CANCER SURVIVAL PREDICTION REPORT

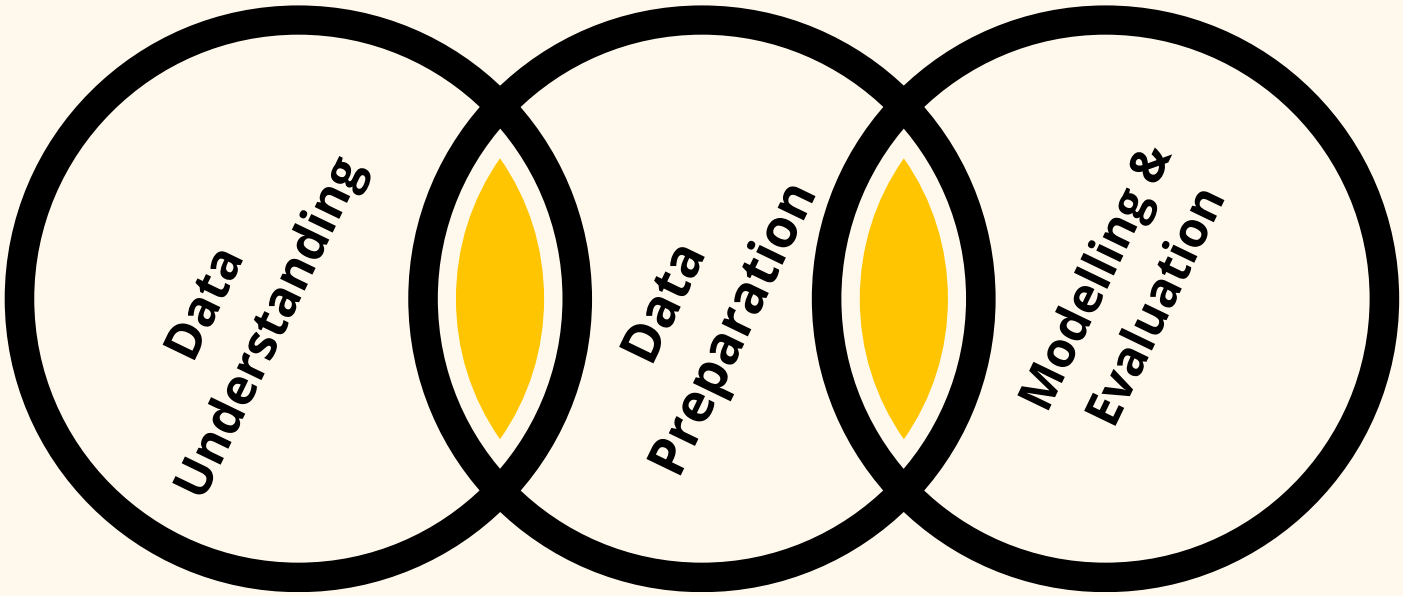
By: Anurag Maheshwari

Problem Statement:

To determine the 7-year survival of prostate cancer patients. A patient survived if they are still alive 7 years after diagnosis. This means that a patient is counted as dead whether or not the death was due to their cancer.

APPROACH:

Following CRSIP-DM framework, this solution compprises mainly 3 steps which are very interconnected and iterative in process.



DATA UNDERSTANDING

Three types of datasets have been provided:

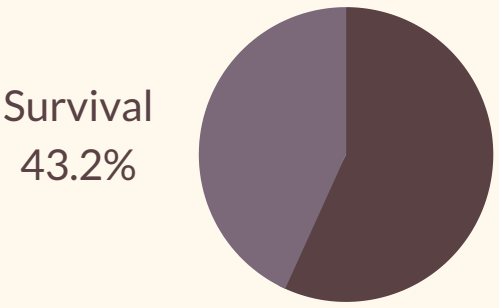
1. Training Set : With dimension of 14385 x 32, this dataset has the details of patients, the state of their cancer at time of diagnosis, and some information about the progression of their disease.

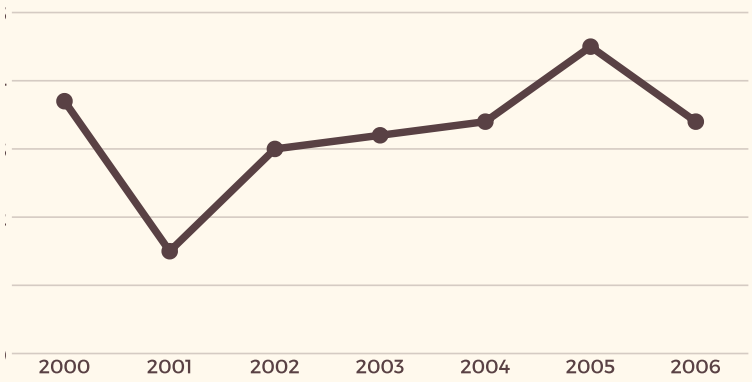
2. Testing Set : With dimension of 1000 x 32, this dataset is provided for the testing purpose. The target variable contain NaN values and should be replaced with the predicted ones.

3. Data Dictionary : This dataset is the metadata for the training data and helps in understanding each attribute in the training data set.

Target Variable Distribution and Trend

43.2% of survival count shows that the dataset is not highly imbalanced.



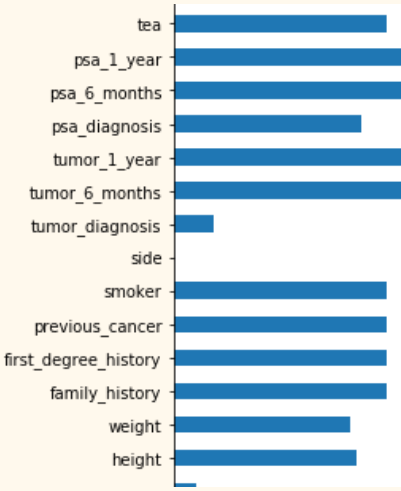


Survival rate trend over Years

DATA PREPARATION

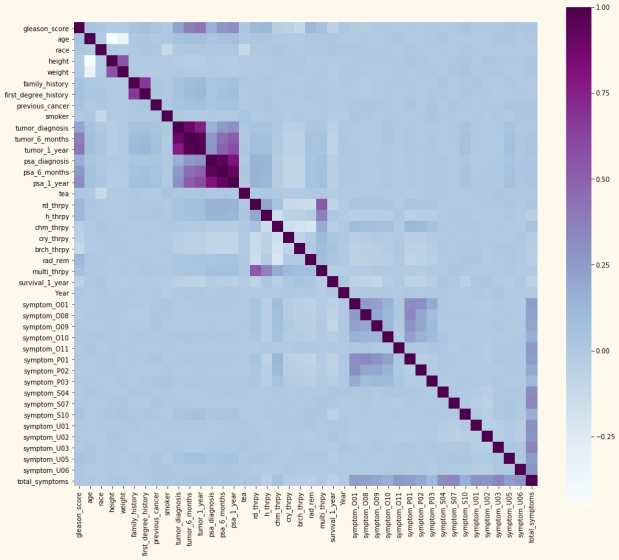
This step mainly divded in two parts:

1. DATA CLEANING



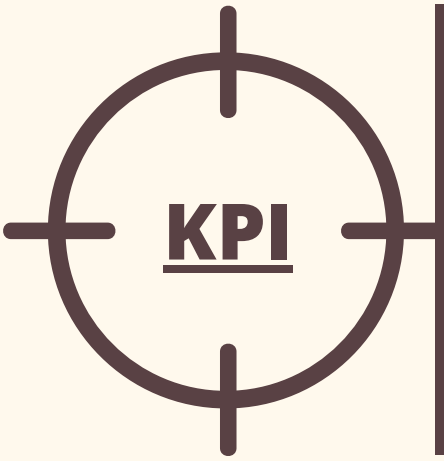
Total 17 attributes constitute missing values and many contain more than 50% missing values. These missing values treated by Multiple Imputation by Chained Equations technique

2. FEATURE ENGINEERING



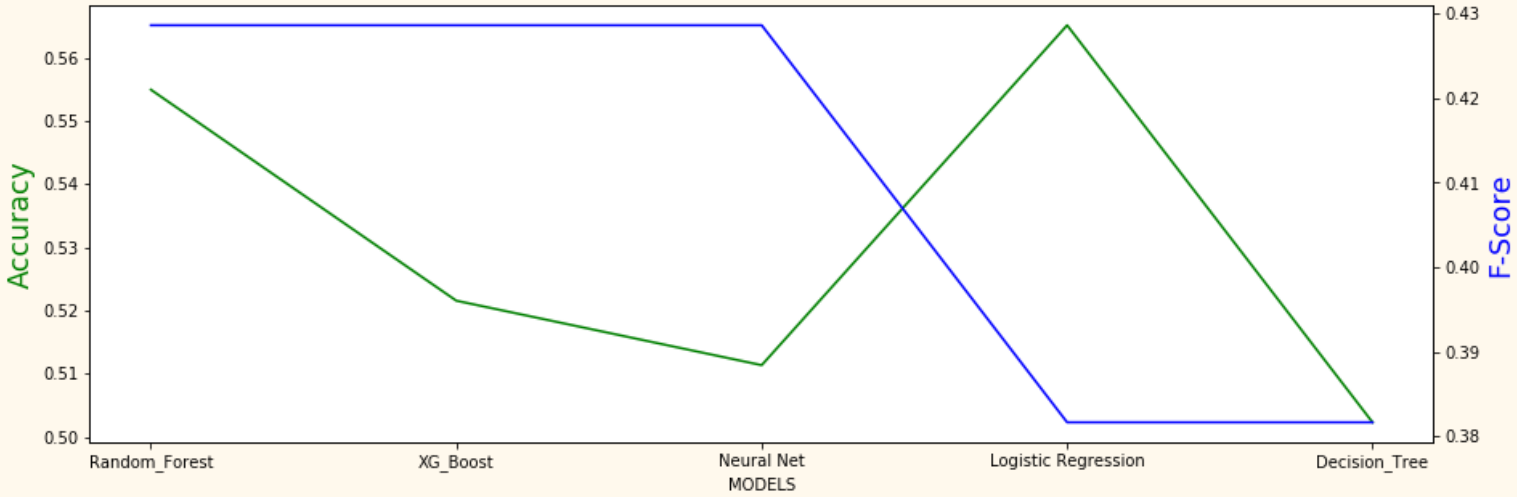
- Using Correlation Matrix, Various correlations in the segemnts like tumor diagnosis, PSA diagnosis, therapies in conjuction, family history etc can be seen and used for the creation of various new attributes like BMI ratio.
- Dummy variables for all the categorical variables have been created.

MODELLING & EVALUATION



As in the problem statement, it is stated that the model will be evaluated on the total number of correct predictions. So giving equal importance to False Positives and False Negatives, We are choosing the KPI as Acuracy & F-Score and will evaluate the models on these two parameters.

Base Performance of Each Model



WINNER: RANDOM FOREST

With highest F1-Score and altogether with very high accuracy we are choosing Random Forest as the first choice for the prediction.