Contents

Chapter 1

Introduction

Prostate cancer (PCa) is one of the most common types of cancer among men. In Ireland, around 1 in 7 men will be diagnosed with prostate cancer in their lifetime [*Prostate cancer* 2021]. However, prostate cancer can be treated with active surveillance and appropriate treatment. It also has one of the highest survival rates among all cancers (>90%) [*Prostate Cancer* 2019; Siegel, Miller, and Jemal 2016]. There are several studies done on diagnosing prostate cancer [*Prostate Cancer Treatment* 2021; Eichler et al. 2006; Madu, and Lu 2010]. The Prostate-Specific Antigen (PSA) level is widely considered to be a key clinical biomarker. An elevated PSA level or an abnormal Digital Rectal Examination (DRE) may be an indicator of prostate cancer [Mohler et al. 2010]. A PSA value of 4 ng/mL or less is considered to be normal; but in fact there is no PSA level below which cancer has not been detected [Mohler et al. 2010]. This issue highlights the need to have more biomarkers which are reliable and can be measured without a biopsy.

However, a definitive diagnosis requires biopsies, where a small sample of tissue is extracted from the prostate and is examined. The pathologist then assigns a primary and secondary Gleason grade to the biopsy specimen. The Pathological Grade of a tumour which is based on the Gleason scoring system, can assume values from 6 to 10, where the former is a low grade cancer and the latter is a high grade cancer. It is measured based on information found during surgery and the laboratory results of the prostate tissue which is removed during surgery [*Prostate Cancer - Stages and Grades* 2021].

This study will focus on developing methods to predict the Pathological Grade of the tumour without requiring surgery. We will also identify suitable biomarkers which may be used to predict the Pathological Grade of the Prostate Cancer.

1.1 Objective

The objective of this work is two-fold. First, we are going to develop a robust framework for predicting the Pathological Grade of the tumour for men diagnosed with prostate cancer. Secondly, we shall also be interpreting the same models to give us information on appropriate biomarkers. This involves using the MSKCC Prostate Cancer dataset to carry out our analysis. This dataset will be discussed in detail in the following sections.

1.2 Prostate Cancer

The prostate is a gland that is a part of the male reproductive system. It lies in front of the rectum and below the bladder and is about the size of a walnut. This gland is responsible for secreting fluid that becomes a part of semen. [*Prostate Cancer Treatment* 2021].

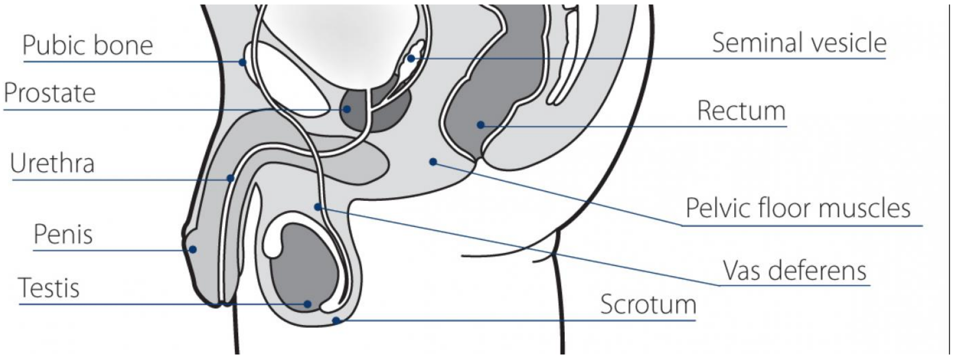


Figure 1.1: Anatomy of the prostate. [*Prostate cancer* 2021]

The type of cancer where malignant cells form in the tissues of the prostate is called prostate cancer. They grow in an abnormal way to form a tumour inside the prostate. In some men, the tumour grows slowly and in others the tumour grows faster and can also spread to other parts of the body.

There are several common symptoms of prostate cancer and are listed below [*Prostate Cancer Treatment* 2021].

- Facing issues when urinating.
- Sudden urges to urinate.
- Frequent urination.
- Discomfort during urination.
- Blood in urine or semen
- Chronic pain in the back, hips or pelvis.
- Shortness of breath, fatigue, fast heartbeat, dizziness or pale skin caused by anemia.

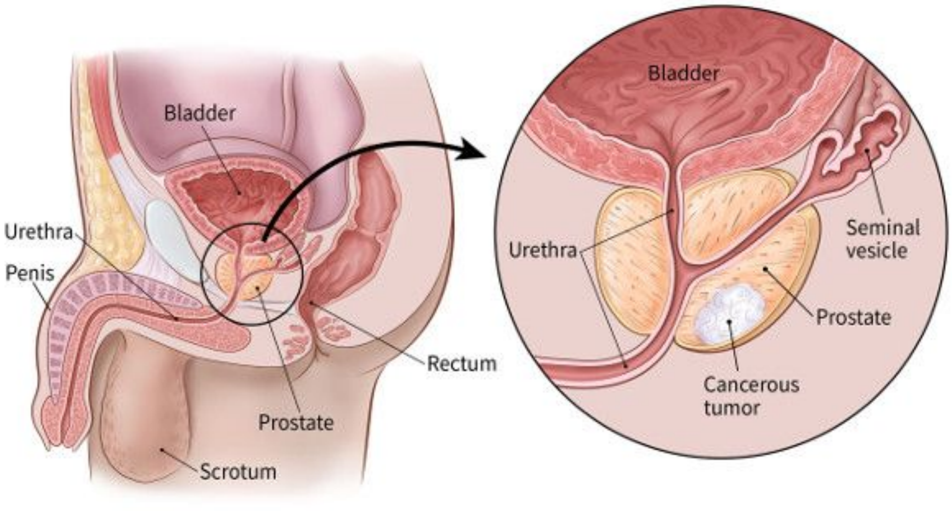


Figure 1.2: Prostate cancer Illustrated.
[*What Is Prostate Cancer?* 2019].

The Pathological Grade of the prostate cancer tumour is determined during surgery using the Gleason Scoring System. The Gleason Scoring System named after Dr. Donald Gleason who developed it in the 1960s, is used to grade cells on a scale from 1 to 5. Grade 1 cells look like normal prostate tissue. Cells higher than 1 and close to 5 are cancerous cells that do not resemble prostate tissue. The pathologist assessing the tissue sample assigns one Gleason grade to the most predominant pattern and a second Gleason grade to the closest following pattern. The two grades are then added together to give a score out of 10 [*Gleason Score and Grade Group* 2021]. However, to be noted is that pathologists never assign scores from 2 to 5 and hence, assigned Gleason scores will always range from 6 to 10. This means that cancers with Gleason score 6 is low grade, 7 is intermediate and higher than 7 is a high grade cancer.

1.3 MSKCC Prostate Cancer Dataset

The MSKCC Prostate Cancer Dataset consists of 232 entries of men diagnosed with prostate cancer. Each sample in the data is described by 39 different features out of which 9 key features are described below.

1. **Type**: The Sample Type, can assume one of two levels, Primary or Metastasis.

2. **PreDxBxPSA**: This indicates the PSA level recorded at diagnosis in ng/mL.

3. **DxAge**: The age of the patient (in years) at diagnosis.

4. **Race**: The race of the patient from one of the following. Black Non-Hispanic, White Non-Hispanic, Black Hispanic, Unknown, Asian, White Hispanic.

5. **BxGG1, BxGG2 and BxGGS**: The primary, secondary and combined Gleason Scores recorded during biopsy.

6. **ClinT_Stage**: This is the Clinical Tumour Stage recorded during biopsy using the TNM Staging system [*Prostate Cancer - Stages and Grades* 2021].

7. **RP_Type**: This refers to the type of Radical Prostatectomy surgery to be conducted, Retropubic Prostatectomy, Salvage Prostatectomy or Laparoscopic Prostatectomy.

8. **Copy-Number Cluster Assignment**: This feature encapsulates information from mRNA data and is grouped into 7 categories.

9. **ERG-fusion aCGH**: This stands for the ETS-related gene (ERG) fusion status determined by copy-number. This is another feature that considers genetic information.

10. **PathGGS**: The Pathological Grade of the tumour. Assumes values from 6 (low-grade cancer) to 10 (very-high grade cancer).

Chapter 2

Machine Learning Methods and Evaluation Metrics

Machine learning (ML) refers to a class of techniques that rely on data to learn patterns and make predictions. We define a model with some parameters and then optimise the parameters by feeding it past data. This is defined as ’training’ the model. The model may be ’predictive’, one that makes predictions in the future, or ’descriptive’, one that describes knowledge present in the data, or both. [Alpaydin 2020].

Machine learning has grown immensely in the last few decades from a prospective technology to ubiquitous commercial use. It has become the preferred method of choice for developing a wide range of solutions, including but not limited to natural language processing, computer vision, robot control, fraud detection systems, etc. [Jordan, and Mitchell 2015]. Most ML methods as built to solve function approximation problems; where the task is represented through a function (e.g., prostate cancer or not prostate cancer) and the challenge is to learn parameters to best approximate solving this problem. Given input and output pairs of data that are labelled as cancer or not, the algorithm must find appropriate parameters via an optimisation procedure. In this age of big data, we have access to huge volumes of data that make this training process easier and the use of ML methods readily available.

The medical field has also greatly benefitted from ML methods. It is not uncommon to have models trained on millions of patient data stored in Electronic Health Records (EHRs) with billions of data points to assist doctors in their medical practices. On the other hand, it is very difficult for a human physician to see more than tens of thousands of patients in their entire career. Since an ML model can learn the patterns in the health trajectories of millions of patients, they are immensely beneficial in prognosis. They can provide insight well beyond the physician’s practical experience using vast amounts of data. Large integrated systems are already using machine learning to identify patients at risk and transfer them to an Intensive Care Unit (ICU) [Escobar et al. 2016]. The risk of disease varies from patient to patient but only the best doctors can accurately diagnose a disease by observing minute changes from a patient’s medical report. If there is an observable pattern that describes the likely onset of a disease, it is an apt problem to be solved by machine learning. These models could help diagnose illnesses that may not occur routinely during the physician's practice. According to a report by the Institute of Medicine, a diagnostic error will occur in the case of nearly every individual patient in their lifetime [McGlynn, McDonald, and Cassel 2015]. In most cases, receiving the right treatment after diagnosis may be a case between life and death.

2.1 Machine Learning Methods

For the purposes of this study, we will focus on four commonly used Machine learning methods. They are,

1. Logistic Regression
2. Random Forest
3. Support Vector Machine
4. Neural Networks

2.1.1 Logistic Regression

Logistic Regression is a method for modelling the probability of a discrete outcome (e.g. Yes/No) given an input variable. The most common logistic regression models a binary outcome, however, multinomial models are capable of modelling multiple outcomes [Edgar, and Manz 2017]. It is a simple and very efficient method for classification problems. Logistic Regression is extensively applied for various problems in the industry due to its simplicity and high interpretability.

Logistic regression will model the probability of an outcome based on the predictor variables. The equation is given below,

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_m x_m$$

where p denotes the probability of the event (e.g., having prostate cancer) and β_i are the regression coefficients associated with the explanatory variables [Sperandei 2014]. The $log(\frac{p}{1-p})$ term is referred to as the logit (log-odds). The variables x_i stand for the explanatory variables used to predict the response. In the case of categorical features, 'dummy' variables may be used in the model. To create dummy variables for a particular feature, one of the levels is used as the reference category and the feature is split into k-1 features (where k is the number of levels that feature can assume). Each dummy variable then assumes 0 or 1 depending on the value of the original categorical feature. We will be creating dummy variables for our analysis while using the logistic regression model.

Logistic regression falls into a larger family of techniques called Generalised Linear Models (GLMs) which can model many different probability distributions. They are called 'linear' as the function of the outcome is modelled using linear predictors, such as the log odds. The following is the structure of the Generalised Linear Model,

$$\text{Random Component}: y_i|x_i \sim \text{Exponential Family Density} \tag{2.1}$$

$$\text{Stochastic Component}: \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \tag{2.2}$$

2.1.2 Random Forest

Random Forest is a machine learning technique that can be applied to classification or regression problems [Breiman 2001]. This ensemble learning technique is comprised of many smaller prediction models called Decision Trees which uses rule-based learning to deliver predictions [James et al. 2013]. Random Forest uses the concept of bagging or bootstrap aggregating. It is where the training data is randomly sampled with replacement, which results in around 66% of the original data [James et al. 2013]. Due to bootstrapping, every decision tree gets a different subset of the training data and is hence unique. Once every decision tree is trained, they cast a vote for classifying an unseen sample and the list of proportions of votes received is treated as a probability vector. The unseen sample is then classified in to the group with the highest probability. The key difference between Random Forests and other bagged methods is that every decision tree being trained gets a randomly selected subset of the predictors referred to as m_{try}. This makes individual tree-building more efficient as we do not use all the predictors but only a small subset of it. For classification problems, Breiman [2001] recommends to keep m_{try} to the square root of the number of predictors.

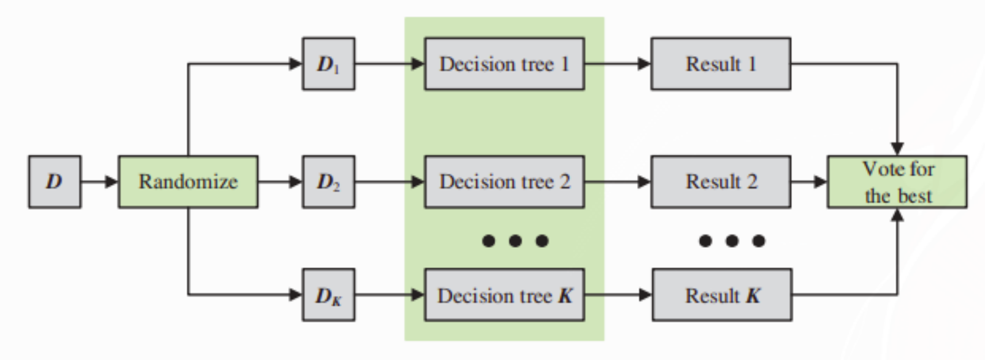


Figure 2.1: Example of a Random Forest. Credits: Ren, Mingchao Li, and S. Han [2019].

The Random Forest, just like its constituent Decision Tree is agnostic to the scale of the data. Hence, there is minimal preparation required before training. Due to its nature as a rule based learning model, it also has feature importance built in. Random Forests have been proven to yield very good results on a range of different classification and regression tasks. Hence, we shall be using this method on our prostate cancer dataset to predict the pathological grade.

2.1.3 Support Vector Machine

Support Vector Machine (SVM) is a technique used in binary classification and regression tasks [James et al. 2013]. It uses the concept of a 'Discriminating Hyperplane' to perform classification. In a p-dimensional space, a hyperplane is a flat subspace of dimension $p-1$ that divides the space into two parts. Given a two-class problem with a number of sample points, this hyperplane is used to discriminate between points on either side of it and is hence called the Discriminating Hyperplane. If the two classes can be perfectly separated by a hyperplane, there could be an infinite number of possible hyperplanes. If the two classes are not separable, it is impossible to build a separator that can perfectly distinguish the two. Hence, SVMs introduce the concept of a 'soft margin'.

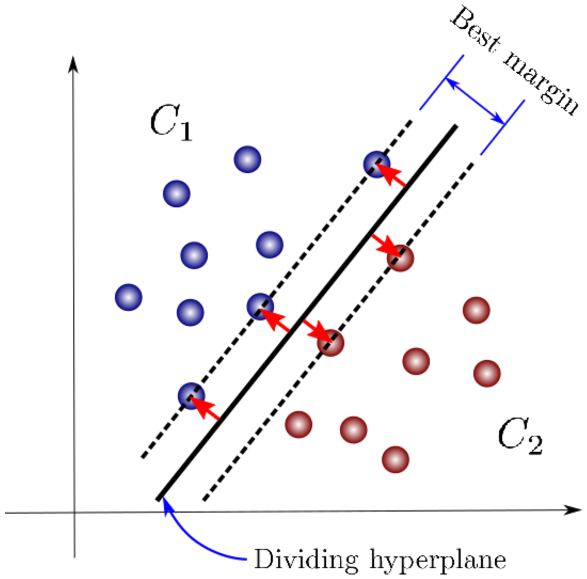


Figure 2.2: Example of an SVM. Credits: Carrasco [2019].

After allowing a soft margin, the SVM allows some samples to be incorrectly classified while trying to get most of the samples on the correct side. This is done by solving the following optimisation problem,

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M} M \tag{2.3}$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1, \tag{2.4}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \tag{2.5}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \leq C \tag{2.6}$$

M is the width of the margin and we wish to make this parameter as large as possible. We do this because the data points nearest to the margin are difficult for the model to categorise and we wish to keep them as far from the margin as possible. The variables $\epsilon_1, \cdots, \epsilon_n$ are called as slack variables which let some samples be misclassified. Equation-2.5 is used to classify the test observation based on the sign of $f(x*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$ [James et al. 2013].

2.1.4 Neural Networks

Neural Networks are a class of algorithms capable of classification or regression. For the scope of this work, we shall be focusing on a particular type of neural network called the Multi-layer Perceptron (MLP) Classifier. The MLP is a kind of feed-forward neural network. It mainly consists of three parts, the input layer, the hidden layer, and the output layer [Abirami, and Chitra 2020]. The input layer receives the training data to be processed.

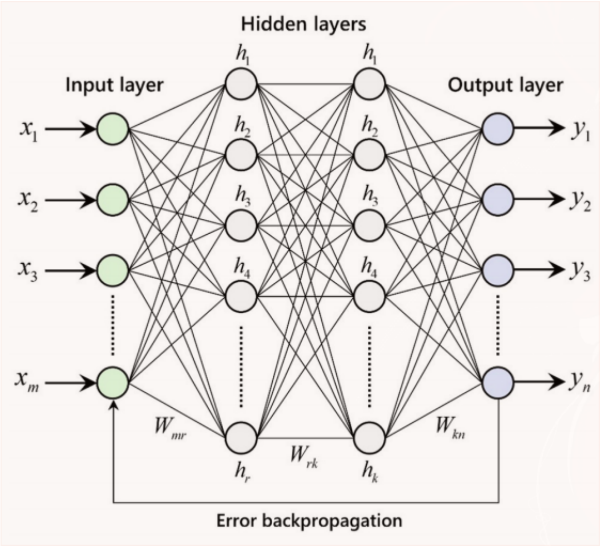


Figure 2.3: Example of an MLP. Credits: Fernández-Cabán, Masters, and Phillips [2018]

The number of input neurons depends on the number of input variables. Various computations are conducted in the hidden layer by introducing non-linear activation functions that transform the data. The simplest network will be to have a single neuron in the hidden layer with no activation function. Finally, the output layer is responsible for conducting the classification or regression task. The number of neurons in the final layer depends on the task at hand; a single neuron is preferred in classification tasks whereas the modeller might choose to have multiple outputs. The choice of activation function is also dependent on the problem to be solved. For a classification problem, the sigmoid activation function (Equation-2.7) may be used. This function outputs a value between 0 and 1, which is used to represent the probability for the test sample belonging to a particular class. The classification is done by selecting a threshold between 0 and 1. For regression problems, a single neuron may be considered with no activation function. In the case of a multi-class classification problem, multiple outputs may be considered alongside the use of the softmax activation function (Equation-2.8).

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.7}$$

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2.8}$$

A strength of the MLP model is being able to set multiple hidden layers between the input and the output layer, which form the heart of the model [Abirami, and Chitra 2020]. Deep Neural Network is a term used to describe networks with a lot of hidden layers. Nowadays, with advances in computing technology, it is possible to train networks with hundreds of layers.

2.2 Resampling Methods

Even in this age of big data, obtaining quality data is expensive. Data is what makes applied machine learning possible and hence each data point must be spent wisely. Resampling methods are used to draw samples from observed data to draw certain conclusions. A single measurement from a statistical learning algorithm is not reliable and we would like to obtain multiple measurements and observe the range. They are a way to effectively use our data to improve the estimate of the population parameter and measure uncertainties of the estimate [Good 2006]. Two popular resampling techniques we will discuss are the Bootstrap and Cross Validation.

2.2.1 The Bootstrap

The Bootstrap is a powerful and widely used statistical tool used for estimating the uncertainty associated with an estimator or statistical learning algorithm [James et al. 2013]. It can be used with a wide range of ML methods, even in those where a measure of the range of variability can be difficult to obtain. This method is used when the target population is unknown and the data is the only available information. The general algorithm is given as follows, Given an initial sample $X_1, \ldots, X_n$, assuming we want to estimate parameter θ,

Algorithm 1 Bootstrapping pseudocode

```
for b=1,...,B do
    Draw a sample $X^{*(b)}$ from $X$with replacement
    Evaluate the estimate $\hat{\theta}^{(b)}$ from $X^{*(b)}$
end for
Deduce the bootstrap estimate of $F_{\hat{\theta}}$ as the empirical distribution of
replicates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$
```

The distribution of these statistics (e.g. The estimate $\hat{\theta}^{(b)}$) is called the Boostrap distribution. This distribution gives us information about the shape, measure of central tendency and spread of the sampling distribution of the statistic.

2.2.2 Cross Validation

Cross Validation (CV) is a very commonly used resampling method to estimate the performance of a statistical learning method [Arlot, and Celisse 2010]. In machine learning, it is used when the data is limited and we need to evaluate model performance on unseen data. We will discuss two types of cross validation procedures used in this study, k-Fold CV and Nested k-Fold CV.

k-Fold Cross Validation

In this type of Cross Validation, the dataset is shuffled and divided into k different parts, or *folds*, of the same size. The first fold is treated as a test set and the method is fitted on the remaining $k-1$ folds. An estimate of performance such as the error rate (ϵ) is measured on the test set and recorded.

This is repeated k times, every time a different fold is used for evaluating the model performance. This process yields k estimates of test error, $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$. Figure-2.4 illustrates the procedure. The k-Fold CV estimate is calculated as follows,

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\epsilon_i \tag{2.9}$$

Nested k-Fold Cross Validation

Using the k-Fold Cross Validation technique is an effective way to evaluate the performance of a machine learning model. However, when we are trying to optimise the set of parameters used to train a machine learning model, we have to run it multiple times on the same data. This is a prime cause for overfitting and may lead to optimistically biased results [Cawley, and Talbot 2010]. Nested k-Fold Cross Validation (Nested CV) helps overcome this issue.

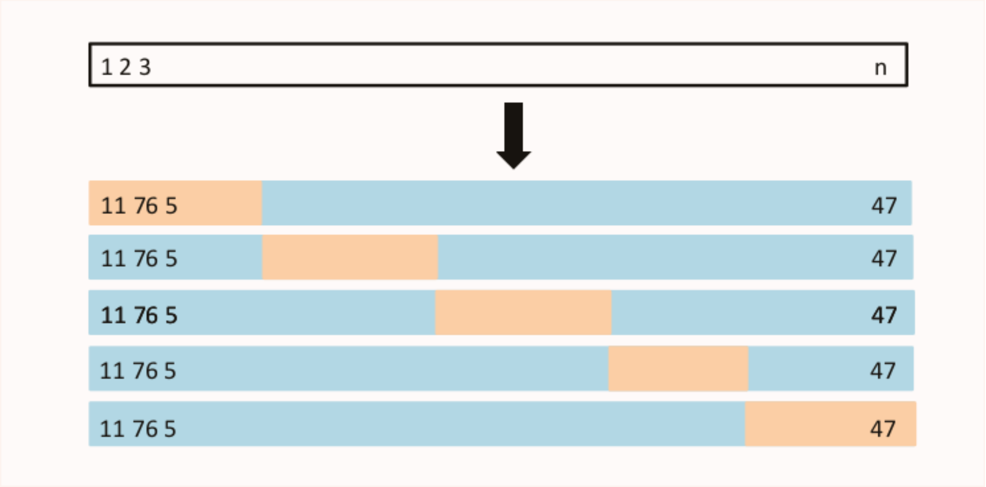


Figure 2.4: 5-Fold CV Illustrated. Credits: [James et al. 2013]

In this technique, model hyperparameter optimisation is treated as part of the model itself and is conducted within a broader k-Fold Cross Validation for evaluating the performance of a model. In other words, hyperparameter optimisation is conducted using a k-Fold Cross Validation procedure which is nested inside another k-Fold Cross Validation procedure conducted for model selection. This is why the technique is called Nested k-Fold Cross Validation or Double Cross Validation and is the preferred technique for model comparison and selection [Cawley, and Talbot 2010].

2.3 Evaluation Metrics

Data preparation, model hyperparameter tuning and even model selection are processes that are guided by the evaluation metric. It acts as a measure of performance which is used to make important decisions while modelling a problem. Evaluation metrics make certain assumptions about what is important in the problem and must be chosen carefully according to it. For e.g., for an imbalanced binary classification problem, using accuracy as the evaluation metric may be highly misleading [Branco, Torgo, and Ribeiro 2015]. Depending on how skewed the data is, simply predicting a single class may give a very high estimate of performance. We will discuss two commonly used performance metrics, Accuracy and Receiver operating characteristic Area Under Curve (ROC AUC).

2.3.1 Accuracy

Perhaps the commonly used performance metric in classification problems, the accuracy score gives the fraction of total number of correct predictions by the total number of predictions. As given in equation,

$$\text{Accuracy} = \frac{\text{Number of Correct Classifications}}{\text{Total Number of Classifications}} \tag{2.10}$$

2.3.2 Receiver Operating Characteristic Area Under Curve

The ROC curve is a diagnostic plot used to visualise a binary classifier's ability to discriminate between two classes. Some models output their test predictions in the form of a probability of belonging to class A or B. By varying the classification threshold, we are able to measure the performance of the model under different thresholds. Each threshold is a point on the plot which forms a diagonal line indicating a model with no discriminatory ability. The results are then plotted to form a curve with the X-axis being the False Positive Rate and the Y-axis being the True Positive Rate.

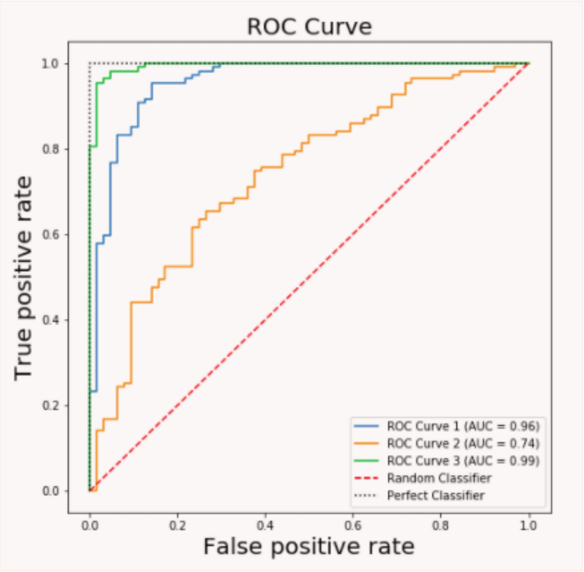


Figure 2.5: Example of the ROC Curve. Credits: A. Kumar [2020]

The Area under the curve can be calculated to give a single metric to summarize the performance of the classifier. The value is called as the ROC AUC.

According to literature [Hajian-Tilaki 2013; Barlow, Mao, and Khushi 2019; R. Kumar, and Indrayan 2011], the ROC AUC is a popularly used metric in the medical field for its proven ability in diagnostic test evaluation. Hence, in this study, this will be the preferred metric to compare models.

2.4 Feature Selection

Feature Selection is the process of choosing a subset of features to be used in predictive modelling. It is helpful in several aspects such as reducing computation requirement, reducing the effect of curse of dimensionality, improving predictive performance etc. [Chandrashekar, and Sahin 2014]. We will discuss two types of Feature Selection strategies, Univariate Feature Selection and Recursive Feature Elimination (RFE).

2.4.1 Univariate Feature Selection

Univariate Feature Selection is used to describe methods that use univariate statistical tests to select features. It is a Filter method, where the relevance of a feature is measured by their correlation with the response variable. In this work, we shall use the ANOVA F-test for numeric features and the Chi-Square Test for categorical features.

ANOVA F-test

The ANOVA F-test is used to test if the means of several groups differ from each other. The formula for the one-way ANOVA F-statistic is given by,

$$F = \frac{\sum_{i=1}^{K} n_i (\bar{Y}_i - \bar{Y})^2/(K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/(N-K)} \tag{2.11}$$

The numerator is referred to as the Explained Variance and the denominator is called the Unexplained Variance. $\bar{Y}_i$ denotes the sample mean in the i^{th} group, n_i is the number of samples in the i^{th} group, $\bar{Y}$ is the mean of the whole data and N and K denote the sample size and number of groups respectively. The F-Statistic follows the F-distribution with degrees of freedom $d_1 = K - 1$ and $d_2 = N - K$ under the null hypothesis.

Chi-Squared Test

The Chi-Squared test is used to test if the observed frequencies for a given categorical variable match the expected frequencies for the second categorical variable. The formula is given by,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2.12}$$

Here, O_i denotes the observed value of the i_{th} observation and E_i refers to the expected value of the i_{th} observation. So in our case, O_i is the categorical predictor and E_i is the response variable. The test statistic here follows a Chi-square distribution.

2.4.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is Wrapper method which uses the feature importance metric of the machine learning method to select features. Initially, a model is fitted with all p predictors and a measure of feature importance is measured that ranks the predictors. The least important feature in this list is then eliminated and the procedure is repeated again with $p-1$ predictors. This procedure can be continued till a certain minimum threshold of predictors are remaining. Amongst all these fitted models, the best performing model is selected as the final model.

Chapter 3

Exploratory Data Analysis and Modelling

3.1 Data Cleaning

In order to create a quality analysis, it is imperative that we wrangle the data appropriately. This includes dealing with missing values, checking for outliers, etc. The MSKCC Prostate Cancer dataset has similar issues which need to be dealt with. Detailed below is the list of steps taken to clean the data and make it usable for our research.

- The features NeoAdjRadTx (Neoadjuvant therapy), MetSite (Site of the metastasis), ChemoTx (Chemotherapy), HormTx (Hormonal therapy) and RadTxType (Radiation) are removed for having a high number of missing values. (>70% Missing values)
- SampleID is dropped since it does not contribute any information.
- The features PathGG1 (Primary Pathological Gleason Score) and PathGG2 (Secondary Pathological Gleason Score) are removed for having a linear dependency with the response variable PathGGS. PreRPPSA (PSA value before RP in ng/mL) is removed for having a high correlation with PreDxBxPSA (PSA value before diagnosis) and because it is measured just before the RP surgery.
- All nomogram features are removed for having many missing values.
- All the following features that are measured after the RP surgery are removed.
  - BCR_FreeTime (Time until Biochemical recurrence)
  - BCR_Event (Recurrence Event)
  - PathStage (Pathological Tumour Stage)
  - Event (Death)

- SurvTime (Overall Survival Time)
- SMS (Surgical margin status)
- ECE (Extra-capsular extension)
- SVI (Seminal vesicle invasion)
- LNI (Lymph node involvement)

As is evident, the data has been reduced considerably from 39 features to 10 features. The reasons for which features were dropped from the dataset here come under 2 categories, missing values and if they were recorded during or after the surgery. This being a relatively small dataset, we would like to drop the fewest number of samples necessary. Some features were dropped because chronologically they are recorded after the point when PathGGS is recorded which makes them redundant.

3.2 Data Exploration

Now that the data has been cleaned, we shall begin the first step of our data analysis by exploring the data. The first step of the modelling process is to understand some important characteristics of the data such as the distributions of the predictors, unusual values within predictors, relationships between predictors, and finally the relationship between each predictor and the response variable. We shall explore each predictor variable in the following sub-sections.

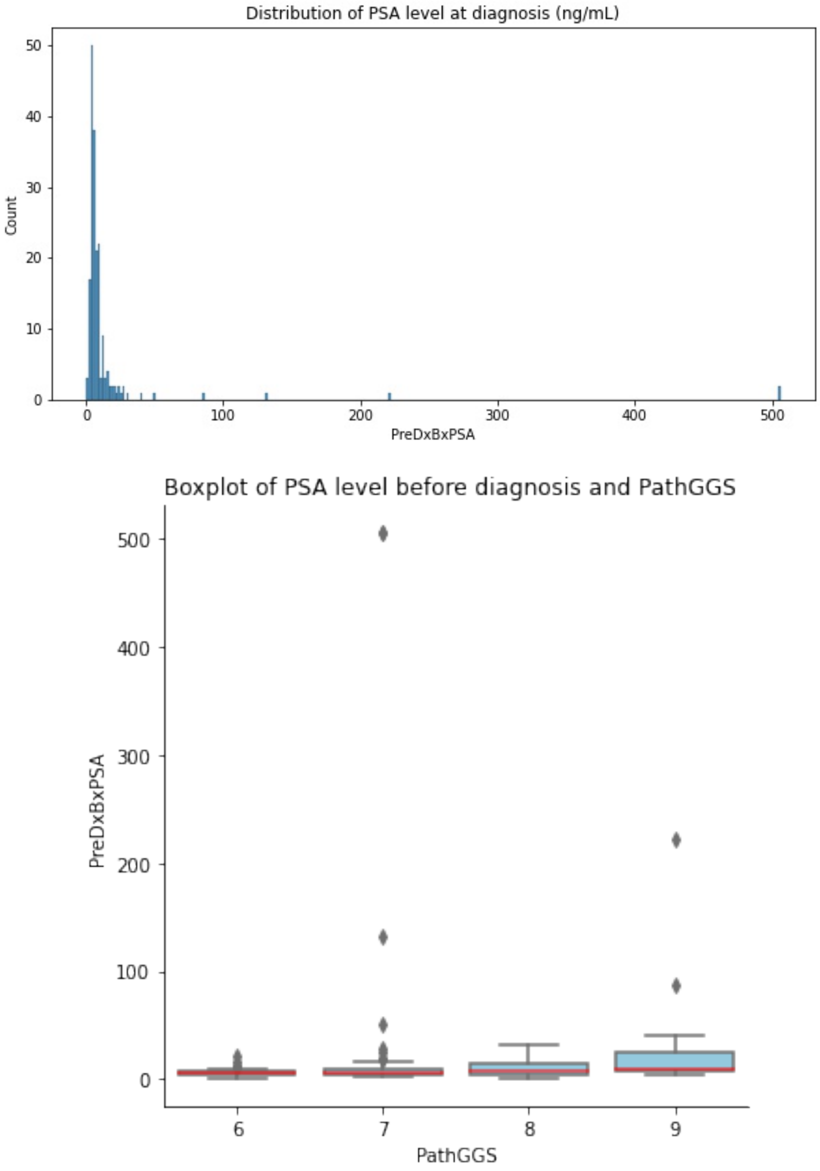
Table 3.1 and Table 3.2 illustrate the summary statistics of the cleaned datatset.

	PreDxBxPSA	**DxAge**	**BxGG1**	**BxGG2**	**PathGGS**
count	190.000000	190.000000	190.000000	190.000000	190.000000
mean	15.414737	58.227971	3.236842	3.373684	6.984211
std	54.468362	6.695967	0.450409	0.584040	0.838721
min	0.200000	37.295800	3.000000	2.000000	6.000000
25%	4.500000	53.365270	3.000000	3.000000	6.000000
50%	6.100000	58.299675	3.000000	3.000000	7.000000
75%	9.200000	63.316210	3.000000	4.000000	7.000000
max	506.000000	72.770890	5.000000	5.000000	9.000000

Table 3.1: Summary statistics of numeric predictors in the dataset.

3.2.1 Exploring Numeric Predictors

Let us take a deep dive into the distribution of the numeric predictors.



Distribution of PSA level at diagnosis (ng/mL)

Boxplot of PSA level before diagnosis and PathGGS

Figure 3.1: PSA level - Histogram and Boxplot with the response.

Feature	Unique Values
Type	’PRIMARY’, ’MET’
Race	’Black Non-Hispanic’, ’White Non-Hispanic’, ’Black Hispanic’, ’Unknown’, ’Asian’, ’White Hispanic’
ClinT_Stage	’T2A’, ’T1C’, ’T2B’, ’T2C’, ’T3A’, ’T3’, ’T3B’, ’T2’, ’T3C’
RP_Type	’RP’, ’SALVRP’, ’LP’
Copy-number Cluster	’1’, ’2’, ’3’, ’4’, ’5’, ’6’, ’flat’
ERG-fusion aCGH	’negative’ ’positive’ ’flat’

Table 3.2: Summary statistics of categorical predictors in the dataset.

PreDxBxPSA (PSA level at diagnosis)

According to studies, a PSA level of over 4 ng/mL is considered to be abnormal and warrants further study [*Prostate-Specific Antigen (PSA) Test* 2021]. We can observe in Figure-3.1 that most of the patients had PSA values under 10 ng/mL. However, here we can see some values much higher than 4, upto 500 ng/mL. The boxplot tells us that the median PSA level steadily increases as the Pathological Grade increases. Hence, we infer that a higher PSA level may indicate an aggressive tumour.

DxAge (Age at diagnosis)

Age appears sufficiently normally distributed. Most men in this dataset are aged around 58, which may indicate that they are at the highest risk of having prostate cancer. From Figure-3.2, we can observe the rising median values of Age as the Pathological Grade increases. We infer that older men tend to be diagnosed with a tumour of a higher grade.

BxGG1 and BxGG2

Here we shall look at the two grades that are assigned to a patient and that represent the Gleason Score during biopsy. Prostate tumours are often constituted of cancerous cells of different grades. A primary grade is assigned to describe the cells of the largest area of the tumour and a secondary grade is assigned to the second largest area [*Gleason Score* n.d.] For the primary biopsy gleason grade, most men have been assigned a 3 and the higher the grade, the lower the number of patients. Regarding the secondary biopsy gleason grade, there is a similar trend but with a small minority having been assigned a 2. When comparing the Biopsy Gleason Grades with Pathological Gleason Grades, we observe a general trend of positive correlation. As the Biopsy Gleason Grade increases, so does the Pathological Gleason Grade.

3.2.2 Exploring Categorical Predictors

Here we shall explore the predictors Sample Type, Race, Clinical Tumour Stage, Type of Radical Prostatectomy, Copy-Number Cluster Assignment and the ERG fusion status

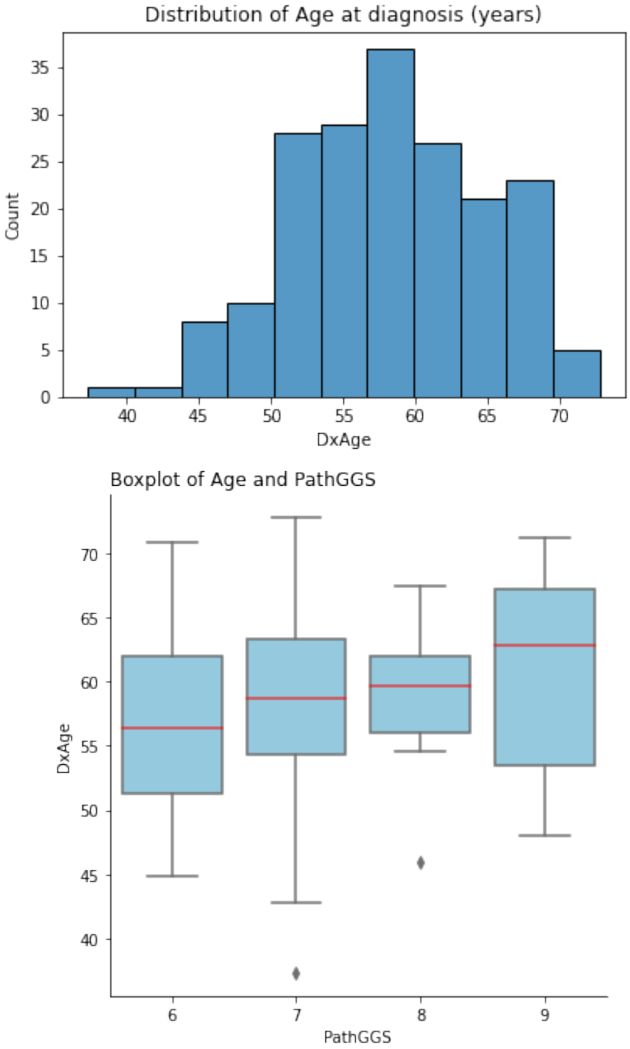


Figure 3.2: Age -Histogram and Boxplot with the response.

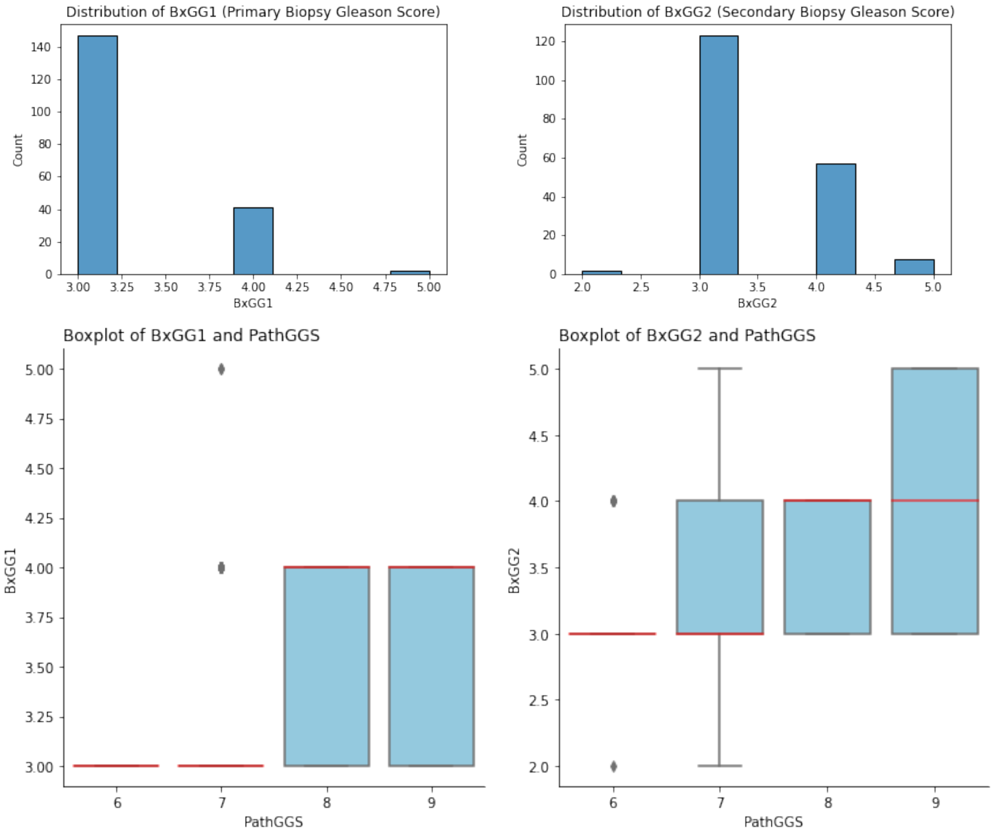


Figure 3.3: Histogram of BxGG1 and BxGG2

determined by copy-number. We shall plot the distribution of the variables.

Sample Type

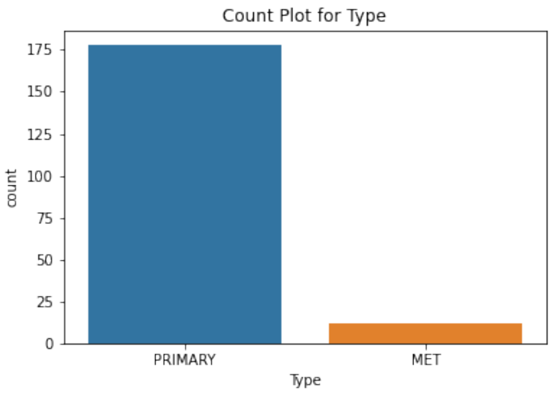


Figure 3.4: Count Plot of Sample Type

The predictor Sample Type describes whether the tumour is the first tumour in the body (Primary) or a cancer that spread to other parts of the body and formed a secondary tumour (Metastasis) [*NCI Dictionary of Cancer Terms* n.d.] Here most samples are of Primary type and the minority is Metastasis. Due to the small size of the dataset, we have not explored techniques to balance the samples by under or over sampling.

Race

This predictor tells us the race of the patient. We see a majority of White non-Hispanic patients followed by Black non-Hispanic and others.

Clinical Tumour Stage

In this barplot, we can observe the distribution of Clinical Tumour Stage of the patients. There are 4 main stages of cancer size in prostate cancer and each stage is divided into multiple subdivisions, e.g. T1A, T2B etc. [*TNM Staging* 2019]. T1C is the most common clinical stage followed by the others.

Type of Radical Prostatectomy

Radical Prostatectomy refers to the surgery conducted to remove the prostate and the surrounding tissues. The three types of Radical Prostatectomy here are Retropubic Prostatectomy, Salvage Prostatectomy and Laparoscopic Prostatectomy. Most surgeries conducted are of type Retropubic Prostatectomy. This may due to the fact that most tumours here are of Primary type, as observed for the predictor Sample Type.

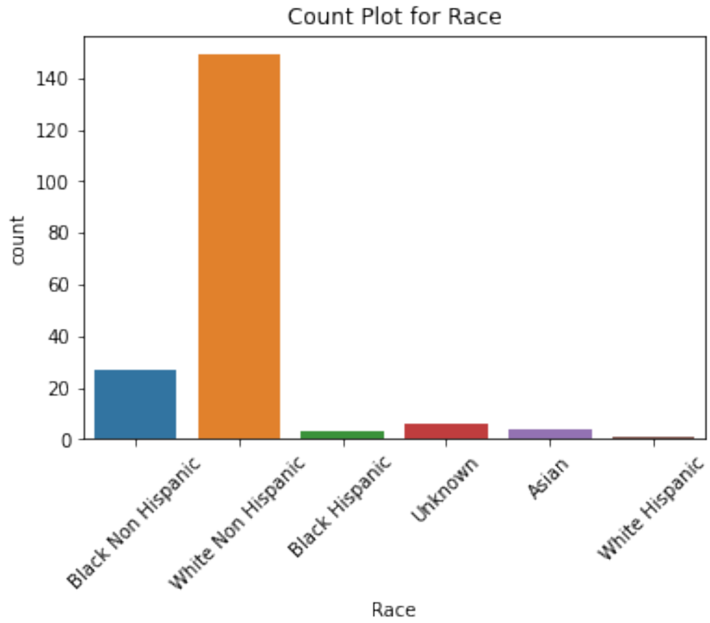


Figure 3.5: Count Plot of Patient Race

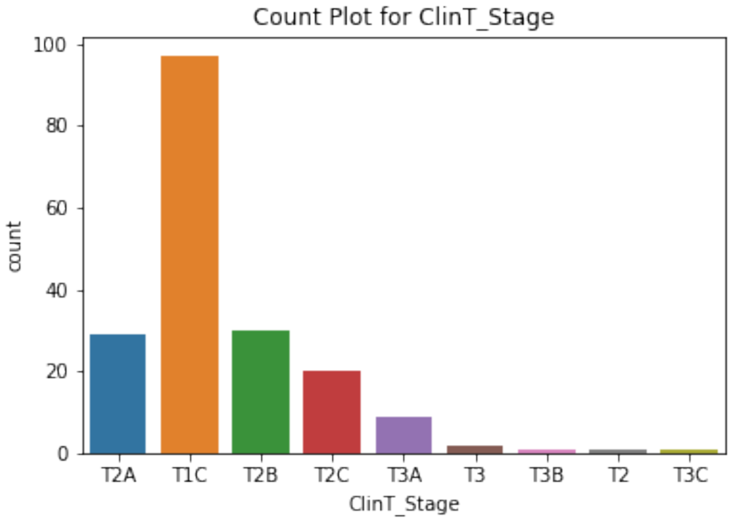


Figure 3.6: Count Plot of Clinical Tumour Stage

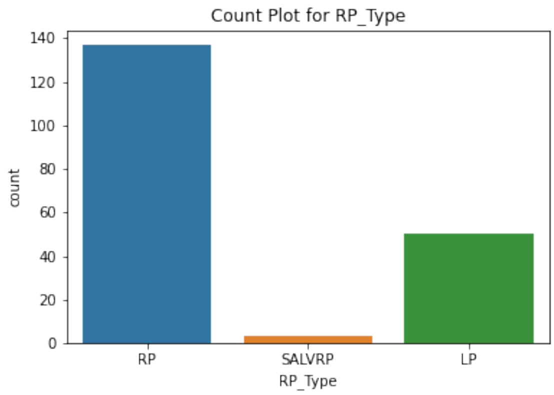


Figure 3.7: Count Plot of Radical Prostatectomy Type

Copy-Number Cluster Assignment

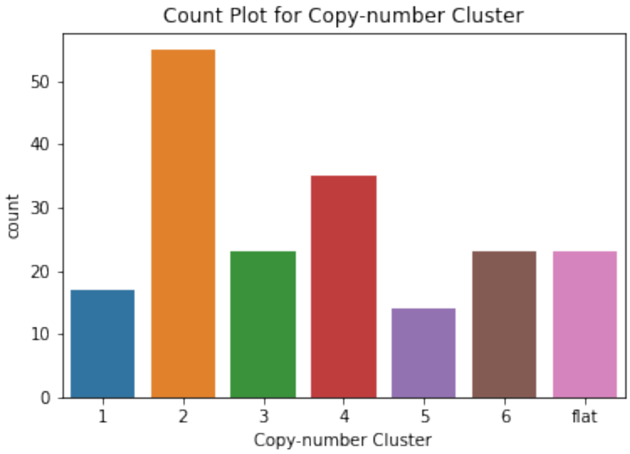


Figure 3.8: Count Plot of Copy Number Cluster Assignment

The Copy-Number Cluster Assignment refers to 7 different categories into which the Copy-number as been grouped into. Copy-number variation refers to a phenomenon observed in the genome. We see a uniform spread of all cluster assignments.

ERG fusion status determined by copy-number

ERG (ETS-related gene) is an oncogene that has become highly associated with prostate cancer in the last decade [Adamo, and Ladomery 2015]. This predictor describes the

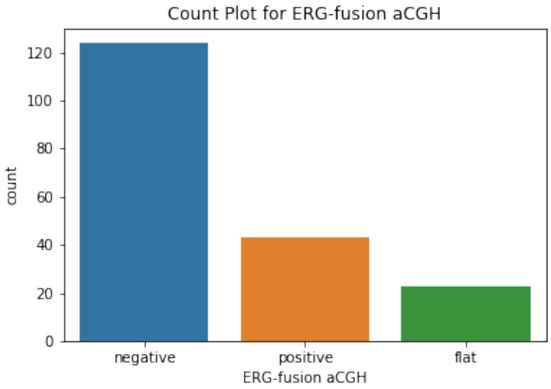


Figure 3.9: Count Plot of ERG fusion status

fusion status determined by the copy number. Most samples are negative followed by positive, and finally the least samples are flat.

3.2.3 Exploring the Response

The response variable PathGGS has 4 levels. Starting from a gleason score of 6 which indicates an low grade tumour to a gleason score of 9 which is a high grade tumour. Due to the apparent imbalance in classes and support from literature [*Understanding Your Pathology Report: Prostate Cancer* 2017], we shall be combining the classes '7', '8' and '9' into '7+' which indicates an aggressive disease. Henceforth, our new binary classification task is to predict an indolent cancer (Gleason Score 6) or an aggressive cancer (Gleason Score 7+).

3.3 Feature Selection

Feature selection techniques are commonly used as a pre-processsing step to model building. A model with fewer predictors may be more interpretable and less costly to build. In this chapter we will explore two feature selection strategies, Univariate Feature Selection and Recursive Feature Elimination (RFE).

3.3.1 Univariate Feature Selection

We shall be using 2 tests to filter predictors to include in our models. For Continuous Features we have used the ANOVA F-value for selecting features. For categorical features we have used the Chi-Squared Test. If the test statistic is deemed statistically significant (p-value <= 0.05), the predictors are included in the model.

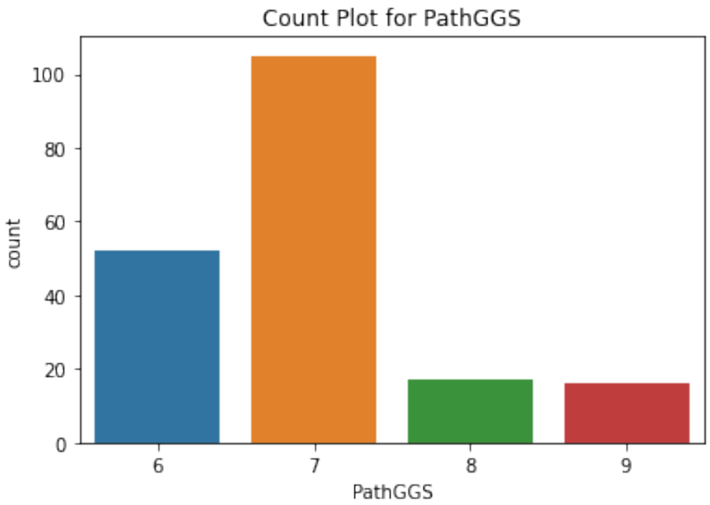


Figure 3.10: Count Plot of Pathological Grade Gleason Score

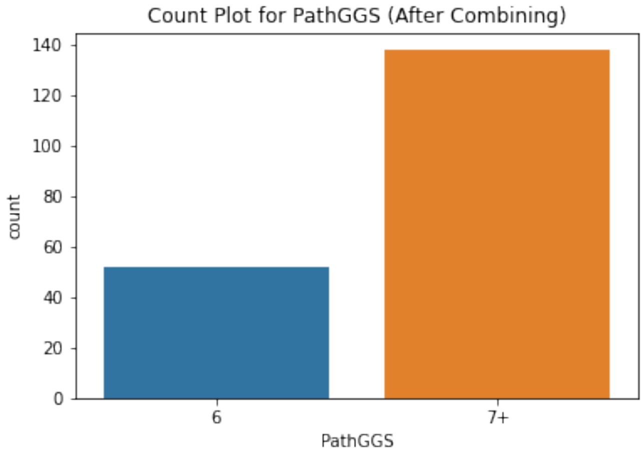


Figure 3.11: Count Plot of Pathological Grade Gleason Score (After combining)

Numeric Predictors

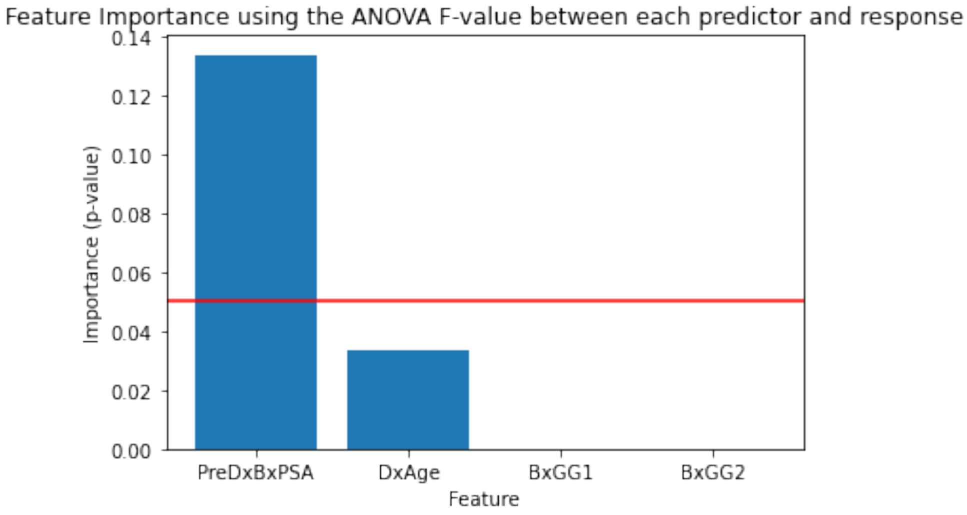


Figure 3.12: Barplot of ANOVA F-test p-values for each numeric predictor. The red line denotes the cut-off at 5%.

Out of the 4 numeric predictors, 3 are selected. Age, Biopsy Gleason Grade 1 and Biopsy Gleason Grade 2 are deemed to be statistically significant and shall be included in our model.

Categorical Predictors

Given in Figure-3.13, only one predictor is deemed statistically significant, Clinical Tumour Stage. The other predictors all show p-values higher than 0.05 and are not included.

Now we shall select an appropriate feature selection strategy by comparing the two. 3 models are used to compare the classification performance without and with the corresponding feature selection strategy applied. Given in Figure-3.14, are the relative performance of the ML models with and without Univariate Feature Selection applied to the data.

With Univariate Feature Selection, we observe a downward trend in the difference in performance before and after feature selection. Only in the SVM, do we see that the performances are similar but the variability has increased.

3.3.2 Recursive Feature Elimination

As discussed in Section-2, RFE is a Wrapper method that uses the feature importance measure of a machine learning model to rank predictors and iteratively eliminate them.

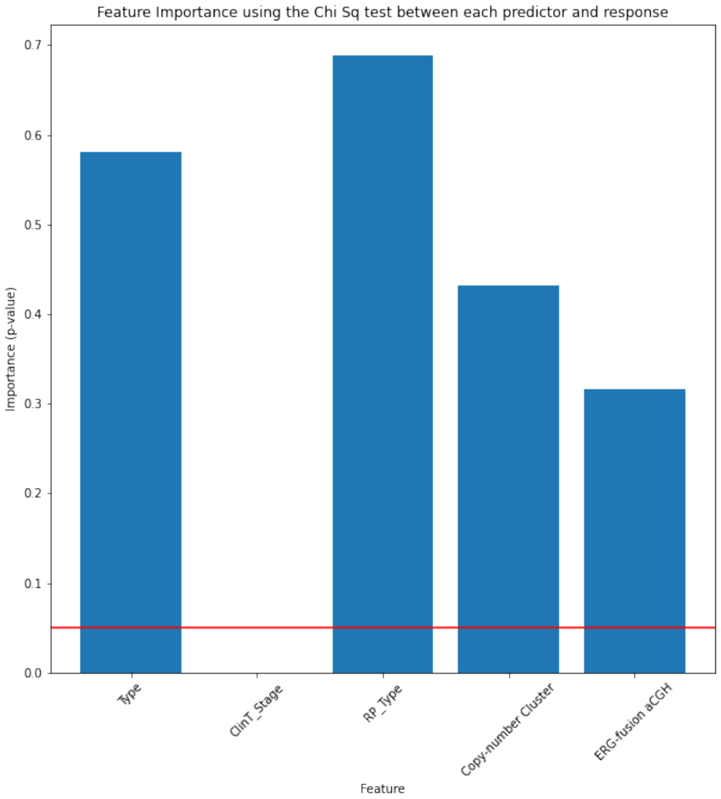


Figure 3.13: Barplot of the Chi-Squared test p-values for each categorical predictor. The red line denotes the cut-off at 5%.

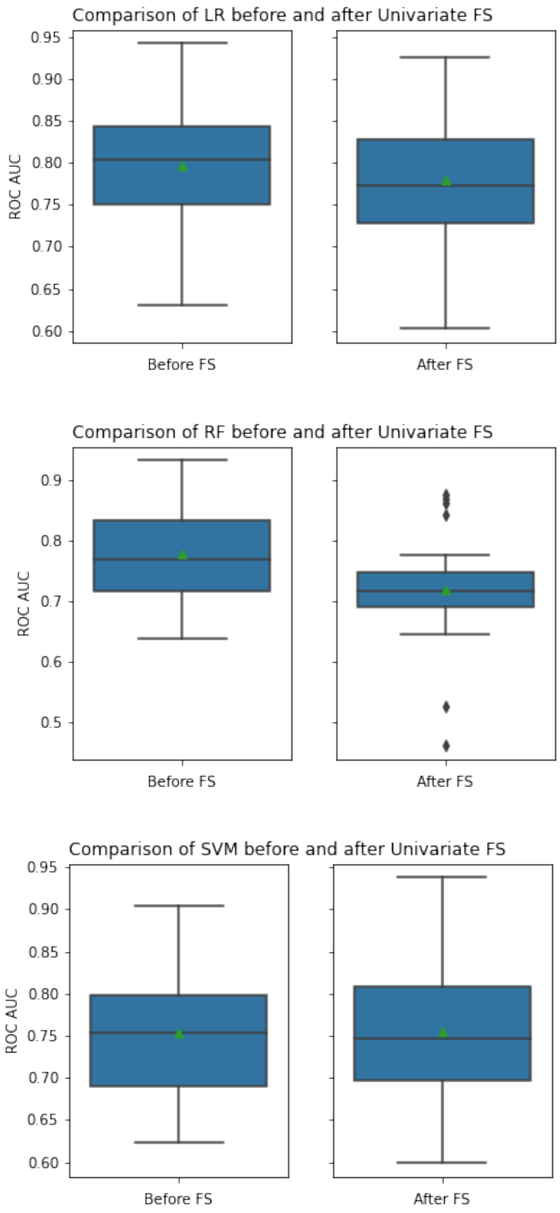


Figure 3.14: LR, RF and SVM compared with and without Univariate Feature Selection.

Here, we shall be using RFE with 3 of our models, namely, Logistic Regression, Random Forest and SVM. RFE is conducted with 5-fold Cross Validation to optimise the number of features. Then, the model is fitted on this optimal subset of features. Figure-3.15 indicate the relative performance of the ML models with and without Recursive Feature Elimination applied to the data.

In Figure-3.15, we notice an upward trend in performance. There is a slight decrease in variability and an improvement in performance in all models. Due to superior performance, we shall select RFE as the feature selection technique of choice.

3.4 Modelling

After exploring our data and relevant feature selection methods, our data is now ready for modelling. We shall be evaluating a range of different machine learning methods namely, Logistic Regression, Random Forest, Support Vector Machine and Neural Networks. All the models discussed here are fit using Nested Cross Validation, a resampling framework that allows hyperparameter tuning as part of the model itself. The data is transformed according to the requirements of the model and the resampling procedure is performed to yield an unbiased estimate of model performance.

3.4.1 Logistic Regression

To fit the requirements of Logistic Regression, the data is dummy encoded (one-hot encoded) so the categorical variables are split into multiple dummy variables. Each dummy variable now contains only 2 levels, 0 or 1.

As discussed in Section 2, Nested Cross Validation has two loops, the inner and outer loop. The inner loop has been configured with a pipeline that performs Recursive Feature Elimination and passes the data with the best features to the Logistic Regression model. The outer loop is configured to run 5-Fold Cross validation and the inner loop runs 3-Fold Cross Validation. Nested Cross validation is then repeated 5 times and the results are plotted in Figure-3.16. LR gives a mean ROC AUC of 0.7829.

3.4.2 Random Forest

Random Forests being a tree-based method, do not require any data scaling. However, the categorical variables are encoded to have numbers instead of text labels.

We set up the inner loop by performing Recursive Feature Elimination followed by a hyperparameter search. The parameter space that was searched through has been listed below.

1. Number of trees: 50, 100 and 200
2. Max Depth: 1, 2 and 3

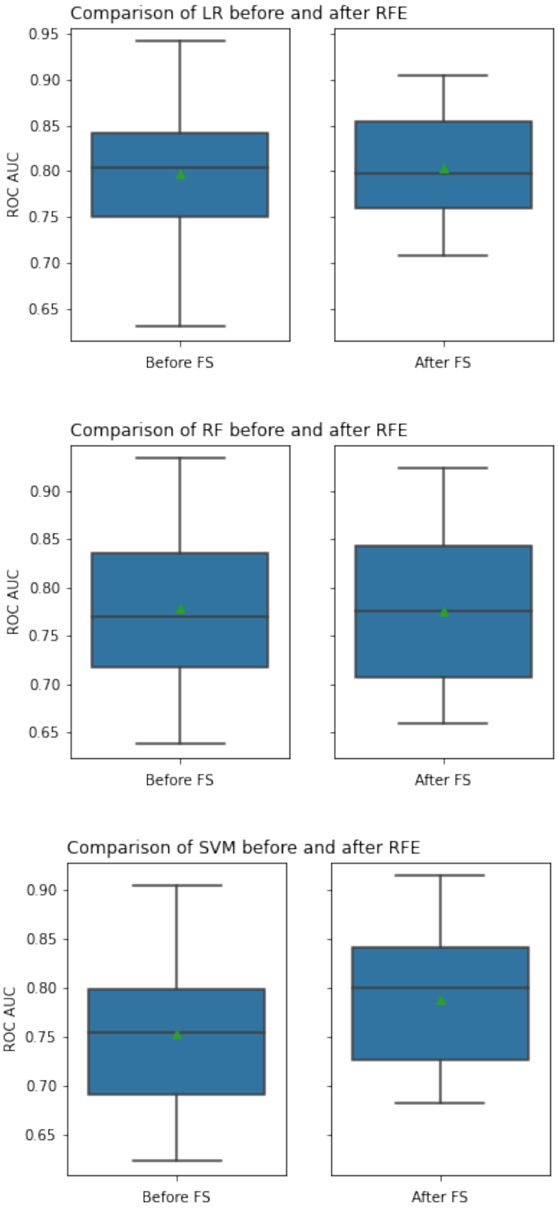


Figure 3.15: LR, RF and SVM compared with and without Recursive Feature Elimination.

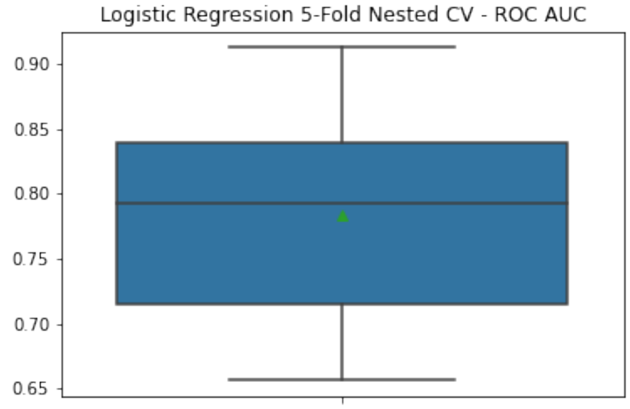


Figure 3.16: Performance of LR using Nested Cross Validation.

3. Max Num of Features: 3, 4, 5 and 6

The outer loop is configured with 5-Fold Cross Validation and the inner loop runs 3-Fold Cross Validation. The procedure is not repeated multiple times due to performance constraints. The results are displayed in Figure-3.17. RF gives a mean ROC AUC of 0.7640.

3.4.3 Support Vector Machine

Support Vector Machines rely on a distance metric to classify data points and hence require the data to be scaled. For the numeric predictors, we subtract the mean value and scale to unit variance. We also transform each predictor to be between 0 and 1 so they are all on the same scale. The categorical variables are dummy encoded with each new predictor having 0 or 1 as possible levels.

The inner loop consists of Recursive Feature Elimination followed by a hyperparameter search. The SVM was configured with a linear kernel and hence only the regularisation parameter was searched with possible values being 0.1, 1, 10 or 100. The outer loop is configured with 5-Fold Cross Validation and the inner loop runs 3-Fold Cross Validation. The procedure is repeated 5 times and the results are displayed in Figure-3.18. SVM gives a mean ROC AUC of 0.7416.

3.4.4 Neural Networks

Similar to Support Vector Machines, Neural Networks also require the data to be scaled and the same data transformation pipeline is applied.

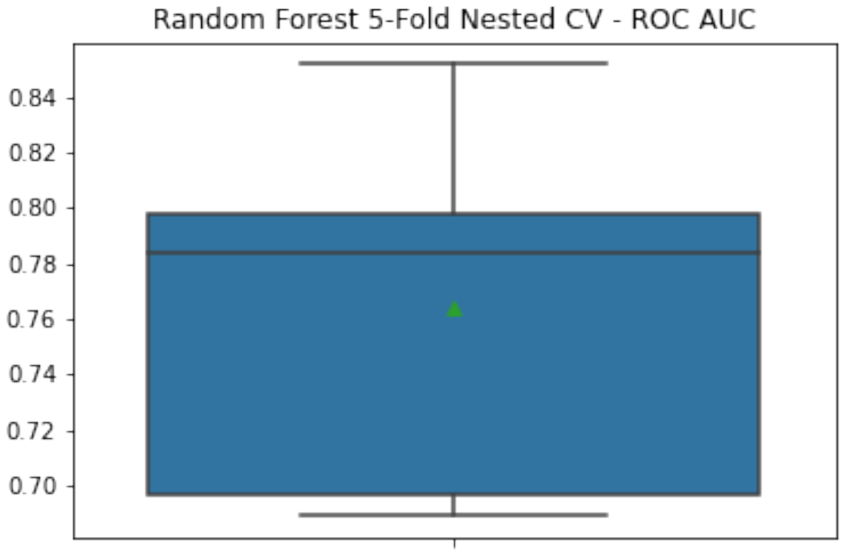


Figure 3.17: Performance of RF using Nested Cross Validation.

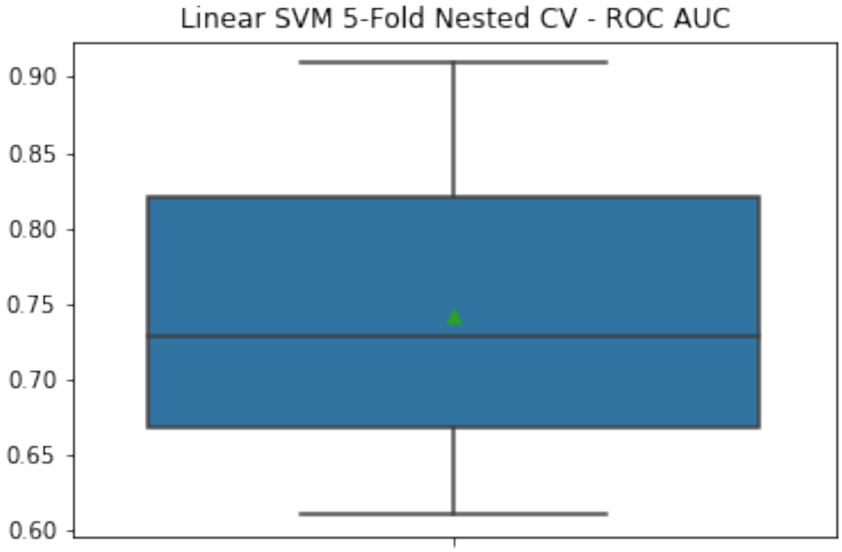


Figure 3.18: Performance of SVM using Nested Cross Validation.

A Multi-layer perceptron Network has two important hyperparameters to be optimised. The size and the learning rate. The following parameter space was searched during the hyperparameter search

1. Hidden Layer Sizes: All possible combinations from one hidden layer to 4 hidden layers with the number of neurons 10, 20, 50 and 70.

2. Learning rate (alpha): 0.0001, 0.001 and 0.01.

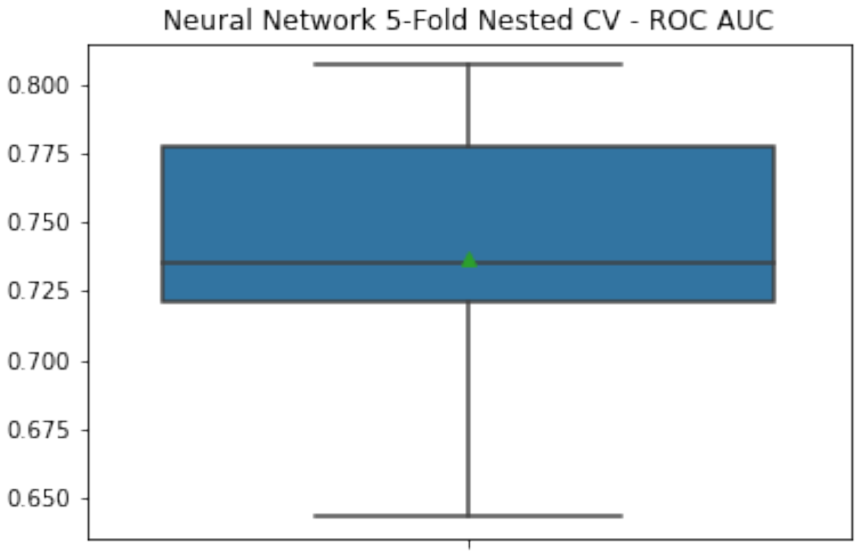


Figure 3.19: Performance of NN using Nested Cross Validation.

The inner loop consists of a step for tuning the Multi-layer Perceptron without RFE. The outer loop is configured with 5-Fold Cross Validation and the inner loop runs 3-Fold Cross Validation. The procedure is not repeated due to performance constraints. The results have been displayed in Figure-3.19. The MLP model gives a mean ROC AUC of 0.7370.

Chapter 4

Results

In the previous chapter we started off by cleaning the data. Then, we explored the data to understand its key characteristics. We also weighed two feature selection methods, namely Univariate Feature Selection and Recursive Feature Elimination. All of these steps culminated in the final step, modelling. This is where we evaluated all our Machine Learning Methods and got measures of their performance. In this chapter we shall look at the results from the modelling stage and interpret them.

4.1 Model Selection

After the modelling stage, it is important that we are able to select an appropriate model for our purpose. Due to the Nested Cross Validation framework used for evaluating each model, we were able to obtain an unbiased estimate of model performance by not allowing any single model to overfit and train it on different sections of our data. The results obtained from this model comparison are displayed in Figure-4.1 and Figure-4.2.

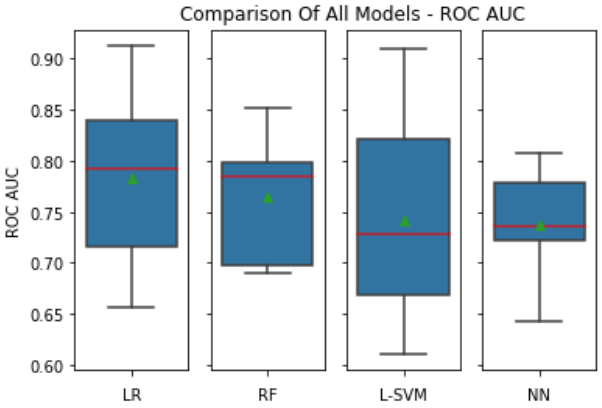


Figure 4.1: Performance of 4 Machine Learning Models (ROC AUC).

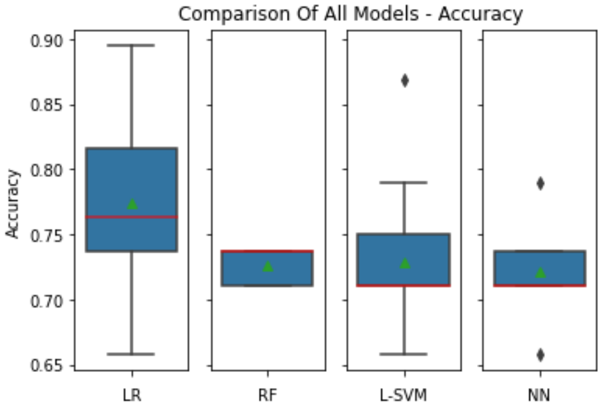


Figure 4.2: Performance of 4 Machine Learning Models (Accuracy).

We observe in both figures that Logistic Regression has the best performance closely followed by Random Forest. It is worth noting that interestingly, the most interpretable models have scored the highest. We shall select the two best performing models, Logistic Regression and Random Forest and interpret them.

4.1.1 Logistic Regression

In order to run this model on the entire dataset, we must repeat the inner loop of nested cross validation. We fit a pipeline of RFE and Logistic Regression using 5-Fold Cross Validation to find the features to be selected for the final model. Illustrated in Figure-4.3 is a plot used to identify the optimal number of features to be included in the final model. The data used to train this model has been dummy encoded and hence has a larger number of features than the original data.

The best 9 features identified are as follows,

1. BxGG1
2. BxGG2
3. Race_Black Hispanic
4. ClinT_Stage_T2A
5. RP_Type_SALVRP
6. Copy-number Cluster_2
7. Copy-number Cluster_5
8. Copy-number Cluster_6
9. Copy-number Cluster_flat

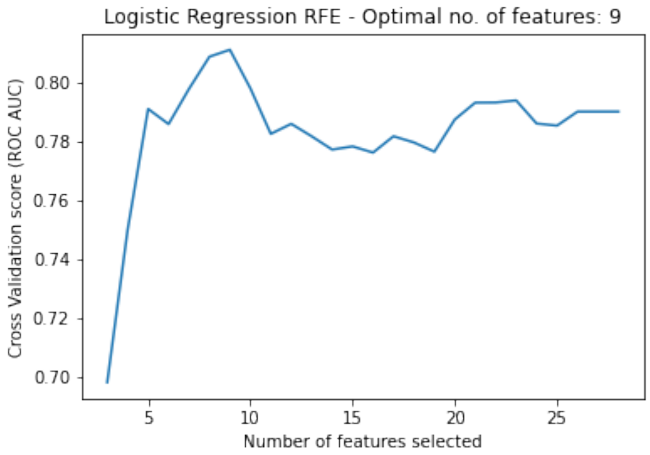


Figure 4.3: Logistic Regression Performance during RFE.

The next step is to fit a Logistic Regression model with the features above. The fit model is given in Table-4.1.1.

We shall interpret the coefficients deemed to be statistically significant (p-value < 0.05).

BxGG2: For a 1-unit increase in Secondary Biopsy Gleason Score, the chances of a patient having an aggressive disease (Tumour of Pathological Grade Gleason Score 7 or higher) increases by (OR = $e^{1.4729}$ = 4.361) 336%. The effect is significant (p-value = 0.0026)

ClinT_StageT2A: Patients with Clinical State T2A are more likely than patients with Clinical stage T1C to have an aggressive disease by (OR = $e^{(}1.8686)$ = 6.479) 548%. The effect is significant (p-value = 0.0086).

Copy Number Cluster 2: Patients with a Copy number cluster assignment 2 are less likely than patients with a Copy number cluster assignment 1 to have an aggressive disease by (OR = $e^{-1.6048}$ = 0.2009) 80%. The effect is significant (p-value = 0.0467).

4.1.2 Random Forest

We rerun the inner procedure of Nested Cross Validation for Random Forest as well. We fit a pipeline of RFE and Random Forest using 5-Fold Cross Validation to get a subset of optimal features. Figure-4.4 describes the performance of Random Forest during RFE, referring which we select 7 features. The optimal features for Random Forest are,

1. PreDxBxPSA
2. DxAge
3. BxGG1

Predictor	**Coefficient Estimate**	**Pr(>|z|)**
(Intercept)	-168.8887	0.98791
RaceBlack Hispanic	50	0.99689
RaceBlack Non Hispanic	28.9185	0.99054
RaceUnknown	29.7599	0.99026
RaceWhite Hispanic	4.2356	0.99981
RaceWhite Non Hispanic	28.6691	0.99062
BxGG1	45.3647	0.98806
BxGG2	1.4729	0.0026
ClinT_StageT2	-25.4804	0.99887
ClinT_StageT2A	1.8686	0.00868
ClinT_StageT2B	0.458	0.41376
ClinT_StageT2C	-0.6582	0.44147
ClinT_StageT3	-3.2336	0.99985
ClinT_StageT3A	12.5506	0.99382
ClinT_StageT3B	-24.9598	0.99889
ClinT_StageT3C	-45.1233	0.99803
RP_TypeRP	0.5206	0.30029
RP_TypeSALVRP	17.8459	0.99831
Copy.number.Cluster2	-1.6048	0.04677
Copy.number.Cluster3	-0.7102	0.43105
Copy.number.Cluster4	-0.4042	0.64859
Copy.number.Cluster5	16.5652	0.99621
Copy.number.Cluster6	0.9438	0.39751
Copy.number.Clusterflat	-1.5976	0.07073
Residual deviance: 141.88 on 166 degrees of freedom		

Table 4.1: Logistic Regression Model Coefficients

4. BxGG2
5. ClinT_Stage
6. Copy-number Cluster
7. ERG-fusion aCGH

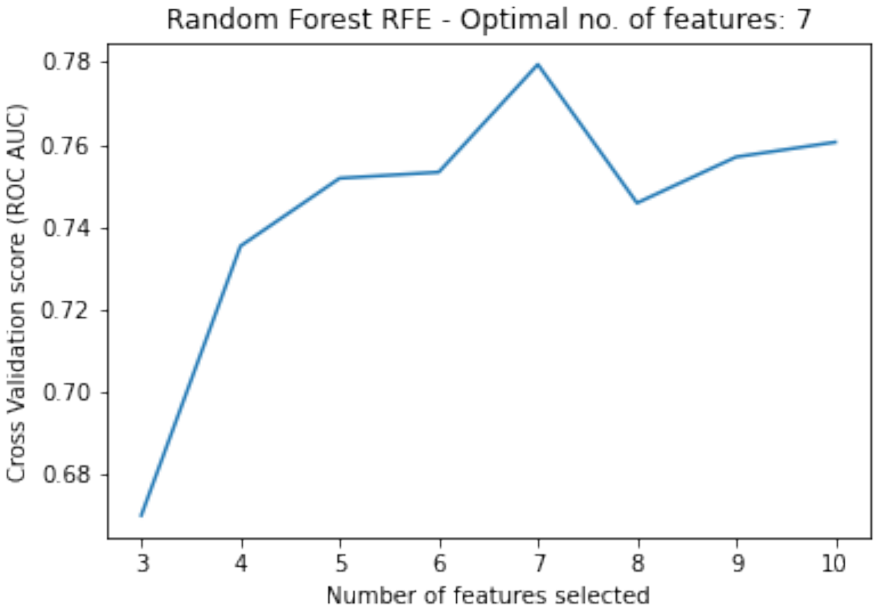


Figure 4.4: Random Forest Performance during RFE

Now that the optimal features have been found by RFE, we shall fit the Random Forest with the parameters optimised during this Cross Validation procedure and fit it on the entire dataset. Figure-4.5 plots the relative importance measures. This importance is calculated internally during the fitting of the Random Forest and is called the Gini Importance or the Mean Decrease in Impurity.

4.2 Discussion

In the previous section we identified Logistic Regression and Random Forest as the two best models. For the Logistic Regression model, we selected 9 optimal features which included dummy variables. We selected the 6 original features that were represented by these dummy variables, fit the LR model and interpreted the model coefficients. We observe that three of these coefficients are statistically significant. The Random Forest

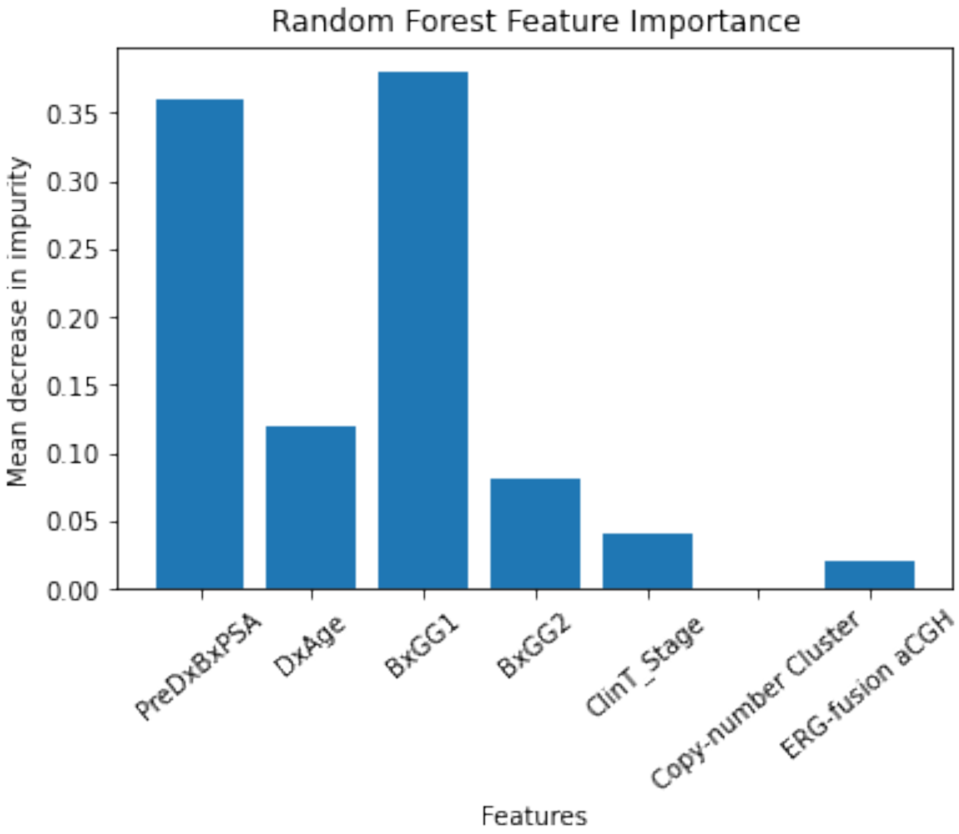


Figure 4.5: Random Forest Feature Importance

model was also evaluated in the same way as above, where RFE was run to select the best 7 features. The model was then fitted using these features on the entire dataset and the feature importance plots were visualised. The PSA level at diagnosis and Primary Biopsy Gleason Score are the two most important features.

The performance measures of Logistic Regression and Random Forest are found to be very competitive. Since we need to select a model for medical purposes, we will select the one with a smaller subset of features and lesser computational power required. Logistic Regression fulfills both these requirements and will be the model of choice.

Chapter 5

Conclusion

In this work we used the MSKCC prostate cancer dataset to predict the Pathological Grade of a tumour. Bearing in mind that the pathological grade is measured during surgery, we completed suitable steps to remove information that is recorded after the surgery. We also removed predictors with many missing values and/or provided inadequate information and got a cleaned version of the dataset.

Data Exploration was a key part of our analysis where we understood every predictor and their respective contribution. Most importantly we visualised the distribution of our response variable, Pathological Gleason Score, and decided to combine the scores '7', '8' and '9' into one class called as '7+'. This new class indicates that the patient has an aggressive disease. We weighed two Feature Selection techniques and selected RFE for its superior performance. During the modelling stage we compared four Machine Learning Models, Logistic Regression, Random Forest, Support Vector Machine and Neural Network. All four of them were fit using Nested Cross Validation to get an estimate of their performance. We assessed that among the four, Logistic Regression and Random Forest performed the best and selected them for interpretation. We found the optimal features for both of the models and fit them on the complete dataset. Finally, we selected the Logistic Regression model to be the model of choice with the features, Primary Biopsy Gleason Score, Secondary Biopsy Gleason Score, Race, Clinical Tumour Stage, Radical Prostatectomy Type and Copy Number Cluster Assignment as the key biomarkers.

5.1 Limitations and Future Work

Here we shall discuss the limitations and possible future steps that could be taken to extend this work. One of the major limitations is that some of the features in the dataset were moderately imbalanced, such as Race, Sample Type, and the Pathological Gleason Score. Oversampling techniques could be explored in the future that may mitigate this issue. Techniques such as SMOTE (Synthetic Minority Oversampling Technique) [Chawla et al. 2002] have been used in recent literature [Abraham, and Nair 2018; Min et al. 2019; Hamzeh et al. 2020].

In the final Logistic Regression model, we find that only 3 explanatory variables are statistically significant. Further diagnostic tests could be conducted to assess the model and steps could be taken to improve the model. Another major limitation is that the dataset was very small. Future work could include collecting a larger and more diverse dataset and hence extending this study.

Bibliography

Abirami, S, and P Chitra [2020]. “Energy-efficient edge based real-time healthcare support system”. In: *Advances in Computers.* Vol. 117. 1. Elsevier, pp. 339–368.

Abraham, Bejoy, and Madhu S Nair [2018]. “Computer-aided diagnosis of clinically significant prostate cancer from MRI images using sparse autoencoder and random forest classifier”. In: *Biocybernetics and Biomedical Engineering* 38.3, pp. 733–744.

Adamo, P, and M R Ladomery [Apr. 2015]. *The oncogene ERG: a key factor in prostate cancer.* URL: `https://www.nature.com/articles/onc2015109`.

Alpaydin, Ethem [2020]. *Introduction to machine learning.* MIT press.

Arlot, Sylvain, and Alain Celisse [2010]. “A survey of cross-validation procedures for model selection”. In: *Statistics surveys* 4, pp. 40–79.

Barlow, Henry, Shunqi Mao, and Matloob Khushi [2019]. “Predicting high-risk prostate cancer using machine learning methods”. In: *Data* 4.3, p. 129.

Branco, Paula, Luis Torgo, and Rita Ribeiro [2015]. “A survey of predictive modelling under imbalanced distributions”. In: *arXiv preprint arXiv:1505.01658.*

Breiman, Leo [2001]. “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

Carrasco, Oscar Contreras [Aug. 2019]. *Support Vector Machines for Classification.* URL: `https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3`.

Cawley, Gavin C., and Nicola L. C. Talbot [2010]. “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation”. In: *Journal of Machine Learning Research* 11.70, pp. 2079–2107. URL: `http://jmlr.org/papers/v11/cawley10a.html`.

Chandrashekar, Girish, and Ferat Sahin [2014]. “A survey on feature selection methods”. In: *Computers & Electrical Engineering* 40.1, pp. 16–28.

Chawla, Nitesh V et al. [2002]. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16, pp. 321–357.

Edgar, Thomas W., and David O. Manz [2017]. “Chapter 4 - Exploratory Study”. In: *Research Methods for Cyber Security.* Ed. by Thomas W. Edgar, and David O. Manz. Syngress, pp. 95–130. ISBN: 978-0-12-805349-2. DOI: `https://`

doi.org/10.1016/B978-0-12-805349-2.00004-2. URL: https://www.sciencedirect.com/science/article/pii/B9780128053492000042.

Eichler, Klaus et al. [2006]. “Diagnostic value of systematic biopsy methods in the investigation of prostate cancer: a systematic review”. In: *The Journal of urology* 175.5, pp. 1605–1612.

Escobar, Gabriel J et al. [2016]. “Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals”. In: *Journal of hospital medicine* 11, S18–S24.

Fernández-Cabán, Pedro L, Forrest J Masters, and Brian M Phillips [2018]. “Predicting roof pressures on a low-rise structure from freestream turbulence using artificial neural networks”. In: *Frontiers in Built Environment* 4, p. 68.

Gleason Score [n.d.] URL: https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score.

Gleason Score and Grade Group [Mar. 2021]. URL: https://www.pcf.org/about-prostate-cancer/diagnosis-staging-prostate-cancer/gleason-score-isup-grade/.

Good, Phillip I [2006]. *Resampling methods.* Springer.

Hajian-Tilaki, Karimollah [2013]. “Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation”. In: *Caspian journal of internal medicine* 4.2, p. 627.

Hamzeh, Osama et al. [2020]. “Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data”. In: *BMC bioinformatics* 21.2, pp. 1–10.

James, Gareth et al. [2013]. *An introduction to statistical learning.* Vol. 112. Springer.

Jordan, Michael I, and Tom M Mitchell [2015]. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245, pp. 255–260.

Kumar, Ajitesh [Sept. 2020]. *ROC Curve AUC Explained with Python Examples.* URL: https://vitalflux.com/roc-curve-auc-python-false-positive-true-positive-rate/.

Kumar, Rajeev, and Abhaya Indrayan [2011]. “Receiver operating characteristic (ROC) curve for medical researchers”. In: *Indian pediatrics* 48.4, pp. 277–287.

Madu, Chikezie O, and Yi Lu [2010]. “Novel diagnostic biomarkers for prostate cancer”. In: *Journal of Cancer* 1, p. 150.

McGlynn, Elizabeth A, Kathryn M McDonald, and Christine K Cassel [2015]. “Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine”. In: *Jama* 314.23, pp. 2501–2502.

Min, Xiangde et al. [2019]. “Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method". In: *European journal of radiology* 115, pp. 16–21.

Mohler, James et al. [2010]. "Prostate cancer". In: *Journal of the National Comprehensive Cancer Network* 8.2, pp. 162–200.

NCI Dictionary of Cancer Terms [n.d.] URL: `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/primary-tumor`.

Prostate Cancer [Nov. 2019]. URL: `https://www.mariekeating.ie/get-men-talking/prostate-cancer/`.

Prostate cancer [Aug. 2021]. URL: `https://www.cancer.ie/cancer-information-and-support/cancer-types/prostate-cancer`.

Prostate Cancer - Stages and Grades [Jan. 2021]. URL: `https://www.cancer.net/cancer-types/prostate-cancer/stages-and-grades`.

Prostate Cancer Treatment [Aug. 2021]. URL: `https://www.cancer.gov/types/prostate/patient/prostate-treatment-pdq#_120`.

Prostate-Specific Antigen (PSA) Test [Feb. 2021]. URL: `https://www.cancer.gov/types/prostate/psa-fact-sheet#what-is-a-normal-psa-test-result`.

Ren, Qiubing, Mingchao Li, and Shuai Han [2019]. "Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives". In: *Big Earth Data* 3.1, pp. 8–25.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal [2016]. "Cancer statistics, 2016". In: *CA: a cancer journal for clinicians* 66.1, pp. 7–30.

Sperandei, Sandro [2014]. "Understanding logistic regression analysis". In: *Biochemia medica* 24.1, pp. 12–18.

TNM Staging [July 2019]. URL: `https://www.cancerresearchuk.org/about-cancer/prostate-cancer/stages/tnm-staging`.

Understanding Your Pathology Report: Prostate Cancer [Mar. 2017]. URL: `https://www.cancer.org/treatment/understanding-your-diagnosis/tests/understanding-your-pathology-report/prostate-pathology/prostate-cancer-pathology.html`.

What Is Prostate Cancer? [Aug. 2019]. URL: `https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html`.