**Project V1 - SoW**

**NANDINI MALHOTRA**

**100768797**

**October 28, 2020**

**TABLE OF CONTENT**

[Executive Summary 2](#_Toc54800291)

[Problem Statement & Rationale Statement 2](#_Toc54800292)

[Data Description and Data Analysis 3](#_Toc54800293)

[Identifying and Justifying Output Variable 4](#_Toc54800294)

[Constraints and Assumptions 4](#_Toc54800295)

[Metrics for evaluating the quality of the model 4](#_Toc54800296)

# Executive Summary

The project aims at detecting Cancer at its initial stage (Stage 1) with the help of features in the dataset. This would involve developing a model that would be trained using the available dataset.

# Problem Statement & Rationale Statement

Princess Margret is concerned with the lack of screening for Stage 1 cancer for the patients who are at potential risk of Cancer. We will present a solution to this problem by developing a model that would identify the class for Stage 1 cancer, based on which the patient would be subjected to different treatments.

The problem statement is

“**To determine whether the patient coming for pre-screening has cancer or not and if he does, then what class (1 or 2) of Stage 1 cancer it is.**”

The problem statement aims at classifying the class of stage 1 cancer. Upon determining the class, the hospital would be able to treat the patient with the exact type of treatment needed. For example, a class 1 type cancer might require 2 sessions of chemo and 1 session of radio however, a class 2 type cancer might

require 4 sessions of chemo and 3 sessions of radiotherapy. After training the model on the dataset, the hospital can use the model for potential patients and determine whether the patient falls in the Class 1 or Class 2 category.

And so, our Rationale Statement would be the reason why we are proposing the problem statement:

“**Upon Determining the class, the patient would be given appropriate treatment by the hospital**”

# Data Description and Data Analysis

The data is collected in Collected from the Princess Margret Hospital and it constitutes only 10% of the current data. It has 10 columns out of which there are 9 independent/predictor variables (features) ranging from V1 to V9 and one dependent/response variable, which indicates the sub-classes of Stage 1 cancer which are:

* 0-Negative
* 1-Stage I(T=1 or 2,N=1 , M= 1)
* 2-Stage I (T=3 or 4, N= 2 or 3, M=1)

Where, T refers to the size of cancer and how far it has spread ranging from 1(small) to 4(large), N is the extent to which it has reached lymph nodes between 0 (no lymph nodes containing cancer cells) and 3 (lots of lymph nodes containing cancer cells) and M is the measure to determine whether cancer has spread to different body parts- it can either be 0 (cancer hasn't spread) or 1 (cancer has spread).

The model will be trained on the training dataset, using a machine learning algorithm, to identify:

* the Class 1- Stage 1 cancer, so that it can be prevented from spreading to further lymph nodes
* the Class 2- Stage 1 cancer, so that it can be prevented from spreading to different body parts

The patient will be treated, depending on the type of class diagnosed:

* the Class 1 type patient can be treated with 2 or 3 chemo sessions
* the Class 1 type patient can be treated with the combination of radio or chemo sessions

# Identifying and Justifying Output Variable

The data is collected in Collected from the Princess Margret Hospital and it constitutes only 10% of the current data. It has 10 columns out of which there are 9 independent/predictor variables (features) ranging form V1 to V9 and one dependent/response variable, which indicates the sub-classes of Stage 1 cancer which are:

* 0-Negative
* 1-Stage I(T=1 or 2,N=1 , M= 1)
* 2-Stage I (T=3 or 4, N= 2 or 3, M=1)

Where, T refers to the size of the cancer and how far it has spread ranging from 1(small) to 4(large), N is the extent to which it has reached lymph nodes between 0 (no lymph nodes containing cancer cells) and 3 (lots of lymph nodes containing cancer cells) and M is the measure to determine whether cancer has spread to different body parts- it can either be 0 (the cancer hasn't spread) or 1 (the cancer has spread).

The **Output Variables** for the model will represent **Multi Class** and they’ll be identified as:

* **Class 0: For the patient who are diagnosed free from cancer**
* **Class 1: For the patients who has Stage 1, Type 1 cancer**
* **Class 2: For the patients who has Stage 1, Type 2 cancer**

# Constraints and Assumptions

The dataset needs to be cleaned before it can be used for training the model. There are few **constraints** in data set like:

1. ‘na’ value
2. Outliers
3. Negative values
4. The column labels are undefined
5. The data should not be re-sized

We also must make some **assumptions** while analysing the data, like:

1. Considering that the data provided by the hospital is accurate
2. The source of data is reliable
3. Depending on the algorithm that we use to define a model, we might have to consider the data to be normally distributed or the features to be independent of each other.
4. The measurements are assumed to be from different patients.

We will create the model **using a machine learning algorithm** and train the model in Jupyter notebook using python.

# Metrics for evaluating the quality of the model

The evaluation metrics that will be used are:

1. Recall: This is used to identify all the positive instances and is calculated as:

**Recall = TP/(TP+FN)**

By measuring the recall, we would be able to identify what percentage of actual positive results are detected. For ex, there are 100 patients with class1 cancer, and the model predicts 95 of them as class 1 and other as class 2, then the recall is 95%. We aim at High Recall for an effective model.

1. F1 Score: This is the mean of precision and recall and is calculated as:

**F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)**

F1 score is the combination of precision and recall hence would consider the wrong prediction, which contributes to the precision i.e., if a model predicts 120 patients to be of Class 1 out of which only 95 were of class 1 then the precision will be 79% (95/120). This metric would consider False- positive and False Negative, which would help us optimize the model.