**PROJECT-V2 | Introduction to AI Algorithm**

**NANDINI MALHOTRA**

**100768797**

**November 28, 2020**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **Index** | **Description** | **Page Number** |
|  |  |  |
| 1 | Problem Statement & Rationale Statement | 2 |
| 2 | Identifying and Justifying Output Variable | 2 |
| 3 | Action Plan for EDA | 3 |
| 4 | Exploratory Data Analysis | 4 |
|  | 4.1 Finding the number of outliers and NA values in the   dataset | 4 |
|  | 4.1.1 Removing the missing and outliers form the   dataset | 5 |
|  | 4.2 Understanding the Descriptive Statistics of the   features | 5 |
|  | 4.3 Interpreting the box plot for the features | 6 |
|  | 4.4 Checking for correlation between the predictor   variables using Heatmap | 6 |
|  | 4.5 Feature Selection | 7 |
|  | 4.6 Data Metric | 7 |

# 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 multiple classes of cancer, based on which the patient would be subjected to different treatments.

“**To determine multiple classes of cancer for a patient appearing for the cancer pre-screening test at Princess Margret Foundation.**”

The problem statement aims at classifying the different classes of cancer. There are broadly 3 classes that the model will identify, namely: Class 0 (No- Cancer), Class 1(Stage 1, Type 1 Cancer) and Class 2(Stage 1, Type 2 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 radio therapy. After training the model on the dataset, the hospital can use the model for potential patients and determine whether the patient falls in 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 an appropriate treatment by the hospital**”

# 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**

# Action Plan for EDA

**Steps for EDA**

1. The statistics of the dataset would be understood using pd.describe() function that summarizes the central tendency(mean) and dispersion of the values of features in the dataset.
2. All the **missing values** in the dataset will be identified using pandas profiling and will be removed.
3. The dataset will be looked in for **outliers using Tukey’s Algorithm**, and the outlier will be removed.
4. The data will be scaled using **Standard Scaler** from sklearn library, in order to scale the Predictor variables.
5. The data will be identified as **balance/imbalanced using confusion matrix** to determine if we have any biasing toward a class. If the dataset is imbalance, it will be balanced using SMOTE Analysis.
6. The **co-relation between the predictor variables will be determined using Variance Inflation Factor** and the variables which have higher VIF can be removed which will help optimize the model.

**Graphs and Analysis types**

1. The **training curve** will be plotted to determine the Variance, which will help us predict if the model is Overfitting.
2. The **Heatmap** will be used to plotted to determine the co-relation between the predictor variables.
3. **ROC curve** will also be plotted to determine the Area Under Curve which will help us determine how well can our model distinguish between the classes.
4. **Permutation Selection** will be performed which will help us determine which predictor variable has the most impact on the response variable. Also, a **box plot** will pictorially represent the Feature Importance Distribution. The model can then be optimized by removing the least impactful feature.
5. **Confusion Matrix** will depict the evaluation metrics that will help us evaluate the performance of the model.

We’ll consider some **assumptions** while analysing the data, like:

1. Considering that the data provided by the foundation is valid and reliable.
2. All the independent variables are assumed to be equally important in predicting the outcome and so none of them can be removed without rationalising their impact on the response variable.
3. The dataset is assumed to be the representation of the whole dataset.
4. 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.
5. Each row is assumed to be from different patients.

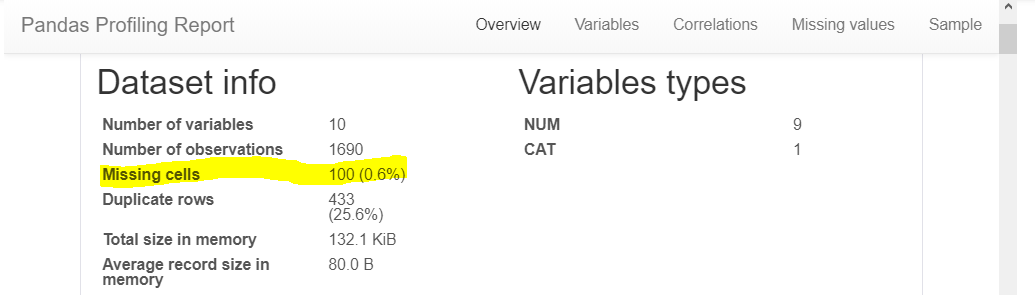
There are few **constraints** in data set like:

1. There are few rows that have ‘na’ value. They must be removed or be substituted with the mean of their respective columns.
2. There are be some Outliers that will be dealt with Tukey’s Algorithm.
3. The data should not be re-sized, i.e. we cannot add new features or data to the dataset.
4. The column labels are undefined hence, we cannot prioritize one feature over the other

# Exploratory Data Analysis

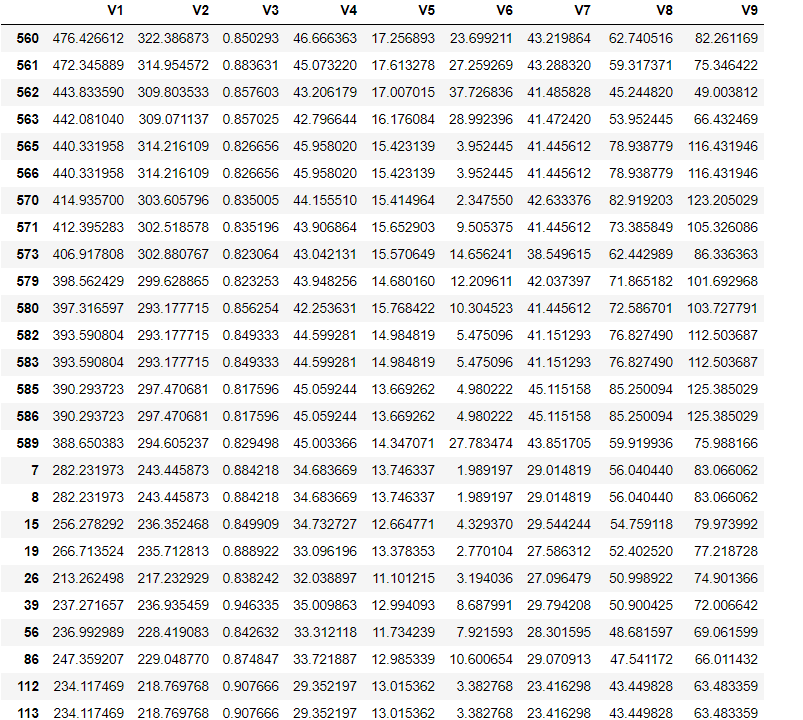
## 4.1 Finding the number of outliers and NA values in the dataset

The NA values can be checked for using pandas profiling:



There are 10 missing values

The Outliers can be identified using Tukey’s Algorithm:

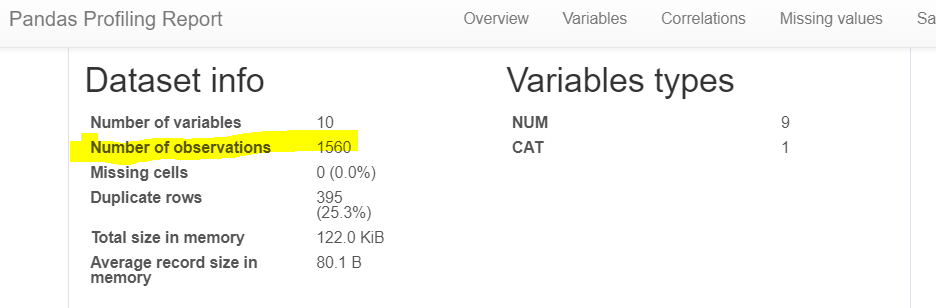


There are **120 outliers** in the dataset.

### Removing the missing and outliers form the dataset

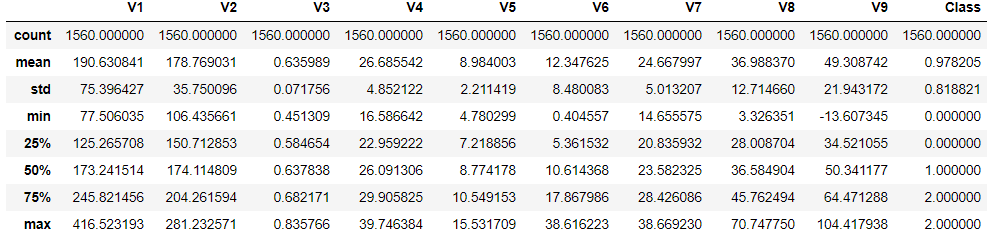
Presence of missing values and outliers can reduce the fit of the model. This can bias the model and make incorrect prediction. As a part of pre-processing of the data, these missing values are dropped for the rows having all values as NA. After removing missing values 1680 observations

are left in the dataset. The Outliers are also removed as they can impact the mean of the features which cam lead to misleading interpretations



## Understanding the Descriptive Statistics of the features

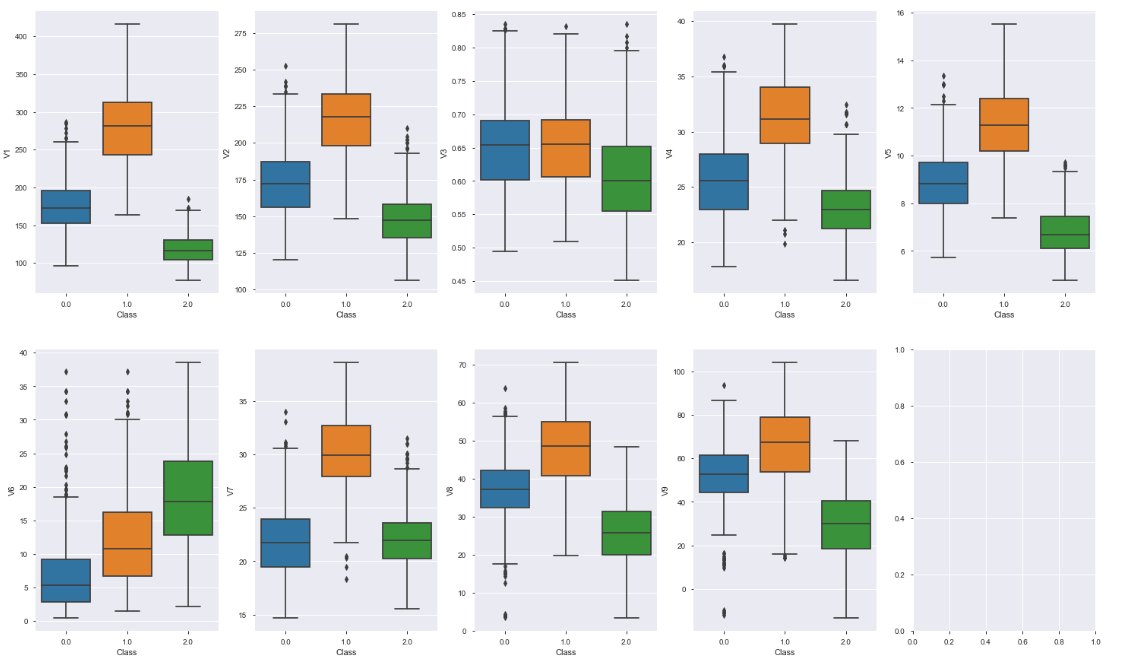
The statistics of the dataset can be interpreted using pd.describe() function.



The describe function helps determine the spread, mean and standard deviation of all the features.

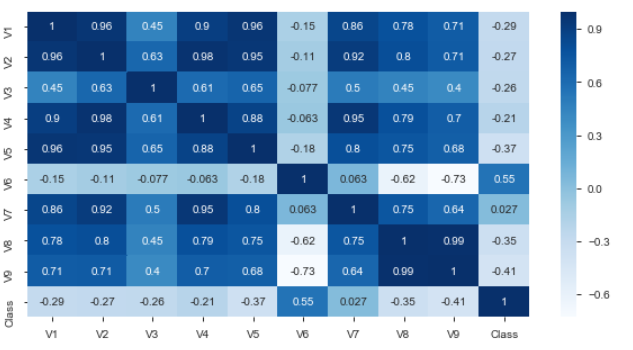
From the mean and std of the feature it can be observed that V1 and V6 have right skewed   
distribution as the deviation is greater than the minimum value, hence most of the values in 65% range, will be to the right of the mean. Upon dividing the std by mean for each predictor variable (V1- V9), a coefficient of variance is obtained. The higher the value of the coefficient of variance the closer the observations are to mean.

## Interpreting the box plot for the features

.

The box plot depicts the graphical presentation of the minimum and maximum value of the features, the median and the outliers. Presence of outliers can skew the data, for example, for V6, the outliers have right skewed the distribution. Too many outliers can limit the type of models that can be used. However, removing them, might introduce bias that can be dealt with bagging and boosting.

## Checking for correlation between the predictor variables using Heatmap

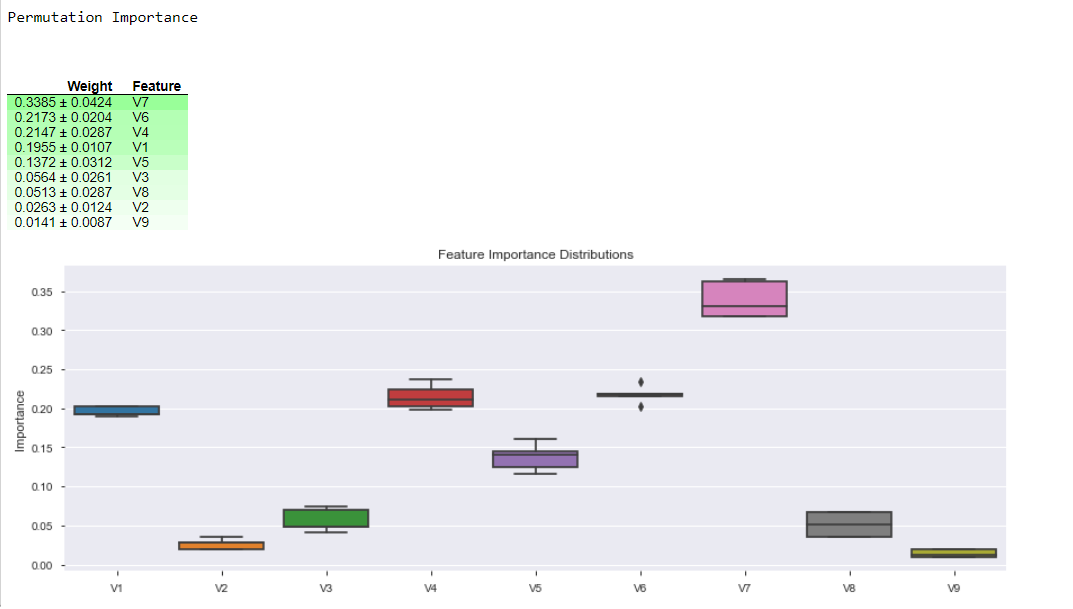


From the heatmap, it can be concluded, all the variable whose correlation factor is greater than 0.8, are strongly correlated. The highly correlated features are listed above.



Since, V1, V2, V5, V4 are strongly correlated as they have high correlation factor, then removing some of the features might improve the performance of the model. The features are removed to reduce the complexity of the model. Also, the lesser the number of features, the lesser will be the dimensions upon which the outcome would depend. The features can be reduced using feature extraction like Linear Discriminant Analysis or Principal Component Analysis.

## 4.5 Feature Selection



V7 has the most impact on the Class of cancer. However, it is observed V9, V2 and V8 have the least impact on the outcome. Hence removing these features might help optimize the model. Also, features with low weight and high correlation can be dropped at the same time. For e.g. V9 and V8 are highly correlated and have low weights. So, removing these features might improve model performance.

## Data Metric

The base model is Logistic Regression Model and the metrics that’ll be used to evaluate the model performance will be:

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.

1. **The Learning Curve**: The curve will depict the recall for different test samples in the training set and test set. This will help determine the variance and bias, which will conclude if the model is an overfit or underfit. A high variance is an overfit and a high bias is an underfit. If the model is an overfit, Ridge regularization can be done to improve the performance.