**AI Algorithm**

**Final Project**

**NANDINI MALHOTRA**

**100768797**

**December 15, 2020**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **Index** | **Description** | **Page Number** |
|  |  |  |
| 1 | Introduction | 2 |
| 2 | Key Questions | 2 |
| 3 | Dataset Summary | 3 |
|  | 3.1 Finding the number of outliers and NA values in the   dataset | 3 |
|  | 3.1.1 Removing the missing and outliers form the   dataset | 4 |
|  | 3.2 Understanding the Descriptive Statistics of the   features | 4 |
|  | 3.3 Interpreting the box plot for the features | 5 |
|  | 3.4 Checking for correlation between the predictor   variables using Heatmap | 5 |
|  | 3.5 Feature Selection | 6 |
|  | 3.6 Graphs and Analysis types | 6 |
|  | 3.7 Data Metric | 7 |
| 4 | Model Analysis | 7 |
| 5 | Results | 17 |
| 6 | Conclusion | 18 |
| 7 | Appendix | 19 |

# Introduction

Canada experiences 30% of it’s death by the hands of Cancer. On an average, around 604 Canadians were diagnosed with cancer in 2019 everyday and 225 of the Canadian lost their lives to Cancer everyday. Breast cancer accounts for about 25% of the cases in women. There’s no control over the origin of cancer, but the growth an death by cancer can be avoided, by diagnosing it in its early stages.

Princess Margret is concerned with the lack of screening for Stage 1 cancer for the patients who are at potential risk of Cancer. 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 which would be trained using the available dataset. A solution to this problem would be presented, 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, the Rationale Statement would be the reason why for proposing the problem statement:

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

# Key Questions

The purpose of the model analysis and the report is to answer the following key questions:

1. **Are we able to distinguish between the patients having cancer from the patient who don’t have cancer?**

The dataset has observations with parameters that determine if the patient has cancer or not. The model will be implemented on the data, so that with the features in the test set, we are able to determine if the patient has cancer or not.

1. **Further, is the model efficient enough to distinguish between the type of cancer if the patient is diagnosed to have cancer?**

The model aims at determining the type of cancer if the patient is diagnosed of having a Stage 1 cancer. Determining the type of cancer will decide how the patient would be treated.

1. **Can this model be used as a definite test for the patients who do not have cancer?**

Not all the patients coming for pre-screening will have cancer. This model will also differentiate the patients who do not have cancer and do not need any treatment.

# Dataset Summary

The dataset initially had 1690 observations before pre-processing.

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

The NA values can be checked for using pandas profiling:

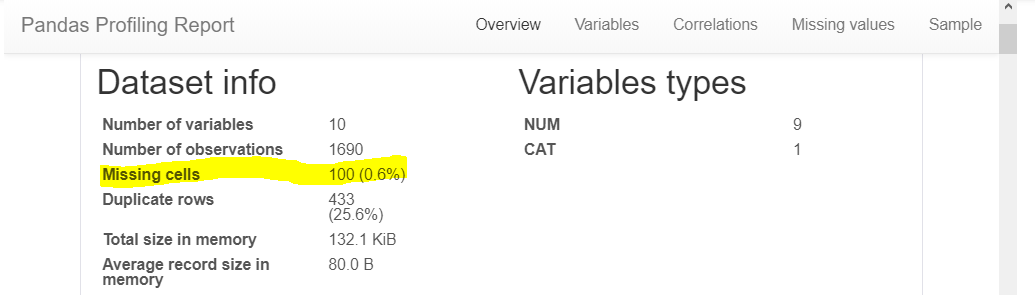


Fig-1

There are **10 missing values**

The Outliers can be identified using Tukey’s Algorithm:

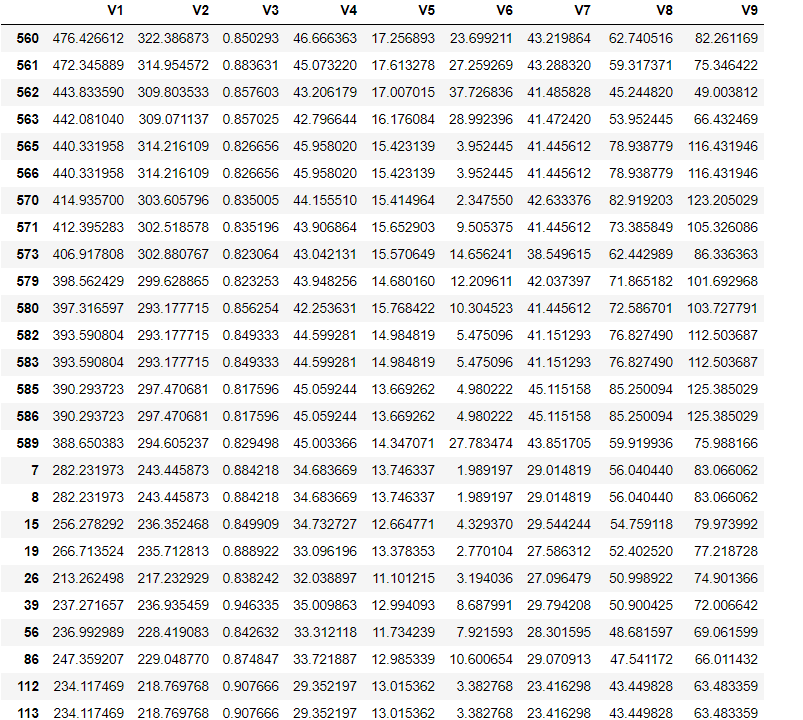


Fig-2

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

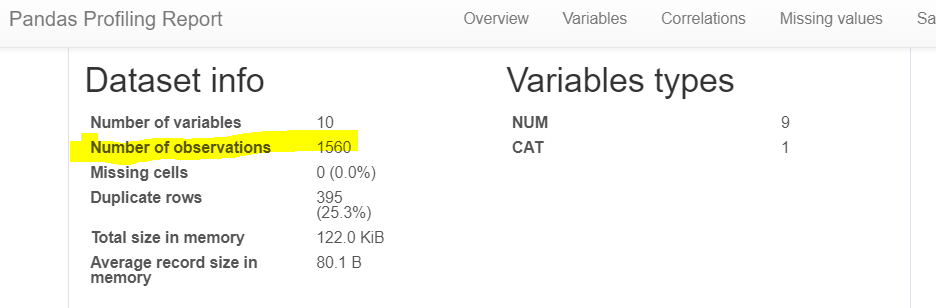


Fig-3

The final dataset after removing NA values and outlier has 1560 observations.

## Understanding the Descriptive Statistics of the features

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

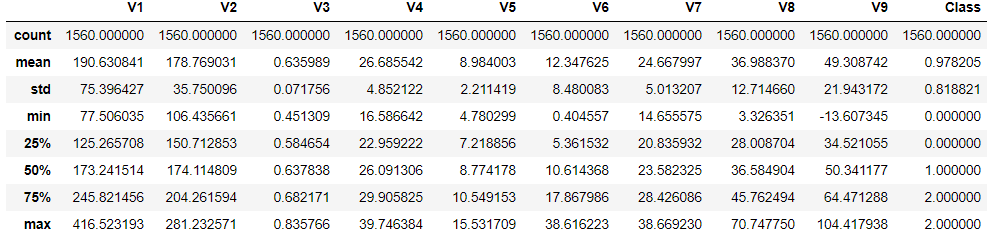


Fig-4

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

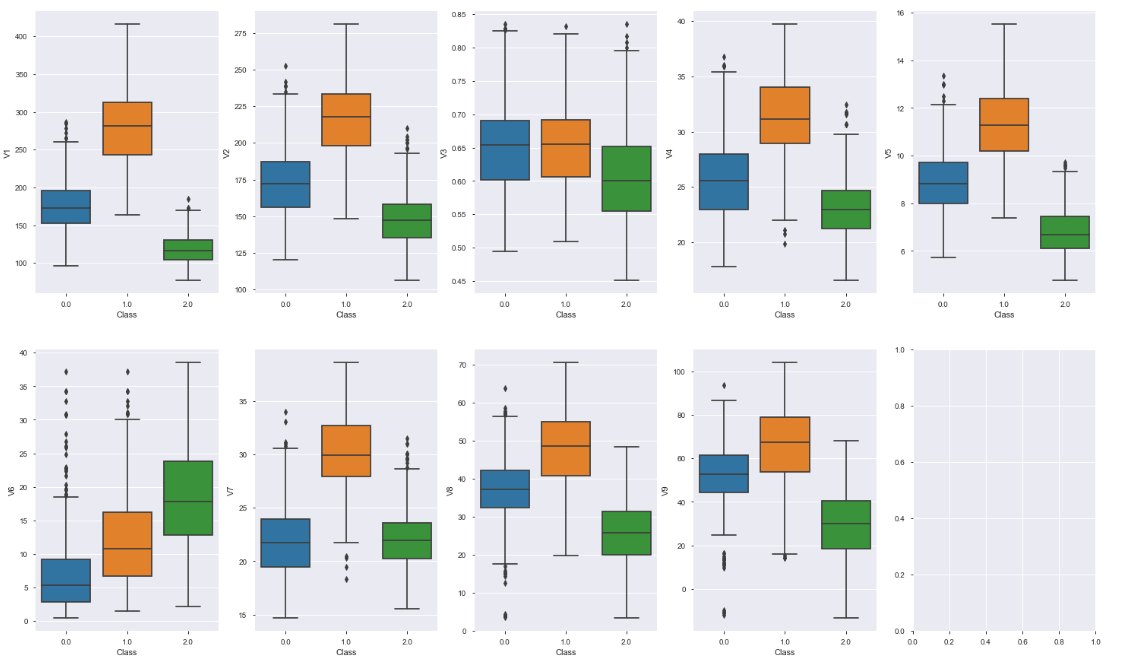
.

Fig-5

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

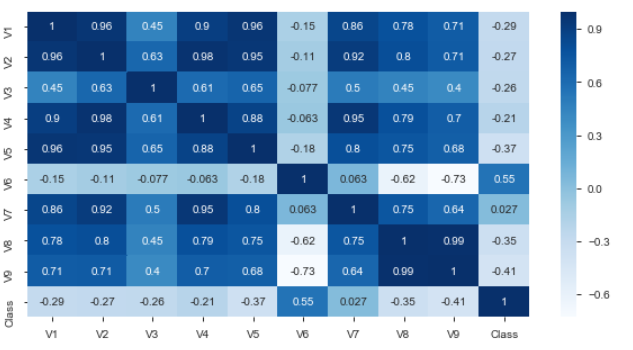


Fig-6

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.



Fig-7

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.

## 3.5 Feature Selection

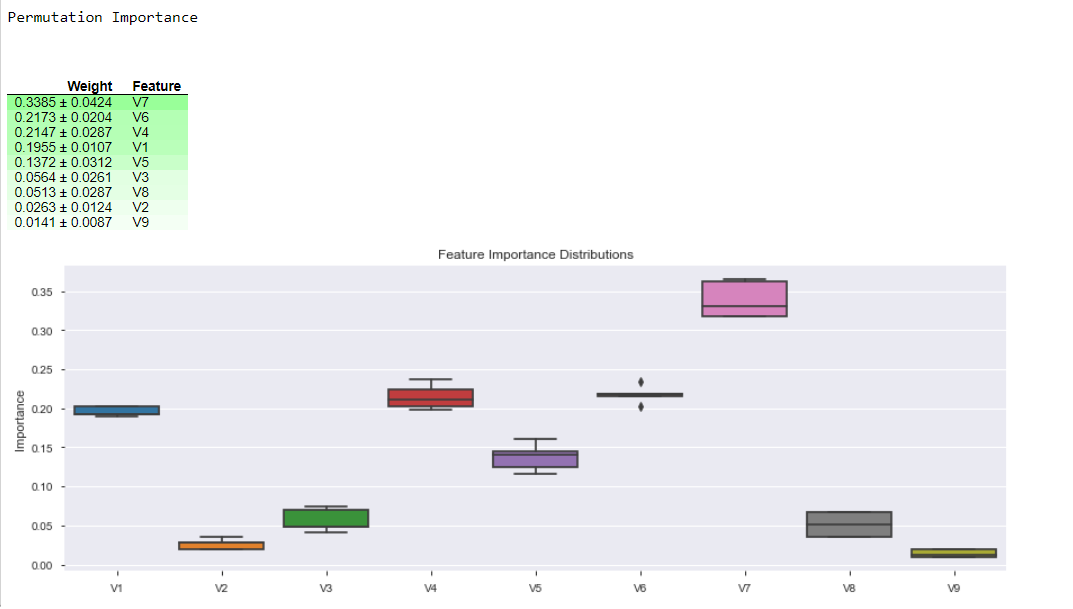


Fig-8

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.

Although, from feature selection we get top 3 features that have the most impact on the outcome, **we will consider all the features for our analysis** as every feature is assumed to be of equal importance for the analysis.

## 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. **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.
4. **Confusion Matrix** will depict the evaluation metrics that will help us evaluate the performance of the model.
5. Grid Search and Cross validation will used to optimize the models.

## 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.

# Model Analysis

**First Algorithm: Logistic Regression(multinomial) as base model**

The multinomial logistic regression generalizes logistic regression to multiple class which in our case are Class 0, Class 1 and Class 2. It is also known as SoftMax Regression.

The analysis was done with the following steps:

**Step 1: The Learning Curve**

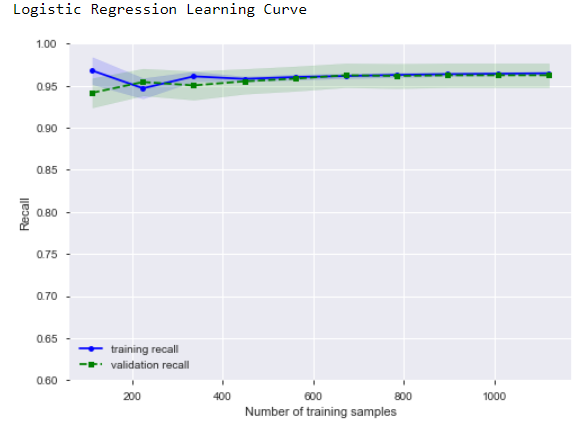
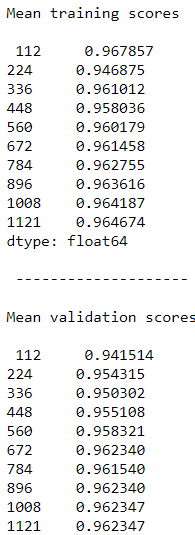
 

Fig-9

The learning and the corresponding values curve illustrate that,

For the training curve, initially the recall decreased, but then eventually it increases, which indicates that the model was trained well, and it was able to predict the outcome successfully.

For the validation curve, initially the recall was lower than the training curve, but as the number of test sample are increased the curve converges with the training curve which indicates low variance.

**Step 2: The Confusion and Classification matrix**

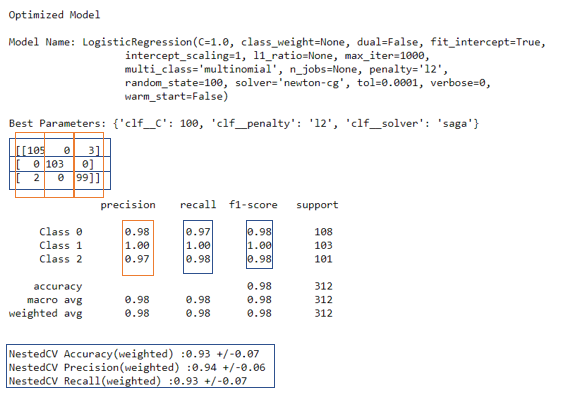


Fig-10

The confusion and classification matrices display the metrics recall and f1 score. It also displays the optimized results for Grid Search using cross validation, the Nested Metrics.

**Results:**

**For precision**

1. For Class 0, 2 out of 107 patients are incorrectly diagnosed (detected as Class 2, when they actually didn’t have cancer). Hence the precision for Class 0 is 105/107, i.e. 98%.
2. From the confusion matrix, it can be concluded that the model correctly identifies Class 1, a type 1-Stage 1 cancer, with a hundred percent precision, recall and f1 score.
3. For Class 2, 3 out of 102 patients are incorrectly diagnosed (detected as Class 0, when they actually Class 2). Hence the precision for Class 2 is 99/102, i.e. 97%.

**For recall**

1. For Class 0, 3 out of 108 patients are incorrectly diagnosed (detected as Class 2, when they actually didn’t have cancer). Hence the recall for Class 0 is 105/108, i.e. 97%.
2. From the confusion matrix, it can be concluded that the model correctly identifies Class 1, a type 1-Stage 1 cancer, with a hundred percent precision, recall and f1 score.
3. For Class 2, 2 out of 101 patients are incorrectly diagnosed (detected as Class 0, when they actually Class 2). Hence the recall for Class 2 is 99/101, i.e. 98%.

**F1 score** is the combination of recall and precision and it considers the wrong predictions. The f1 score for the three classes is 98%, 100% and 98% resp. It indicates that the model successfully identifies most of the outcomes correctly

The results from Grid Search are - NestedCV Accuracy(weighted) :0.93 +/-0.07, Precision(weighted) :0.94 +/-0.06 Recall(weighted) :0.93 +/-0.07.

**Second Algorithm: Ridge (L2) Regularisation**

If a dataset has large number of features in comparison to the number of observations, then the model ‘learns’ the training dataset and the error in predicting the outcome reduces gradually. Now when this model is applied on a test dataset then the prediction error increases. This is the sign of overfitting, which is also high variance, i.e. bigger difference between training and test error on the learning curve.

To deal with overfitting, we can regularise the model that minimizes the complexity of the model by optimizing the value of model parameter.

The Ridge Regularisation has the following mathematical representation:

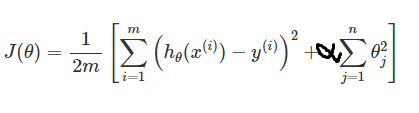


Fig-11

The analysis was done with the following steps:

**Step 1: The Learning Curve**

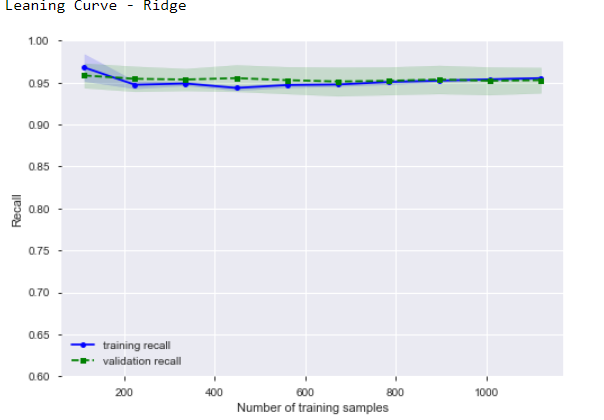
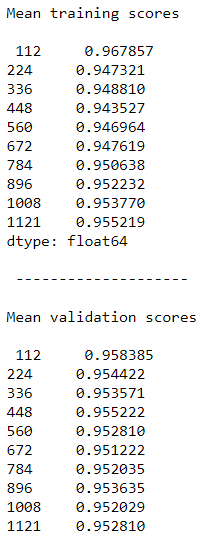
 

Fig-12

The learning and the corresponding values curve illustrate that,

For the training curve, initially the recall decreased, but then eventually it increases, which indicates that the model was trained well, and it was able to predict the outcome successfully.

For the validation curve, initially the recall was lower than the training curve, but as the number of test sample are increased the curve converges with the training curve which indicates low variance. The recall for validation curve has reduced, which means that the model should be trained further for a perfect prediction.

**Step 2: The Confusion and Classification matrix**

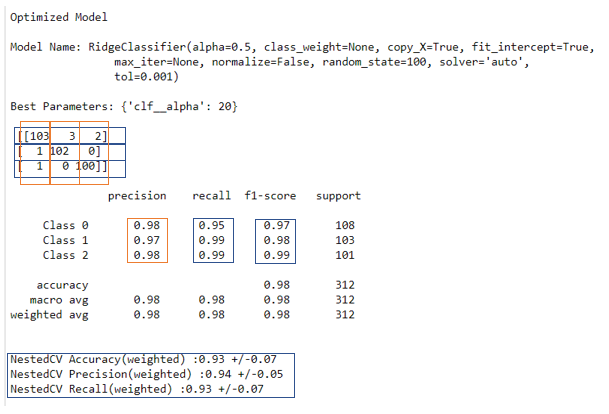


Fig-13

The confusion and classification matrices display the metrics recall and f1 score. It also displays the optimized results for Grid Search using cross validation, the Nested Metrics.

**Results:**

**For precision**

1. For Class 0, 2 out of 105 patients are incorrectly diagnosed (one each detected as Class 1 and Class 2, when they actually didn’t have cancer). Hence the precision for Class 0 is 103/105, i.e. 98%.
2. For Class 1, 2 out of 105 patients are incorrectly diagnosed (one each detected as Class 1 and Class 2, when they were actually Class 1). Hence the precision for Class 1 is 103/105, i.e. 98%.
3. For Class 2, 2 out of 102 patients are incorrectly diagnosed (detected as Class 0, when they actually Class 2). Hence the precision for Class 2 is 100/102, i.e. 98%.

**For recall**

1. For Class 0, 5 out of 108 patients are incorrectly diagnosed (3 detected as Class 1 and 2 detected as Class 2, when they actually didn’t have cancer). Hence the recall for Class 0 is 103/108, i.e. 95%. **The recall has reduced** in comparison to Logistic Regression.
2. For Class 1, 1 out of 103 patients are incorrectly diagnosed (detected as Class 0, when they were actually Class 1). Hence the recall for Class 1 is 102/103, i.e. 99%. It is less than that for Logistic Regression
3. For Class 2, 1 out of 101 patients are incorrectly diagnosed (detected as Class 0, when they actually Class 2). Hence the recall for Class 2 is 100/101, i.e. 99%. The recall has increased in comparison to LR.

**F1 score** is the combination of recall and precision and it considers the wrong predictions. The f1 score for the three classes is 97%, 98% and 99% resp. It indicates that the model successfully identifies most of the outcomes correctly. But the f1 score is less than that for Logistic Regression.

The results from Grid Search are - NestedCV Accuracy(weighted) :0.93 +/-0.07, Precision(weighted) :0.94 +/-0.05 Recall(weighted) :0.93 +/-0.07.

**Third Algorithm: Support Vector Machine (SVM)**

Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error.

The core idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

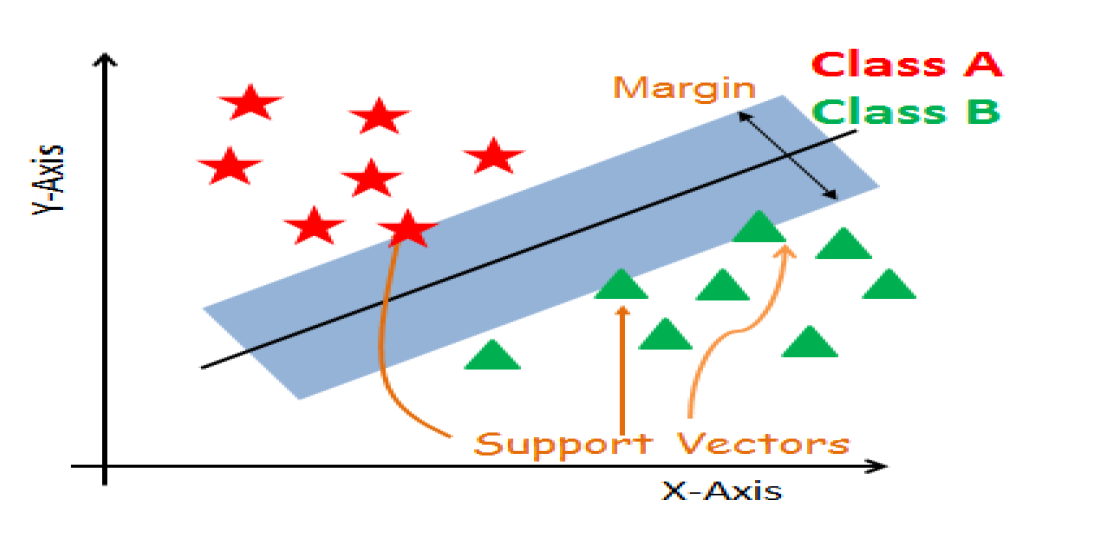


Fig-17

**Step 1: The Learning Curve**

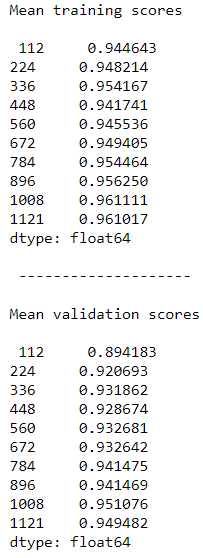
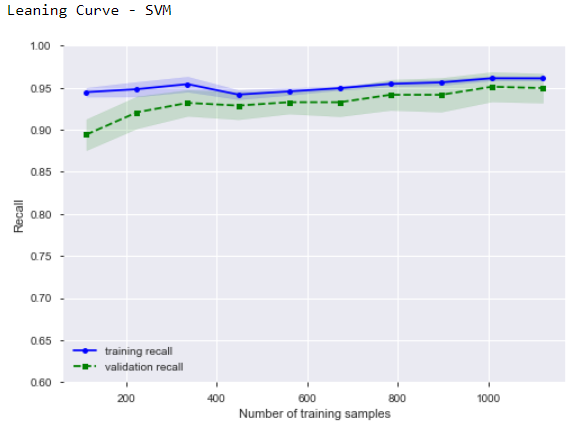


Fig-18

The learning and the corresponding values curve illustrate that,

For the training curve, eventually it increases, which indicates that the model was trained well, and it was able to predict the outcome successfully.

For the validation curve, initially the recall was lower than the training curve, but as the number of test sample are increased the curve converges with the training curve and the recall is increases, which indicates low variance. But the variance is higher than the earlier three models.

**Step 2: The Confusion and Classification matrix**

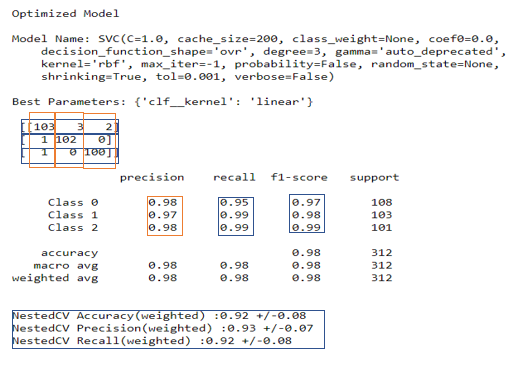


Fig-19

The confusion and classification matrices display the metrics recall and f1 score. It also displays the optimized results for Grid Search using cross validation, the Nested Metrics.

**Results:**

**For precision**

1. For Class 0, 2 out of 105 patients are incorrectly diagnosed (1 detected as Class 1 and 1 detected as Class 2, when they actually didn’t have cancer). Hence the precision for Class 0 is 103/105, i.e. 98%.
2. For Class 1, 3 out of 105 patients are incorrectly diagnosed (detected as Class 0, when they were actually Class 1). Hence the precision for Class 1 is 102/105, i.e. 97%.
3. For Class 2, 2 out of 102 patients are incorrectly diagnosed (detected as Class 0, when they actually Class 2). Hence the precision for Class 2 is 100/102, i.e. 98%.

**For recall**

1. For Class 0, 5 out of 108 patients are incorrectly diagnosed (3 detected as Class 1 and 2 detected as Class 2, when they actually didn’t have cancer). Hence the recall for Class 0 is 103/108, i.e. 95%.
2. For Class 1, 1 out of 103 patients are incorrectly diagnosed (detected as Class 0, when they were actually Class 1). Hence the recall for Class 1 is 102/103, i.e. 99%.
3. For Class 2, 1 out of 101 patients are incorrectly diagnosed (detected as Class 0, when they actually Class 2). Hence the recall for Class 2 is 100/101, i.e. 99%. The recall has increased in comparison to LR and is same as ridge.

**F1 score** is the combination of recall and precision and it considers the wrong predictions. The f1 score for the three classes is 97%, 98% and 99% resp. It indicates that the model successfully identifies most of the outcomes correctly.

The results from Grid Search are - NestedCV Accuracy(weighted) :0.92 +/-0.08, Precision(weighted) :0.93 +/-0.07 Recall(weighted) :0.92 +/-0.08. **These values are the lowest among all the 4 models.**

# Results

Model Evaluation for all 4 models

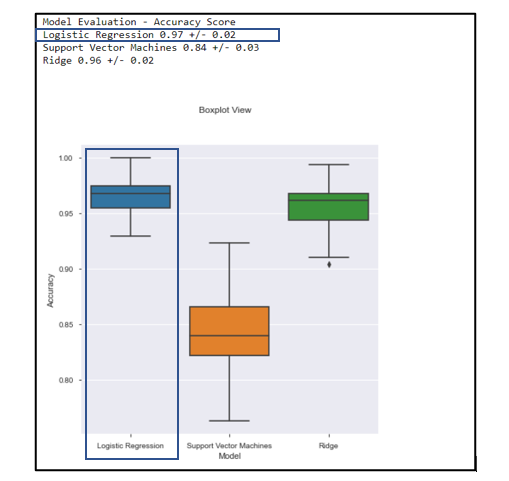


Fig-20

Comparison of the metric for each class.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | | | **Ridge Regularization** | | | **SVM** | | |
| **Class** | **0** | **1** | **2** | **0** | **1** | **2** | **0** | **1** | **2** |
| **Precision** | 0.98 | 1.00 | 0.97 | 0.98 | 0.97 | 0.98 | 0.98 | 0.97 | 0.98 |
| **Recall** | 0.97 | 1.00 | 0.98 | 0.95 | 0.99 | 0.99 | 0.95 | 0.99 | 0.99 |
| **F1** | 0.98 | 1.00 | 0.98 | 0.97 | 0.98 | 0.99 | 0.97 | 0.98 | 0.99 |

Table-1

Form the Model evaluation and the comparison of the metrics, it can be concluded that **Logistic Regression is the best model.** The insights and findings form this model are:

1. The patient with Stage 1- Type 1 is diagnosed with a 100 % precision and recall , which means that if a patient is diagnosed to have a Type 1 cancer than there is no further pre-screening required, he can be directly treated with the appropriate amount of radio/chemo therapy.
2. For the Class 2, the model correctly predicts 99 out of 101 patients to be at Stage 1 – Type 2 cancer. For these patients’ further test might be required as the model predicts hem free of cancer, which is a matter of concern.
3. For the patients who don’t have cancer, the model predicts 3 of them to have class 2 cancer.

Although, it is observed that a total of 5 observations are incorrectly predicted, our major concern would be the patients having Type 2 cancer who are wrongly diagnosed.

# Conclusion

Upon implementing Logistic Regression on the processed dataset, the 3 key questions can be answered:

1. The patients who don’t have cancer are correctly identified with the precision of 98%.
2. This model is a definite test for Type 1 cancer, i.e. the patients who are diagnosed to have Type 1 cancer should be immediately given treatment so that it does not escalate to Type 2.
3. For the patient diagnosed with type 2 cancer must be treated immediately as it is 98% correctly diagnosed if the patient has a Type 2 cancer.

# Appendix

## 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**

## High Level 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. (Added)
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.

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