

University of Westminster

# Machine Learning and Data Mining

5DATA002W.2

Analyses Report of Two Case Studies, Predicting Cancer Patients' Mortality Status and Survival Months

Muhammed Rizwan Ahamed Rizlan

Uow Number: W1987516  
 IIT Number: 20223256

# **Table of contents**

[**Machine Learning and Data Mining 0**](#_ecd546uz66d0)

[**Table of contents 1**](#_4i8x9uysoxxr)

[**Q1. Does machine learning have the potential to assist doctors in predicting those who will survive breast cancer or not? 2**](#_7ztfn4u1u3w7)

[Task (1) – Domain Understanding: Classification 2](#_wmvke87jnja8)

[Task (2) – Exploring and Understanding Your Dataset 3](#_fz05di11he4e)

[Task (3a) – Data Preparation: Cleaning and Transforming your data 5](#_xm87x16oq8rc)

[Task (3b) – Data Preparation: Cleaning and Transforming your data 6](#_ll7b9nah8sv)

[Task (4a) – Classification Modelling of Cancer Patients' Mortality Status 8](#_2pjg3e7pajnq)

[Task (4b) – Classification Modelling of Cancer Patients' Mortality Status 8](#_rvmg5tw21vl)

[( i ) 8](#_bhk2t0t5pt6x)

[( ii ) 8](#_odjrzs6m6p5)

[( iii ) 8](#_ctfrodt50ycb)

[Task (5a) – Evaluating your Cancer Mortality Status Classification Models 9](#_maufwswpzy7i)

[Task (5b) – Evaluating your Cancer Mortality Status Classification Models 10](#_d6crgsk5ma0l)

[Task (5c) – Evaluating your Cancer Mortality Status Classification Models 10](#_3ynupu0n2og)

[Task (5d) – Evaluating your Cancer Mortality Status Classification Models 11](#_cg0oey5barsu)

[( i ) 11](#_r752avtm34t5)

[( ii ) 11](#_d4tfv612mlo9)

[Task (5e) – Evaluating your Cancer Mortality Status Classification Models 12](#_u1luualbq6iq)

[Task (5f) – Evaluating your Cancer Mortality Status Classification Models 12](#_6vaeib4om7jl)

[( i ) 12](#_15msq7syh2sk)

[( ii ) 13](#_m7uhfs73xyrr)

[( ii ) 13](#_xk4bpy6kzj5z)

# 

# **Q1**. **Does machine learning have the potential to assist doctors in predicting those who will survive breast cancer or not?**

## **Task (1) – Domain Understanding: Classification**

| **Variable Name** | **RETAIN or DROP** | **Brief justification for retention or dropping** |
| --- | --- | --- |
| Patient ID | DROP | This is just a Unique Identifier, won't help for Prediction |
| Month of Birth | DROP | Breast cancer survival is unlikely to depend on birth month |
| Age | RETAIN | Age directly depends on mortality (National Breast Cancer Foundation, 2024). |
| Sex | RETAIN | For females in the UK, breast cancer is the 2nd most common cause of cancer death, and 85 men every year so Sex also plays a major role |
| Occupation | DROP | Since Occupation has more Missing values, we have to drop it |
| T Stage | RETAIN | T Stage refers to the size of the cancer cell; the larger the cells more vulnerable they are. |
| N Stage | RETAIN | N Stage is the measure of the spread of cancer cells around the lymph, so this also directly affects the severity |
| 6th Stage | RETAIN | This Breast Imaging Reporting and Data System, or BI-RADS This also depends |
| Differentiated | RETAIN | This shows the growth and appearance of cancer cells when compared to normal cells |
| Grade | RETAIN | This is a grading system given to cancer cells (Nottingham Grading System) |
| A Stage | RETAIN | This shows the spread of cancer cells. This also affects the severity |
| Tumour Size | RETAIN | This refers to the size of the Tumour cells |
| Estrogen Status | RETAIN | This would indicate the availability of estrogen hormone receptors |
| Progesterone Status | RETAIN | Indicate the availability of Progesterone hormone receptors |
| Regional Node Examined | RETAIN | This gives a count of examined regional lymph nodes for cancer spread |
| Regional Node Positive | RETAIN | Count of cancer-positive regional lymph nodes to contain metastases |
| Survival Months | DROP | This can be confusing, and the Model would learn falsely that a low survival month is death. |
| Mortality Status | RETAIN | This is the Feature to be Tested - Target |

## **Task (2) – Exploring and Understanding Your Dataset**

| **The basic descriptive stats and variable scale type For Numeric Variables** |
| --- |
| **The basic descriptive stats and variable scale type for Object Variables** |
| **Distribution of Mortality Status** |

**The descriptive stats and variable scale type For Variables**

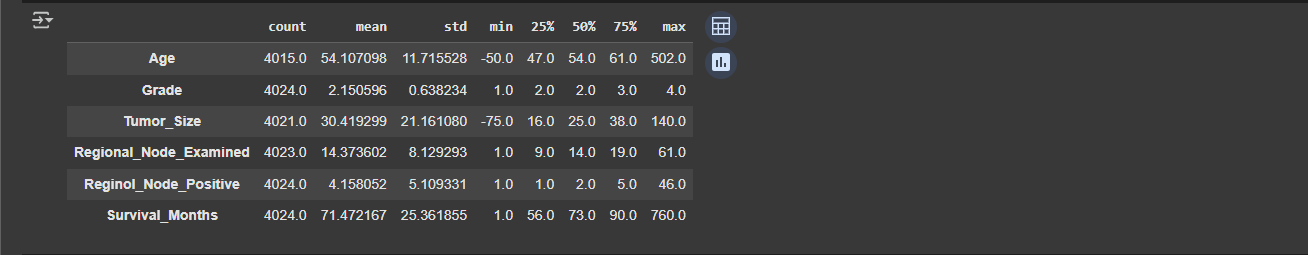
| **Variable Name** | **Scale Type** | **Scale Type** |
| --- | --- | --- |
| Age | Numerical | This is a Continuous variable |
| Sex | Nominal (Binary) | This is important because women have a higher death rate than men |
| T Stage | Ordinal | This refers to the Size of Tumours |
| N Stage | Ordinal | This Measures the Spread of Tumours |
| 6th Stage | Ordinal | This is a Breast Imaging Reporting and Data System |
| Differentiated | Ordinal | This shows how they look and are growing compared with normal cells |
| Grade | Ordinal | This is the score of the Nottingham Grading System |
| A Stage | Ordinal | This categorizes the distance of the Spread |
| Tumour Size | Numerical | Tumor size measured in millimeters |
| Estrogen Status | Nominal (Binary) | This tells whether the cells have estrogen receptors or not |
| Progesterone Status | Nominal (Binary) | This tells whether the cells have Progesterone receptors or not |
| Regional Node Examined | Numerical | Count of examined regional lymph nodes for cancer spread |
| Regional Node Positive | Numerical | Count of cancer-positive regional lymph nodes to contain metastases |
| Mortality Status | Nominal (Binary) | Any patient who dies after the follow-up cut-off date is recorded as alive  as of the cut-off date. If the date of last contact > study cutoff date, vital  status recoded = alive. |

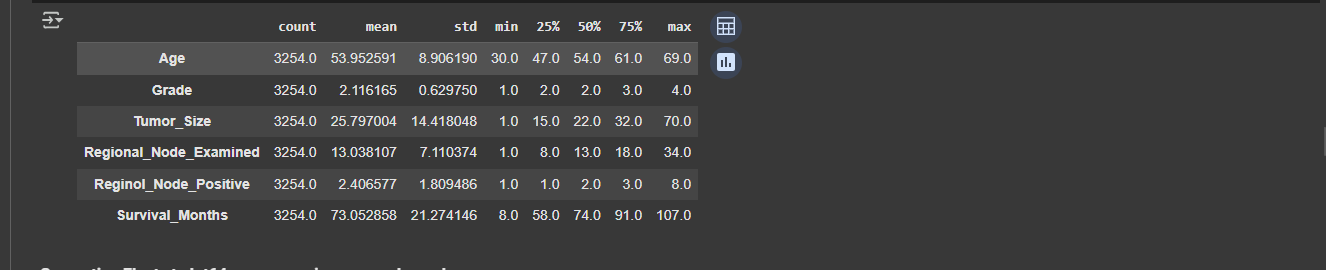
## **Task (3a) – Data Preparation: Cleaning and Transforming your data**

| **Variable Name** | **Issues Found** | **Proposed Fix** | **Justification for use fix method** |
| --- | --- | --- | --- |
| Age | 1. There were values and extreme values 2. Null Values | 1. Remove the Outliers 2. Fill nulls with mean values | 1. Outlier removal will remove the extreme values 2. We can have more data since nulls are imputed |
| Sex | 1. Inconsistent values 2. Null Values | 1. Assumed non-null all values as Males 2. Removed the Null values | 1. We can have data for both men and women 2. Since I can't predict the gender |
| Tumor\_Size | 1. There were values and extreme values 2. Null Values | 1. Remove the Outliers 2. Fill nulls with mean values | 1. Outlier removal will remove the extreme values 2. We can have more data since nulls are imputed |
| Regional\_Node\_Examined | 1. Null Values | 1. Fill Null values with the Mean | 1. We can have more data since nulls are imputed |
| Mortality\_Status | 1. Duplicated values in Alive and Death Values | 1. Did a function to change the values into Alive and Death | 1. The Duntion changes Duplicated Spellings to correct Spellings |

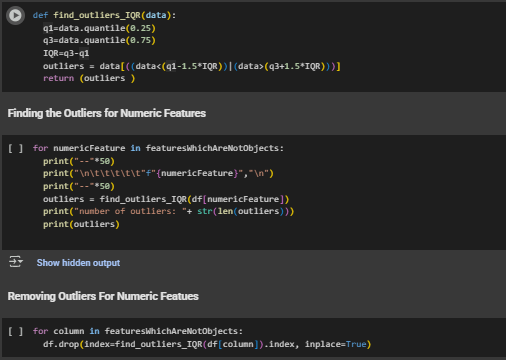
## 

## **Task (3b) – Data Preparation: Cleaning and Transforming your data**

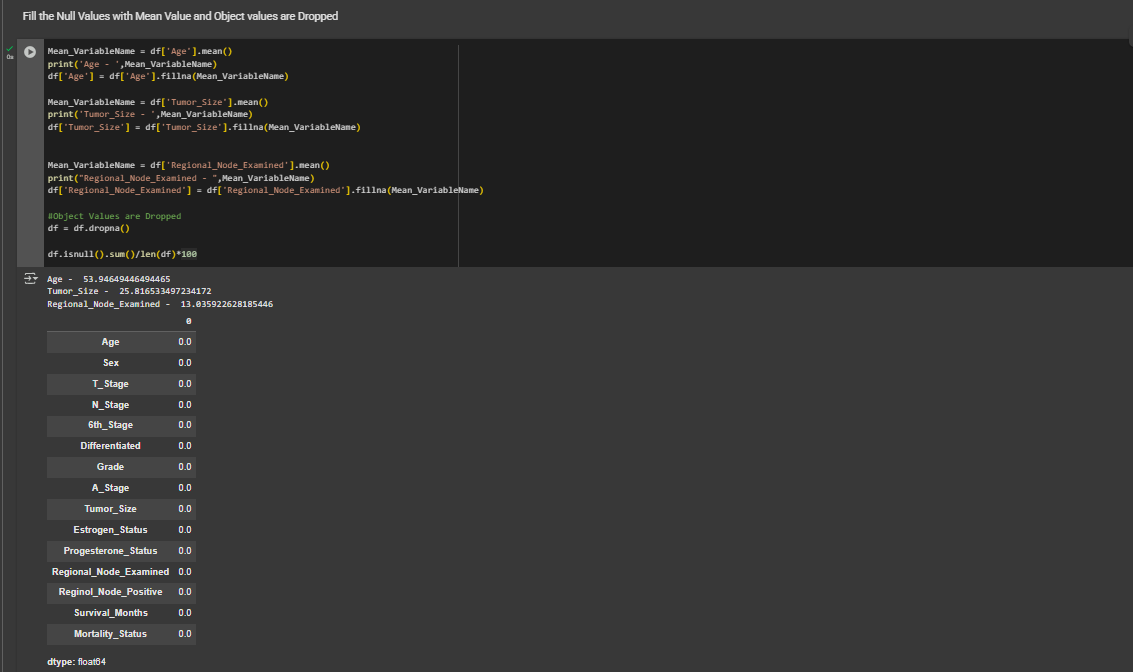


From the Above, you can see that the 50% value and Max have a big difference for Age, Tumor Size, Regional Node examined, Regional Node Positive, Survival Months also Negative values in Age, Tumor Size 

In this image, the extreme values are fixed, and no big difference between the 50% value and Max, No Negative values in Age, Tumor Size



The above issues are fixed by finding and removing the outliers from the dataset, The implementation is above



This image on the left shows null values as a count of the dataset. Here, we can divide the null-valued features into two

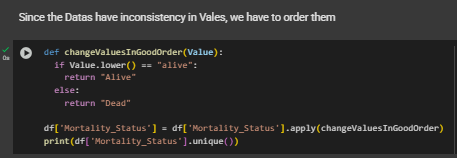
1. Numeric - These are Filled with the Mean
2. Object - These are dropped

The Image on the right shows how the Nulls are removed, and Imputing mean values for Numeric types, and Dropping nulls in object Features



The above images show that the Mortality Status and Sex have inconsistencies in their values, that is fixed by checking the available data and turning the duplicated values into a single value in Mortality Status

For Sex, the Nan are removed and 1s are considered as Males. The Implementation below



## **Task (4a) – Classification Modelling of Cancer Patients' Mortality Status**

| **Algorithm Name** | **Algorithm Type** | **Learnable Parameters** | **Strategic Hyperparameters** |
| --- | --- | --- | --- |
| NB | parametric | priors | priors [0.5, 0.5] - set manually to handle imbalance in the data |
| LR | parametric | Cofficient  Intercept | Class\_weight = Balanced This would handle the class weight due to the imbalanced data  C, Penalty ,solver |
| KNN (N=5) | Non-parametric | - | 1. n\_neighbors (1 to 34)- No of neighbours to consider 2. weights:'uniform','distance'How to weigh the neighbours 3. metric: 'euclidean',  'manhattan', 'minkowski' |

## **Task (4b) – Classification Modelling of Cancer Patients' Mortality Status**

### **( i )**

### **( ii )**

The original dataset was split into 70% for training and 30% for testing. This allows the model to learn from a wide range of patient cases, avoid overfitting, and become more generalised and robust. The remaining 20% is used to evaluate how well the model performs on unseen data, providing a more realistic testing scenario. This type of split is commonly used and widely accepted in machine learning (Gunkurnia, 2023).

### **( iii )**

| X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, **stratify=y, random\_state=42**) |
| --- |

The above code line, **random\_state=42**, proves that the split to train and test data is split in a reproducible manner. Every time the data is split in a defined place, this helps with Consistency in Model Evaluation, so it would be the same every time. The code line,  **stratify=y,** would maintain the same ratio between the Alive and Death classes since we have an imbalance in the dataset(Gupta, 2021). Since these are equally distributed and defined split, this would prevent the model from being trained in a non-biased way

## **Task (5a) – Evaluating your Cancer Mortality Status Classification Models**

1. **Naive Bayes (NB)**

|  |  |  |
| --- | --- | --- |

1. **Logistic Regression (LR)**

|  |  |  |
| --- | --- | --- |

1. **K-Nearest Neighbours ( KNN )**

|  |  |  |
| --- | --- | --- |

## **Task (5b) – Evaluating your Cancer Mortality Status Classification Models**

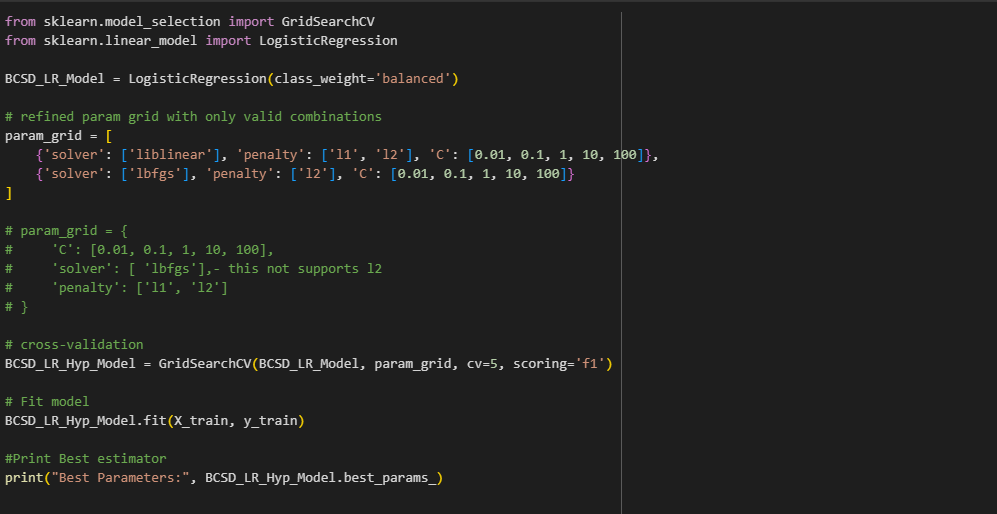
| **Metrics** | **Use or  Not Use** | **Justification for choosing “USE” or “DO**  **NOT USE” in relation to the success criteria** | **Model Name** | **Score** |
| --- | --- | --- | --- | --- |
| Accuracy | Not Use | Accuracy won't distinguish between alive and dead; instead, this shows how correctly our model predicts from all the predictions. This can be misleading in an imbalanced dataset  **Accuracy = Correct Predictions / All Predictions** | NB | 0.73 |
| LR | 0.64 |
| KNN n=5 | 0.89 |
| Recall | Use | Recall helps to understand how the model identifies True Positives from all actual Positives  E.g. - How many people are correctly predicted as dead out of **actually dead people**  **Recall = True Positive /**  **(True Positive + False Negative)** | NB | 0 = 0.49 1= 0.76 |
| LR | 0 = 0.64 1= 0.64 |
| KNN n=5 | 0 = 0.00 1= 1.00 |
| Precision | Use | This is the measure of how many are True predictions out of all true predictions  E.g. - How many people are dead out of the people model **predicted as Dead**  **Precision = True Positives /  (True Positive + False Positive)** | NB | 0 = 0.21 1= 0.92 |
| LR | 0 = 0.18 1= 0.93 |
| KNN n=5 | 0 = 0.00 1= 0.89 |
| F-Score | Not Use | This is the harmonic mean of both Recall and Precision  The higher the F1 well-balanced Model  Usually, F1 below 0.5 is considered not good  **F1 score = 2 \* Precision \* Recall /  (Precision + Recall)** | NB | 0 = 0.29 1= 0.84 |
| LR | 0 = 0.29 1= 0.76 |
| KNN n=5 | 0 = 0.00 1= 0.94 |
| AUC-ROC | Use | This plots the graph between the **True Positive Rate** and the **False Positive Rate**  Means Probability of differentiating alive and Dead values | NB | 0.67 |
| LR | 0.68 |
| KNN n=5 | 0.64 |

## **Task (5c) – Evaluating your Cancer Mortality Status Classification Models**

When we select the best model, we should focus more on Death(0), where accuracy is less important since we are focusing on distinguishing alive and the dead. Precision, Recall, F1 Score, and AUC play a major role here, when comparing to them, other than KNN, LR and NB share equal chance, but with the comparison of AUC and Recall (0.64>0.49)(0.68>0.67), **Logistic Regression would be the best model for our classification**

## **Task (5d) – Evaluating your Cancer Mortality Status Classification Models**

### **( i )**

  
The Best Parameters from the Model

|  |
| --- |

### **( ii )**

| **Logistic Regression (LR) Before Hyperparameterization** | **Logistic Regression (LR) after Hyperparameterized** |
| --- | --- |
|  |  |

The Hyperparameterization didn't give any positive improvements, instead, it gave a slightly lower value than non-hyperparameterized logistic regression. This may be because of the imbalance

## **Task (5e) – Evaluating your Cancer Mortality Status Classification Models**

Logistic regression (class\_weight=’balanced’) is chosen as the best model, and it performed better than the Hyperparameterized model

1. Model Critique & Limitations
2. Not a very good performance on the minority class (class = 0 Dead)

The model has a low F1 score value for the minority class, so the model struggles to predict dead people accurately

1. Imbalanced dataset

The Class distribution is very bad, class 1 has more data when compared to class 0. Even though the weight is balanced small number of samples of class 0 limits the generalizability

1. Logistic regression is a simple algorithm

Since logistic regression is a linear model, it may not capture complex relationships

1. The hyperparameter had no Impact

Hyperparameterization didn't help to enhance the model, maybe it's because of the imbalanced data

1. Ethical Considerations
2. False Negatives can be harmful

False prediction of an alive person as dead is very harmful

1. Bias and Fairness

Survival also depends on ethnicity and the environment the patient lives and etc

## **Task (5f) – Evaluating your Cancer Mortality Status Classification Models**

### **( i )**

**from sklearn.naive\_bayes import GaussianNB**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.ensemble import VotingClassifier**

**base\_learners=[('GaussianNB',GaussianNB(priors=[0.5,0.5])), ('LogisticRegression',**

**LogisticRegression(class\_weight='balanced'))]**

**ensemble\_learner = VotingClassifier(base\_learners, voting='soft')**

### 

### **( ii )**

1. **Naive Bias**

|  |  |  |
| --- | --- | --- |

1. **Logistic Regression (LR)**

|  |  |  |
| --- | --- | --- |

1. **Ensembled Model**

|  |  |  |
| --- | --- | --- |

**Both NaiveBayes and Logistic Regression share favourable characteristics with KNN, and NaiveBayes has a higher Recall and lower precision mean, while Logistic Regression Lower Recall and Higher Precision in LR, so I thought combining these two would help us in the detection of Dead Patients**

### **( ii )**

The ensemble model combining NB and LR has a slight better over NB on recall and Precision, also it does not outperform LR in Recall, which is critical for us. The F1 score remains the same for all 3, and AUC-ROC remains the same as LR, But Still LR with Balanced class weight remains the best out of these 3

Also, it maintains better Recall and ROC than others, so **LR remains the best model for Mortality Prediction**

# **Q2**. **Does machine learning have the potential to assist doctors in predicting survival months for patients who are not going to survive breast cancer?**

## **Task (1) – Domain Understanding: Classification**

| **Dimension of the Dataset** |  |
| --- | --- |
| **Features for Regression Modelling** |  |

## **Task (2a) – Modelling: Build Predictive Regression Models**

A Decision Tree is a flowchart-like structure that internal nodes represent a test on a feature, also in the point of healthcare, its predictions are clear and easy to understand. DTs can capture complex relationships between features with survival months. Meanwhile, DTs are tree-like structures that doctors can follow each step of the tree and see how survival is measured, and can explain them to the relevant people. This makes them a strong choice for predicting survival months in breast cancer patients, where clear and trusted results are very important.Also these DTs don't need much cleaning, work well with numeric and objects

## **Task (2b) – Modelling: Build Predictive Regression Models**

### **( i )**

| **Imports For Decision Trees** | |
| --- | --- |
| **Fitting and Train the Model DT-1** | **Fitting the Model DT-2** |

### **( ii )**

The Decision Tree is pruned to a max depth of 4 using hyperparameter Gridsearchcv. This would help the model to be away from overfitting, by restricting the tree growth to 4. We have made the model simpler and generalised, since this is a health-related data, small contexts may depend, so the model can overfit, and overfitting can give unreliable and false predictions. The pruned tree is beneficial, more generalised, and has good variance, but as a drawback, the model can underfit as well. important patterns are missed, which are on the deeper levels of the tree, possibly lowering the accuracy

## **Task (2c) – Modelling: Build Predictive Regression Models**

| **Fully Grown Tree** | **Pruned Tree** |
| --- | --- |

**Task (3a) – Evaluating your Cancer Survival Months DT Regression Models**

| **Metrics** | **Use/Do Not Use** | **Justification in relation to the**  **success criteria** | **Model Name** | **Test**  **Score** |
| --- | --- | --- | --- | --- |
| MSE |  |  | DT-1 (Fully Grown DT) |  |
| DT-2 (Pruned DT) |  |
| MAE |  |  | DT-1 (Fully Grown DT) |  |
| DT-2 (Pruned DT) |  |
| R-Square |  |  | DT-1 (Fully Grown DT) |  |
| DT-2 (Pruned DT) |  |