PRINCIPLES OF DATA SCIENCE

**ASSIGNMENT-1**

MANOJ KUMAR SUGGALA

**(16357169)**

**Task-1:**

According to given problem statement, we need to collect the data from document, and we need to perform the three stages of reproduction workflow:

As per given document in the lecture we can say that the three stages of reproducibility are:

1. Data Input
2. Data Pre-processing
3. Data Analysis

**Step 1: Data Input**

* Initially we need to collect the data from the available resource which is present in the given document.
* Collect all the data and then input them into a file as shown below and save it as .csv file for easy accessing and processing it in the future.

A screenshot of a computer

Description automatically generated

Once the raw data is collected and it is imported on to the file we have to make it get useful for future data pre-processing.

**Step-2 Data Pre-processing:**

1. Here we need to clean the data.
2. Check for any null values and remove them accordingly.
3. To perform data cleaning we can use any python notebook of our choice like jupytheror vs code or google colab. we are using Google Colab python notebook as shown below:

A screenshot of a computer

Description automatically generated

1. As we can clearly see in the data, they are no null values or NA values, so we can go to next step and we can perform analysis of the data.
2. If it is necessary, you can also rename the column names by removing the spaces if required.

**Step 3: Data Analysis:**

1. At last, we need to perform some data analysis on it.
2. Perform some data visualizations on this data. These visualization helps on getting the data understand easily.
3. As shown below the visualizations are performed on the below data we can some output as shown below:

A screenshot of a computer

Description automatically generated

At last, we can see some folders locations and their path:

Step 1:

|— Raw data

|— task1\_raw.csv

Step 2:

|— raw data

|— task1\_raw.csv

|— Processing

|— ctask1\_clean.csv

|—clean.py

Step 3:

|— raw data

|— task1\_raw.csv

|— Processing

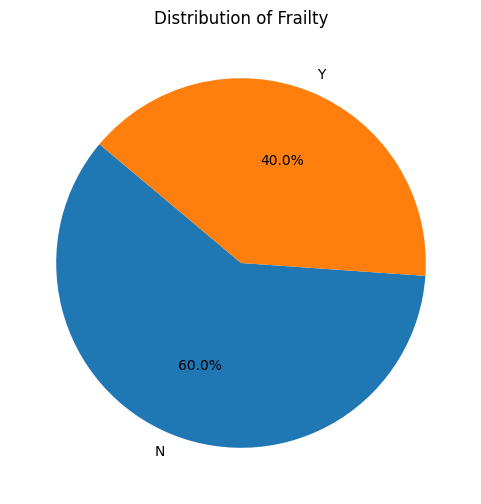
|— ctask1\_clean.csv

|—clean.py

|— results

|— output.png

|—analysis.py



By seeing the data we can clearly see that he ratio of frailty Yes to No is 40.0% to 60.0%. In this way we can clearly understand the type of data by doing visualization.