UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BITI 2513: INTRODUCTION TO DATA SCIENCE

MULTIPLE DISEASE PREDICTION
Task 2

Python Team

NAME:	1) Mohd Hariz Bin Abdul Malek (B031810296) 2) Shivedhassen a/l Balasinggam (B031810360) 3) Muhamad Nur Bin Muhamad Zaidi (B031810314) 4) Agilan a/l Manivasagam (B031810281)
CLASS:	2 BITI S1G1
COURSE:	BITI 2513
LECTURER NAME:	AP. Dr. Sharifah Sakinah Syed Ahmad

DATA MANAGEMENT PLAN FOR MULTIPLE DISEASE PREDICTION

Multiple disease prediction project is a project that can predict the disease that has in the human body that gives a sign like a symptom of the disease. This project involves the data analysis on how to predict using data mining techniques. Data mining techniques are used for a variety of applications. In the healthcare industry, data mining plays an important role in predicting diseases. Data mining is the process of selecting, discovering and modeling huge amounts of data. This process has become an increasingly insidious activity in all areas of medical science research.

0. Project Name
Multiple Disease Prediction
1. Description of the data
1.1 Type of study The study of this data is about the multiple type of disease with their symptoms which are collected from the kaggle website dataset **1.2 Type of data** The type of the data that is being recorded is qualitative that deals with characteristics of the symptoms of the disease which only can be observed but not to be measured. The data type for the dataset is object. **1.3 Format and scale of the data** The data format for this dataset is more to string which is the symptom and the type of disease. Most of the data scale is nominal scales or could be simply called as labels for the symptoms for the disease.
2. Data collection/generation
Data collection is one of the important processes of gathering and measuring the information. The method that is being used for the collection of the data should be one of the major concerns because the data that is being collected must be precisely perfect for the future use. **2.1 Methodology for data collection/generation** This is the primary data collection which is qualitative data that is being recorded based on the observation and research.

3. Data management, documentation and curation
The dataset provides the column containing symptoms of the disease and the type of the disease. This dataset can be obtained via kaggle website. This website allows you to make any documentation and get the details for the symptom, precaution, disease type and the weights. This data is placed in the kaggle for the long-term storage and preservation with no limit to the retention period.

The plan after this process is about getting the best cleaned data by the process of cleaning, restructuring and enriching the raw data available into a more usable format. It is easier to call it a data wrangling process. This will help in the process of decision making, and thus get better insights in less time. Data wrangling, like most data analytics processes, is an iterative one – the practitioner will need to carry out these steps repeatedly in order to produce the results desired.

Data Wrangling

Dataset that we obtained from kaggle is basically already cleaned and there is no need to do data wrangling. The data is already in appropriate and tidy form that is easy to analyze and manipulate. However the dataset contains null values which there might be a little cleaning to do with the data like switching the null values with something that holds value like ‘none’ instead of ‘?’. We are using RapidMiner Studio to do the data wrangling. We use the operator ‘Read CSV’ to read the csv file of the disease dataset. Then, we use the operator ‘Replace Missing Values’ to replace all the null values or ‘?’ in the dataset to ‘none’. Below is the operator used.

Below is the result of the data wrangling.

Row No.	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Sympto
1	Fungal infecti...	itching	skin_rash	nodal_skin_e...	dischromic _...	?	?	?	?
2	Fungal infecti...	skin_rash	nodal_skin_e...	dischromic _...	?	?	?	?	?
3	Fungal infecti...	itching	nodal_skin_e...	dischromic _...	?	?	?	?	?
4	Fungal infecti...	itching	skin_rash	dischromic _...	?	?	?	?	?
5	Fungal infecti...	itching	skin_rash	nodal_skin_e...	?	?	?	?	?
6	Fungal infecti...	skin_rash	nodal_skin_e...	dischromic _...	?	?	?	?	?
7	Fungal infecti...	itching	nodal_skin_e...	dischromic _...	?	?	?	?	?
8	Fungal infecti...	itching	skin_rash	dischromic _...	?	?	?	?	?
9	Fungal infecti...	itching	skin_rash	nodal_skin_e...	?	?	?	?	?
10	Fungal infecti...	itching	skin_rash	nodal_skin_e...	dischromic _...	?	?	?	?
11	Allergy	continuous_s...	shivering	chills	watering_fro...	?	?	?	?
12	Allergy	shivering	chills	watering_fro...	?	?	?	?	?
13	Allergy	continuous_s...	chills	watering_fro...	?	?	?	?	?
14	Allergy	continuous_s...	shivering	watering_fro...	?	?	?	?	?
15	Allergy	continuous_s...	shivering	chills	?	?	?	?	?
16	Allergy	shivering	chills	watering_fro...	?	?	?	?	?

Before replacing missing values.

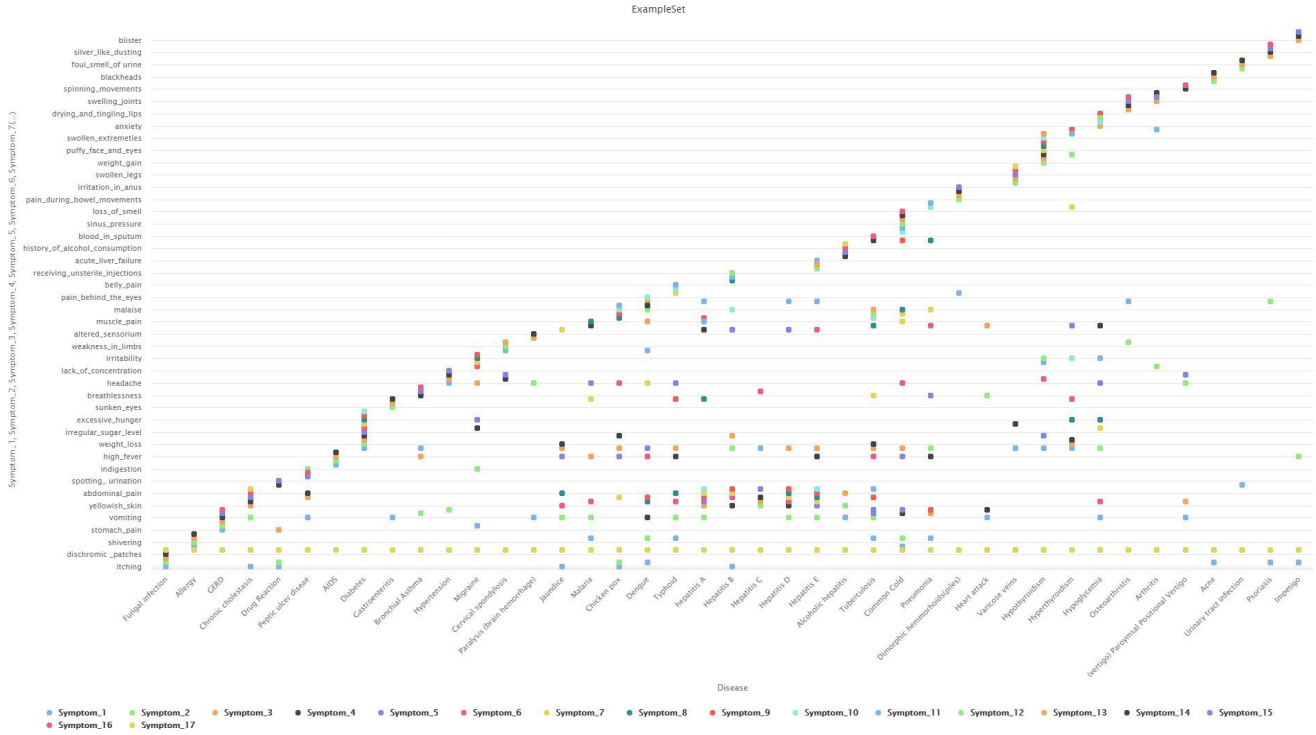
Row No.	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Sympto
1	Fungal infecti...	itching	skin_rash	nodal_skin_e...	dischromic _...	none	none	none	none	none
2	Fungal infecti...	skin_rash	nodal_skin_e...	dischromic _...	none	none	none	none	none	none
3	Fungal infecti...	itching	nodal_skin_e...	dischromic _...	none	none	none	none	none	none
4	Fungal infecti...	itching	skin_rash	dischromic _...	none	none	none	none	none	none
5	Fungal infecti...	itching	skin_rash	nodal_skin_e...	none	none	none	none	none	none
6	Fungal infecti...	skin_rash	nodal_skin_e...	dischromic _...	none	none	none	none	none	none
7	Fungal infecti...	itching	nodal_skin_e...	dischromic _...	none	none	none	none	none	none
8	Fungal infecti...	itching	skin_rash	dischromic _...	none	none	none	none	none	none
9	Fungal infecti...	itching	skin_rash	nodal_skin_e...	none	none	none	none	none	none
10	Fungal infecti...	itching	skin_rash	nodal_skin_e...	dischromic _...	none	none	none	none	none
11	Allergy	continuous_s...	shivering	chills	watering_fro...	none	none	none	none	none
12	Allergy	shivering	chills	watering_fro...	none	none	none	none	none	none
13	Allergy	continuous_s...	chills	watering_fro...	none	none	none	none	none	none
14	Allergy	continuous_s...	shivering	watering_fro...	none	none	none	none	none	none
15	Allergy	continuous_s...	shivering	chills	none	none	none	none	none	none
16	Allergy	shivering	chills	watering_fro...	none	none	none	none	none	none

After replacing missing values.

Changing the null values into certain values will help us build a model for the prediction of the data in the future.

Exploratory Data Analysis (EDA)

After cleaning and transforming the data into an appropriate form by replacing the missing values, we decided to view and observe the data for what we can or might find in the data if there is a pattern or important information in it. Again we are using RapidMiner Studio to visualize the data that has been cleaned. For the visualization, we are using scatter plots to visualize the data into a graph. Below is the result of the visualization.



The graph above shows that the x-axis is the disease and the y-axis is the symptoms. From our observation and view, we can see that the symptoms belong to certain diseases. A disease can have a small and large amount of symptoms. From the graph, we can see that there are diseases that have small and large amounts of symptoms. The diseases that have large amounts of symptoms show that the disease is easy to diagnose since the patient or the victim of the disease experience those symptoms. However for the diseases that have less symptoms shown, it indicates that the disease might be difficult to identify or diagnose. From a medical perspective, it is reasonable that most diseases in early stages are difficult to identify and diagnose like cancer for example. Hence the graph helps us identify which disease has less symptoms that make it difficult to diagnose.

WHICH VARIABLES TO EXPLAIN AND WHICH ONES TO PREDICT?

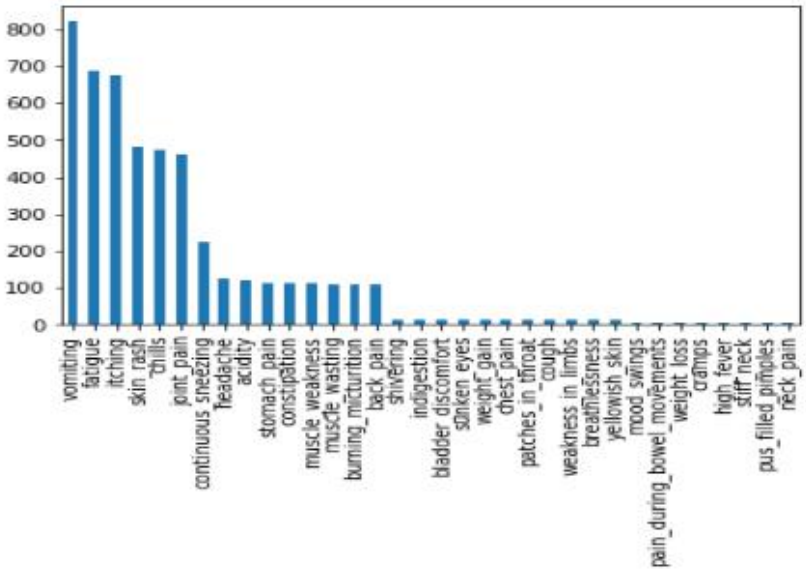
The data of diseases and their respective symptoms that are obtained has two main attributes; the name of the disease and the symptoms of the disease. The same symptoms will occur for several diseases. But as explained earlier in EDA part, the disease with less symptoms will be hard to be predicted as the symptoms might also fall into other diseases. And the more the symptoms shown for a disease the chances of predicting the disease will be much greater and more accurate. Therfore, symptoms will be the variables that are to be explained and disease is the variable to be predicted.

Based on the symptoms entered by the patient or a person, the system will predict the disease of the patient or the person according to the symptoms. And as said earlier, if the symptoms entered are enough (estimated more than five symptoms) the prediction of the disease will be precise and acccurate but will not be preise if the symptoms entered are less (one to four symptoms).

Data Visualization

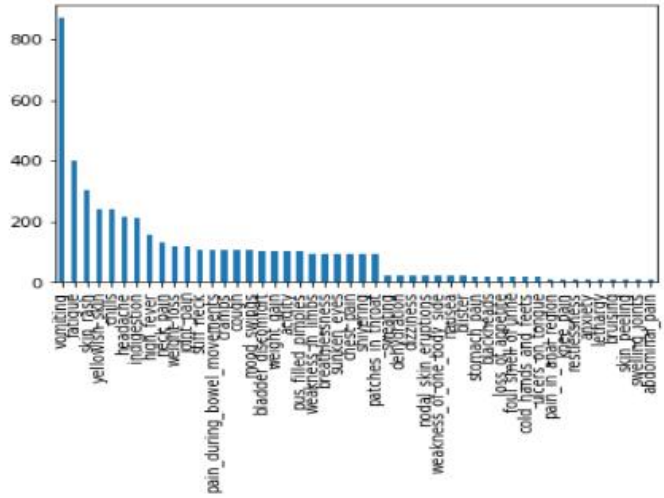
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends and patterns in data.

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv(r"C:\Users\User\Downloads\disease_dataset.csv")

df['Symptom_1'].value_counts().plot(kind='bar')
plt.show()
```



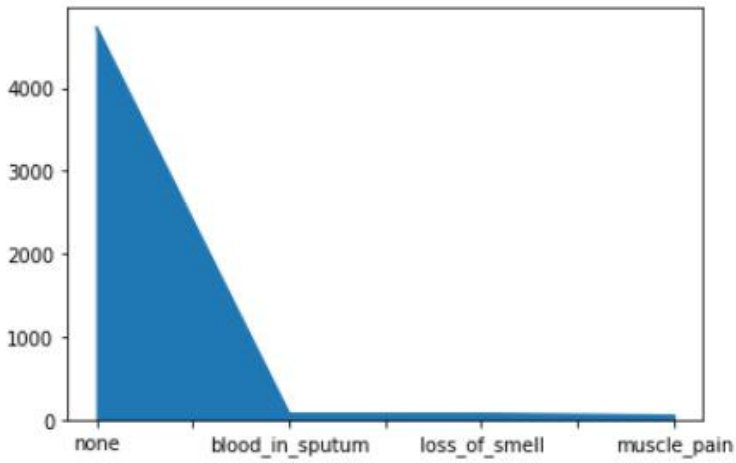
The figure above shows that majority of the people have been vomiting as their symptom for Symptom 1 followed by being fatigue, itching ,skin rash and so on. Symptoms such as shivering, indigestion and chest pain are less experienced by the people.

```
df['Symptom_2'].value_counts().plot(kind='bar')
plt.show()
```



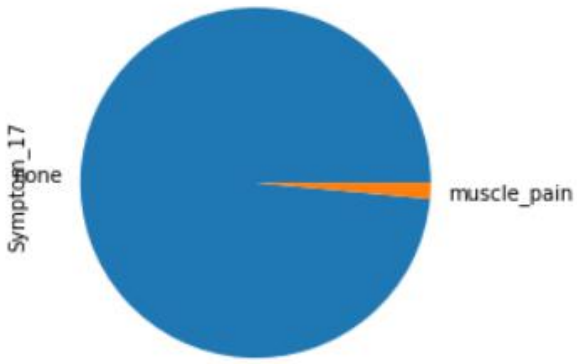
The figure above shows that majority of the people have been also vomiting for Symptom 2 followed by fatigue,skin rash,yellowish skin and so on. They also experienced very less symptoms such as swelling joints and abdominal pain.

```
df['Symptom_16'].value_counts().plot(kind='area')
plt.show()
```



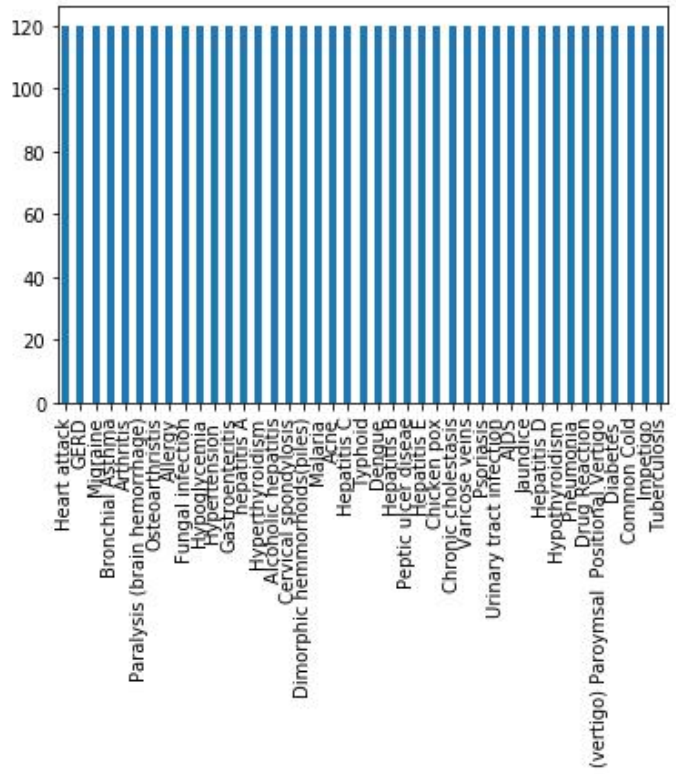
The figure above shows that more than 4000 didn't had Symptom 16 and very less people had blood in sputum, loss of smell and muscle pain as the symptoms.

```
df['Symptom_17'].value_counts().plot(kind='pie')
plt.show()
```



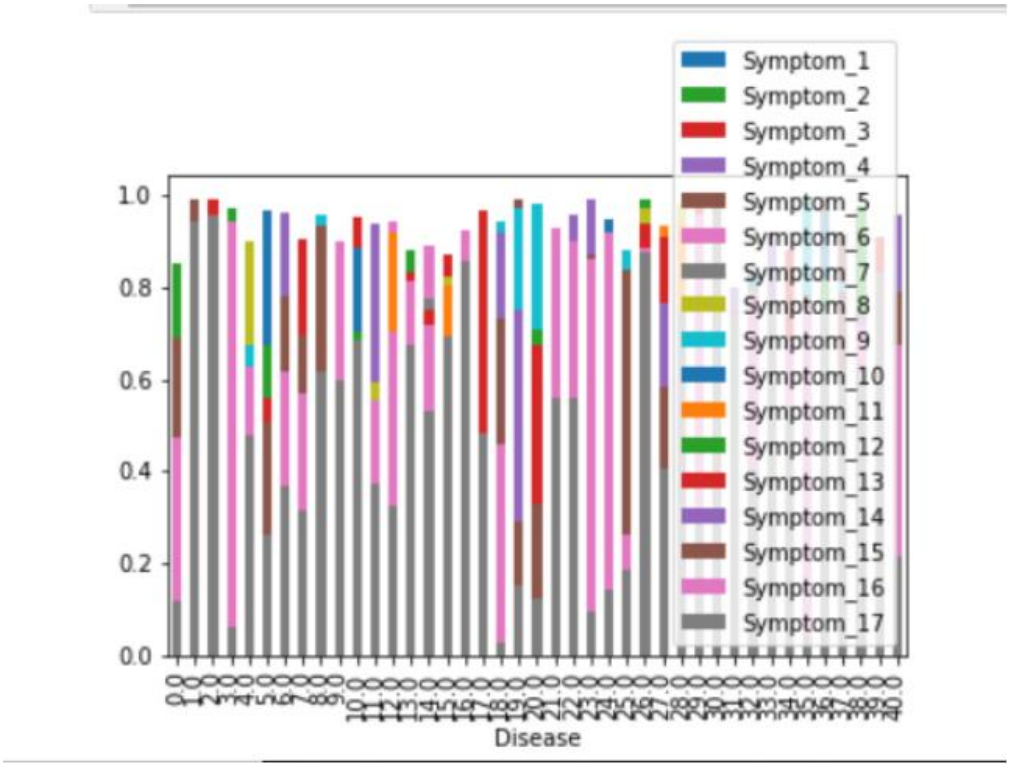
The figure above shows that majority had no symptom 17 and very few people had muscle pain as their 17th symptom.

```
df['Disease'].value_counts().plot(kind='bar')
plt.show()
```



The figure above shows the list of diseases the people may experience based on the symptoms mentioned previously.

```
import matplotlib.pyplot as plt
import pandas as pd
import matplotlib.pyplot as plt
from IPython import get_ipython
get_ipython().run_line_magic('matplotlib', 'inline')
import numpy as np

df = pd.read_csv(r"C:\Users\User\Downloads\disease_dataset.csv")

y = np.random.rand(41,18)
y[:,0]= np.arange(41)
df = pd.DataFrame(y, columns=["Disease", "Symptom_1", "Symptom_2", "Symptom_3","Symptom_4","Symptom_5",
                              "Symptom_6","Symptom_7","Symptom_8","Symptom_9", "Symptom_10","Symptom_11",
                              "Symptom_12","Symptom_13","Symptom_14","Symptom_15","Symptom_16","Symptom_17"])

ax = df.plot(x="Disease", y="Symptom_1", kind="bar")
df.plot(x="Disease", y="Symptom_2", kind="bar", ax=ax, color="C2")
df.plot(x="Disease", y="Symptom_3", kind="bar", ax=ax, color="C3")
df.plot(x="Disease", y="Symptom_4", kind="bar", ax=ax, color="C4")
df.plot(x="Disease", y="Symptom_5", kind="bar", ax=ax, color="C5")
df.plot(x="Disease", y="Symptom_6", kind="bar", ax=ax, color="C6")
df.plot(x="Disease", y="Symptom_7", kind="bar", ax=ax, color="C7")
df.plot(x="Disease", y="Symptom_8", kind="bar", ax=ax, color="C8")
df.plot(x="Disease", y="Symptom_9", kind="bar", ax=ax, color="C9")
df.plot(x="Disease", y="Symptom_10", kind="bar", ax=ax, color="C10")
df.plot(x="Disease", y="Symptom_11", kind="bar", ax=ax, color="C11")
df.plot(x="Disease", y="Symptom_12", kind="bar", ax=ax, color="C12")
df.plot(x="Disease", y="Symptom_13", kind="bar", ax=ax, color="C13")
df.plot(x="Disease", y="Symptom_14", kind="bar", ax=ax, color="C14")
df.plot(x="Disease", y="Symptom_15", kind="bar", ax=ax, color="C15")
df.plot(x="Disease", y="Symptom_16", kind="bar", ax=ax, color="C16")
df.plot(x="Disease", y="Symptom_17", kind="bar", ax=ax, color="C17")

plt.show()
```



The figure above shows a Stacked Bar Chart. In the stacked bar plot, the bars at each index are literally “stacked” on top of one another. It shows the visual representation of the total symptoms each disease has, and the breakdown of each symptom. Many of the disease share several symptoms. Based on this stacked bar chart ,majority of the people experience Symptom 16 and Symptom 17 which means they experience no symptoms according to the previously mentioned pie chart.

References

1. https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed
2. https://www.kaggle.com/itachi9604/disease-symptom-description-dataset?select=dataset.csv
3. https://pstblog.com/2016/10/04/stacked-charts
4. https://www.bmc.com/blogs/matplotlib-stacked-bar-chart/
5. https://www.youtube.com/watch?v=2f_RZFDU_Os [Data Pre-processing using Rapidminer]
6. https://docs.rapidminer.com/latest/studio/operators/cleansing/missing/replace_missing_values.html