**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission.*

|  |  |
| --- | --- |
| **Module Title:** | *Statistical Techniques of Data Analysis*  *Data Preparation*  *Machine Learning* |
| **Assessment Title:** | *CA2\_STDA\_DP\_ML\_HDip\_Lvl8* |
| **Lecturer Name:** | *Marina Soledad Iantorno, James Garza, Muhammad Iqbal* |
| **Student Full Name:** | Ariel Goldman |
| **Student Number:** | sbs23073 |
| **Assessment Due Date:** | 24/05/2023 |
| **Date of Submission:** | 24/05/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Contents

[Introduction 1](#_Toc135845664)

[Dataset summary 1](#_Toc135845665)

[Technology Used 1](#_Toc135845666)

[Models 1](#_Toc135845667)

[Libraries 2](#_Toc135845668)

[Data Preparation 3](#_Toc135845669)

[Exploratory Data Analysis 7](#_Toc135845670)

[Statistical Techniques 14](#_Toc135845671)

[Central Tendency Metrics for Attrition 14](#_Toc135845672)

[Measures of Dispersion: Variance and Standard Deviation for Attrition 15](#_Toc135845673)

[ANOVA 16](#_Toc135845674)

[T-Test, one population 18](#_Toc135845675)

[Confidence Interval 20](#_Toc135845676)

[Correlated Heatmap 21](#_Toc135845677)

[Introduction to Machine Learning Models 22](#_Toc135845678)

[Modelling 22](#_Toc135845679)

[Machine Learning models to be analysed. 23](#_Toc135845680)

[Support Vector Machine 24](#_Toc135845681)

[Grid Search to Find Optimal Hyperparameters (KERNEL='RBF') 24](#_Toc135845682)

[Building and Evaluating the Final Model (KERNEL='RBF') 24](#_Toc135845683)

[Grid Search to Find Optimal Hyperparameters (KERNEL='LINEAR') 25](#_Toc135845684)

[Building and Evaluating the Final Model (KERNEL='LINEAR') 26](#_Toc135845685)

[Performance using Grid Search with RBF Kernel Vs Linear Kernel. 26](#_Toc135845686)

[Random Forest 27](#_Toc135845687)

[Hyperparameter Tuning 27](#_Toc135845688)

[Grid Search to Find Optimal Hyperparameters 28](#_Toc135845689)

[Classification Report 28](#_Toc135845690)

[Features Importances 29](#_Toc135845691)

[Evaluation Metrics SVM Vs RF 30](#_Toc135845692)

[Principal Components Analysis 31](#_Toc135845693)

[Linear Discriminant Analysis 32](#_Toc135845694)

[References 33](#_Toc135845695)

# Introduction

In the current times of remote work and home office, the workface is changing rapidly, adults and more specifically young adults are seeking a new type of job or employment contract. This is because employees who want to leave a company has become one of the main concerns for companies as young adults are looking for more flexibility and work-life balance in their jobs.   
After the COVID-19 pandemic, there has been a significantly increase in remote work and employees do not need to be in a specific location. In other words, they are looking for jobs that allow them to work from anywhere while they can still maintain high levels of productivity and job satisfaction.

Interestingly, according to [BasuMallick](https://www.spiceworks.com/user/about/chiradeepbasumallick) (2021) there five types of employee attrition:

1. Attrition due to retirement
2. Voluntary attrition
3. Involuntary attrition
4. Internal attrition
5. Demographic-specific attrition

With that being said, my objective is to build a predictive machine learning model that can accurately classify employees either churn or not churn based on these features, among other techniques I will apply.

In summary, the idea of this project is to provide insights and recommendations to help IBM to reduce employee attrition, increase employee retention, and improve its overall performance.

## Dataset summary

The dataset from IBM contains 1,470 observations and 35 features, with features indicating the employee’s attrition. Specifically, there are 1233 employees currently working in the company while there are 237 employees who left the company.

# Technology Used

## Models

For my project I have implemented two Machine Learning models, one of them is Support Vector Machine (SVM) and the other one Random Forest to build predictive models for customer churn. I used hyperparameters such as Kernel RBF and Linear and applied cross-validation, classification reports and confusion matrices to evaluate the performance of the models.

I have also performed the Hyperparameter Tuning applied for Random Forest model by specifying the number of folds for ‘k-fold’ and adjusting the parameters to plot the accuracies of ‘max\_depth’.

## Libraries

Different libraries have been used for the purpose of performing the analysis of the dataset which is being implemented in Jupyter Notebook.

The following libraries are crucial for data analysis:

* Warnings will supress the errors that would normally be displayed.
* Scikit-learn for machine learning tasks.
* Seaborn and Matplotlib for visualizations.
* Pandas
* NumPy
* Encoders to encode categorical variables.
* data\_profiling which provides a way to quickly generates an overview of the dataset.
* Imbelear.over\_sampling.SMOTE for oversampling imbalanced datasets
* Missigno for analysing missing data.
* Scipy.stats

# Data Preparation

In this stage, I will develop a deep understanding of the data I have.

By Importing ProfileReport library, it generates a detailed report that summarised the statistical measures and visualizations of the data and it allows a quick overview of the data we are dealing with.

A screenshot of a computer

Description automatically generated with low confidence

After performing .head() and .tail() we can see that the dataset contains 35 columns and 1,470 observations (0 to 1,469) in the dataset.

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

After executing the function .info() there are:

* 1470 observations (0 to 1,469)
* 26 attributes contain integer numbers.
* 9 attributes contain string values.
* Total columns 35.

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, font, white, typography

Description automatically generated

Importing the library ‘Missigno’ it is more visual to analyse whether there are missing values or not. In this case, there are no missing or null values in the dataset.

A picture containing line, rectangle, pattern, grate

Description automatically generated

Removing the unnecessary features from the dataset which are not relevant or add redundancy as they might cause biased to the model.

A screenshot of a computer

Description automatically generated with medium confidence

The aim of performing the function .nunique() is to find unique values for each column to understand which column is categorical and which one is continuous. In other words, if the number of unique values is less than 20 then the variable is likely to be categorical otherwise continuous.

A screenshot of a computer

Description automatically generated with low confidence

# Exploratory Data Analysis

In this phase we need to prepare the data to perform a deep analysis through Exploratory Data Analysis (EDA).

It is clear that the number of existing employees (‘No’) is much higher than the number of employees who left IMB (‘Yes’) in the target variable, which means that we are dealing with class imbalance.

With that being said, Tara Boyle (2019) has stated that most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error.

I will address imbalanced data before applying the Machine Learning models chosen.

A screenshot of a graph

Description automatically generated with low confidence A picture containing text, circle, diagram, font

Description automatically generated

The younger employees are more likely to leave the company compared to the older employees. Analysing the boxplot, the median age for attrition is around 36 years old.   
There may be various reasons why younger employees want to leave the company such as lack of career growth or career opportunities, preference for flexible hours and working fixed hours rather than working for objectives.

A picture containing screenshot, diagram, rectangle, design

Description automatically generated

The majority of the employees have studied Life Sciences and Medical fields, which means that IMB has a strong human resources and experts working in areas such as healthcare, pharmaceuticals, and biotechnology in the Department of Research & Development where the staff is more concentrated.

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, rectangle

Description automatically generated

The majority of employees are based within a range of 2 to 10 kilometres from the company. Although there is no correlation to suggest that employees who live closer to IMB will be more satisfied, it is important to consider this factor such as commute times and transportation options.

A picture containing text, plot, screenshot, diagram

Description automatically generated

It is clear that the histogram shows the majority of the employees have a tenure of up to five years, with a rapid decrease in the data distribution after ten years working at IBM. In contrast, the distribution of Years in Current Role seems to have ups and downs. The majority of the employees have worked in the same role for two years, with a decrease until five years, after which the distribution line goes up until seven years before rapidly decreasing again.   
Emphasizing on the Years in Current Role, this may indicate a lack of career growth or opportunities within the company suggesting employees either leaving or switching to different roles.

A picture containing screenshot, plot, text, line

Description automatically generatedA picture containing text, diagram, screenshot, plot

Description automatically generated

There is a strong relationship observed at levels 2 and 3 between Satisfaction and Job Involvement with Work-life Balance for both the attrited employees and those who are still employed. However, there seems to be weak relationship between Job Involvement and Work-life Balance Job Involvement at levels 1 and 4.

A screenshot of a graph

Description automatically generated with low confidence A picture containing line, plot, rectangle, diagram

Description automatically generated

There are no clear clusters for both scatterplots, however, by analysing the dots in ‘Attrition-No’ the employees who are still employed have a higher monthly income with a total of at least 20 years of work and wages starting from 15,000. On the other hand, the most densely populated cluster has a monthly income within a range of 2,500 to 10,000 with a total of working years between 1 and 20.

In contrast, among employees marked as ‘Attrition-Yes’, the majority of the employees had worked between 1 and 10 years. Particularly, there were only two employees who had worked 40 years and also three employees with more than 30 years of service at IMB, assuming that they had retired.

A graph showing an attrition and an attrition

Description automatically generated with low confidence A picture containing text, screenshot, line, plot

Description automatically generated

Given the boxplots, it is evident that the Managers and Research Directors are the highest paid roles at 20,000 monthly, however, Managers have a higher median income than Research Director, at 17,500 per month and roughly 16,000 per month respectively. Followed by Healthcare Representatives, who earn a bit more than Manufacturing Directors. Research Scientists and Laboratory Technician are equally paid roles. Finally, the lowest paid role is Sales Representatives.

A screenshot of a graph

Description automatically generated with medium confidence

The job Levels at IBM are named by numbers, being the lowest level paid 1 and highest level paid 5. It is clear that Job Level 5 is the highest level paid, it suggests that the Managers and Research Directors are in this position and the lowest level paid is 1, assuming that Sales Representatives are in this role.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Performance Rating and Percentage Salary hike are strongly correlated. By analysing at the same time, the countplot, scatterplot and histoplot, it is evident that the higher the Performance Rating, the higher the Percentage Salary Hike is. The majority of the employees (around 200 obtained paid rise between 12% and 14%) have a Performance Rating of 3, they are given a paid rise between 11% and 19%, In contrast, the minority of the employees (around 50 obtained paid rise between 22% and 25%) who have a Performance Rating of 4, they are given a paid rise between 20% and 25%.

A picture containing text, screenshot, rectangle, diagram

Description automatically generatedA picture containing text, screenshot, number, plot

Description automatically generatedA picture containing text, plot, diagram, screenshot

Description automatically generated

# Statistical Techniques

## Central Tendency Metrics for Attrition

## 

These statistical measures can be useful for understanding the characteristics of the two groups (Attrition ‘Yes’ and Attrition ‘No’) to identifying trends that may be relevant for the company.

A screenshot of a computer

Description automatically generated with medium confidence

The outcomes suggest that employees who left the company have worked, on average, for fewer years and have lower monthly incomes compared to employees who stayed, based on the median. This measure can be more representative and is not influenced by extreme values. Additionally, employees who left the company tend to be younger than those who stayed, based on the mode of their age.

To sum up, it is evident that younger employees with lower salaries tend to leave IBM earlier than the other group.

## Measures of Dispersion: Variance and Standard Deviation for Attrition

The variance and standard deviation are measures which explain how spread the dataset is from the mean.

It seems that there is a higher variation in Total Working Years and Monthly Income for employees who left the company compared to those who stayed. This could suggest that employees who leave IBM are dissatisfied with their salary or career progression.

On the other hand, the STD for age is not significantly different between the two groups, which could indicate that age is not the main factor in employee attrition.

A screenshot of a computer

Description automatically generated with low confidence

**Conclusion**

After evaluating measures of central tendency and measures of dispersion, I would suggest to IMB consider taking actions to retain young employees, such as offering career opportunities, trainings, flexible working hour, competitive salaries, and benefits.

## ANOVA

ANOVA can be used to determine if there is a significant difference in the average ‘Age’ between different Departments.

We are analysing numerical and categorical variables (Age is numerical and Department is categorical).

1. After plotting the numerical variable ‘Age’, the histogram is not exactly symmetric, however, if it is a bit skewed, we can proceed with the analysis. Then I plotted the Q-Q Plot to compared it with the histogram and analyse the same information with another visualization.

A picture containing screenshot, diagram, plot, electric blue

Description automatically generated A picture containing text, plot, screenshot, line

Description automatically generated

2. I will use Shapiro Wilk Test to determine the normality of the numerical variable ‘Age’ across three Departments: Research & Development, Sales, and Human Resources.

A picture containing text, screenshot, font, line

Description automatically generated

In this case, only one out of three categories (Human Resources) are normally distributed according to the Shapiro Wilk Test results.

p-value for Human Resources Department is the only category greater than 0.05 (significance level), which means that there is no strong evidence to reject null hypothesis.

3. Levene Test

In this step, I will perform the Leneve Test to determine if the variances of ‘Age’ variable are equal across the different Departments and the data is cantered.

H0 : The variances between the department are equal.

H1: The variances between the departments are not equal.

A picture containing text, screenshot, font, line

Description automatically generated

In this case, p-value is greater than alpha, then we accept the null hypothesis and therefore we can say that the variances are equal.

4. ANOVA ONE-WAY

With the one-way ANOVA, I will compare the means across the three departments. I have conducted an F-test on the data and calculate the p-value in Excel and Python.

Excel results:

A picture containing text, screenshot, font, number

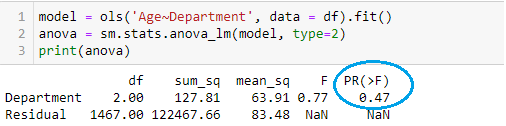
Description automatically generated A picture containing text, screenshot, font, number

Description automatically generated

F < Fc AND p-value > 0.05 then I accept null hypothesis.

Python results:

I have also performed one-way ANOVA in Python to find the F of the test and the p-value.



At 5% significance level, there is no evidence to say that there are differences in the ‘Age’ variable across the three Departments. In other words, I accept null hypothesis, therefore there is no evidence to say that the means are not equal.

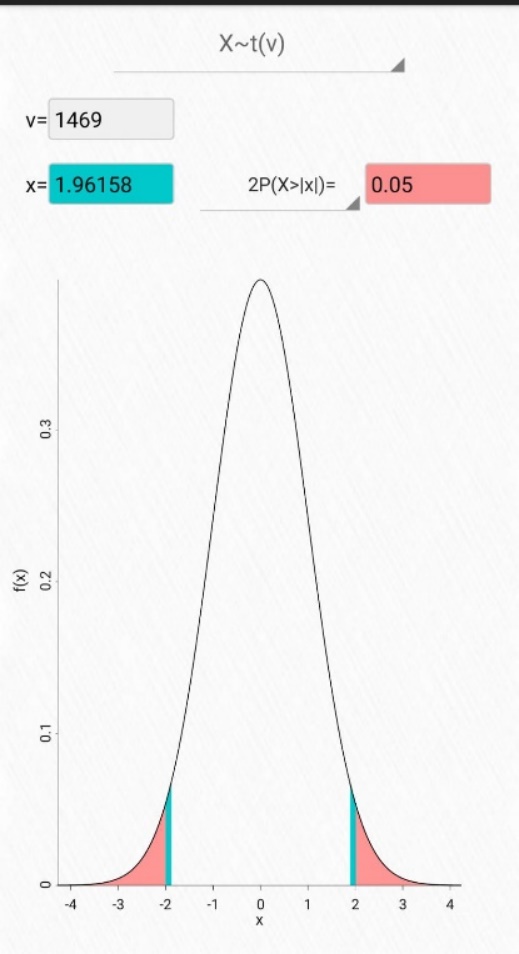
## T-Test, one population

I will intend to determine the average salary of the employees at IMB. As a data analyst, I need to know if the Monthly Income is, in average 6200, in order to suggest the management whether to provide a pay rise or not to the staff. For that reason, I will take the Monthly Income for all the employees to perform a test and prove the claim at 5% significance level.

A) **Hypothesis Statement**

H0: μ = 6200  
H1: μ =! 6200

B) **Critical Values**



C) **Conclusion**

I reject null hypothesis and conclude that there is enough evidence to support that the average Monthly Income is not 6,200 at 5% significance level. Therefore, I would suggest to the management a pay rise might be necessary for the staff by taking into account other factors before making a final decision.

Continuing with the analysis, I also want to know if the salary is, in average, greater than 6,000 per month. The significance level is calculated at 5%.

H0: μ = 6000  
H1: μ > 6000

A picture containing text, screenshot, font, line

Description automatically generated

**Conclusion**

I reject null hypothesis and conclude that there is enough evidence to suggest that the average income at IBM is greater than 6,000 per month calculated at 5% significance level.

## Confidence Interval

The existence of the gender pay gap is a worldwide issue for many individuals and organisations. It refers to the difference in the average earnings between men and woman in a company.

As a data analyst, I intend to determine whether there is any gender gap or not at IBM, and if so, which gender earns more on average and what could be the reasons.

I will apply a statistical test called Confidence Interval to support my findings, which is basically find a range between the monthly income for males and females.

The interval Monthly Income for Males at IBM is between 13,601.24 and 14,542,97 .

A screenshot of a computer

Description automatically generated with low confidence

The average Monthly Income for Females at IBM is between 14,099.74 and 15,249.46.

A picture containing text, font, screenshot, line

Description automatically generated

Female employees at IBM earn on average 4.28% more than male employees.

A close-up of a computer code

Description automatically generated with low confidence

**Conclusion**

Based on my analysis, there is a small percentage difference of 4.28% between gender earnings interval with a confidence interval of 95%. Therefore, I would suggest to the management to take actions and consider this gender pay gap. However, I would also advise them to consider other factors such as responsibilities, performance rating, job levels and job roles.

## Correlated Heatmap

Given the fact that the Attrited Employees are only 16% of the sample, from my point of view is a good practice to analyse the employees as a whole and not only the ones who have left IMB, as I consider the employees who remain in the company are likely to churn in the future as well. Therefore, I have decided to perform a correlated heatmap to analyse the features accurately and identify potential patterns or trends in the data.

It is evident that there is a strong positive correlation between ‘JobLevel’ and ‘MonthlyIncome’ as the more Job Level an employee is the more will earn, while the features highlighted in orange are also positively correlated.

A picture containing screenshot, pattern, diagram, text

Description automatically generated

# Introduction to Machine Learning Models

The initial question we need to address is how we can improve, reduce errors, and avoid bias when applying Machine Learning Model?

There are steps that I have implemented to improve the accuracy of the models:

1. Encoding Categorical variables into numerical variables.

We have selected the categorical features to transform them into numerical variables using the Ordinal Encoder.

2. Data normalization

As I do not have negative values in our dataset, I rescaled the continuous variables applying MinMaxScaler in our data to a range between 0 and 1. Particularly, this method is useful to see all the variables from the same lens (same scale), in this way we will bring all values into the range [0,1].

When we Encode Categorical Data, we turned string variables into numerical variables, when we did that, we do not have to scale or normalized that data, it is not recommended to normalize or scale them because they are no longer continuous variables.

3. Separate and define the dataset into X (input features) and y (target variable) and then split the data into independent and dependent variable.

The main variable for predicting an attrited employee or not is the target variable (dependant variable) ‘Attrition’, which is a binary classification. Then the model evaluates ‘y’ depending on the features from the employees, such as age, marital status, monthly income, education field and years at the company.

## Modelling

In this phase, I have built and trained the Machine Learning models across three different splits (30%/20%/10%). I have split the dataset into training and testing set rather than the entire dataset. In other words, I have scaled the data after splitting. Thus, I was able to scale the train and test set separately to prevent data leakage (Jason Brownlee, 2016)

# Machine Learning models to be analysed.

I have chosen two Machine Learning models to analyse: Support Vector Machine and Random Forest. The decision was made as both models can be used for binary classification.

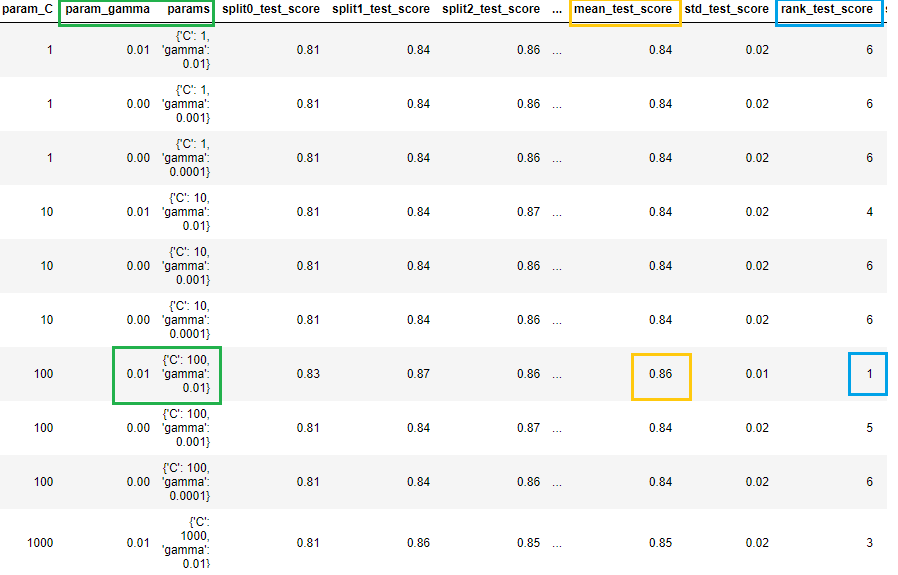
Support Vector Machine and Random Forest are supervised learning models which can be used for both classification or regression challenges. However, it is mostly used in classification problems where the data is sparse (easy to classify). Pragati (2020).

With that being said, considering that the data is easy to classify (Attrition ‘Yes’ / ‘No’), both SVM and Random Forest models can effectively handle classifications task.

I will be splitting the models across three different splits (30%/20%/10%) by using Confusion Matrices, K-Fold Cross-Validation and Hyperparameters.

# Support Vector Machine

## Grid Search to Find Optimal Hyperparameters (KERNEL='RBF')



The best test score is 0.8587 corresponding to hyperparameters {'C': 100, 'gamma': 0.01}

A picture containing text, diagram, screenshot, plot

Description automatically generated

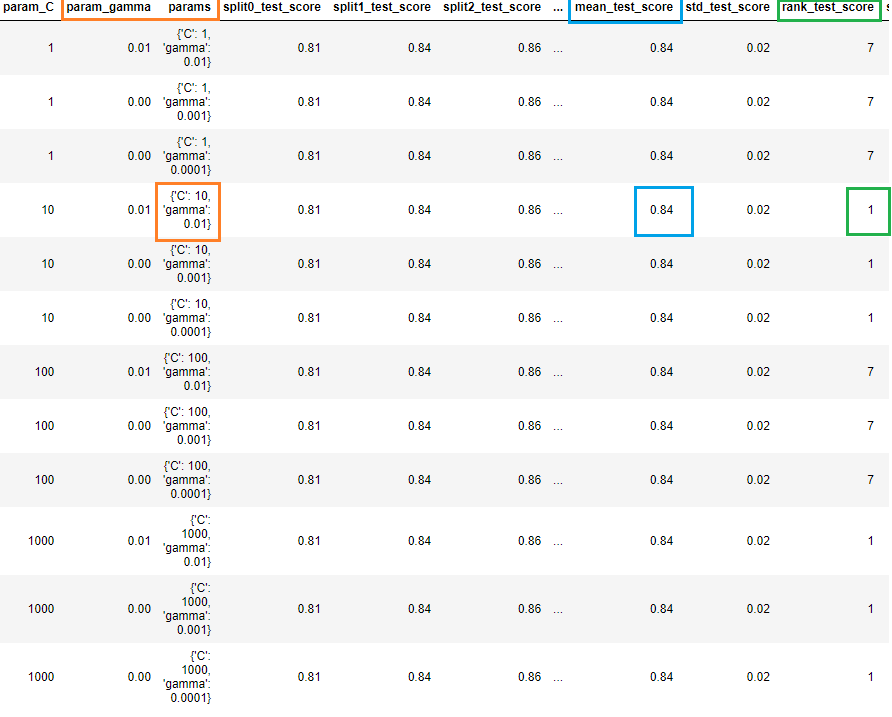
## Building and Evaluating the Final Model (KERNEL='RBF')

A screenshot of a computer code

Description automatically generated with low confidence

The model has high accuracy, however, did not correctly identify any positve cases, resulting in a precision and recall of 0.0. This means all the predicitons made by the model were negative. In other words, the model has failed in identifying any positive cases correctly and also is not effectively distinguishing between attrited and non attrited employees.

## Grid Search to Find Optimal Hyperparameters (KERNEL='LINEAR')



The best test score is 0.8405 corresponding to hyperparameters {'C': 10, 'gamma': 0.01}

A picture containing text, screenshot, plot, line

Description automatically generated

## Building and Evaluating the Final Model (KERNEL='LINEAR')

A picture containing text, font, white, algebra

Description automatically generated

The model has high accuracy and precision, but a low recall indicating the model is missing considerable number of attrited employees.

## Performance using Grid Search with RBF Kernel Vs Linear Kernel.

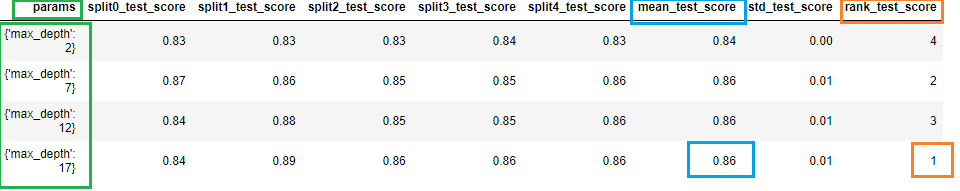
The Linear Kernel model performed better than the RBF kernel model in terms of precision and recall. However, RBF Kernel model had a higher train accuracy (around 92%) compared to the Linear Kernel model (around 85%).

For both cases, the test accuracy is similar. However, it appears that the model may be slightly overfitting the training data. The train accuracy of 92% (RBF Kernel ) suggests that the model is able to work quite well. But on the other hand, the test accuracy of 84% (Linear Kernel) is lower than the train accuracy, indicating that the model may not generalize well to new, unseen data.

# Random Forest

## Hyperparameter Tuning

I have performed the Hyperparameter Tuning applied for Random Forest model by specifying the number of folds (5), which means that the dataset will be split into 5 and also, I adjusted the parameters to plot the accuracy of ‘max\_depth’ in a range of (2, 20, 5), then I used the classification report to assess the model’s performance.

A picture containing diagram, plot, text, line

Description automatically generated

After fitting the model with the values of ‘max\_depth’, the above plot indicates that the test accuracy 86% is high when ‘max\_depth’ is set in a range of 2 to 20 while splitting the model into 5.

## Grid Search to Find Optimal Hyperparameters

We can now find the optimal hyperparameters using GridSearchCV.

We can get accuracy of 0.8316 using {'max\_depth': 4, 'max\_features': 5, 'min\_samples\_leaf': 100, 'min\_samples\_split': 200, 'n\_estimators': 100}

## Classification Report

A picture containing text, screenshot, font, number

Description automatically generated

The overall accuracy of the model is 0.87, wich means that it correctly classified roughly 87% of all ocurrencies in the dataset. Moreover, the model seems to perform well for class 2 (non-attrited employees) but has poor performance for class 1 (attrited employees). Therefore, further investigton is needed to improve its performance for attrited employees.

## Features Importances

Feature Importance refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the “importance” of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. (Terence Shin, 2021).

It is clear that ‘OverTime’ is the highest importance feature in Random Forest Model, it means that it is the most important feature for predicting the target variable. It appears that this feature is an important predictor of employee attrition. One possible interpretation would be that the employees who work overtime are more likely to experience burnout, which could lead to higher levels of job dissatisfaction and as a result, they will probably leave the company.

Plotting the feature importance for Top 10 most important columns

A picture containing text, screenshot, number, font

Description automatically generatedA picture containing text, screenshot, software, web page

Description automatically generated

# Evaluation Metrics SVM Vs RF

A screenshot of a computer

Description automatically generated with low confidence

Similarities:

Both SVM and RF models show similar performance across the three splits. For both models, the accuracy scores remain close, with only minor variations. A higher accuracy score indicates the model has better performance.

Contrasts:  
  
RF model’s performance may vary more depending on how the data is split for cross-validation, while SVM model’s performance remains more stable.

Confusion Matrix SVM and RF

It seems the models have performed well across all test sets. Where type I error means the models incorrectly predicted that the employees are attrited, but they are actually not attrited. Type I error in this case is relatively high in the test set 30%. Conversely, type II error means the model incorrectly predicted that the employees are not attrited, but they are actually attrited.

# Principal Components Analysis

it can be observed that the Principal Components Analysis describe a significant amount of variance in the data, while the amount of variance explained by each successive Principal Component decreases rapidly. In this scenario, it seems that with 3 Principal Components we can retain most of variance. In other words, by specifying 3 components in the PCA model, means that we have retained the most important information in those components while reducing the dimensionality of the dataset.

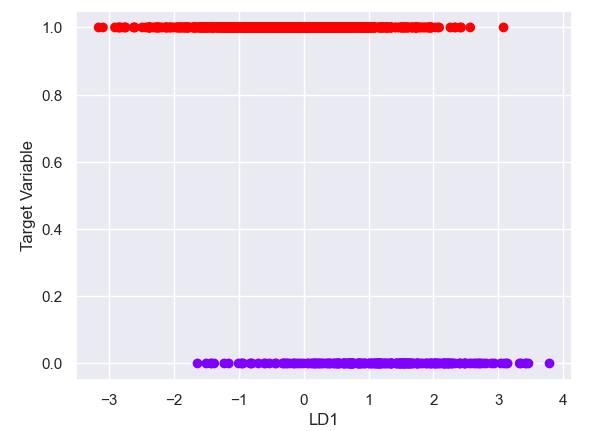
A picture containing text, screenshot, line, plot

Description automatically generated

# Linear Discriminant Analysis

The maximum number of components that LDA can find is equal to the number of classes minus one in the classification dataset. (Rukshan Pramoditha, 2022)

Since the dataset has two class labels [‘Yes’ ‘No’] I was not able to perform an appropriate LDA due to the limitation of having only one component available as the dataset consisting of two class labels.



# References

Boyle, Tara. “Methods for Dealing with Imbalanced Data.” *Medium*, Towards Data Science, 3 Feb. 2019, towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18.

BasuMallick, Chiradeep . “What Is Employee Attrition? Definition, Attrition Rate, Factors, and Reduction Best Practices.” *Spiceworks*, 11 Mar. 2021, www.spiceworks.com/hr/engagement-retention/articles/what-is-attrition-complete-guide/#\_005. Accessed 12 May 2023.

Brownlee, Jason. “Data Leakage in Machine Learning.” *Machine Learning Mastery*, 15 Aug. 202AD, machinelearningmastery.com/data-leakage-machine-learning/.

Pragati21. “SVM and RANDOM FOREST: A Case Study.” *Medium*, 27 Jan. 2020, medium.com/@pandeypragati2112/svm-and-random-forest-a-case-study-6213da5be02f.

Shin, Terence. “Understanding Feature Importance and How to Implement It in Python.” *Medium*, 26 Feb. 2021, towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285#:~:text=Feature%20Importance%20refers%20to%20techniques.

Pramoditha, Rukshan. “LDA Is More Effective than PCA for Dimensionality Reduction in Classification Datasets.” Medium, 4 Jan. 2023, towardsdatascience.com/lda-is-highly-effective-than-pca-for-dimensionality-reduction-in-classification-datasets-4489eade632.

Trehan, Daksh. “Determining Perfect Fit for Your ML Model.” *Medium*, 27 July 2020, medium.datadriveninvestor.com/determining-perfect-fit-for-your-ml-model-339459eef670. Accessed 16 May 2023.