Predictive Analytics on Employees Promotion using Machine Learning

this one

by

Huda Hannani binti Yahaya

17002493

Dissertation submitted in partial fulfilment of

the requirements for the

Bachelor of Information Technology (Hons)

SEPTEMBER 2021

Universiti Teknologi PETRONAS

32610 Seri Iskandar

Perak Darul Ridzuan

Malaysia

CERTIFICATION OF APPROVAL

Predictive Analytics on Employees Promotion using Machine Learning

By

Huda Hannani binti Yahaya
17002493

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
BACHELOR OF INFORMATION TECHNOLOGY (Hons)

Approved by,

Ts Dr Norshakirah Ab. Aziz
Senior Lecturer
Computer Information & Sciences Department
Universiti Teknologi PETRONAS

(Dr. Norshakirah bt A Aziz)

UNIVERSITI TEKNOLOGI PETRONAS
BANDAR SERI ISKANDAR, PERAK
September 2021

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own expect as specifies in the reference and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

HUDA HANNANI BINTI YAHAYA

ABSTRACT

Human resource refers to the people who work for a firm or organisation, as well as the department in charge of handling personnel issues. Human resource gathers a large amount of data on all aspects of employee activity. Throughout the years, the data gathered keep growing and it will be no use if it did not provide any insights. Thus, human resource analytics is one of the approaches to convert the big data into a smart data. Machine learning has become one of the main components in human resources. The problem with the current promotion is time-consuming because of the various steps involved in the promotion procedure. Thus, it caused a delay in promotion and directly affected the transition of the employee into their new role. Therefore, it could be more efficient if the human resource department could predict which employee is more eligible and deploy them with a new job description, salary and others. The goal of this study is to propose predictive analytics model on employee's promotions using supervised machine learning method that could predict which employees that could get promoted based on their past performance. The data visualization using Power BI will be utilised to obtain the most accurate prediction model. The results from the study show that Prediction Model using logistic regression gives 93.4% accuracy as compared to k-nearest neighbour and decision tree prediction models.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful. All the praises and thanks are to Allah. With His guidance, I am able to complete this final year project report. With His love and blessing, I can complete my degree study in Universiti Teknologi Petronas.

Next, I would like to thank my supervisor throughout doing this project, Dr. Norshakirah A Aziz. She is a nice person and a good supervisor. I would like to also thank coordinator for both FYP I and FYP II for the smooth process throughout the semester. My completion of this project could not have been accomplished without the continuous support from my friends and classmates.

Lastly, I would like to express my gratitude and deep regards to my family. Their words of support during difficult times were much appreciated and recorded. Thank to everyone who has supported me in this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background

Human resource can be defined as the people who work for a company or organization and its department that is responsible for managing matters related to employees. Tripathi and Sharma (2018) stated that, human resource management is an ongoing process that aims to maximise employee's potential. The implemented policies and systems are necessary in order to handle employees effectively. The purpose of human resource management is to maximise organizational productivity through an effective use of resources and human capability. Human resource gathers a large amount of data on all aspects of employee activity. It collects data from various sources throughout the organization including employee surveys, telemetric data, attendance records, rating value reviews, employees promotion history, job history, employees data base and others (Daash, 2020).

Throughout the years, the data gathered keep growing and it will be no use if it did not provide any insights. Huselid and Minbaeva (2019) stated that most of human resource managers knew that they have a lot of data but were not sure the use of the data gathered. It is no longer about how big data is but how smart the data can be used to provide insights to an organization. Thus, human resource analytics is one of the approaches to convert the big data into a smart data. Through data analytics it can improvise the decision-making capability in order to achieve organization goals. Machine learning has become one of the main components in human resources.

1.2 Problem Statement

The current process of employee's promotion is not being carried out effectively due to the complexity in the promotion procedure. The process begin with a set of employees will be identified by human resource department based on their past work. Next, the selected employees will go through certain evaluation and test in various phases. The final results of the promotion will be announced after the final evaluation. The problem with this promotion process caused a delay of transition of promoted employee into their new role. The difference between the existing research and the proposed study is past research only deals in predicting one attribute. For example, findings by Ameer et al (2020), covers on prediction of employee turnover rate and other research done by Thorström (2017) that focus on predicting the key performance indicator (KPI) of employees. This research plan will take all the attributes that are related to employee's performance and make them as a parameter for the prediction model. It could be better if the human resource department could predict which employee that will get promoted for easier management in terms of the job description, salary and others. Thus, the problem statement is the delay in the transition of promotion could lead to business processes getting delayed. Chang and Xue (2020) reported that an effective employee evaluation mechanism is of great significance for improving the overall competitiveness of the organization.

1.3 Objectives

The objective of this project are as follows:

a) To identify the factor in evaluating the eligibility of employees getting promoted.
b) To determine a high accuracy model that could predict the employee that is likely to be promoted.
c) To develop data visualization using Power BI for the proposed prediction model.

1.4 Scope of Study

This study focuses on the data analytics related with the promotion of staff in human resource management. It will identify few factors that could be the attribute in deciding which employee will get promoted. Besides, this study is also focus in determining a prediction model that could predict the employee that could get promotion based on their past performance. Moreover, this study will also develop a data visualization dashboard for the analysis by using Power BI to illustrates the results of the analysis.

CHAPTER 2
LITERATURE REVIEW

2.1 Employees promotion

Employee promotion refers to the upgrade of employee to a higher position in a company. This could involve an increase in their salary, position, job scope, and benefits. In an organization, by promoting their employees, it acts as a reward and the way the organization appreciating their hard-working employees. This is due to the fact that the well-being of employees will greatly affect the enterprises or organization. Long et al (2018) suggested that by promoting employees, organization could improve the competitiveness and give award to the employees with outstanding achievement.

Most of organization promote their employees based on a few different factors. One of the factors will be the employee's key performance indicator. Organization use key performance indicators (KPI) to track whether they have achieve the organization goals or not (Aksu et al., 2019). In the context of an employee, KPI could be used as an indicator to measure the employee performance whether they are underperforming or others. Through this KPI system, it could create the awareness among employees in giving the outcomes or outputs that meets their employer expectations.

Secondly, organization could consider promoting their staff based on the length of service of the employee. The length of service is the year the employee has dedicated to work in an organization or company. The length of services also indicates the long experiences and skills that has been developed by employee in the specific position (Adenuga, 2015). Employees that have such long experience should be rewarded by organization. This is an excellent way for organization to express their gratitude for the value and dedication of the individuals. One of the ways to reward them is by promoting their position into a higher rank or increasing their salary.

Thirdly is the number of training or length of training that employee have undergo. Employee training program is important because it creates a chance for employees to learn and improve a new skill in the workplace. By offering a training program, employees will feel appreciated an do their best in work in order to achieve business goals (Elnaga & Imran, 2018). Furthermore, employees who are advancing into higher roles and taking a greater responsibility can benefit from training programmes. This programme will assist students in learning the skills they need for their new roles.

2.2 Key Performance Indicator (KPI)

Key performance indicator refers to a collection of quantitative metrics used to assess a business's overall long-term performance. It is being used to evaluate a company's strategic, financial and operational accomplishments, particularly in comparison to those of other organisation in the same industry. In this study, the focus is mainly in measuring the employees KPI. In the context of an employee, KPIs can be used to determine whether a person is performing well or poorly. Through this KPI system, it is possible to raise employee knowledge about the importance of delivering results or outputs that satisfy their employer's expectations.

Each organisation has KPI that are relevant with their industry. The metrics for the KPI may include the overall performance of the organisation depends on the size of the organisation. It could include specific activities associated with each department such as marketing, sales, customer service, and finance. As stated by Mohamed (2014), one of the examples in determining employees KPI across different industries in terms of finance is by looking at the profit. A project's success could be measure more accurate by developing KPIs. Thereby, all project members could keep in track in doing their task and would have a clear direction for the project. According to De Andrade and Sadaoui (2017), board of directors in a company would employ KPIs to conduct an audit of the company's recent state and develop a new action plan in the event that the measurements indicates a poor future situation.

Other than that, KPIs could also provide extra information that could help in understanding business growth.

For these past few years, machine learning has been applied in developing KPIs since machine learning could assist in detecting hidden patterns from the dataset. In the study that has been conducted by Mohamed (2014), fuzzy logic algorithm has been used to group KPIs values and predict its future values. Fuzzy logic algorithm is a method of variable processing that enables the processing of numerous possible truth values in the same variable. Apart from this, deep neural network model has also been used in predicting the employee's productivity. Deep neural network is a machine learning model in which the model employs multiple layers of nodes to extract high-level functions form the input data. This involves converting the data into a more abstract and creative component. The experimental results in predicting employee's productivity indicates that deep neural network model could accurately predict the employees; actual productivity and greatly improves the prediction performance with the mean absolute error is smaller than the baseline performance score (Imran et al., 2019).

2.3 Human resource analytics

Human resource analytics is the way to know the insights of a large amount of data. Kakulapati et al. (2020) stated that human resource analytics is an approach to understand the behavioural of employees in order to improve organization productivity. It is a process of collecting and analysing human resource data into a useful data. Its major goal is to find employees that could contribute to organization goals where organization could maximize its profit by taking into account several characteristics that would help in giving meaningful information through predictive analytics (Ameer et al., 2020).

Nowadays an abundant amount of data is no longer relevant to the organization. A raw and unorganized data could not provide any useful information towards the organization. According to Dahlbom et al. (2019), a company could have a numerous

data that is associated with their employees performance and their workforce, and it the data could be analysed together, it may provide a useful insights about the organization with the correct analysis tool. If all the data could be combined, it may offer some insights about business performance. Human resource analytics could offer insights related to organizational issues such as turnover of employees and seeking suitable candidates in recruitment. Thus, it is clear that human resource analytics could improve in decision making and it would be waste of time and storage if the data are not being used for analytics.

One of the approaches in human resource analytics is by using machine learning. One of branches in artificial intelligence is machine learning where it could be taught based on past history and no need interruption or help from outside (Dianah et al., 2021). By using machine learning, human resource management now can apply the prediction model in their business case. For example, machine learning could help in predicting employee attrition, tracking a candidate journey throughout interview process, predict employee performance and others.

2.4 Machine learning

Machine learning is a branch of artificial intelligence (AI) that focuses on using data and algorithms to replicate the way humans learn and improve its accuracy over time. It is a critical component of the growing field of data science. Algorithms are trained to develop a classification or predictions using statistical approaches. As a result, it could discover the hidden critical information in data mining projects. Business usually use this information to make important decision that will act as a key growth indicator for the businesses. Thus, as the years goes by, many company and organization will need their own data scientist team to assist the main expert in solving their business problems.

There are two different concepts in machine learning which are the supervised and unsupervised learning. The use of labelled datasets to train algorithms for accurately classifying data or predicting its results is referred to as supervised machine learning.

A training set is necessary to train the algorithm while a testing is used to determine the accuracy of the algorithm. The testing set is a subset of the total dataset that contains the predicted values. The predicted values then will be compared to the target or true values from the dataset measure how well the algorithm performs (Thorström, 2017). The model's weights are adjusted when input data is entered until the model is appropriately fitted. This is done as part of the cross-validation process to see if the model is overfitting or underfitting. Organizations can use supervised learning to handle a variety of real-world problems such as spam classification in a separate folder from the inbox. Several methods such as neural networks, naive bayes, linear regression, logistic regression, random forest, and support vector machine (SVM) are utilised in supervised learning.

Next is the unsupervised learning. Unsupervised learning is a technique in which the training set contains data but no solutions. This means that the computer must discover the pattern on its own. This type of machine learning will analyse and clusters the unlabelled datasets using the suitable algorithm. Unsupervised machine learning will be the most suitable for case study such as cross-selling techniques, consumer segmentation and image and pattern recognition. This is due to the capacity of the algorithms in identifying similarities and contrasts in the dataset. One of the most popular approach is by using the k-means clustering. K-means clustering will group the dataset in a way that each observation is closest to the mean of its cluster (Louridas & Ebert, 2016). Besides that, hierarchical clustering, gaussian mixture models and dimensionality reduction are also a part of clustering approaches that are available.

In this research, supervised machine learning will be used. This is because one of the objectives for this research is to classify the employees into two groups which are to predict whether employee will get promotion or not. Various criteria will be included as parameter in determining which employee is eligible for a promotion. Parameter that is relevant such as the average training score, length of service, KPI score will be used to train the models. Since these problems is under classification problem within two binary class, there are few models that are suitable to be used.

Firstly, is the logistic regression model. Logistic regression is the simplest and most efficient model because one or more independent variables will be used to predict the outcomes. Secondly, is the naive bayes model. Bayes's theorem is being applied where it assumes the dependency among the predictors. This classifier makes assumption that the existence of one feature in a class is unrelated with another feature. Next, is the k-nearest neighbour algorithm. It classifies the dataset based on the class of majority of the datapoints. Besides that, support vector machine algorithm (SVM) often being used in solving a classification problem. This classifier will classify the datapoints based on identifying the hyperplane that best separates the two classes. Lastly, is the decision tree modelling. It will incrementally breaks down a dataset into smaller and smaller sections while also developing an associated decision tree.

CHAPTER 3
METHODOLOGY

3.1 Project methodology

The overall project methodology is by adopting agile methodology. Agile methodology is a constant methodology where changes could be made at each stage to ensure the project will achieve its objectives. In agile methodology project management, it consists of few phases which include planning the requirements, design and development, testing and deployment. After the development phases, the project will go through a continuous cycle and amendments could be done when necessary. Other than that, agile methodology has been chosen because it is flexible to work out if there is a need for any changes. As a result of a constant testing in each cycle, in general the outcome will have a better quality.

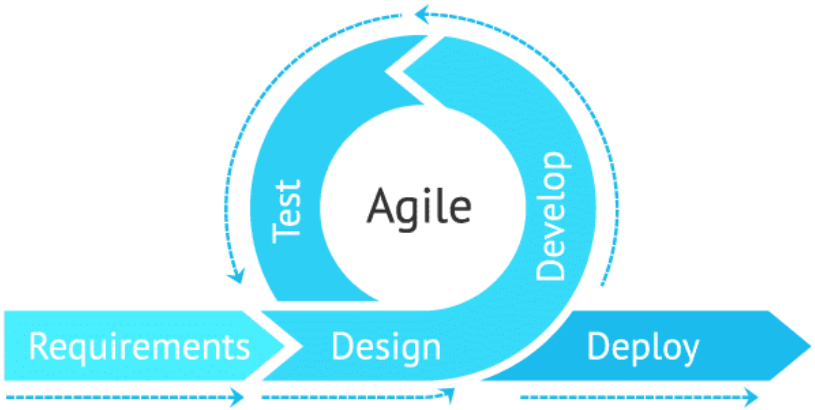

Figure 1. Agile methodology

In phase one which is the requirements phase, a thorough planning should be done. Planning is a crucial phase in every project management. Before any design and development can start, a solid understanding and purpose of this project is needed. This is to ensure that all the project activities later will not be out of track and stays in the scope of study. In this phase, the purpose and needs of the predictive model is

discussed so that it could solve the problem that has been stated earlier. Then, it should be followed by design and development phase. In this phase is where the technical part will start which include design for the machine learning framework and the coding. Next, the project will go through testing phase after each development. Changes could be made if the model need any changes. Finally, the project will go through the deployment phase.

Figure 2 shows the general work flow for this study. It specifies all the tasks needed in order to achieve all objectives. In the first step, research and literature review are studied on related topics to identify the possible factors of employees getting promoted. This method will contribute to achievement of objective one in this study. Next, the machine learning process will be executed. It covers the end-to end pipeline in machine learning framework which include collecting the data sets and understanding it, preparing the data, develop and evaluate the prediction model. Once the model is completed, it will be tested until the maximum accuracy is obtain which contributed to achievement of objective two. Finally, a dashboard will be generated to visualize the proposed predicted model by using Power BI.

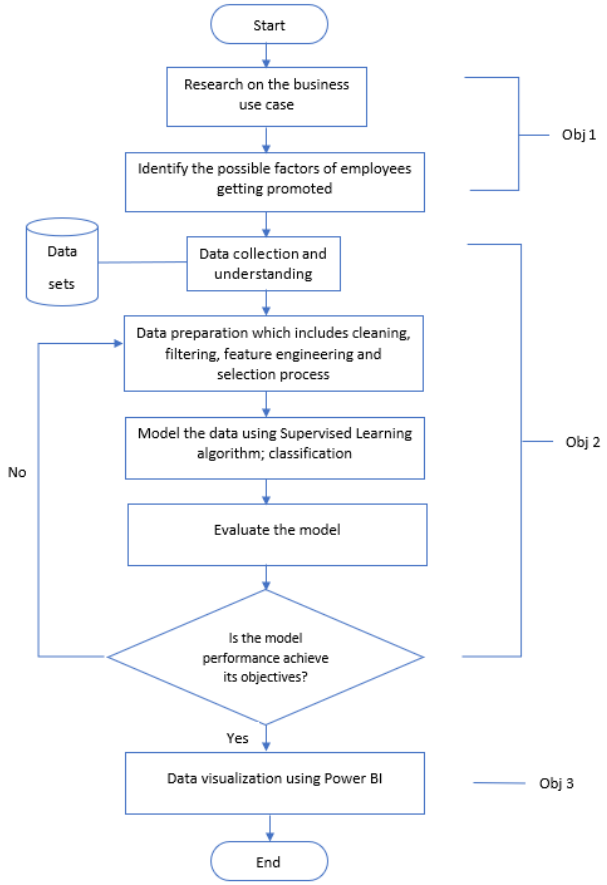

Figure 2. General work flow

3.2 Tools and software

The tools required for this study is python language. Python is an object-oriented programming language that is being used the most in machine learning. One of the reasons is because python language offers a wide range choice of library that could be used to access, handle and transform the data. In addition, python programs are also easier to debug because it will display an exception when there is an error occurs. Secondly, is machine learning. This is due to the fact that this project is focusing in developing a prediction model where it involved machine learning algorithm. The third tools and software needed in this project is Microsoft Power BI. Microsoft Power BI is a business intelligence software where dashboard could be created for data visualization purposes. With data visualization, a company or an organization could make a better decision-making and could gain more insights about their company.

3.3 Project activities

Figure 3 shows the phases in CRISP-DM. In this project, I have chosen CRISP-DM (cross-industry standard process for data mining) as the data mining frameworks that will be used. Nowadays, most of organization are using this framework in their machine learning project. This is because CRISP-DM framework provide a systemic approach in handling a data mining project. There are 5 phases that is crucial in this framework including the business understanding, data understanding, data preparation, modelling and evaluation, and followed by deployment phase at the end.

Figure 3. CRISP-DM framework

3.3.1 Business understanding

During this phase, the focus is to understand the overall project objectives and its expectations. Machine learning or data mining problem will be defined based on the objectives and a course of action will be created. This is where the requirements, assumptions, and constraints will be listed out. In this study, the current process of employees' promotion will start off by identifying a set of employees based on their past performance. This past performance will include few criteria's such as whether they have achieved a certain KPI, their previous year training score, their length of service and others. Next, the selected employees will go through certain evaluation and test in each phase. The final results of the promotion will only be announced after the final evaluation. Thus, the objective in this project is to make a prediction which employee that will get promoted based on the criteria's by using prediction

model. Through this prediction, human resource department could speed up the promotion process for easier management in terms of job description, salary and others.

3.3.2 Data understanding

In this phase, the data has been retrieved from internet. The dataset contains the list of potential employees that could get promoted along with multiple attributes and employees past performance. It consists of 54808 rows and 14 attributes.

```
#to see how many instances in the dataset
print('there are ' + str(len(data)) + ' rows of data in the dataset')
print("Total Feature", data.shape[1])
data.head(10)
```

```
there are 54808 rows of data in the dataset
Total Feature 14
```

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	awards
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	1	
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	
5	58896	Analytics	region_2	Bachelor's	m	sourcing	2	31	3.0	7	0	
6	20379	Operations	region_20	Bachelor's	f	other	1	31	3.0	5	0	
7	16290	Operations	region_34	Master's & above	m	sourcing	1	33	3.0	6	0	
8	73202	Analytics	region_20	Bachelor's	m	other	1	28	4.0	5	0	
9	28911	Sales & Marketing	region_1	Master's & above	m	sourcing	1	32	5.0	5	1	

Figure 4. The raw dataset

Table 1. Data description

Attributes	**Description**
employee_id	An employee ID that is unique for each employee
Department	Employee’s department
Region	Employee’s region
Education	Employee’s education level
Gender	Employee’s gender
recruitment_channel	Employee’s channel of recruitment
no_of_trainings	Employee's total number of trainings completed
Age	Employee’s age
previous_year_rating	Employee’s evaluations from the previous year
length_of_service	Length of service in years
KPI_met>80%	Employee’s key performance indicator
awards_won	Total number of awards won by the employee
avg_training_score	The average training score of the employee
is_promoted	Whether employee could get promotion or not

3.3.3 Data preparation

Data preparation stage is important because this is where we need to select which data, we are going to use for the prediction model. This stage includes a few tasks such as data processing and wrangling processes, feature extraction and engineering processes and the feature scaling and selection processes. Cleaning, manipulating, and translating data from one form to another in order to use it for a specific activity is known as data wrangling. The data that has been collected might be incomplete, missing and noisy. Thus, the data need to be structured properly through data preparation before fitting it in a machine learning model.

i) Data pre-processing and wrangling

Understanding and taking a short look at the dataset, number of entries, attribute names, and data types is the initial stage in data pre-processing. This process will help in understanding and get more information about the data. By understanding the data types of the attributes, it confirms that the information is collected in an appropriate format. In this dataset, it shows 3 types of data types which is object that refers to text or mixed numeric values, float64 for floating point numbers and int64 for integer numbers. The next stage is to clean up the dataset. Cleaning the dataset involves tasks such as removing and handling the missing values, handling outliers, and standardizing attribute column details.

a) Handling missing values

As shown in Appendix 1, there are two columns that have missing values which are ‘education’ column with 2,409 values and ‘previous_year_rating’ column with 4,124 missing values. For ‘education’ column, I have replaced the missing values with statistics measure which is mode. Mode indicates the number that occurs the most often in a collection of data. In this case, the mode is the ‘Bachelor’s’ category as shown in figure 5. For ‘previous_year_rating’, I have utilized the fillna() method from pandas to fill these values with median value from the data.

```
#check which education is the most frequent
data['education'].value_counts()
```

```
Bachelor's          36669
Master's & above    14925
Below Secondary       805
Name: education, dtype: int64
```

Figure 5. Handling missing values

b) Handling duplicates

In order to have a cleaner data, the existence of duplicates value need to be checked. In identifying duplicates, duplicated () utility can be used on the whole data frame. For this study, I have checked if there are any duplicates value especially in 'employee_id' attributes since each employee ID should be unique to each other. Figure 6 shows there are no duplications in this data frame.

```
#check for any duplicated values
data.duplicated().sum()
data.duplicated(subset=['employee_id'])
```

```
0        False
1        False
2        False
3        False
4        False
         ...
54803    False
54804    False
54805    False
54806    False
54807    False
Length: 54808, dtype: bool
```

Figure 6. Check duplicates values

ii) Data summarization

Data summarization is the process of preparing a compact representation of raw data in hand. This process is helpful for data visualization, compressing raw data, and better understanding of its attributes. Since this project dealing with employee promotion, I have calculated the percentage of employees that have been promoted before and those who are not based on the attribute 'is_promoted'. This attribute displays the value equal to 1 if employee has received promotion and value equal to 0 if employee has not received any promotion. Based on this data records, it shows the percentage of employees who have not received any promotion are 91.48% and 8.52% employees have received promotion.

```
#to see the count for 'is_promoted' attributes
data_clean['is_promoted'].value_counts()
```

```
0    50140
1     4668
Name: is_promoted, dtype: int64
```

```
#display how many employees from the list have received promotion and not receive
print('1. Employees who have not received a promotion::' + str(round((50140/54808)*100,2)) + '%')
print('2. Employees who get promoted::' + str(round((4668/54808)*100,2)) + '%')
```

```
1. Employees who have not received a promotion::91.48%
2. Employees who get promoted::8.52%
```

Figure 7. Percentage of employees who have received promotion

3.3.4 Modelling

The modelling phase is the fourth phase of the machine learning framework. This phase will determine which model is most appropriate for achieving this project's objectives. For this study, supervised machine learning will be used. Supervised machine learning includes the use of labelled datasets to train algorithms on how to accurately classify or predict outcomes. The dataset will be split into 80% for training and 20% for testing the model. Since the objective of this study is to predict which employee that will get promoted, it will be categorized as a classification problem. This is because the dataset needs to be classified into two groups which are being promoted or not. For solving classification problem, there are a few models that I will used to test the data. The models are:

a) Logistic regression

Logistic regression is a statistical method in predicting binary classification problem with only two possible classes. The probability of employee getting promoted or not will be based on the employee's performance. Based on this study, the first class in this model will be getting promoted and the model should predict the probability of employee getting promoted given an employee's performance level.

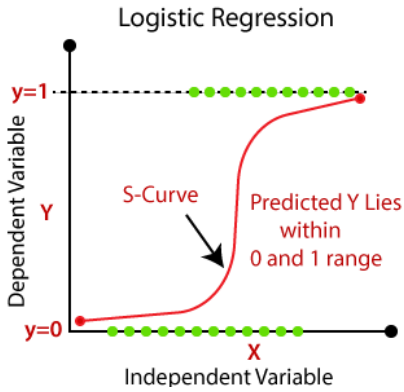

Figure 8. Logistic regression

b) K-Nearest Neighbour (KNN)

Secondly, is the k-nearest neighbour classifier model. K-nearest neighbour is a straightforward method for supervised learning that saves all known cases and classifies new ones using a similarity metric. A case is simply assigned to the class of its neighbours, with the case being allocated to the class that is most frequent among its K nearest neighbours and it is determined by a distance function. If K equals to 1, then the instance is simply classified and belongs to the class of its nearest neighbour.

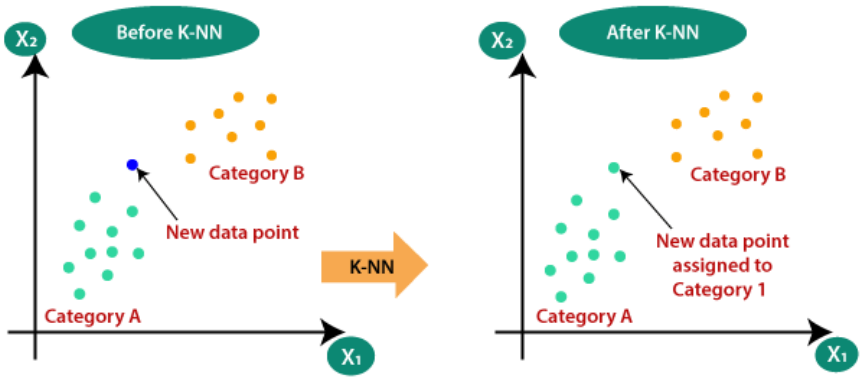

Figure 9. KNN model

c) Decision tree classifier

Thirdly, is the decision tree classifier model. A decision tree structure is used to construct classification or regression problems. While developing a classification tree, the dataset will be broken down into smaller and smaller sections. As a

result, a tree is formed with leaf nodes and a decision node. A leaf node implies a categorization or decision, while a decision node usually has two or more branches. The root node in a tree refers to the best predictor and is at the top decision node. In decision tree modelling, both quantitative and qualitative data can be used.

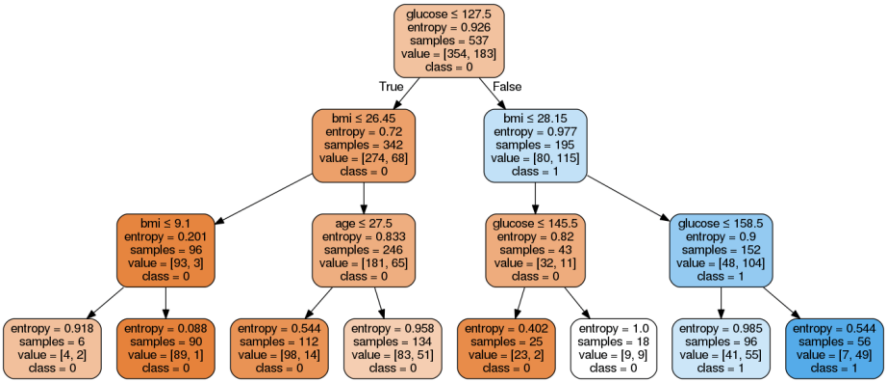

Figure 10. Decision tree model

3.3.5 Evaluation

Early evaluation steps considered aspects such as the model's accuracy and predictive validity. This step will evaluate the model's alignment with the business objectives that have been specified during the business understanding phase. Moreover, evaluation process also includes a review of any other data mining results that have been generated. The outcomes of the data mining will include the models that are necessary in achieving the initial business objectives and might as well identify additional information or patterns. After evaluating all the test models, the models that meet the evaluation metrics and the business criteria will be selected.

In classification problems, there are a few evaluation metrics that could be done to assess the model. The most common evaluation metrics in classification problems is by looking at the accuracy score of the model. Accuracy is defined as the number of correct predictions divided by the total number of predictions produced given a dataset. Accuracy is the best metric when the target class is well balanced but is a poor choice when the target class is unbalanced. For example, in this study, assuming we had 100 employees and 95 out of the 100 employees got promoted. Only 5 employees did not get promoted in the training data. As result, our model would always predict employees getting promoted by giving the 95% accuracy score. When in fact, in reality, the data is always skewed. Thus, to fully understand the model

evaluation, additional metrics like recall score and precision score should also be examined.

Other than that, classification models could also be assess by using the confusion matrix. The performance of a classification model is determined by the numbers of test records predicted correctly and incorrectly by the model. The confusion matrix will display a more comprehensive picture, revealing not just the predictive model's performance, but also which classes predicted correctly and incorrectly, as well as the type of errors made. Four classification metrics such as true positive, false positive, false negative and true negative are generated in the confusion matrix table below. True positive (TP) occurs when the actual value is positive and the model predicts it to be the same while false negative (FN) occurs when the actual value is positive but the model predicts it to be negative. When the actual value is negative and the model predicts it to be negative, it is called true negative (TN), while when the actual value is negative but the model expects it to be positive, it is called false positive (FP). In this case, precision and recall score could be used in determining whether the model that has been generated is suitable or not according to the business objective.

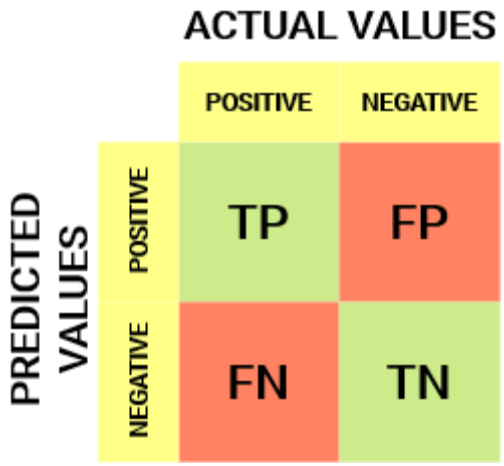

Figure 11. Confusion matrix

3.3.6 Deployment

During deployment phase, evaluation results will be used to develop strategy for their rollout. It is appropriate to examine the deployment methods and techniques

during the business understanding phase too as deployment critical to the project's success. This is where predictive analytics will truly help the organisation enhance its operational efficiency. The real method of deployment for a predictive analytics project can take numerous forms and different organisations will employ different deployment depending on their situation. In fact, the main objective of an analytical output and intended usage can change greatly depending on the operational situation. For example, in this study, the predictive analysis will be visualized into a dashboard and present it to the human resource department (HR). The dashboard will display insightful data and suggest which employee that are qualified to get a promotion. HR department then can see which employee that will get promoted based on the success criteria and prepare for the promotion procedure.

3.4 Gantt chart

Task	W1	W2	W3	W4	W5	W6	W7	W8	W9	W 10	W 11	W 12	W 13	W 14	W 15	W 16	W 17	W 18	W 19	W 20	W 21	W 22	W 23	W 24
Project Proposal																								
Data Acquisition																								
Research relevant literature																								
Proposal Defence																								
Data Understanding																								
Data Wrangling																								
Draft Interim Report Submission																								
Model development (FYP2)																								
Evaluation (FYP2)																								
Data Visualization (FYP2)																								

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Exploratory data analysis

With the use of summary statistics and graphical representations, exploratory data analysis refers to the crucial process of doing first investigations on data in order to uncover patterns, spot anomalies, test hypotheses, and check assumptions. The dataset for this study contains the list of potential employees that could get promoted along with multiple attributes and employees past performance. It consists of 54 808 rows and 14 columns. In total, this dataset has 767 312 entries. One out 14 attributes is dependent variable while the other 13 attributes are independent variables. The independent variables consist of various attributes that could identify how employees work performance during the previous year. This dataset has integer, float and object as their datatypes based on the figure 12 below. In this study, 3 separate datasets have been created from the original dataset. The first dataset is called ‘data’ which represents the dataset from the original source. Secondly is called as ‘data_clean’ which represents the data that has been cleaned and modified from the missing values and duplicated issues. Thirdly is the ‘df_pre’ that has gone through pre-processing steps and being used in the modelling phase.

```
In [86]: #to see how many columns and rows in the dataset
         print("Number of rows::",data.shape[0])
         print("Number of columns::",data.shape[1])
         print("Entries:",data.size)
         print("\n")

         print("Column Names::",data.columns.values.tolist())
         print("\n")

         print("Column Data Types::\n",data.dtypes)
```

```
e', 'previous_year_rating', 'length_of_service', 'KPIs_met >80%', 'awards_won?', 'avg_training_score', 'is_promoted']


Column Data Types::
 employee_id             int64
department              object
region                  object
education               object
gender                  object
recruitment_channel     object
no_of_trainings          int64
age                      int64
previous_year_rating   float64
length_of_service        int64
KPIs_met >80%            int64
awards_won?              int64
avg_training_score       int64
is_promoted              int64
dtype: object
```

Figure 12. Dataset datatypes

Since this dataset contains both numerical and categorical data, the data is separated into their respective group. The pandas describe() function is extremely useful for obtaining various summary statistics. The count, mean, standard deviation, minimum and maximum values, and quantiles of the data are returned by this function. Based on the figure below, it returned the summary statistics for the numerical data. The 50% in index column represent the median for each data column. These quantitative measures act as a useful guidance in determining which values should be used in case of replacing any missing entries.

```
data[numerical].describe()
```

	employee_id	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	awards_won?	avg_training_score	is_promoted
count	54808.000000	54808.000000	54808.000000	50684.000000	54808.000000	54808.000000	54808.000000	54808.000000	54808.000000
mean	39195.830627	1.253011	34.803915	3.329256	5.865512	0.351974	0.023172	63.386750	0.085170
std	22586.581449	0.609264	7.660169	1.259993	4.265094	0.477590	0.150450	13.371559	0.279137
min	1.000000	1.000000	20.000000	1.000000	1.000000	0.000000	0.000000	39.000000	0.000000
25%	19669.750000	1.000000	29.000000	3.000000	3.000000	0.000000	0.000000	51.000000	0.000000
50%	39225.500000	1.000000	33.000000	3.000000	5.000000	0.000000	0.000000	60.000000	0.000000
75%	58730.500000	1.000000	39.000000	4.000000	7.000000	1.000000	0.000000	76.000000	0.000000
max	78298.000000	10.000000	60.000000	5.000000	37.000000	1.000000	1.000000	99.000000	1.000000

Figure 13. Summary statistics of numerical data

Next, is the categorical data. Categorical data is the data that has a range of possible values. For example, in this dataset, the employees come from 9 different department which are technology, finance, procurement, analytics, operations, sales and marketing, research and development, human resource and lastly is the legal

department. Based on the statistics below, most of the employees are from sales and marketing department, from region 2, have education level of bachelor's, male and recruited from another channel.

```
# assign the categorical data into categorical object
categorical = ['department','region','education','gender','recruitment_channel']
```

```
data[categorical].describe()
```

	department	region	education	gender	recruitment_channel
count	54808	54808	52399	54808	54808
unique	9	34	3	2	3
top	Sales & Marketing	region_2	Bachelor's	m	other
freq	16840	12343	36669	38496	30446

Figure 14. Summary statistics of categorical dataset

4.1.1 Data visualization

Graphical representation has been done with each parameter to see in details how each parameter affects the possibility of employees getting promotion.

a) Distribution of promoted employees based on age

Based on the figure above, the average age of promoted employees is 35 years old while the median age of promoted employees is 33 years old.

Figure 15. Data visualization of promoted employees based on age

b) Promotion probability based on KPI score

Based on the figure shown above, employees who have achieved KPIs score more than 80% have higher chance of getting promoted which is 16.91% better than employees who did not achieve KPIs of more than 80%. The result was achieved by grouping the number of employees who has fulfil the KPIs for more than 80% and get promoted.

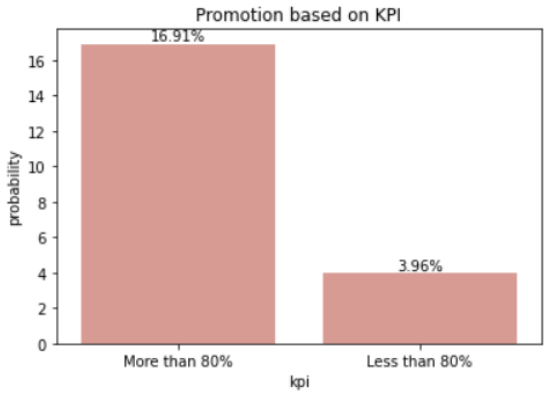

Figure 16. Promotion probability based on KPI

c) Promotion probability based on previous year rating

Previous year rating has 5 rating from 1 until 5. 1 indicates the lowest rating while 5 indicates the highest rating employees could achieve. Based on figure 17, employees who have a previous year rating of 5 have the highest possibility of getting promotion by 16.36% from a total of 11 741 employees who got the same rating.

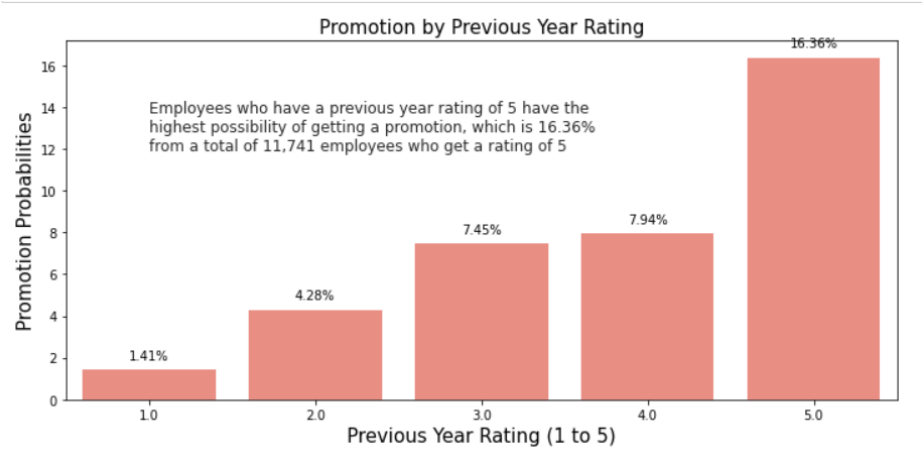

Figure 17. Promotion based on previous year rating

d) Promotion probability based on awards won by employees

This attribute shows if the employee has received an award previously during his service in the company. Based of figure 18, it shows that employees who have won an award have a 44.02% chance of being promoted than those who never won an award only have probability of 7.67%.

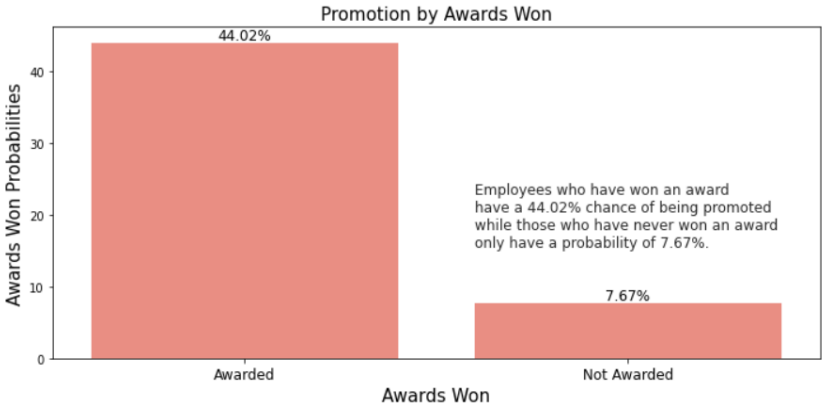

Figure 18. Promotion based on awards won by the employee

e) Promotion probability based on employee's length of service

Length of service indicates how long the employees have worked in the company. The shortest length of service is 1 year and 37 year is the longest. Based on the graph, employees with 34 years of services have a 24% probability of being promoted compared to other employees.

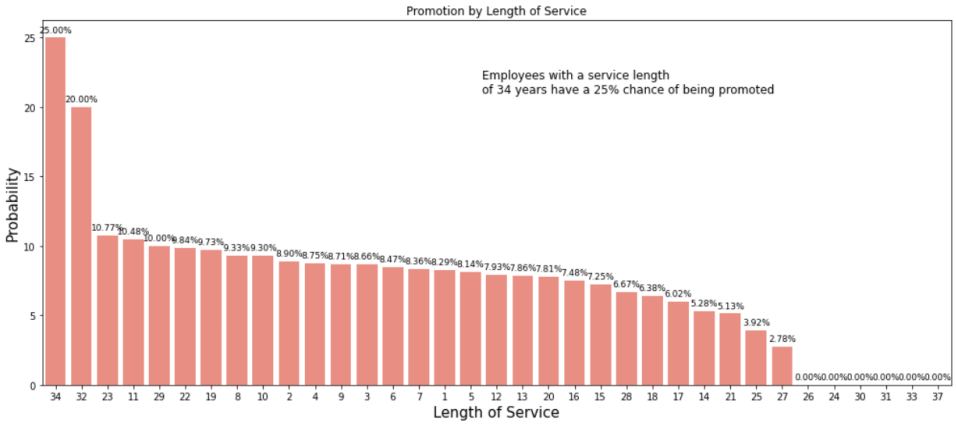

Figure 19. Promotion by length of service

f) Promotion probability based on employee's number of trainings

Number of trainings indicates how many trainings course that has been participated by the employees. Based on the observation below, most of promoted employees only attend 1 training with the highest probability of 8.81%.

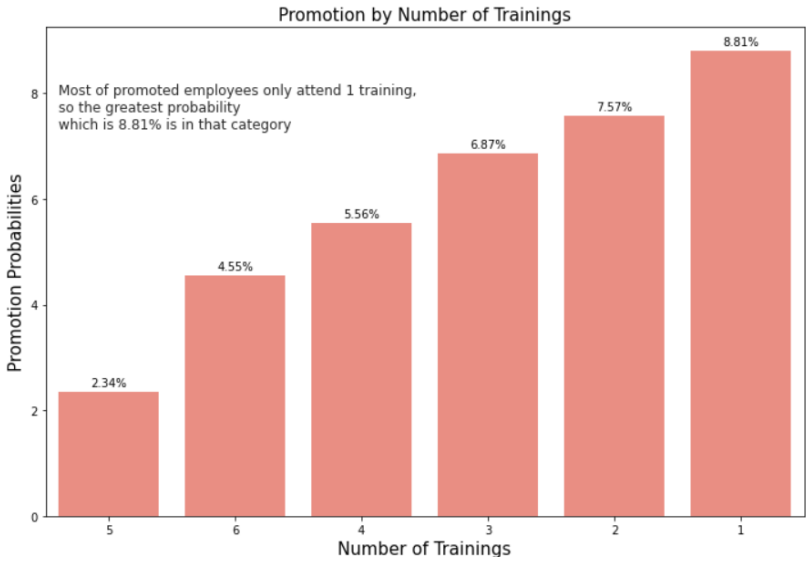

Figure 20. Promotion based on number of trainings

g) Promotion probability based on employee's average training score

Average training score is the score for each employee during the current evaluations. Employees with an average training score of greater or equal to 90 have higher chance of getting promoted which is 76.83% compared to those employees who have average training score of less than 90.

Figure 21. Promotion based on average training score

h) Promotion probability based on employee's region

Based on the Figure 22, employees who have a large opportunity of being promoted with probability more than 10% comes from region 4,17,25,28,23,22,3 and 7.

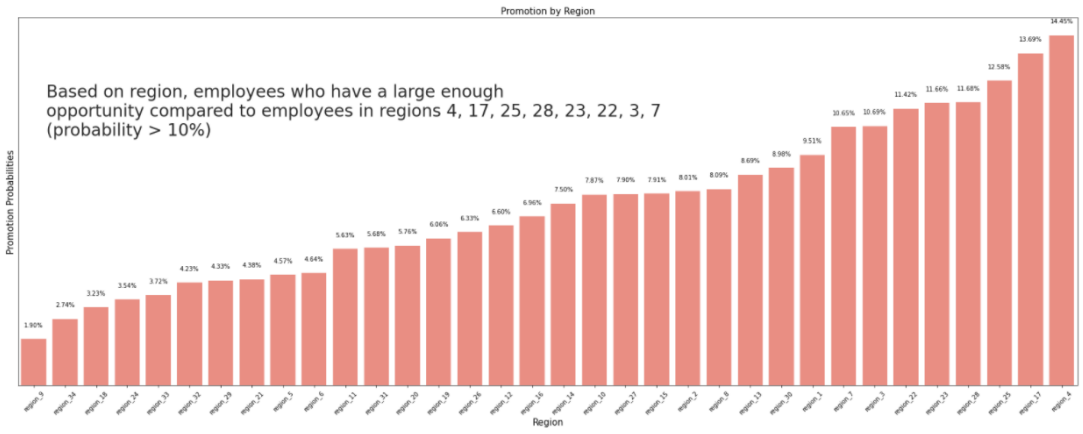

Figure 22. Promotion by region

i) Promotion probability based on employee's department

According to the dataset, employees that are from technology department have a chance of 10.76% to be promoted, followed by employees from procurement and analytics department with probability value more than 9%.

In contrast, employees from legal department will have the least chance with probability value of 5.10%.

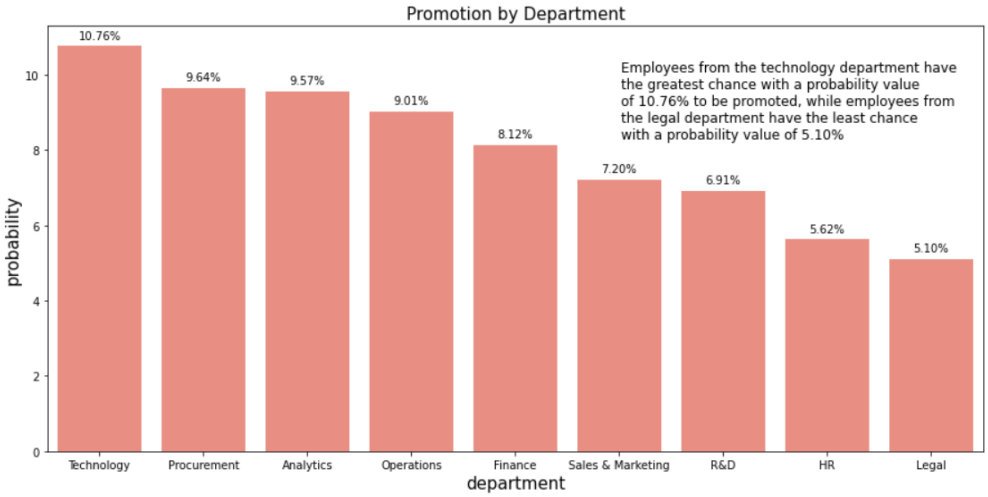

Figure 23. Promotion by department

j) Promotion probability based on employee's education level

As mentioned before in the summary statistics, most of the employees comes from bachelor's degree education background. However, according to figure 24, employees with a master's education level or above has a chance to be promoted by 9.86%.

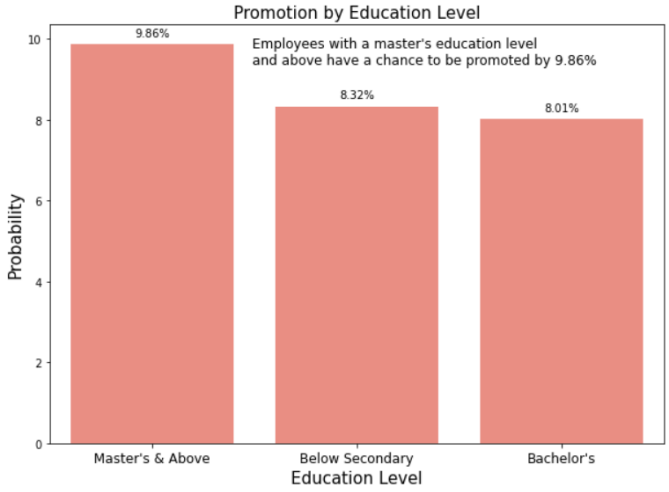

Figure 24. Promotion based on education level

k) Promotion probability based on employee's gender

Based on figure 25, there is not much difference in the promotion probability based on the gender. It shows that 8.99% of female employees are likely to get a promotion.

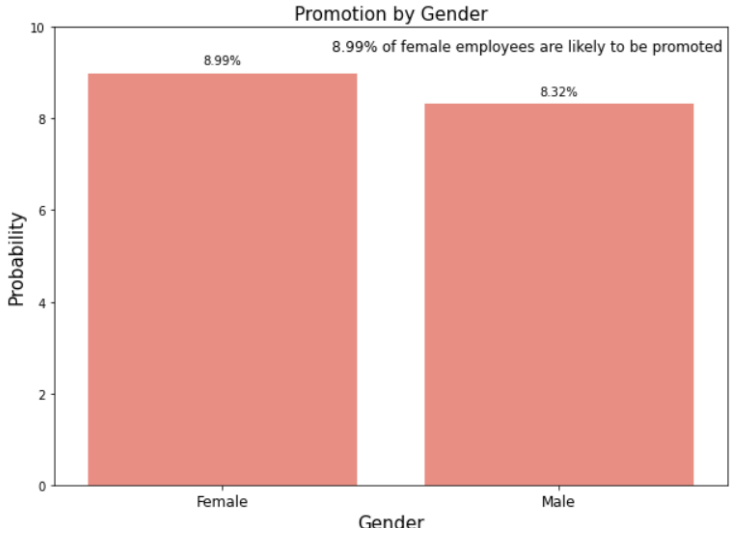

Figure 25. Promotion based on gender

l) Promotion probability based on potential region

Next, additional features like potential region have been created. Potential feature indicates the region that have the most probability to get promotion. Based on the information that have been identified earlier in promotion probability based on region, there are 8 potential regions which are region 4,17,25,28,23,22,3 and 7. This region is being arranged in descending order according to their probability value. Thus, any employee who is located in any of this region, the potential region value will '1', otherwise '0'.

```
#feature - potential region
#create 'potential region' where the region that have most probability to get promortion is region 4, 17, 25, 28, 23, 22, 3, 7
#if employees is located in these region-1, else 0
data_clean['potential_region'] = np.where(data_clean['region'].isin(['region_4','region_17','region_25','region_28','region_23',
                                                                     'region_3','region_7']),1, 0)
data_clean.head()
```

Figure 26. Potential region features

Based on the figure above, it proved that employees who are located in the potential region will have a higher chance to get promotion than the employees that are located out of the potential region.

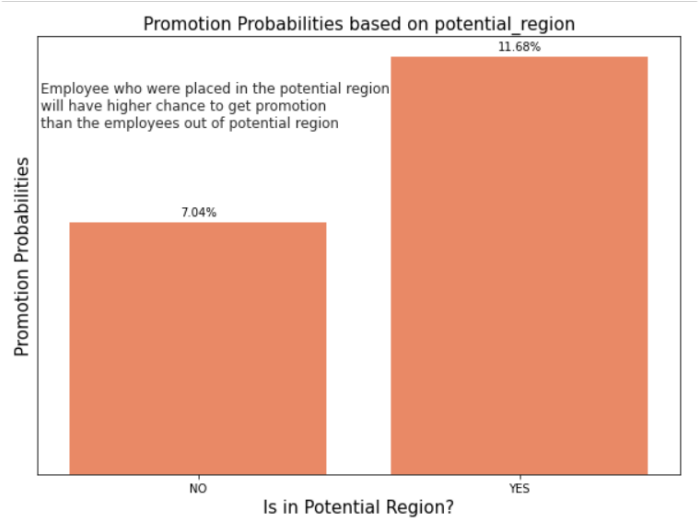

Figure 27. Promotion based on potential region

m) Promotion probability based on employee's performance level

The second additional features that has been created is the performance level features. This feature will combine the values from 'previous_year_rating', 'KPIs_met>80%', and 'awards_won'. Based on the results that have been obtained from the 3 attributes, it implies that employees will get a higher chance of being promoted if an employee met 3 conditions as follows:

i) An employee who has achieved KPI more than 80% have higher chance of getting promoted than employee who did not met the criteria

ii) An employee who got rating of 5 in the previous year rating has 16% probability of getting promoted compared to others employees who got a lower rating

iii) An employee who has won an award in the past year has higher chance to be promoted than other employees.

	performance_level	not_promoted	promoted	total_employees	promotion_probs
0	Excellent	114	125	239	52.301255
1	Good	5203	1232	6435	19.145299
2	Best	9460	1953	11413	17.112065
3	Low	35363	1358	36721	3.698156

Figure 28. Performance level count

Performance level will have 4 values range from 1 which is low, 2 indicates good, 3 is best and 4 is excellent, with the conditions as follows:

i) If KPIs>80% = 1 & previous_year_rating = 5 & awards_won? = 1, then performance level = ‘Excellent’ else,

ii) If KPIs>80% = 1 & previous_year_rating = 4 or 5 & awards_won? = 1 or 0, then performance level = ‘Best’ else,

iii) If KPIs>80% = 1 & previous_year_rating = 3 & awards_won? = 1 or 0, then performance level = ‘Good’,

iv) Else performance level = ‘Low’

Based on the performance level, it shows that employees in who has excellent performance level have the highest possibility which is 52.30% compared to employees who is in another performance level.

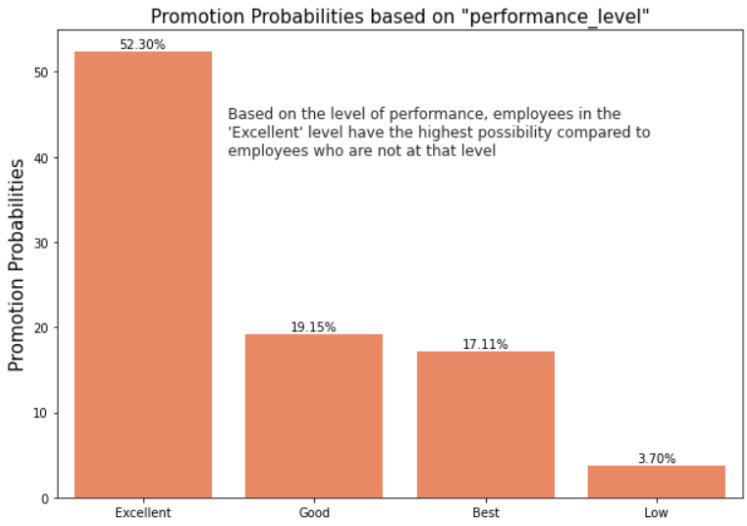

Figure 29. Promotion based on performance level

4.2 Model I: Prediction using logistic regression

Before the model being developed, since the classification problem in this study is included under supervised machine learning, the dataset needs to be split. The dataset will be split into 80% training data and 20% testing data. All the features of the dataset will be separated into independent variable (X) and dependent variable (Y). In supervised machine learning, there are two types of variables that are important. Firstly, is the independent variable. Independent variable is the features or the predictors variable. In this case, the independent variables are all of the attributes except the last column which is the 'is_promoted'. Secondly, is the dependent variable. Dependent variable (Y) is the predicted or the outcomes value. In this study, the dependent variable is the is_promoted column where the outcome will be a binary value. '0' indicates that the employees did not get the promotion while '1' shows that the employee receive a promotion.

```
: #split the feature into independent(x) and dependent variable(y)

x = data_pre[['nor_no_of_trainings', 'nor_age', 'nor_previous_year_rating',
       'nor_length_of_service', 'nor_KPIs_met >80%', 'nor_awards_won?',
       'nor_avg_training_score', 'nor_potential_region',
       'nor_Dept_Analytics', 'nor_Dept_Finance', 'nor_Dept_HR',
       'nor_Dept_Legal', 'nor_Dept_Operations', 'nor_Dept_Procurement',
       'nor_Dept_R&D', 'nor_Dept_Sales & Marketing', 'nor_Dept_Technology',
       "nor_Bachelor's", 'nor_Below Secondary', "nor_Master's & above",
       'nor_other', 'nor_referred', 'nor_sourcing', 'nor_male',
       'nor_High_Avg_Tscore', 'nor_Excellent', 'nor_Best', 'nor_Good',
       'nor_Low']]

y = data_pre['nor_is_promoted']
```

```
: #split into train-80% and test-20%

from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest = train_test_split(x,y,test_size=0.2, random_state=40)
```

Figure 30. Split train and test data

The first model to predict employee's promotion is by using logistic regression. Training data will be fitted into the model. The model then will predict the 'y' value.

```
#modelling using logistic regression
from sklearn.linear_model import LogisticRegression

#train the model
log_reg = LogisticRegression(random_state=42, solver = 'liblinear').fit(xtrain, ytrain)
```

```
#test the model
y_predict = log_reg.predict(xtest)
```

Figure 31. Logistic regression

For model evaluation, confusion matrix and classification report has been used. According to the confusion matrix, there are 10 030 results of true negative and 205 result of true positive. This indicates that a total of 10 235 correct predictions and 727 incorrect predictions.

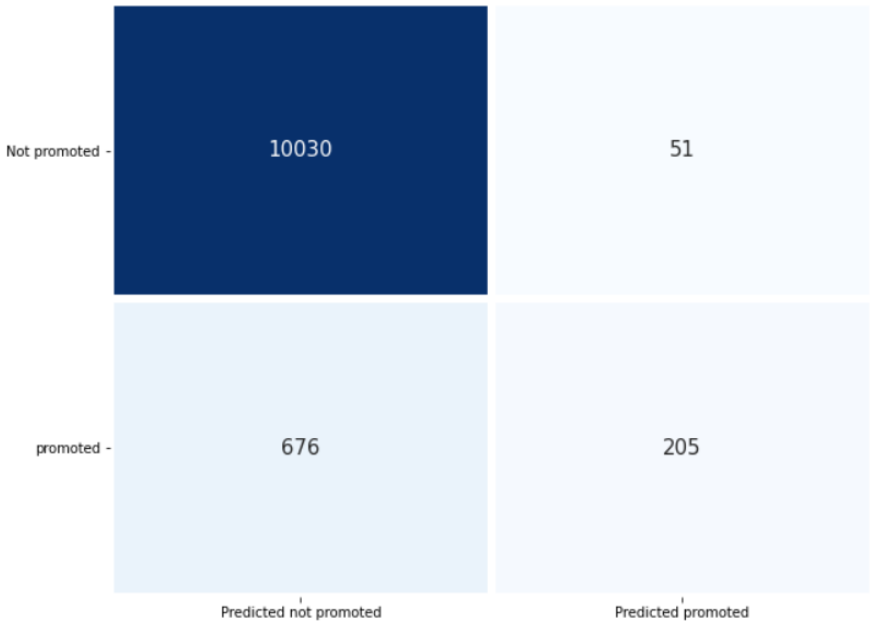

Figure 32. Confusion matrix for logistic regression

Based on the classification report, this logistic regression predictions have accuracy score of 93.4% and precision score of 80.1%.

	precision	recall	f1-score	support
0.0	0.94	0.99	0.97	10081
1.0	0.80	0.23	0.36	881
accuracy			0.93	10962
macro avg	0.87	0.61	0.66	10962
weighted avg	0.93	0.93	0.92	10962

Accuracy = 93.4 %
Precision = 80.1 %

Figure 33. Classification report for logistic regression

4.3 Model II: Prediction using k-nearest neighbour

The second model for prediction is by using k-nearest neighbour classifier (KNN). KNN will simply assigned case to the class of its neighbours, with the case being allocated to the class that is most frequent among its K nearest neighbours and it is determined by a distance function. Figure 34 shows there are 10 143 correct prediction and 819 incorrect predictions.

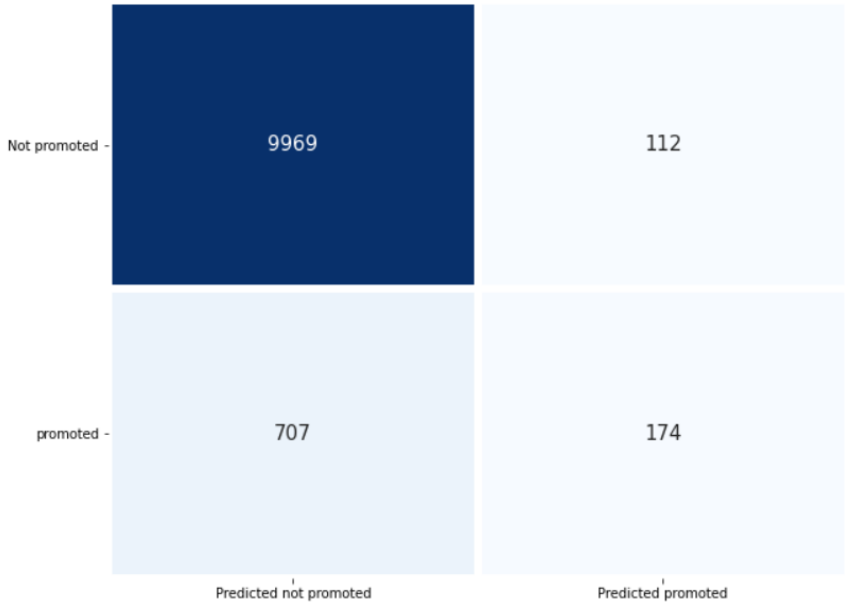

Figure 34. Confusion matrix for KNN

Based on the classification report, this KNN predictions have accuracy score of 92.5% and precision score of 60.8%.

	precision	recall	f1-score	support
0.0	0.93	0.99	0.96	10081
1.0	0.61	0.20	0.30	881
accuracy			0.93	10962
macro avg	0.77	0.59	0.63	10962
weighted avg	0.91	0.93	0.91	10962

Accuracy = 92.5 %
Precision = 60.8 %

Figure 35. Classification report for KNN

4.4 Model III: Prediction using decision tree

The last model in this study is by using decision tree modelling to predict employee's promotion. This decision tree will break down the dataset into a smaller section until a tree with leaf nodes and decision node are formed. Figure 36 indicates that there are 9 807 correct predictions and 1 155 incorrect predictions.

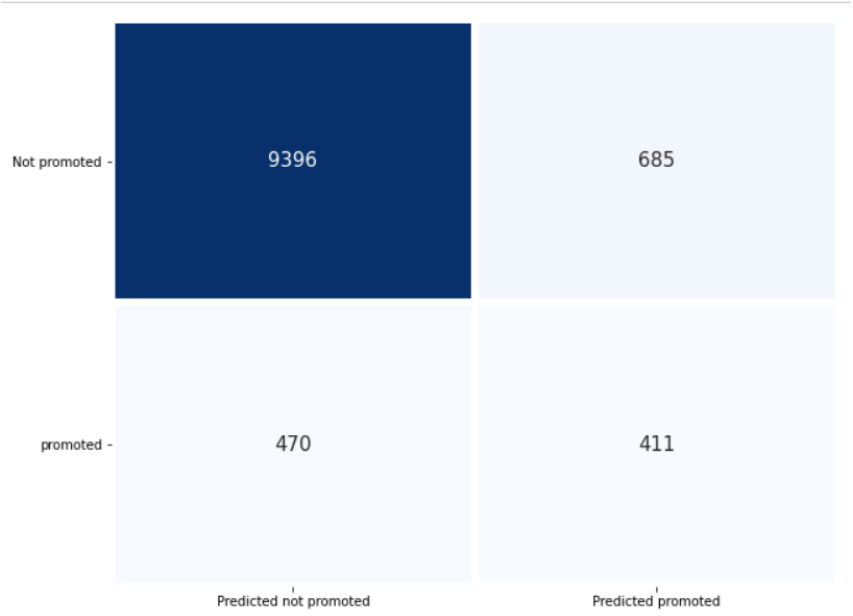

Figure 36. Confusion matrix for Decision tree

Besides that, decision tree classification report shows the model accuracy of 89.5 and precision score of 37.5%.

	precision	recall	f1-score	support
0.0	0.95	0.93	0.94	10081
1.0	0.38	0.47	0.42	881
accuracy			0.89	10962
macro avg	0.66	0.70	0.68	10962
weighted avg	0.91	0.89	0.90	10962

Accuracy = 89.5 %
Precision = 37.5 %

Figure 37. Classification report for decision tree

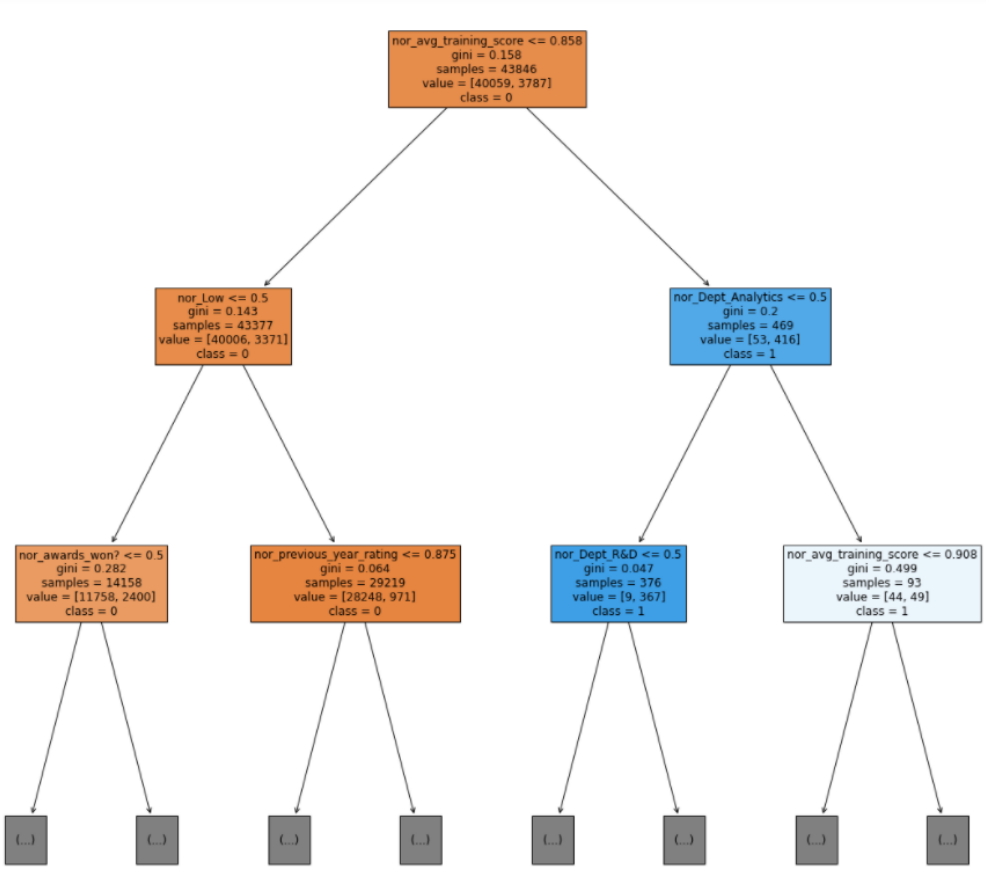

Figure 38. Decision tree

4.5 Data visualization in Power BI

Data visualisation aids in the telling of stories by transforming data into a more understandable format and showing trends and outliers. A good visualisation tells the story by reducing noise from data and emphasising the most important facts. Power BI is a Microsoft cloud-based business analytics application that allows anyone to visualise and analyse data more quickly and efficiently. It's a versatile and effective tool for connecting to and analysing a wide range of data. For data-science-related employment, many businesses consider it indispensable. The fact that Power BI features a drag-and-drop interface contributes to its ease of use. This functionality makes it simple and quick to accomplish operations like sorting, comparing, and analysing.

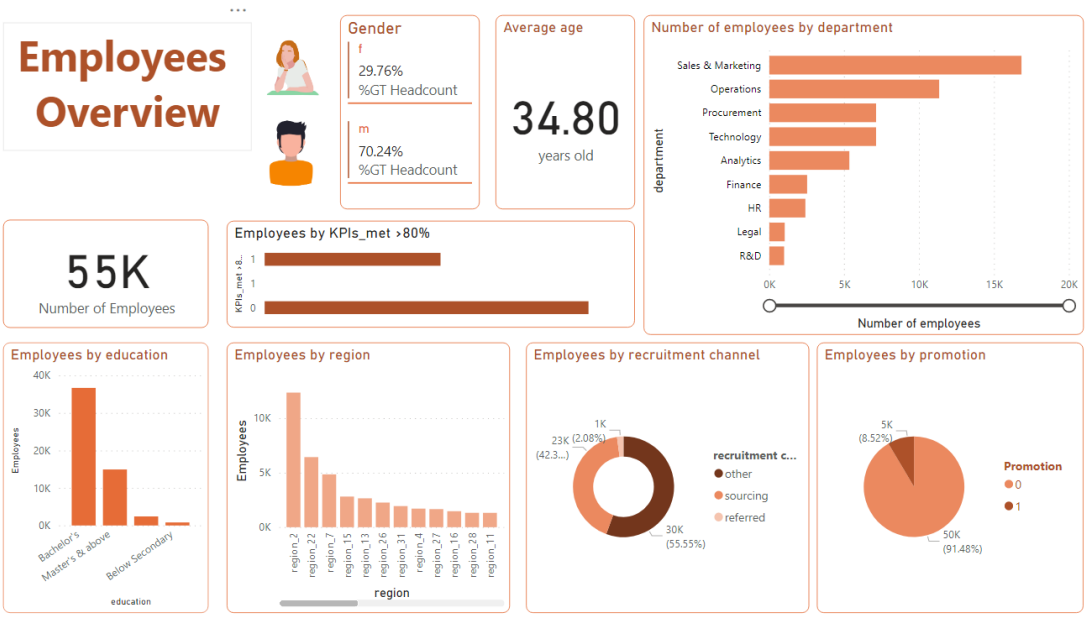

Figure 39. Data visualization

CHAPTER 5
CONCLUSION AND RECOMMENDATIONS

In this study, three algorithms models which are logistic regression, KNN, and decision tree has been tested to predict the which employees will get a promotion based on certain attributes. To determine the accuracy of the models, the results of predictions will be compared to the actual values. In Table 2, shows that logistic regression (LR) model performed better than other algorithms with the highest accuracy rate of 93.4%. In contrast, the decision tree shows an 89.5% accuracy score and a huge difference in precision score with only 37.5%. Thus, this comparison shows that LR is the most suitable model to be choose dur to the high accuracy and precision score. Not only that, LR model is simple to build and, in some situations, delivers excellent training efficiency. Because of these factors, training a model with this technique does not necessitate a lot of computing resources.

Table 2. Comparison of models' performance

Model	Accuracy (%)	Precision (%)
Logistic regression	93.4	80.1
K-nearest neighbour (KNN)	92.5	60.8
Decision tree	89.5	37.5

In this study, a few attributes that contributes to the decision making in employee promotion system has been identified. Promotion issues are related with both organization and their staff. This study has concluded that promotions are affected by many factors like employee key performance indicator, length of service of the employee and number of training that employee have undergo through. Besides that, logistic regression model has been developed with an accuracy score of 93.4% and precision score of 80.1% that could predict which employee will get promoted. Through the prediction, human resource management could prepare beforehand

which employee will promoted. Thus, this could improve the transitioning of the employee into their new position especially for human resource management when it is related with paperwork such as promotion letter, job description, salary and others. Moreover, the results for the prediction and descriptive analysis will be visualized by using data visualization. Hence, human resource management could see the trends and use these insights as future reference. Nevertheless, through this study, it could provide an insight and improve the prediction analysis related to employee promotion. In the future, other related attributes could be added in order to provide a more accuracy in the prediction model.

REFERENCES

Adenuga, O. A. (2015). *Employee ' s Salary , Gender , Length of Service and Job Involvement as Determinants o f Employees ' Performance in Nigerian Breweries Plc* . *1*(1), 20–25.

Aksu, Ü., Schunselaar, D. M. M., & Reijers, H. A. (2019). Automated Prediction of Relevant Key Performance Indicators for Organizations. *Lecture Notes in Business Information Processing*, *353*(May), 283–299. https://doi.org/10.1007/978-3-030-20485-3_22

Ameer, M., Rahul, S. P., & Manne, S. (2020). Human Resource Analytics using Power Bi Visualization Tool. *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, *Iciccs*, 1184–1189. https://doi.org/10.1109/ICICCS48265.2020.9120897

Chang, X., & Xue, J. (2020). Research on the Evaluation and Promotion of Employees from the Perspective of Competency. *Open Journal of Social Sciences*, *08*(02), 99–108. https://doi.org/10.4236/jss.2020.82009

Daash, A. (2020). IMPORTANCE OF HR ANALYTICS IN THE ERA OF 2020 POST COVID-19 Annjaan Daash ABSTRACT. *Journal of Natural Remedies*, *21*(3), 13–24.

Dahlbom, P., Siikanen, N., & Sajasalo, P. (2019). *Big data and HR analytics in the digital era*. https://doi.org/10.1108/BJM-11-2018-0393

De Andrade, P. R. M., & Sadaoui, S. (2017). Improving business decision making based on KPI management system. *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, *2017-January*, 1280–1285. https://doi.org/10.1109/SMC.2017.8122789

Dianah, S., Bujang, A., Selamat, A., & Krejcar, O. (2021). *A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning*. *1051*, 1–9. https://doi.org/10.1088/1757-899X/1051/1/012005

Elnaga, A., & Imran, A. (2018). The effect of training on employee performance. *International Journal of Recent Technology and Engineering*, *7*(4), 6–13.

https://doi.org/10.36555/almana.v4i3.1477

Huselid, M., & Minbaeva, D. (2019). Big data and human resource management. *Big Data and Human Resource Management. Sage Handbook of Human Resource Management. Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE Publications Ltd.* https://doi.org/10.4337/9781788112352.00008

Imran, A. Al, Amin, M. N., Islam Rifat, M. R., & Mehreen, S. (2019). Deep neural network approach for predicting the productivity of garment employees. *2019 6th International Conference on Control, Decision and Information Technologies, CoDIT 2019*, 1402–1407. https://doi.org/10.1109/CoDIT.2019.8820486

Kakulapati, V., Chaitanya, K. K., Chaitanya, K. V. G., & Akshay, P. (2020). Predictive analytics of HR - A machine learning approach. *Journal of Statistics and Management Systems*, *23*(6), 959–969. https://doi.org/10.1080/09720510.2020.1799497

Long, Y., Wang, T., Liu, J., Fang, M., & Jiang, W. (2018). *Prediction of Employee Promotion Based on Personal Basic Features and Post Features*. 5–10.

Louridas, P., & Ebert, C. (2016). *Machine Learning*.

Mohamed, A. (2014). *Design of Prediction System for Key Performance Indicators in Balanced Design of Prediction System for Key Performance Indicators in Balanced Scorecard*. *April*. https://doi.org/10.5120/12512-6016

Thorström, M. (2017). *Applying machine learning to key performance indicators*. 1–64.

Tripathi, S., & Sharma, A. (2018). *Human Resource Management : Machine Learning Perspective*. *V*(44557), 23–28.