shall be in 2 decimal places. **‘cmap=’Blues’’** is used to customize the color of the heatmap visualizations. **‘plt.savefig’** is used to save the confusion matrix figures under the current working directory.

y\_pred = rf.predict(X\_test)

logreg\_y\_pred = logreg.predict(X\_test)

svm\_y\_pred = svm.predict(X\_test)

from sklearn.metrics import confusion\_matrix

import seaborn as sns

confusion\_mtx = confusion\_matrix(y\_pred, y\_test)

sns.heatmap(confusion\_mtx, annot=True, fmt='.2f', cmap='Blues')

plt.xlabel('Predicted class')

plt.ylabel('True class')

plt.title('Random Forest/Logistic Regression/SVM')

plt.savefig('Random\_Forest', ‘Logistic Regression’,’SVM’)

plt.show()

### **ROC Receiver Operating Characteristics**

**‘Roc\_auc\_score’** is used to generate the area under the ROC curve. The following study generates 3 different curves such that **‘logreg’** for Logistic Regression, **‘rf’** for Random Forest, and **‘svm’** for Support Vector Machine. Followed by the usage of **‘fpr, tpr, thresholds = roc\_curve(y\_test, logreg.predict\_proba(X\_test)[:,1]’** where **‘fpr’** stands for False positive rate **‘tpr’** stands for True Positive Rate, **‘thresholds’** known as the decision threshold to determine the true positive class from the prediction, whereas **‘predict\_proba(x\_test)[:,1]’** usage is to return the matrix of predicted values into the first class for each samples.

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve, roc\_auc\_score

logit\_roc\_auc = roc\_auc\_score(y\_test, logreg.predict(X\_test))

fpr, tpr, thresholds = roc\_curve(y\_test, logreg.predict\_proba(X\_test)[:, 1])

rf\_roc\_auc = roc\_auc\_score(y\_test, rf.predict(X\_test))

rf\_fpr, rf\_tpr, rf\_thresholds = roc\_curve(y\_test, rf.predict\_proba(X\_test)[:, 1])

svm = SVC(probability=True)

svm.fit(X\_train, y\_train)

svm\_roc\_auc = roc\_auc\_score(y\_test, svm.predict\_proba(X\_test)[:, 1])

svm\_fpr, svm\_tpr, svm\_thresholds = roc\_curve(y\_test, svm.predict\_proba(X\_test)[:, 1])

plt.figure()

plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit\_roc\_auc)

plt.plot(rf\_fpr, rf\_tpr, label='Random Forest (area = %0.2f)' % rf\_roc\_auc)

plt.plot(svm\_fpr, svm\_tpr, label='Support Vector Machine (area = %0.2f)' % svm\_roc\_auc)

plt.plot([0, 1], [0, 1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('ROC')

plt.show()

## **Synthetic Minority Oversampling Technique (SMOTE)**

Smote is imported from **‘imblearn.over\_sampling’** function where the random state is set to 42 to reuse the samples that were initially tested and trained in the study. New sets of training data/ new sets of functions are being introduced for the usage of SMOTE **‘smote\_x\_train, smote\_y\_train = smote.fit\_resample(X\_train, y\_train)’. ‘Fit\_resample’** allows Smote to create new sets of data to solve the issue of imbalances in minority classes via implementing an over-sampling method.

Usage of **‘Counter’** was used to ensure that it shows the number of samples before and after it was balanced. The following SMOTE test is then proceeded to run through the process of all the evaluation metrics such as K-Fold Cross Validation, Precision & Recall, and ROC AUC.

from imblearn.over\_sampling import SMOTE

smote = SMOTE(random\_state=42)

smote\_X\_train, smote\_y\_train = smote.fit\_resample(X\_train, y\_train)

X\_smote\_test, y\_smote\_test = smote.fit\_resample(X\_test, y\_test)

!pip install collections

from collections import Counter

print('Before sampling class distribution:-',Counter(y\_train))

print('After sampling class distribution:-',Counter(smote\_y\_train))

## **Feature Importance analysis**

The following feature importance analysis arrays the significant variables that were classified by the recursive feature elimination. The variables such as **'promoted', 'review', 'tenure', 'satisfaction','department\_admin','department\_finance','department\_engineering','department\_operations','department\_sales','department\_support’ were** used to calculate its feature importance scorings by using random forest model. **‘importance.argsort()’** sorts the most important variables in ascending order. The printed format shall be generated in percentile form with 2 decimal places with the usage of **‘{}-{:.2f}%’.**

feature\_labels = np.array(['promoted', 'review', 'tenure', 'satisfaction', 'department\_admin','department\_finance','department\_engineering','department\_operations','department\_sales','department\_support'])

importance = rf.feature\_importances\_

feature\_indexes\_by\_importance = importance.argsort()

for index in feature\_indexes\_by\_importance:

print('{}-{:.2f}%'.format(feature\_labels[index], (importance[index] \*100.0)))

# **Solution Output**

**Splitting Dataset into Training and Testing**

When the variables for this business study had been obtained, the next step was building a model. To build a model, it must first undergo training and testing using sci-kit-learn to split the dataset into training and testing sets. This step is important as it evaluates the performance of the machine learning models.



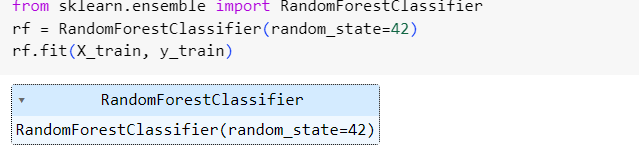
From this coding, the **‘test\_size = 0.2’** argument indicates that 20% of the data will be used for testing whereas the remaining 80% will be used for training. The **‘random\_state=42’** argument specifies that the data is spilled in a reproducible manner.

**Model Training**

After the datasets have been trained and tested, the next process is to train the model. As explained earlier, this business study will be using three types of models and from these models, the highest number of accuracies will be used for the study.

The diagram below shows the coding that was used to train and test the three models: -

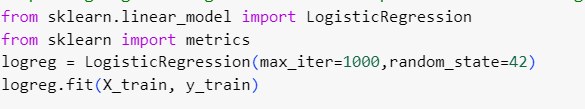
1. Random Forest Classification





The accuracy for this model is 0.864.

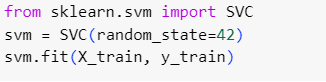
1. Logistic Regression Model





The accuracy for this model is 0.734.

1. Support Vector Machine





The accuracy for this model is 0.707.

From these three models, Random Forest Classification given the highest accuracy for the model-predicted outcome of the tested data.

**K-Fold Cross-Validation**

To obtain a more robust estimate of the model’s performance, the K-Fold Cross Validation was used in this study. Compared to a single train-test split, it helps to mitigate the issues related to variability in performance metrics due to the randomness in data splitting.

The coding use for K-Fold Cross Validation is shown below: -

A screenshot of a computer code

Description automatically generated

The breakdown of the codes is as follows: -

* **‘model\_selection.KFold(n\_splits10, shuffle=True)’:** This line created a 10-fold cross-validation iterator (**‘kfold’**) with shuffling enabled. This code would randomly shuffle the data before splitting into 10 folds.
* The three models that were used i.e. **‘RandomForestClassifier’**, **‘LogisticRegression’**, and **‘SVC’** were classified as **‘model1CV1’,** **‘model1CV2’,** and **‘model1CV3’** respectively, each with the appropriate parameters. These three models will be evaluated using cross-validation.
* **‘model\_selection.cross\_val\_score’:** This function was used to perform cross-validation. It had taken the model, training data **‘x\_train’, ‘y\_train’**, cross-validation iterator **‘kfold’,** and scoring metric **‘scoring’** as arguments. It returned an array of scores obtained for each fold.
* **‘results1’**, **‘results2’,** and **‘results3’** stored the cross-validation scores for Random Forest, Logistic Regression, and Support Vector Machine models respectively.
* **‘mean():** This was used to calculate the average accuracy for each model on the results obtained from cross-validation.

From the K-cross Validation, the average accuracy calculated for the 10-fold cross-validation are as follows: -

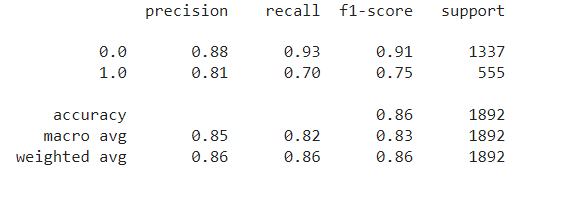
* Random Forest Classification: - 0.86 / Smote: - 0.88
* Logistic Regression Model: - 0.72 / Smote: - 0.67
* Support Vector Machine: - 0.71/ Smote: - 0.51

However, the model building must undergo further analysis as it does not show the accuracy of the dataset of the employees that remained and the employees that have already left. Thus, further analysis needs to be made to have the proper model for this business study.

**Precision and Recall**

Precision and Recall are commonly used metrics to evaluate the performance of models trained on an imbalanced dataset. For this study, the model built must be accurate in testing the employee who stayed and the employee who left. Thus, this analysis will examine further the accuracy of each test. The binary numbers **‘0.0’** and **‘0.1’** will be used for the employee remaining and the employee left respectively.

1. Random Forest Classification



The classification of the report is as follows: -

* **Precision: -**The precision for class 0.0 (non-leave) is 0.88, which means that among all instances predicted as non-leave, 88% were employees that remained. For class 1.0 (leave), the precision is 0.81% which means 81% of instances predicted as employee left.
* **Recall: -** The recall for 0 is 0.93 which means that among all actual non-leave instances, 93% were correctly classified as employees remained. For class 1, the recall is 0.70 indicating that only 70% of actual leave instances were correctly classified as employees leaving.
* **F1-scores: -** The F1-scores are the harmonic mean of precision and recall. It balances both metrics and is useful when classes are imbalanced. The F1 scores for 0.0 and 1.0 are 0.91 and 0.75 respectively.
* **Support: -** Support represents the number of actual occurrences of each class in the dataset where 0.0 has a support of 1337 and 1.0 has a support of 555.

The "macro avg" and "weighted avg" rows provide the average values of precision, recall, and F1-score across all classes. "Macro avg" calculates the unweighted mean of these metrics, while "weighted avg" considers the number of instances for each class, providing higher weight to classes with more instances.

The heatmap was then constructed for the Random Forest Classification: -

A graph with numbers and a number of classes

Description automatically generated with medium confidence

Based on the heatmap visualization of the confusion matrix, Random Forest Classification retrieved 1248 out of 1337 employees remaining and 387 out of 555 employees left which translates to 93% and 69% respectively.

1. Logistic Regression

A screenshot of a computer screen

Description automatically generated

The classification of the report is as follows: -

* **Precision: -**The precision for class 0.0 (non-leave) is 0.74, which means that among all instances predicted as non-leave, 74% of employees remained. For class 1.0 (leave), the precision is 0.68% which means 68% of instances predicted as employee leaving.
* **Recall: -** The recall for 0 is 0.97 which means that among all actual non-leave instances, 97% were correctly classified as employees remained. For class 1, the recall is 0.17 indicating that only 17% of actual leave instances were correctly classified as employees leaving.
* **F1-scores: -** The F1-scores are the harmonic mean of precision and recall. It balances both metrics and is useful when classes are imbalanced. The F1 scores for 0.0 and 1.0 are 0.84 and 0.28 respectively.
* **Support: -** Support represents the number of actual occurrences of each class in the dataset where 0.0 has a support of 1337 and 1.0 has a support of 555.

The heatmap was then constructed for the Logistic Regression: -

A graph showing a logistic regression

Description automatically generated

Based on the heatmap visualization of the confusion matrix, Logistic Regression retrieved 1291 out of 1337 employees remaining and 97 out of 555 employees left which translated to 96% and 17% respectively.

1. Support Vector Machine

A screenshot of a computer screen

Description automatically generated

The classification of the report is as follows: -

* **Precision: -**The precision for class 0.0 (non-leave) is 0.71, which means that among all instances predicted as non-leave, 71% were employees that remained. For class 1.0 (leave), the precision is 0 which means 0% of instances predicted as employee leaving.
* **Recall: -** The recall for 0 is 1.0 which means that among all actual non-leave instances, 100% were correctly classified as employees remained. For class 1, the recall is 0 indicating that only 0% of actual leave instances were correctly classified as employees leaving.
* **F1-scores: -** The F1-scores are the harmonic mean of precision and recall. It balances both metrics and is useful when classes are imbalanced. The F1 scores for 0.0 and 1.0 are 0.83 and 0 respectively.
* **Support: -** Support represents the number of actual occurrences of each class in the dataset where 0.0 has a support of 1337 and 1.0 has a support of 555.

The heatmap was then constructed for the Support Vector Machine: -

**A blue squares with numbers and a number of numbers

Description automatically generated with medium confidence**

Based on the heatmap visualization of the confusion matrix, Support Vector Machine retrieved 1337 out of 1337 employees remaining and 0 out of 555 employees left which translates to 100% and 0% respectively.

Based on the three models tested, Random Forest Classification obtained the most accurate f1-score for both remaining (0.0) and left (1.0). However, the models show an imbalance of data due to the nature of the dataset of having more employees remaining (0.0) compared to employees left (1.0).

The next step is to generate the Receiver Operating Characteristics to have a better depth in determining the right model for this business study.

**Receiver Operating Characteristic (ROC)**

The use of ROC is useful when the datasets are imbalanced. It provides an insight into the performance of the binary classification model (0.0 – employee remained, 1.0 – employee left). The closer the curve to the upper left corner indicates a better model performance.

Below is the graph for the ROC:-

A graph of a logistic and logistic regression

Description automatically generated with medium confidence

Based on the graph of the ROC, the Random Forest Classification curve graph is the closest to the upper left thus indicating the model to be better compared to others. The Area Under Curve (AUC) for Random Forest Classification is 0.82 which indicates that the model could separate the classes more effectively as it goes higher.

The Random Forest Classification seems to have a better model compared to others for this business study. Although it shows a positive result compared to the other two models, the dataset shown by the model was imbalanced. Thus, it could affect the accuracy of the data and can show incorrect results.

**Synthetic Minority Oversampling Technique (SMOTE)**

SMOTE is used to address class imbalance in machine learning datasets, particularly in binary classification tasks where one class is significantly underrepresented compared to the other.

1. SMOTE Random Forest Classification

A screenshot of a computer

Description automatically generated

The classification of the report is as follows: -

* **Precision: -**The precision for class 0.0 (non-leave) is 0.82, which means that among all instances predicted as non-leave, 82% were employees that remained. For class 1.0 (leave), the precision is 0.85% which means 85% of instances predicted as employees that left.
* **Recall: -** The recall for 0 is 0.86 which means that among all actual non-leave instances, 86% were correctly classified as employees that remained. For class 1, the recall is 0.81 indicating that only 81% of actual leave instances were correctly classified as employees that left.
* **F1-scores: -** The F1-scores are the harmonic mean of precision and recall. It balances both metrics and is useful when classes are imbalanced. The F1 scores for 0.0 and 1.0 are 0.84 and 0.83 respectively.
* **Support: -** Support represents the number of actual occurrences of each class in the dataset where both 0.0 and 1.0 have a support of 1337.

The "macro avg" and "weighted avg" rows provide the average values of precision, recall, and F1-score across all classes. "Macro avg" calculates the unweighted mean of these metrics, while "weighted avg" considers the number of instances of employee class, providing higher weight to classes with more instances.

The heatmap was then constructed for the SMOTE Random Forest Classification: -

A graph with numbers and a number in a row

Description automatically generated with medium confidence

Based on the heatmap visualization of the confusion matrix, SMOTE Random Forest Classification retrieved 1154 out of 1337 employees remaining and 1072 out of 1337 employees left which translated to 86% and 80% respectively.

1. SMOTE Logistic Regression

A screenshot of a computer

Description automatically generated

The classification of the report is as follows: -

* **Precision: -**The precision for class 0.0 (non-leave) is 0.69, which means that among all instances predicted 69% of employees remained. For class 1.0 (leave), the precision is 0.67% translates to 67% of instances predicted for employee leaving.
* **Recall: -** The recall for 0 is 0.65 which means that among all actual non-leave instances, 65% were correctly classified as employee remained. For class 1, the recall is 0.72 indicating that only 72% of actual leave instances were correctly classified as employees that left.
* **F1-scores: -** The F1-scores are the harmonic mean of precision and recall. It balances both metrics and is useful when classes are imbalanced. The F1 scores for 0.0 and 1.0 are 0.67 and 0.69 respectively.
* **Support: -** Support represents the number of actual occurrences of each class in the dataset where both 0.0 and 1.0 have a support of 1337.

The heatmap was then constructed for the SMOTE Logistic Regression: -

A graph with numbers and a number in blue squares

Description automatically generated

Based on the heatmap visualization of the confusion matrix, SMOTE Logistic Regression retrieved 865 out of 1337 employees remaining and 949 out of 1337 employees left which translates to 64% and 71% respectively.

1. SMOTE Support Vector Machine

A screenshot of a computer

Description automatically generated

The classification of the report is as follows: -

* **Precision: -**The precision for class 0.0 (non-leave) is 0.55, which means that among all instances predicted as employees that remained, 55% were employees that remained. For class 1.0 (leave), the precision is 0.54% which means 54% of instances predicted as employee leaving.
* **Recall: -** The recall for 0 is 0.46 which means that among all actual non-leave instances, 46% were correctly classified as employees remained. For class 1, the recall is 0.63 indicating that only 63% of actual leave instances were correctly classified as employees that left.
* **F1-scores: -** F1-scores are the harmonic mean of precision and recall. It balances both metrics and is useful when classes are imbalanced. The F1 scores for 0.0 and 1.0 are 0.51 and 0.58 respectively.
* **Support: -** Support represents the number of actual occurrences of each class in the dataset where both 0.0 and 1.0 have a support of 1337.

The heatmap was then constructed for the SMOTE Support Vector Machine: -

A blue squares with numbers and a number on it

Description automatically generated with medium confidence

Based on the heatmap visualization of the confusion matrix, SMOTE Support Vector Machine retrieved 617 out of 1337 employees remaining and 837 out of 1337 employees left which translates to 46% and 63% respectively.

Based on the three models tested, SMOTE Random Forest Classification obtained the most accurate f1-score for both remaining (0.0) and left (1.0). However, unlike the Random Forest Classification, the SMOTE Random Forest Classification had shown desirable f1-scores of employees leaving at 0.83, compared to 0.75. Thus, the accuracy of this model in predicting employee turnover is much better.

**SMOTE Receiver Operating Characteristic (ROC)**

A graph of a logistic and logistic

Description automatically generated

The graph above shows the SMOTE ROC for the three models i.e. SMOTE Logistic Regression, SMOTE Random Forest Classification, and SMOTE Support Vector Machine. In this graph, the SMOTE Random Forest Classification curve graph is the closest to the upper left thus indicating the model to be better compared to other SMOTE models. However, compared to the default ROC curve, the adjustment for the SMOTE Random Forest Classification and SMOTE Logistic Regression does not seem to have major or rather obscure changes except for the SMOTE Support Vector Machine. Thus, the graph for both default and SMOTE indicates similar results for the Random Forest Classification.

After comparing all the models above, the default Random Forest Classification stood to have the highest accuracy although the SMOTE had been included. It can be concluded that the most suitable model for this business study is the default Random Forest Classification.

**Feature Importance Analysis**

Based on the dataset given, the following percentage represents the importance of features used in building the model: -

1. **Department\_operations:** This feature has a percentage of 0.88% which is the least important among the other features.
2. **Department\_engineering and Department\_sales :** Both of these features have a percentage value of 0.91 % which is relatively low.
3. **Department\_Admin:** Department\_Admin has a percentage value of 1.03%.
4. **Department\_Support:** This feature has a slightly higher percentage compared to the above features which is 1.18%.
5. **Review:** Review has a percentage value of 2.16%, suggesting it has a relatively high importance or contribution towards the outcome being studied.
6. **Tenure:** Tenure has a high level of percentage value of 10.52% and it has relatively importance to contribute into the study of turnover
7. **Promoted:** Promoted has a high level of percentage value of 25.73% and it contributes a high impact towards the outcome of the study.
8. **Satisfaction:** Satisfaction has the second highest percentage value of 27.12%, indicating the higher impact of contribution on the outcome of the study.
9. **Department\_Finance:** obtained the highest percentage value of 29.26%. This percentage shows that the finance department has the highest importance among the features mentioned.

# **Recommendations**

The recommendations that can reduce employee turnover issues include promoting a supportive working environment. In this business study, the finance and support department have a higher turnover rate and may benefit from this initiative as it can improve job satisfaction. Some improvements that the company can make include providing additional resources, support systems, and adequate training as these can enhance job performance and satisfaction.

The company with a high turnover should be able to improve in managing employee salaries, especially in the departments where the number of employees resigning is high. Higher salaries would directly influence employee satisfaction and their perception of the working environment. Besides competitive salaries, recognizing and rewarding employees for their tenure and loyalty to the company will also improve employee satisfaction. These methods will build a sense of appreciation and loyalty among employees, hence increasing the longevity and commitment to the company.

Employees with outstanding achievements should be given promotion opportunities. Employees who feel they have opportunities for growth and advancement with the company are less likely to leave. However, the company should play a vital role by providing adequate training and development programs to increase the ability of the employees to perform better in their jobs. For instance, if an individual does not possess any certification of certain degrees but is determined and keen to sacrifice his or her time to learn in the lights on contributing better to the industries. Companies can offer certification programs allowing individuals to achieve a higher dream and a higher position in the workplace. In return, companies could benefit from retaining one of the talented employees and at the same time improving the overall turnover rate as opportunities are higher in the workplace now.

Creating a positive working environment must be a priority for the company, making sure that all employees feel supported and valuable. Departments that experience high turnover rates will benefit from initiatives focused on enhancing the working environment. Constant feedback sessions with employees could provide valuable information on areas for improvement, allowing the company to combat weaknesses and cultivate a more positive and conducive workplace atmosphere. By implementing a supportive environment and actively seeking employee input, the company would increase job satisfaction and ultimately decrease the turnover rates.

# **Future Work Recommendation**

There are some improvements that can be made to improve the accuracy of the building model to measure the employees to be more accurate. The first step is to explore ensemble methods where it could enhance the performance of machine learning by having more robust prediction. This can be done by experimenting with different combinations of base models, i.e. Random Forest Classification, SVM and Logistic Regression and testing them by stacking or by combining them to produce more accurate predictions.

The next improvement that can be made is by integrating a new data stream into the existing data source. This can help the machine to have the ability to make an in-depth analysis of employee turnover based on more variables given. One of the variables that can be tested is the employee sentiment analysis where the data can be extracted from social media or the performance review.

The interpretability of the model in predicting employee turnover is very important as it reflects the accuracy of the model built. Thus, improving the interpretability of the models is crucial and some techniques that can enhance the interpretation of the models are Feature Importance Analysis, Partial Dependence Plots or SHAP (Shapley Additive Explanations) value. These techniques help the model to explain model predictions and gain insights into the factors driving employee turnover.

One of the good methods to have in evaluating employee satisfaction towards the working environment is to embed the facial expression AI-detector. This AI will have the ability to detect real-time emotions where the data of the employees’ expressions can be used to analyze the employee's emotions. This can help in the detection of early stress and discontent from the employees and this can help the company to take early action. The early detection made by facial recognition can help to improve employee working experience.

# **Conclusion**

In conclusion, effectively predicting and managing employee turnover is unavoidable in ensuring any company’s success in this competitive society. Through utilizing these machine learning algorithms, companies can gain valuable insights into the factors that cause employee turnover and implement effective strategies to mitigate and reduce the rate of turnover. The vast evolution of machine learning tools creates great opportunities to achieve better accuracy in the prediction of employee turnover, ensuring that companies can predict with more precision. By fully utilizing and understanding the capabilities of machine learning, companies can make well-informed decisions and mitigation strategies that are specially tailored to individual employee needs. In summary, integrating data-driven and machine-learning approaches can easily create an environment that fosters employee satisfaction, longevity, and the success of the company. Through consistent evolution and innovation in employee retention strategies by using machine learning, companies can position themselves better in the competitive marketplace ultimately maximizing the success and the profit of the company.

# 

# **References**

* + - 1. Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. *Academy of management Perspectives*, *24*(2), 48-64.
      2. Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (n.d.). The “K” in K-fold Cross Validation. <http://www.i6doc.com/en/livre/?GCOI=28001100967420>.
      3. Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14(1). <https://doi.org/10.1186/1471-2105-14-106>
      4. Buckland, M., & Gey, F. (1994). The Relationship between Recall and Precision.
      5. Candy, J. V, Breitfeller, E. F., Candy, J. V, Breitfeller, E. F., & Livermore, L. (2013). Receiver Operating Characteristic (ROC) Curves: An Analysis Tool for Detection Performance Receiver Operating Characteristic (ROC) Curves: An Analysis Tool for Detection Performance Analysis.
      6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. <https://doi.org/10.1613/jair.953>
      7. Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), 102798.<https://doi.org/10.1016/j.ipm.2021.102798>
      8. Do, K. T., Wahl, S., Raffler, J., Molnos, S., Laimighofer, M., Adamski, J., Suhre, K., Strauch, K.,Peters, A., Gieger, C., Langenberg, C., Stewart, I. D., Theis, F. J., Grallert, H., Kastenmüller, G.,& Krumsiek, J. (2018). Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*, *14*(10).<https://doi.org/10.1007/s11306-018-1420-2>
      9. Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*(3), 463-488.
      10. Komarek, P., Moore, A., Em, U., Ed U Committee, A., Calvet, A., Frieze, B., & Nichol. (2004). *Logistic Regression for Data Mining and High-Dimensional Classification*.
      11. Lee, T. W., & Mitchell, T. R. (1994). An alternative approach: The unfolding model of voluntary employee turnover. *Academy of Management Review, 19*(1), 51-89.
      12. Liu, L., Akkineni, S., Story, P., & Davis, C. (2020). Using HR analytics to support managerial decisions: A case study. ACMSE 2020 - Proceedings of the 2020 ACM Southeast Conference, 168–175. <https://doi.org/10.1145/3374135.3385281>
      13. Mehrolia, S., Alagarsamy, S., & Solaikutty, V. M. (2020). Customers' response to online food delivery services during COVID‐19 outbreak using binary logistic regression. *InternationalJournal of Consumer Studies*, *45*(3), 396–408. <https://doi.org/10.1111/ijcs.12630>
      14. Pratibha, G., & Hegde, N. P. (2022). HR Analytics: Early Prediction of Employee Attrition using KPCA and Adaptive K-means based Logistic Regression. Proceedings - 2022 2nd International Conference on Interdisciplinary Cyber-Physical Systems, ICPS 2022, 11–16. <https://doi.org/10.1109/ICPS55917.2022.00010>
      15. Ren, R., Wu, D. D., & Wu, D. D. (2019). Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Systems Journal, 13(1), 760–770. <https://doi.org/10.1109/JSYST.2018.2794462>
      16. Shaw, J. D., Gupta, N., & Delery, J. E. (2005). Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. *Academy of Management Journal,48*(1), 50-68.
      17. Tian, Y., Shi, Y., & Liu, X. (2012). RECENT ADVANCES ON SUPPORT VECTORMACHINES RESEARCH. *Technological and Economic Development of Economy*, *18*(1), 5–33.<https://doi.org/10.3846/20294913.2012.661205>
      18. Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., & Alharbey, R. (2021). An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction. Scientific Programming, 2021, 1–12. <https://doi.org/10.1155/2021/6621622>
      19. Wong, T.-T., & Yeh, P.-Y. (2020). Reliable Accuracy Estimates from *k*-Fold Cross Validation.*IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586–1594.<https://doi.org/10.1109/tkde.2019.2912815>
      20. Ongori, H. (2007). A review of the literature on employee turnover. *African Journal of Business Management*, 49–054. <http://www.academicjournals.org/ajbm>
      21. Tyler Jadah. (2023, November 30). *27 Employee Turnover Statistics That Might Surprise You*.Applauz Resources. <https://www.applauz.me/resources/employee-turnover-statistics>