**Module 1 - Statement of Work**

**Name: Maviya Shaikh**

**Student number: 100766785**

**Date: January 24th,2020**

**Table of Contents**

|  |
| --- |
| 1. Executive Summary……..………………………………………………………………………….3 |
| 1. Problem Statement………………………………………………………………………………….3 |
| 1. Analytics Rationale Statement…………………………..…………………………………….3 |
| 1. Data ………………………………………..………………………………………………………………3 |
| 1. Data Analysis Approach ………….………………………….……………………………………4 |
| 1. Project Plan.………………………....…………………………………………………………………4 |

1. **Executive Summary**

The purpose of this document is to provide the reference guide in achieving the milestone to develop the model that identifies whether the patient has cancer or not. The objective is to provide solution to Princess Margret Hospital who can alert the patient at potential cancer risk.

1. **Problem Statement**

*“The analysis for the test of TNM stage 1 pre-screening is conducted to identify whether the patient has cancer or not”.* The purpose of doing the thorough research on this problem is to help an individual who is diagnose with the stage 1 cancer in which cancer is relatively small and contained within the organ it started in.

**Evaluation metrics:**

The number ofevaluation metrics that will be used to analyse the model are as follows:

* **Accuracy**: It is the ratio of number of correct predictions to the total number of predictions made. Thus, accuracy represent the percentage of predictions that model got right. This metric will give the correct result only if there is equal distribution of the class
* **Precision**: It tells how often the model is correct when it makes the prediction
* **Recall**: It is the ratio of correctly classified the positive instances or negative instances from the total number of positive instances or negative instances respectively. This metric will identify all the relevant instances from the retrieved instances.
* **F1-score**: It is the harmonic mean of precision and recall. If there is an uneven distribution of the class then this metric is considered for evaluation of the algorithm.

1. **Analytics Rationale Statement**

*“The model is developed to identify the patient has stage 1 cancer or not”*. This analysis is carried out because Princess Margret is concerned with the lack of screening for cancer during routine yearly examinations for patients that might be at risk. So, the early detection of the cancer in patient will help them to recover from this disease after it is too late.

1. **Data**

The data has been originated from Princess Margret Hospital. It is the Clinical Research Unit for Cancer. So, it is assumed that data source is reliable, and all the data provided is correct. Below is the brief overview of the dataset:

* The dataset has 10 variables in which 9 variables are independent and represent the structure of the cell in numerical values, whereas 1 variable is dependent and represent class with categorical value (i.e. 0 describe that patient is not detected with cancer, while 1 and 2 describe that patient is detected with cancer)
* The total number of observations in the dataset is 1690 from which 10 observations (or 100 cells) data are missing
* It is evident that nine independent variables are highly correlated with each other as all the cells for particular observation came from the same body

1. **Data Analysis Approach**

The analysis approach used to identify the patient who has cancer or not are as follows:

* The machine learning algorithms will be used to develop the model in the backend. The software tool used to run the algorithms is Jupyter Notebook (Anaconda 3) using python programming or RStudio using R programming
* For the visualization of the data in terms of various graphs and charts, tableau or power bi will be used in the frontend to display the analysis of the data

1. **Project Plan**

The project plan includes all the deliverables with the date. This includes all the task that will be carried out during this project which is as described below:

|  |  |  |
| --- | --- | --- |
| **ID** | **Milestone** | **Submit Date** |
| 1 | Exploratory Data Analysis and Data Cleaning | February 14,2020 |
| 2 | Review and Chose algorithm(s) | March 6,2020 |
| 3 | Train model and Validate model | March 27,2020 |
| 4 | Develop finalized model(s) and final report | April 17, 2020 |

Table 6.1: Project Plan