Task 1

Hazim Fitri Bin Ahmad Faudzi (P152419)

Data Science (STQD6014) Project 1

Dr Nor Hamizah Binti Miswan

---

1. What do you understand about data science?

Data Science is a discipline that is a combination of primarily consist of Math Stat, Computer Science, and Business Knowledge. A data scientist must be able to collect, prepare, analyze, visualize, manage, and preserve a large combination of data. A data scientist also should be able to extract information from raw data and turn it into actionable insight for stakeholders to make decision. Current data consists of 80% unstructured data and 20% of structured data that is available for further analysis. Thus, in order to generate insight from it, data scientists can use a lot of tools to complete their work. For example, some of the tools for data management and data analysis is R, SQL, and Python while some famous tools for visualization is Power BI and Tableau.

Data science process usually involves several stages. The first step is collecting the necessary data to do the next step. These data come from various sources such as websites, databases, survey, and so on. Data collected often messy since it is an input from human with different train of thought. In order to tackle this problem, data scientists will need to do some data pre-processing and data cleaning before proceed to the next step.

The next step in data science is performing the Exploratory Data Analysis (EDA) to fully understand the structure of the data. During this step, visualization and some statistical skills will be utilized to deep dive into the data set. After this step is completed, data scientist will be able to identify the ouliers, patterns, and trends of the data.

Now, data scientist will be able to proceed to the next step which is to apply the suitable machine learning or predictive modelling to the data. This step is the most

important in a data science and it requires a good understanding of stat, machine learning, and business. Some common concept that is needed such as shallow learning, deep learning, supervised and unsupervised machine learning.

After that, data science also include communicating to the stakeholders. This can be done through strorytelling using data visualization, reports, or dashboards. By doing this, stakeholders without any technical knowledge can make decisions based on the findings. Lastly, the data collected must be preserved in a highly resuable manner for future research.

2. **Assuming you are a department manager, and would like to investigate the customer's preference on three types of your company's products. Hence, give a bit introduction about your company and what types of products you want to investigate. Next, explain the active roles of data scientist for this task.**

Assuming I am a department manager in a company that sells home electronic appliances. This company sells a variety of home appliances from a small appliances such as rice cooker, hair dryer, table fan, and blender to large appliances such as washing machine, dryer, freezer, and refrigerator. In order to generate sales and maintain relevant in the market, I wish to investigate what are the customers' preference and what are the things that customers would prioritize when deciding to purchase a home electronic appliances.

Data scientists play a vital role in identifying the customers' preference. The first role of data scientist is to build a good data architecture so that data will be properly routed and organized. A good management on an organization's data can help with the efficiency of storage, access, and analysis. This will ensures the quality, security and consistency of data.

The next active role of a data scientist is to collect data for analysis. Data scientist needs to gather data from various sources such as databases, web scraping, survey, and so on. The collected data then needs to be pre-processed and cleaned before further analysis can be done to ensure smooth investigation on the customer preferences.

Next, data scientists will perform data analysis to the data. This was done by doing some EDA first to further understand the data. Then, data scientist can apply machine learning algorithms such as market basket analysis to find out which items are usually bought together so that I can suggest item that are compatible to the customer. Mining data text can also be used by data scientist to see at the emotion behind the feedback given by customers.

The findings then will be converted into story-telling visualization so that it can be used by others to make decision. Graphics like charts, graph, and maps can be used

during the data visulization process. A precise and easy-to-understand visualization will help me to understand better the customer preferences by seeing its patterns, trends, and relationships that I might missed in text-based data.

Lastly, data scientist will play a vital role in archiving data to preser it so that it will be highly reusable if I want to use for other purposes in the future. Data scientist needs to make sure that the data is secure against loss of data and degradation over time.

3. **Based on notes week two of "Basic of Algorithms", find/create one problem. You may refer to example of "Direction of numbered NYC streets algorithms" or "Class average algorithms".**
    a. **State the problem, input, processing and output.**

Problem: Determining employee performance level based on several factors such as performance review score, years of experience, number of projects completed, employee's age, and employee engagement survey score.

Input:

- Performance review score (out of 100)
- Years of experience
- Number of projects completed
- Employess engagement survey score (out of 100)
- Number of training completed

Processing: Multiply the performance review score and times with its weightage (30%); multiply the years of of experience with it weithage (20%); Multiply the number of projects completed with its weightage (15%); Multiply the employee's engagement survey score with it weightage (25%); Multiply the number of training completed with its weightage (10%); Calculate the total score; Categorize it into performance level (High: > 80, Medium: 60 – 80, Low: < 60)

Output: Employee performance level ('Low', 'Medium', 'High')

    b. **From the problem specified in part a) above, create three popular program design tool (flowcharts, pseudocode, hierarchy charts). For flowchart, please inclcude as many symbols as possible.**

Pesudocode:

Program: Determine employee performance level.

Multiply the performance review score with 0.3 and save it into variable ‘prs’

Multiply the years of experience with 0.2 and save it into variable ‘ye’

Multiply the numbers of projects completed with 0.15 and save it into variable ‘np’

Multiply the employee engagement survey score with 0.25 and save it into variable ‘eess’

Multiply the number of training completed with 0.1 and save it into variable ‘nt’

Sum ‘prs’, ‘ye’, ‘np’, ‘eess’, ‘nt and assign it to ‘score’

If ‘score’ greater than 80

    Emp_perf_lvl = ‘High’

If ‘score’ between 60 and 80

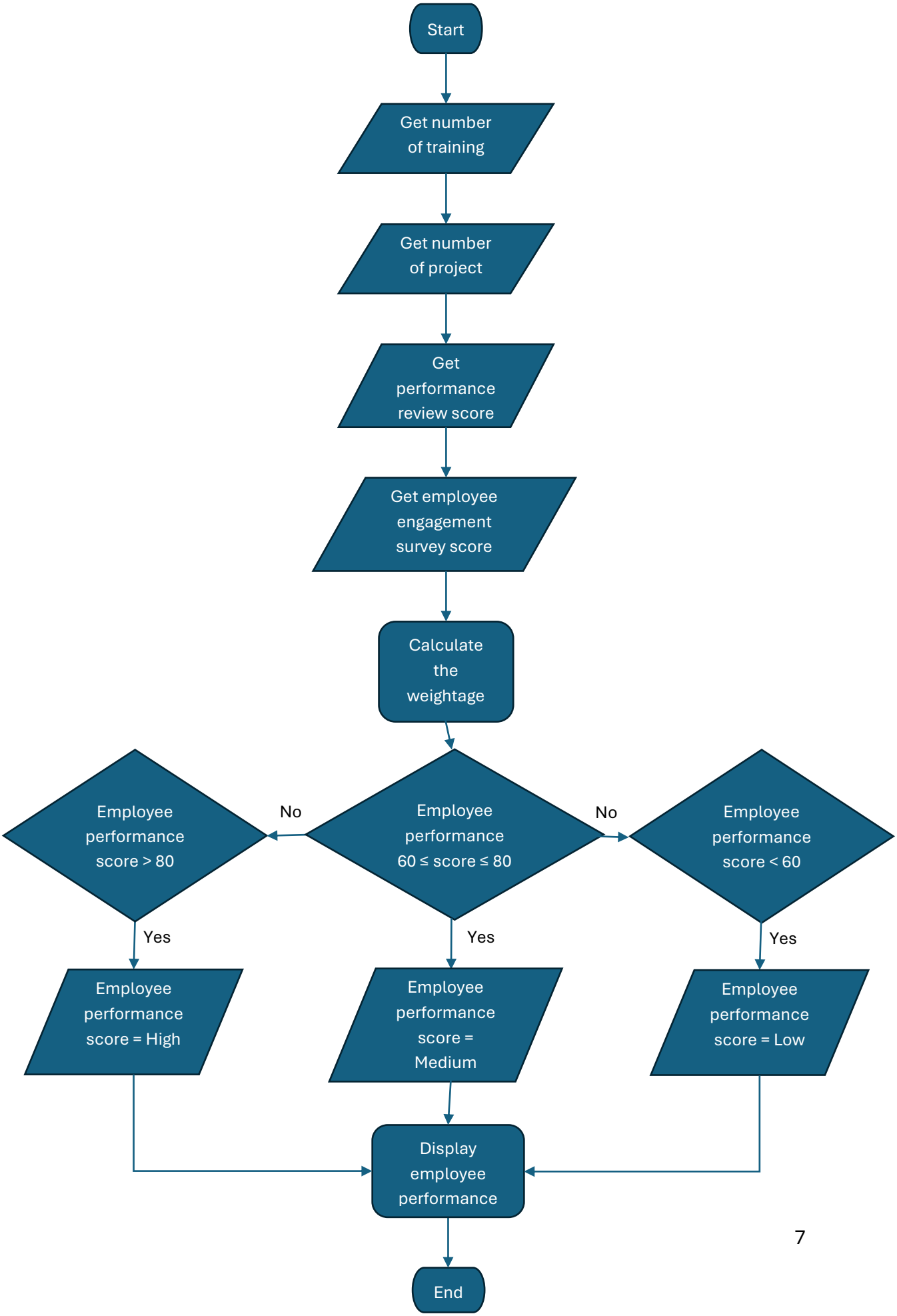
    Emp_perf_lvl = ‘Medium’

If ‘score’ less than 60

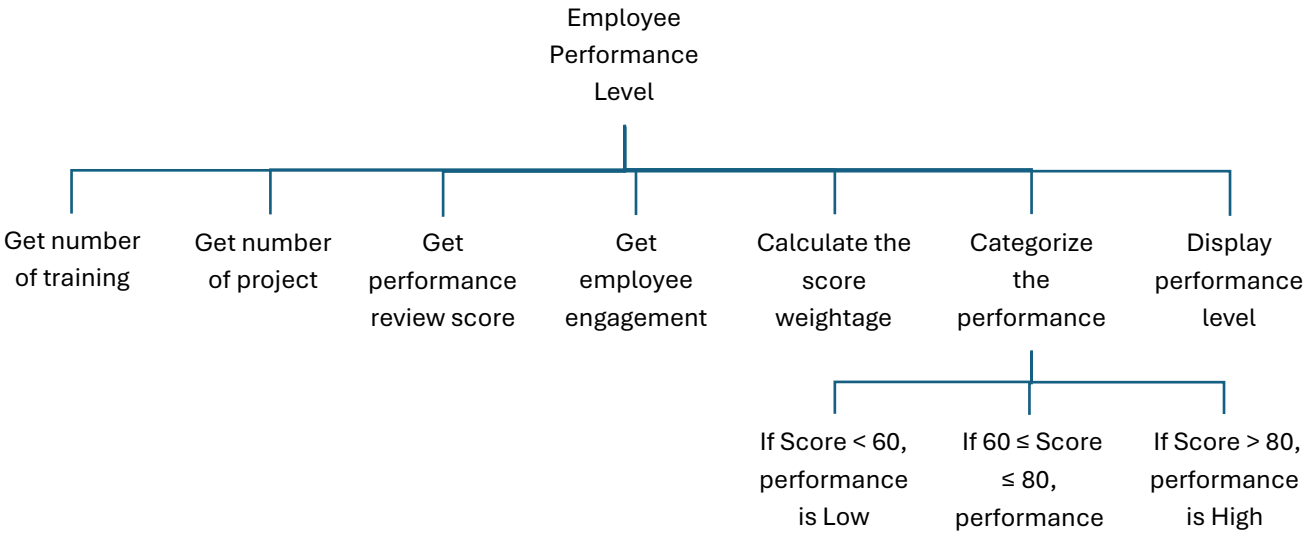
    Emp_perf_lvl = ‘Low’

Print ‘Emp_perf_lvl’ onto the console

Flowchart:


Start
Get number of training
Get number of project
Get performance review score
Get employee engagement survey score
Calculate the weightage
Employee performance score > 80
No
Employee performance 60 ≤ score ≤ 80
No
Employee performance score < 60
Yes
Yes
Yes
Employee performance score = High
Employee performance score = Medium
Employee performance score = Low
Display employee performance
End

Hierarchy Chart:



4. Give two examples of current data technology and its explanation.

First example of current data technology is **Fraud Detection Systems** that is being used mainly by banking and finance industry. This system use machine learning learning algorithms such as decision trees, random forest, and neural networks to identify unusual patterns in financial transactions. For example, if a customer often makes small, local purchases but suddenly attempts to make a large, international trnsactions, system will marks it as suspicious.

Next example is **Precision Algiculture** also known as smart farming. This method use multiple machine learning algorithms such as supervised learning (regression & classification), unsupervised learning (clustering), and deep learning (convolutional neural networks & recurrent neural networks. For example, drones, sensors, and satellite can be used to gather information on soil quality, moisture levels, and crop health. Then, this data is being used to identify areas with low nutrient levels or area prone to drought. This will be able to increase yields and reduced environmental impacts.