**Module 2 (Part I) – EDA Action Plan**

**Name: Maviya Shaikh**

**Student number: 100766785**

**Date: February 14th,2020**

**Table of Contents**

[1. Problem Statement and Analytical Statement 3](#_Toc32074938)

[2. Identification and Justification of Output Variable Class Structure 3](#_Toc32074939)

[3. Exploratory Data Analysis (EDA) Approach 3](#_Toc32074940)

[4. Assumptions 3](#_Toc32074941)

[5. Constraints 4](#_Toc32074942)

1. Problem Statement and Analytical Statement

**Problem Statement**: The analysis for the test of TNM stage 1 pre-screening is conducted to identify whether the patient has cancer or not.

**Analytical Statement:** The model is developed to identify the patient has stage 1 cancer or not.

1. Identification and Justification of Output Variable Class Structure

* The class structure is *binomial* because the analysis will conduct for detecting whether the patient has cancer or not
* The naming convention that is used to represent the binomial problem is *No Cancer* and *Cancer*

1. Exploratory Data Analysis (EDA) Approach

The following steps will be taken to understand the dataset:

* **Basic statistics** is calculated to understand the number of observations and spread of the data
* **Fixing the missing values** by removing or fillingthe data to clean the dataset
* **Detecting outliers** to know if the value is much deviated from the other data points. It can be removed if it is biasing the model
* **Checking whether the dataset is balanced or not,** if it is not balanced then apply the strategies to balance the dataset so that model will not give biased predictions
* **Scaling the data** will help independent variables to be in the certain range that will normalize the data and speed up the algorithms

The following graphs will be used for the analysis of the dataset:

* **Boxplot**: The boxplot shows the significance of each variable and the distribution of the data around the mean. It also helps in identifying the outliers in the dataset
* **Histogram**: The histogram will help identify whether the data is normal, right skewed or left skewed. It also represents the occurrences of the specific value in the dataset and also detect the outliers in the dataset
* **Heatmap**: The heatmap will determine the intercorrelation between the variables that will help in selecting the model

1. Assumptions

* The dataset is valid and come from the reliable source
* The dataset provided is legible and comprehensible
* All independent variables in the dataset are useful to predict the outcome variable
* The necessary software and hardware tools require to complete the project are available

1. Constraints

* The additional data cannot be added in the dataset
* The additional feature or attribute cannot be added in the dataset
* All the independent variables are cell structures which is highly corelated because it is coming from the same body
* The final model to detect whether patient has cancer or not must be built to meet the deadline which is April 17, 2020

All the constraints above are hard constraints because it is not legitimate to add the data or new feature to decrease the correlation between independent variables. Also, the milestone must be achieved within the deadline.