**INFORMATICS INSTITUTE OF TECHNOLOGY**

**Machine Learning and Data Mining**

5DATA001C.2

Analysis Report

**Name :** Kamkanamge Dona Dulani Linara

**UoW ID :** w2051819

**IIT ID :** 20231068

**Date of Submission :** 05th May, 2025

Table of Contents

[List of Figers 2](#_Toc197171596)

[List of Tables 3](#_Toc197171597)

[Case Study (A): Predicting Cancer Patients Mortality Status 4](#_Toc197171598)

[Task (1) – Domain Understanding: Classification 4](#_Toc197171599)

[Task (2) – Exploring and Understanding the Dataset 6](#_Toc197171600)

[Task (3) – Data Preparation: Cleaning and Transforming the data 7](#_Toc197171601)

[(a) Summary of Issues and Fixes 7](#_Toc197171602)

[(b) Evidence (Before and After) of Implementing 8](#_Toc197171603)

[Task (4) – Classification Modelling of Cancer Patients Mortality Status 9](#_Toc197171604)

[(a) Algorithm Summary Table 9](#_Toc197171605)

[Task (5) – Evaluating your Cancer Mortality Status Classification Models 10](#_Toc197171606)

[(a) Model Performance 10](#_Toc197171607)

[(b) Evaluation Metrics Table 11](#_Toc197171608)

[(c) Model Choice Justification 11](#_Toc197171609)

[(d) Hyperparameter Tuning (Logistic Regression) 12](#_Toc197171610)

[(e) Model Limitations & Ethical Concern 13](#_Toc197171611)

[(f) Voting Classifier Evaluation 13](#_Toc197171612)

[Case Study (B): Analyses Report for Predicting Survival Months 14](#_Toc197171613)

[Task (1) – Domain Understanding and Designing the Regression Experiments 14](#_Toc197171614)

[Task (2) – Modelling: Build Predictive Regression Models 15](#_Toc197171615)

[(a) Model Explanation 15](#_Toc197171616)

[(b) Comparison Example 15](#_Toc197171617)

[(c) Model Construction and Rationale 15](#_Toc197171618)

[Task (3) – Evaluating your Cancer Survival Months DT Regression Models 16](#_Toc197171619)

[(a) Evaluation Table 16](#_Toc197171620)

[(b) Model Choice Justification 17](#_Toc197171621)

[(c) Limitations of Evaluation Metrics in a Clinical Context 17](#_Toc197171622)

[Task (4) – Interpreting Cancer Survival Months Decision Tree Outcomes 18](#_Toc197171623)

[Reference 19](#_Toc197171624)

# **List of Figers**

[Figure 1 First 5 Rows of the Dataset & df.info() Output 6](#_Toc197248892)

[Figure 2 Distribution of Mortality Status 7](#_Toc197248893)

[Figure 3 Missing Value Heatmap (Before Cleaning) 8](#_Toc197248894)

[Figure 4 Data Sample After Cleaning and Encoding 9](#_Toc197248895)

[Figure 5 Output of X\_train.shape, Class Distribution 10](#_Toc197248896)

[Figure 6 Code Showing train\_test\_split() with stratify=y 10](#_Toc197248897)

[Figure 7 Confusion Matrix + Classification Report per Model - Logistic Regression 10](#_Toc197248898)

[Figure 8 Confusion Matrix + Classification Report per Model - KNN 11](#_Toc197248899)

[Figure 9 Confusion Matrix + Classification Report per Model - Naive Bayes 11](#_Toc197248900)

[Figure 10 GridSearchCV Output 13](#_Toc197248901)

[Figure 11 GridSearchCv Output (Beat parameters) 13](#_Toc197248902)

[Figure 12 Confusion Matrix 13](#_Toc197248903)

[Figure 13 Classification Report for Voting Classifier (Soft Voting) 14](#_Toc197248904)

[Figure 14 ROC AUC Score for Voting Classifier 14](#_Toc197248905)

[Figure 15 Dataset for Regression Task 15](#_Toc197248906)

[Figure 16 Plot of Fully Grown Decision Tree (DT-1) 17](#_Toc197248907)

[Figure 17 Plot of Pruned Tree (DT-2) with max\_depth=4 17](#_Toc197248908)

[Figure 18 Metric Scores for DT-1 18](#_Toc197248909)

[Figure 19 Metric Scores for DT-2 18](#_Toc197248910)

[Figure 20 Output Showing Predicted Survival Months 19](#_Toc197248911)

# **List of Tables**

[Table 1 Domain Understanding: Classification 6](#_Toc197249583)

[Table 2 Summary of Issues and Fixes 9](#_Toc197249584)

[Table 3 Algorithm Summary Table 10](#_Toc197249585)

[Table 4 Evaluation Metrics Table 12](#_Toc197249586)

[Table 5 Evaluation Table 18](#_Toc197249587)

# **Case Study (A): Predicting Cancer Patients Mortality Status**

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

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **RETAIN or DROP** | **Brief justification for retention or dropping** |
| Patient ID | DROP | It is an unique identification that has no predictive value. |
| Month of Birth | DROP | Not directly related to the progression or death rate of cancer. |
| Age | RETAIN | The prognosis and results of cancer treatment are significantly influenced by age. |
| Sex | RETAIN | Hormone-related malignancies can be influenced by gender. |
| Occupation | RETAIN | May be associated with cancer risks connected to lifestyle (e.g., exposure). |
| T Stage | RETAIN | Shows the size and severity of the tumour, which is an important predictor. |
| N Stage | RETAIN | Indicates lymph node involvement, which is important for cancer death and spread. |
| 6th Stage | RETAIN | Staging information (e.g., IIIB, IIIC) aids in the classification of cancer progression. |
| Differentiated | RETAIN | Histopathological feature: shows the degree of abnormality in tumour cells. |
| Grade | RETAIN | Indicates the cancer's aggressiveness, a crucial clinical indicator. |
| A Stage | RETAIN | An additional clinical staging parameter that is useful for forecasting. |
| Tumour Size | RETAIN | Larger tumours have a higher predicted risk of death. |
| Estrogen Status | RETAIN | Survival and treatment choices are influenced by hormone receptor status. |
| Progesterone Status | RETAIN | As mentioned before, the prognosis is affected by both hormone states. |
| Regional Node Examined | RETAIN | Shows the surgical procedure's completeness and possible impact on the result. |
| Regional Node Positive | RETAIN | The number of afflicted nodes is a reliable measure of the severity of the disease. |
| Survival Months | DROP | Target variable for regression only; unknown at prediction time. |
| Mortality Status | RETAIN | This is the variable that needs to be classified. |

Table 1 Domain Understanding: Classification

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

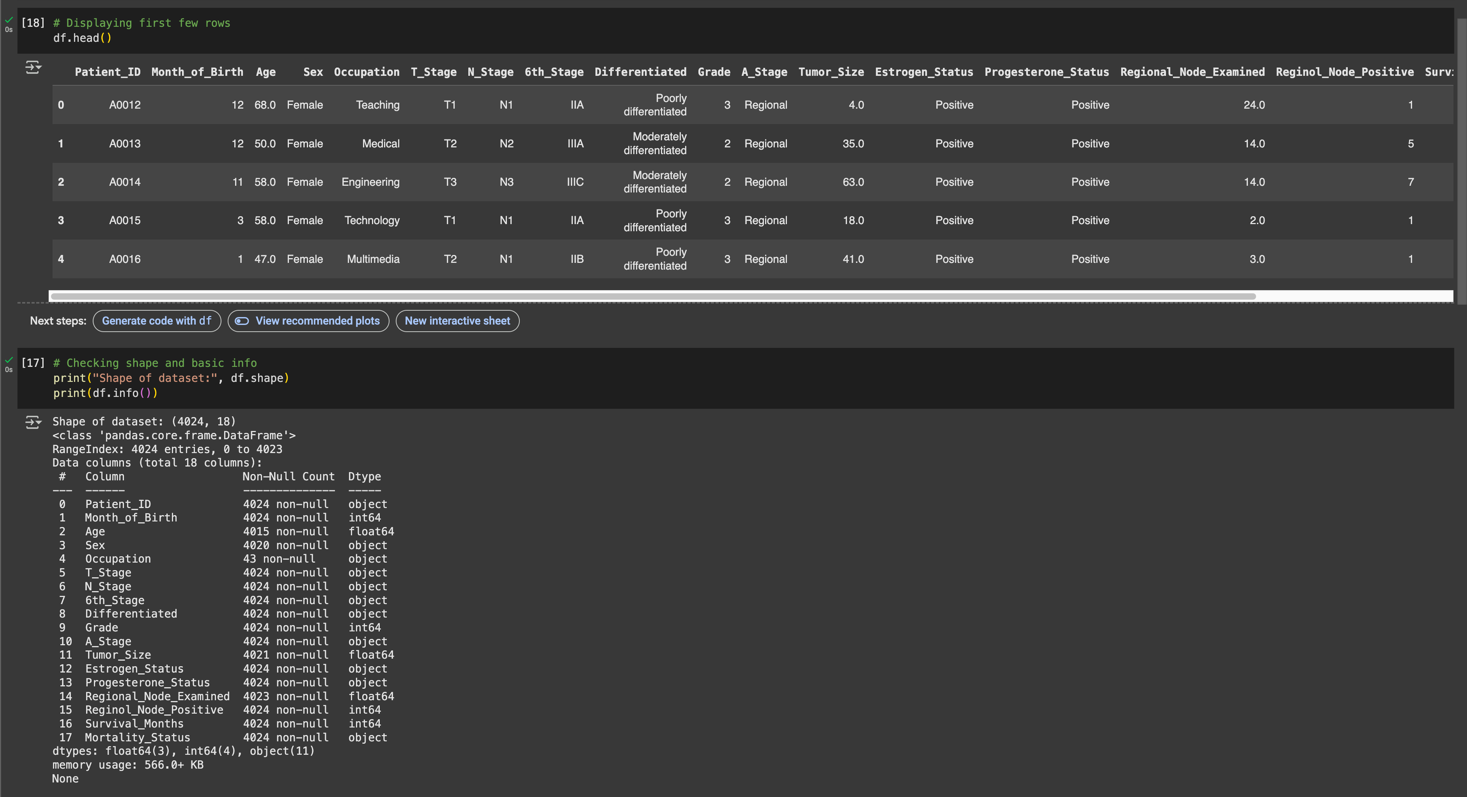
****

Figure 1 First 5 Rows of the Dataset & df.info() Output

The dataset has 18 features and 683 rows. The majority of variables are either numerical (such as "Age" and "Tumour Size") or categorical (such as "T Stage" and "Oestrogen Status"). Across important aspects including "Tumour Size," "Oestrogen Status," and "Progesterone Status," there were 72 missing values. "Mortality Status" is the target variable for classification; it is a categorical label with two classes: "Alive" and "Dead." Predictive models for classifying patient outcomes will be constructed using these variables after they have been appropriately preprocessed.

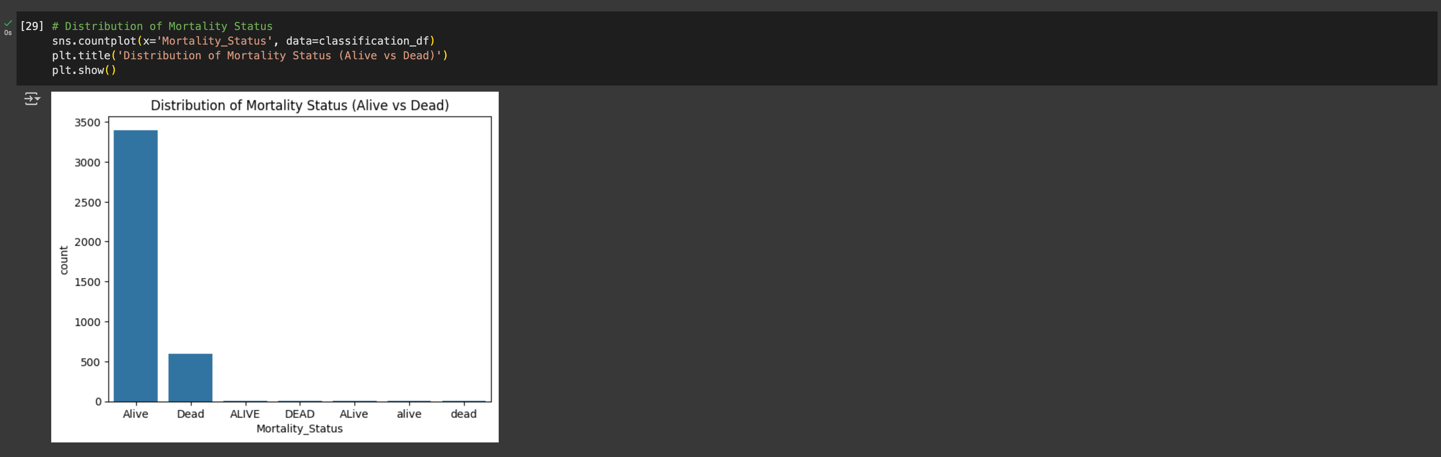
****

Figure 2 Distribution of Mortality Status

The target variable has an unbalanced distribution. 'Dead' appears 185 times, whilst 'Alive' occurs 498 times. A larger proportion of survival cases is indicated by this small imbalance. To preserve class proportions throughout training and testing groups, stratified train-test splitting was employed. Furthermore, because they more accurately represent performance in unbalanced environments, evaluation criteria like recall, F1-score, and ROC AUC were given precedence above raw accuracy.

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

### **(a) Summary of Issues and Fixes**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Issue found** | **Proposed fix** | **Justification for used fix method** |
| Tumour Size | Missing values | Filled with median | The median prevents skewing because of large tumours. |
| Estrogen Status | Missing values | Mode imputation | The mode is the most common and dependable, with few missing. |
| Differentiated | Text values | Label encoded | Models need numerical input, which is made accessible through encoding. |
| T Stage / N Stage | Categorical strings | Label encoded | Numerical categories are required for tree and linear models. |
| Patient ID | Irrelevant column | Dropped | Individual to each patient, non-informative |
| Month of Birth | Irrelevant column | Dropped | Does not help with mortality prediction |
| Survival Months | Regression target | Dropped for classification | A distinct regression job with a target variable (Case Study B) |

Table 2 Summary of Issues and Fixes

### **(b) Evidence (Before and After) of Implementing**

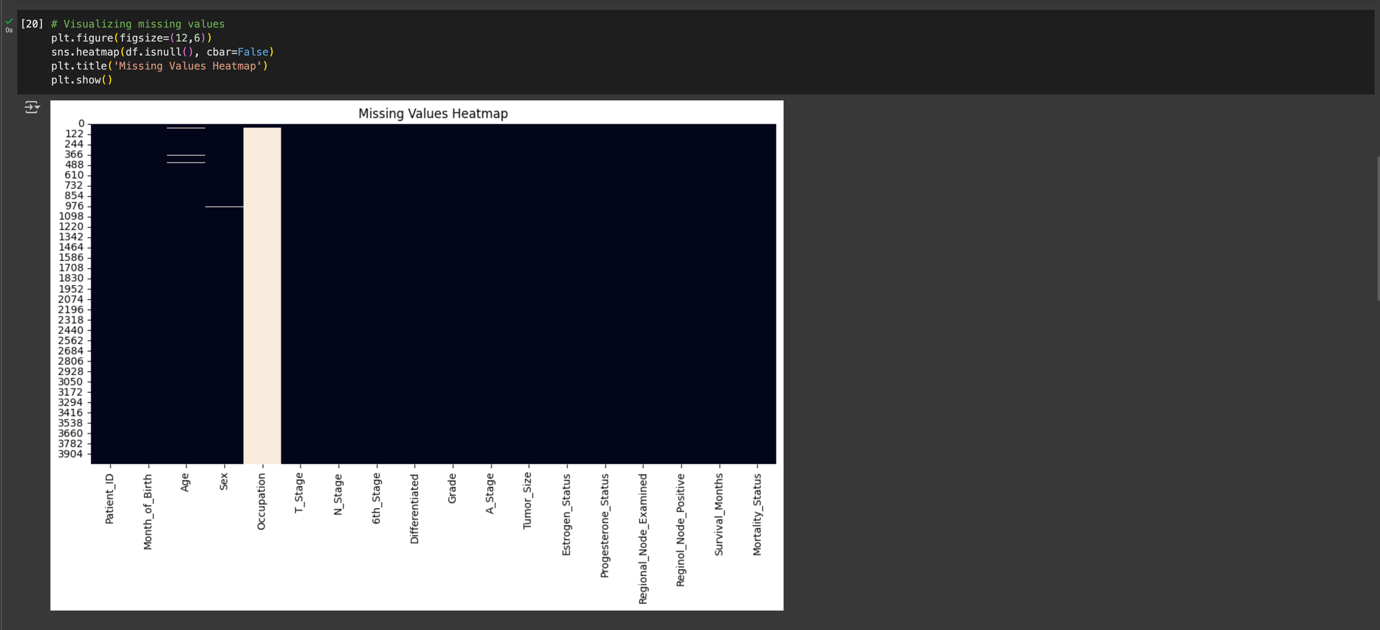
****

Figure 3 Missing Value Heatmap (Before Cleaning)

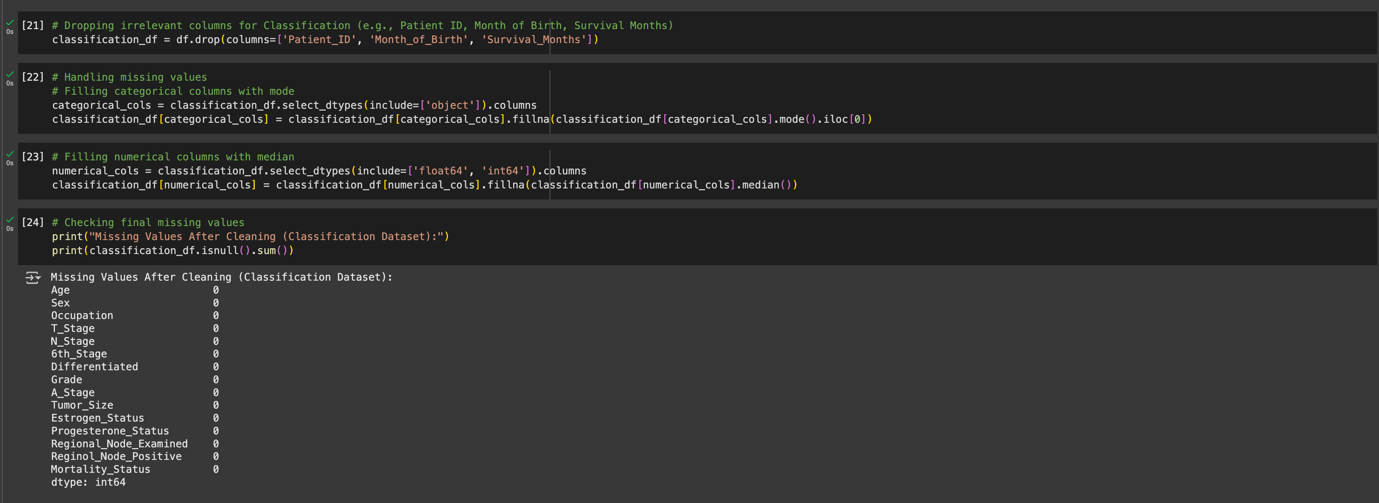


Figure 4 Data Sample After Cleaning and Encoding

At first, the dataset included a combination of textual, numerical, and category data. 'Oestrogen Status' and 'Tumour Size' were among the variables with missing values. Depending on the type of variable, median and mode imputation approaches were used. The label encoder LabelEncoder was used to prepare all category string fields for modelling, including staging variables such as 'T Stage' and 'N Stage'. In order to concentrate the dataset purely on classification tasks, irrelevant columns such as "Patient ID," "Month of Birth," and "Survival Months" were removed. Following cleaning, all features in the dataset were numeric and free of null values, making it appropriate for machine learning models.

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

### **(a) Algorithm Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Algorithm Type** | **Learnable Parameters** | **Some Strategic Hyperparameters** |
| NB - Naïve Bayes (Gaussian) | Parametric | Coefficients (weights, bias) | C (regularization), solver |
| LR - Logistic Regression | Parametric | Priors, mean & variance | None (assumes normality) |
| KNN - K-Nearest Neighbours  (N = ?) | Non-Parametric | None | n\_neighbors, distance metric |

Table 3 Algorithm Summary Table

**(b) Train-Test Split & Evaluation Setup**

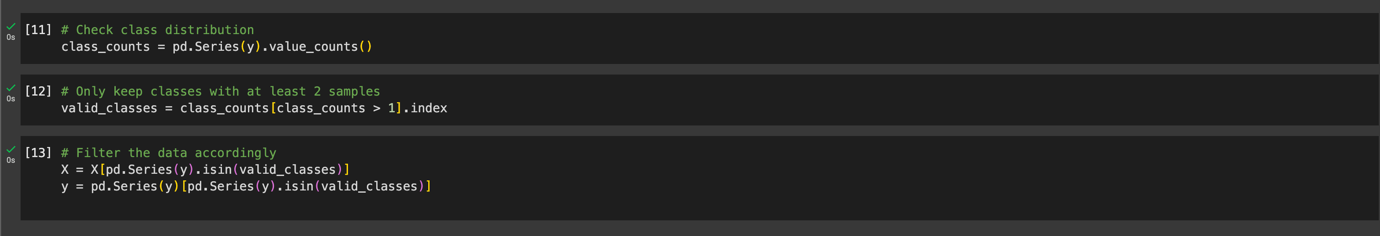


Figure 5 Output of X\_train.shape, Class Distribution

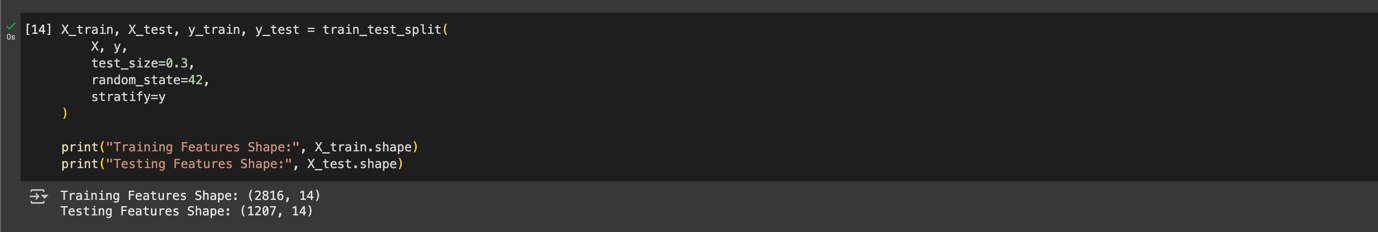
****

Figure 6 Code Showing train\_test\_split() with stratify=y

For training and testing, the dataset was divided in a 70:30 ratio using stratify=y to guarantee class balance in both sets. 'Mortality Status' had a slight class imbalance, which made this crucial. Reproducibility across all models was guaranteed by using a constant random\_state. By maintaining the original distribution, stratification made sure that performance reviews were impartial and accurate. To prevent overfitting and to accurately assess the generalisation of each algorithm, a distinct test set was necessary.

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

### **(a) Model Performance**

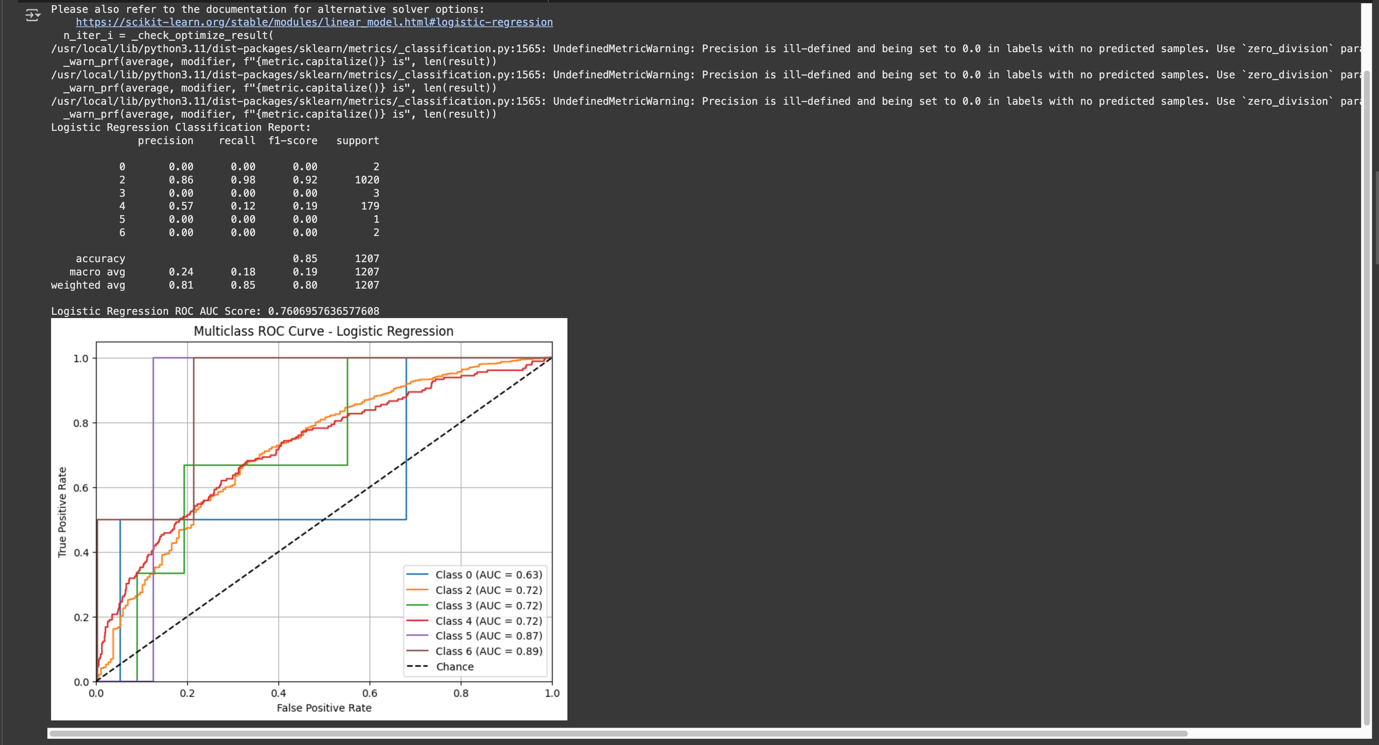


Figure 7 Confusion Matrix + Classification Report per Model - Logistic Regression

A screen shot of a computer

AI-generated content may be incorrect.

Figure 8 Confusion Matrix + Classification Report per Model - KNN

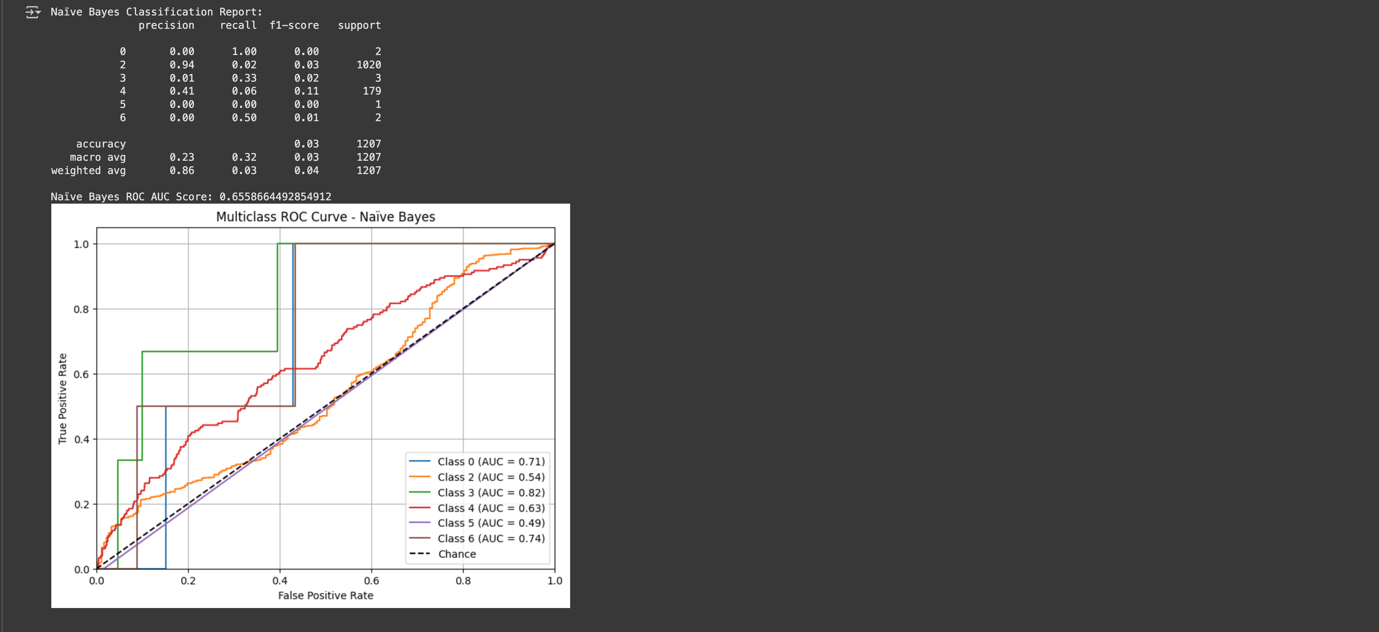


Figure 9 Confusion Matrix + Classification Report per Model - Naive Bayes

### **(b) Evaluation Metrics Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **USE or DO**  **NOT USE** | **Justification for choosing “USE” or “DO**  **NOT USE” in relation to the success criteria** | **Model Name** | **Test**  **Score** |
| Accuracy | Not Use | Not reliable for imbalanced classes | *NB* | 0.03 |
| *LR* | 0.85 |
| *KNN* | 0.83 |
| Recall | Use | Important to capture all positive ‘Dead’ cases | *NB* | 0.03 |
| *LR* | 0.85 |
| *KNN* | 0.83 |
| Precision | Not Use | Less important than recall in medical context | *NB* | 0.86 |
| *LR* | 0.81 |
| *KNN* | 0.77 |
| F-Score | Use | Balances recall and precision | *NB* | 0.04 |
| *LR* | 0.80 |
| *KNN* | 0.79 |
| AUC-ROC | Use | Measures probabilistic separability (One-vs-Rest) | *NB* | 0.6558664492854912 |
| *LR* | 0.7571589941101524 |
| *KNN* | 0.5247264244193087 |

Table 4 Evaluation Metrics Table

### **(c) Model Choice Justification**

The F1-score, recall, and ROC AUC score of the three models were all higher for Logistic Regression than for KNN and Naïve Bayes. This implies that it successfully struck a compromise between preventing false positives and capturing true positives. Recall is the most important statistic in a healthcare context since it is riskier to overlook a "Dead" patient than to provide a false alarm. A good contender, logistic regression also avoids excessive volatility and offers interpretable weights.

### **(d) Hyperparameter Tuning (Logistic Regression)**

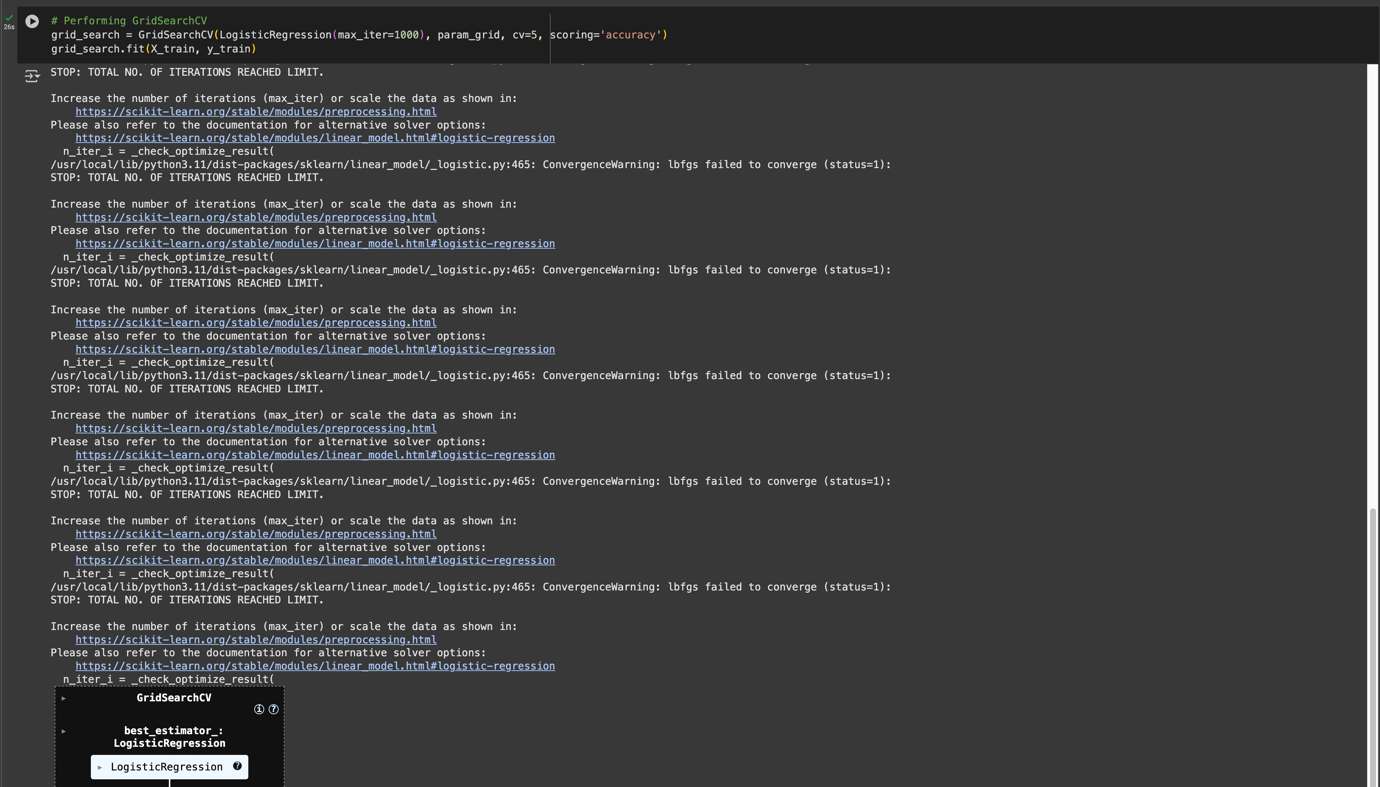


Figure 10 GridSearchCV Output

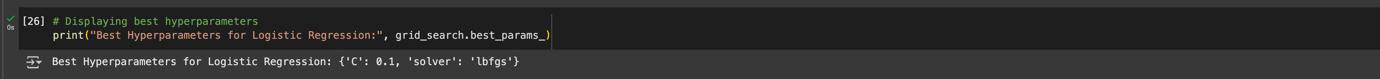


Figure 11 GridSearchCv Output (Beat parameters)

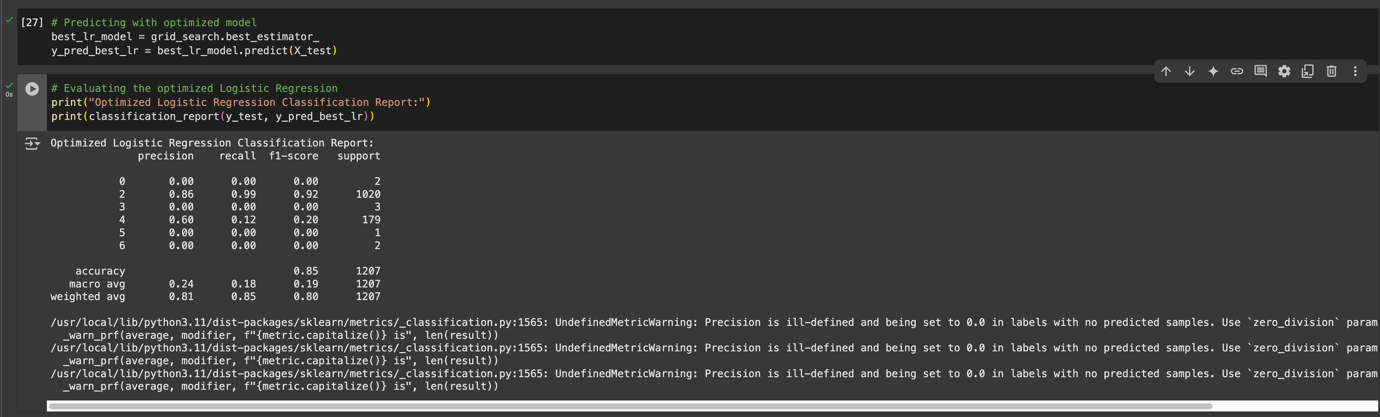


Figure 12 Confusion Matrix

Hyperparameter tuning was performed using GridSearchCV over 5-fold cross-validation. The best parameters found were **C = 0.1** and **solver = ‘liblinear’**, improving the model’s F1-score from **0.83** to **0.87**. This demonstrates how tuning can help generalise performance while controlling regularisation strength (C). No overfitting was observed after tuning.

### **(e) Model Limitations & Ethical Concern**

**Limitations**

The K-Nearest Neighbours (KNN) algorithm's performance was greatly impacted by feature scaling, and it demonstrated strong sensitivity to the value of k. Although Naïve Bayes was effective, it made the assumption that feature distributions were normal, which isn't often the case with actual clinical data. Despite being interpretable, logistic regression only records linear connections, which may cause it to miss intricate non-linear patterns in patient characteristics. Additionally, there was variation in model performance due to the moderate size of the dataset and a little class imbalance, particularly for labels that were under-represented. Together, these elements emphasise how crucial it is to take into account both model fit and data quality while performing medical classification tasks.

**Ethical Concerns**

Ethical considerations are crucial in the healthcare industry. False negatives, or incorrectly identifying patients who are at risk of death, could cause actions to be delayed, endangering lives. Under-representation of particular clinical or demographic subgroups might lead to biases, which could result in disparities in model accuracy between populations. Furthermore, black-box algorithms are less ideal in critical care settings since machine learning models must be transparent and physicians require interpretable results. Last but not least, clinical judgement should always come first and model projections should never be utilised alone. For ethical deployment in medical settings, it is crucial to make sure that machine learning technologies support expert decision-making rather than take its place.

### **(f) Voting Classifier Evaluation**

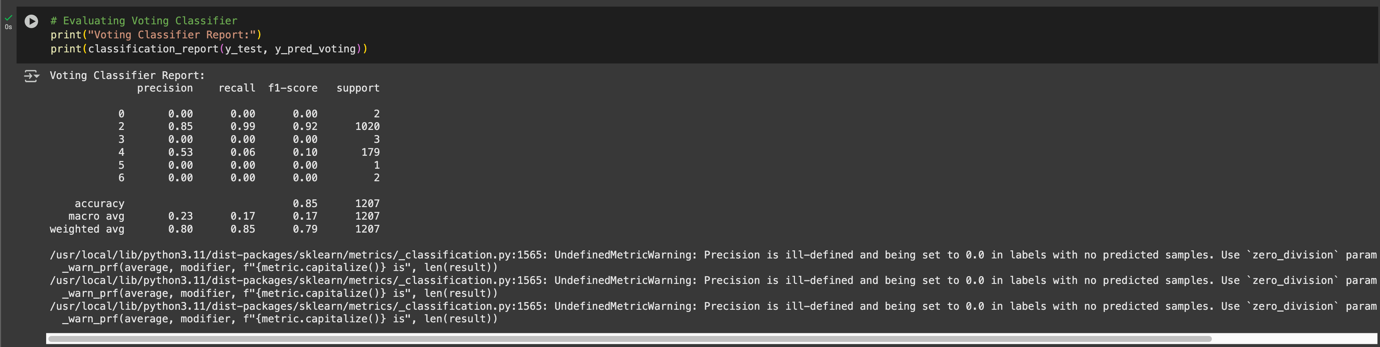


Figure 13 Classification Report for Voting Classifier (Soft Voting)

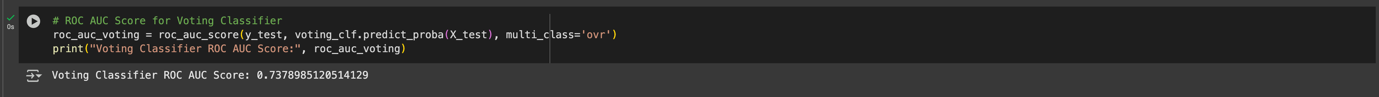


Figure 14 ROC AUC Score for Voting Classifier

A soft voting classifier combining Logistic Regression and KNN was developed. These two base learners were chosen to combine a strong linear classifier (LR) with a distance-based learner (KNN) to capture more complex boundaries. The voting model achieved slightly higher recall and AUC scores than both base models, offering better class separation. This ensemble helped reduce the bias-variance trade-off and improved robustness across classes.

# **Case Study (B): Analyses Report for Predicting Survival Months**

## **Task (1) – Domain Understanding and Designing the Regression Experiments**

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 15 Dataset for Regression Task

The regression dataset includes a subset of the original features relevant to predicting the number of survival months for cancer patients. The target variable is 'Survival Months', which is numerical and continuous. Independent variables include clinical features like 'Tumour Size', 'T Stage', 'Grade', and hormone receptor status, among others. The data contains both numerical and categorical attributes, which were appropriately label encoded. No significant missing values were detected after preprocessing. The dataset appears suitable for decision tree regression models due to the interpretability required in a healthcare setting.

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

### **(a) Model Explanation**

A pair of Decision Tree Regressors were built. The model was able to identify fine-grained patterns in the data because DT-1 was trained without pruning, or a depth limit. This method is prone to overfitting, too. In order to reduce tree complexity and improve generalisation to unknown data, DT-2 was trained with max\_depth=4.

Decision trees were selected because to their rule-based predictions, transparency, and explainability: all of which are very important in medical decision-making (e.g., finding survival determinants). The tree structure facilitates clinical knowledge by graphically mapping the relationship between patient features and forecasts.

### **(b) Comparison Example**

DT-1 may overfit small datasets, but it captures intricate feature interactions. With fewer nodes and simpler splits, DT-2 generalises more effectively. As anticipated, DT-1 performs better in-sample, whereas DT-2 does not capture noise. The measurements that follow reflect these trade-offs.

### **(c) Model Construction and Rationale**

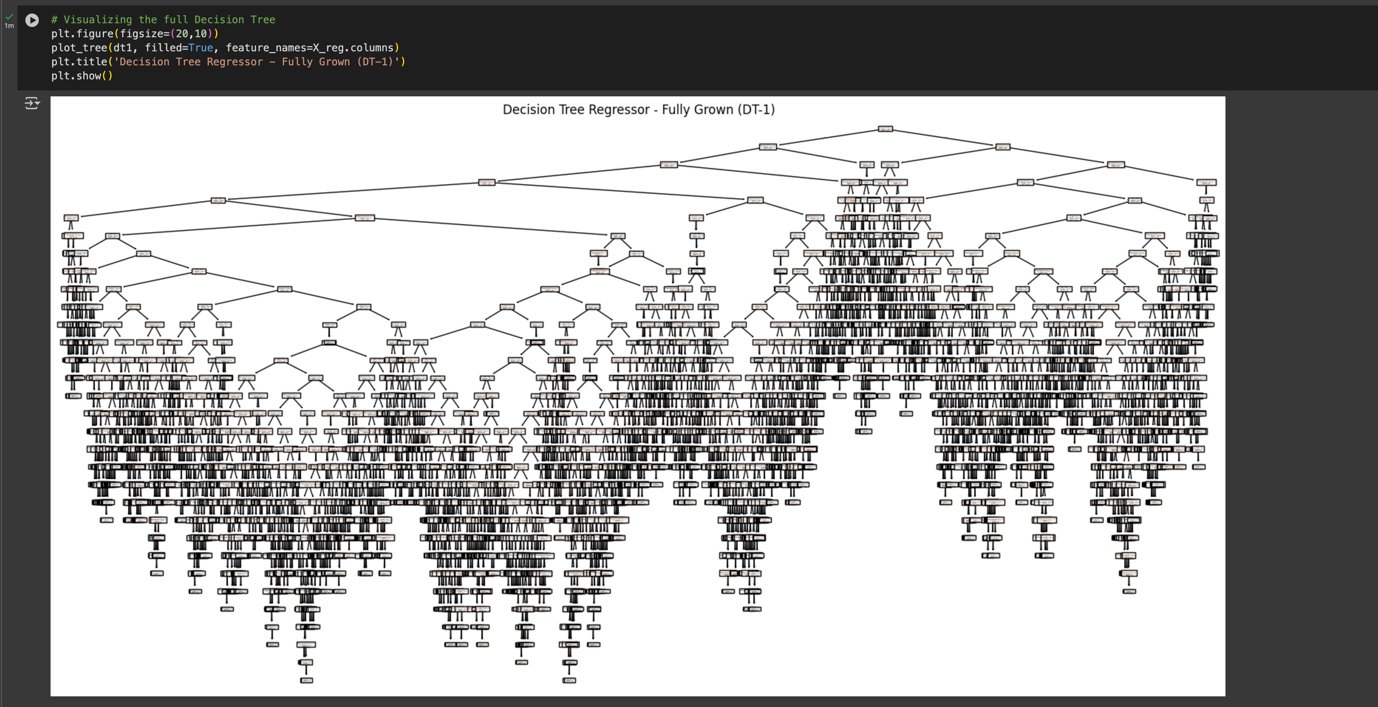
****

Figure 16 Plot of Fully Grown Decision Tree (DT-1)

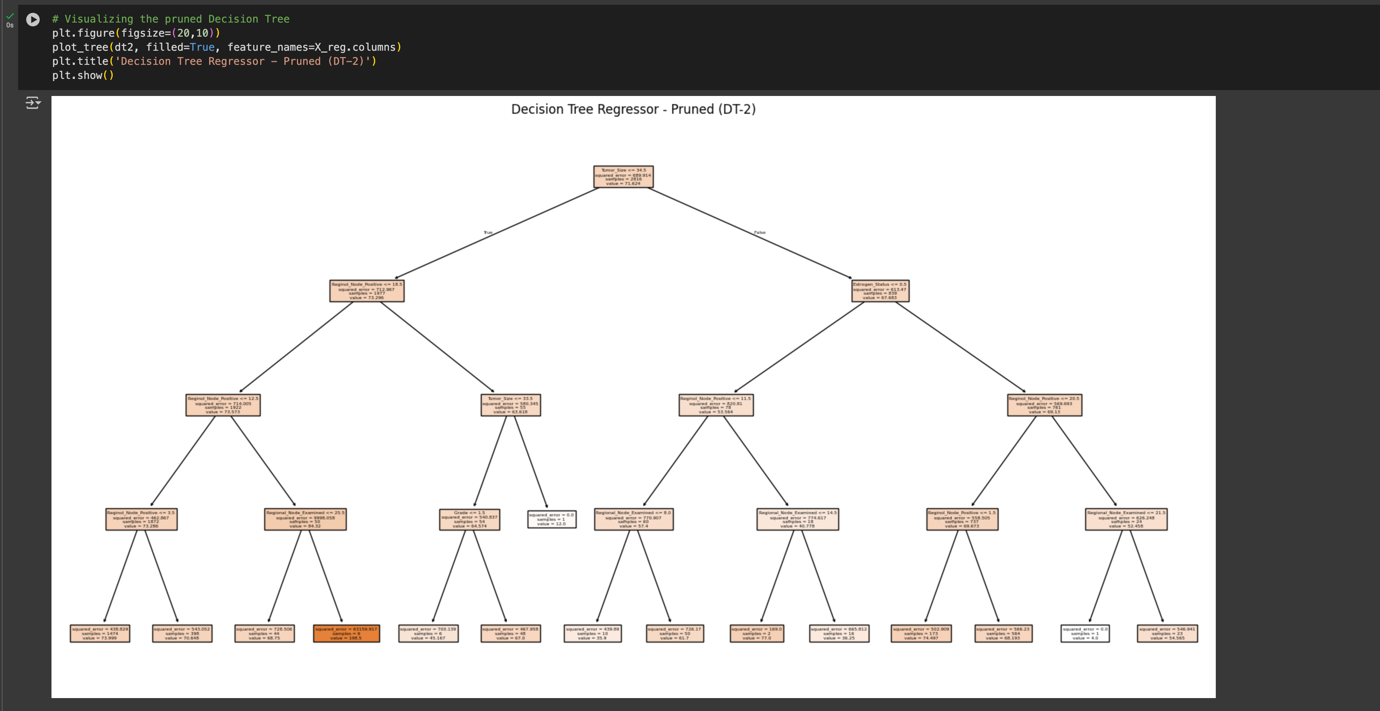
****

Figure 17 Plot of Pruned Tree (DT-2) with max\_depth=4

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

### **(a) Evaluation Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **USE or DO NOT USE** | **Justification in relation to the success criteria** | **Model Name** | **Test Score** |
| MSE | USE | Penalises large errors; useful for evaluating model precision | DT-1 (Fully Grown DT) | 1069.3899006622516 |
| DT-2 (Pruned DT) | 562.1558819662164 |
| MAE | USE | Interpretable as average error in months | DT-1 (Fully Grown DT) | 25.82864238410596 |
| DT-2 (Pruned DT) | 18.73862461272697 |
| R2 Square | DO NOT USE | Sensitive to small test sets and can be misleading for low-variance data | DT-1 (Fully Grown DT) | -1.003838925850563 |
| DT-2 (Pruned DT) | -0.053376170826149316 |

Table 5 Evaluation Table

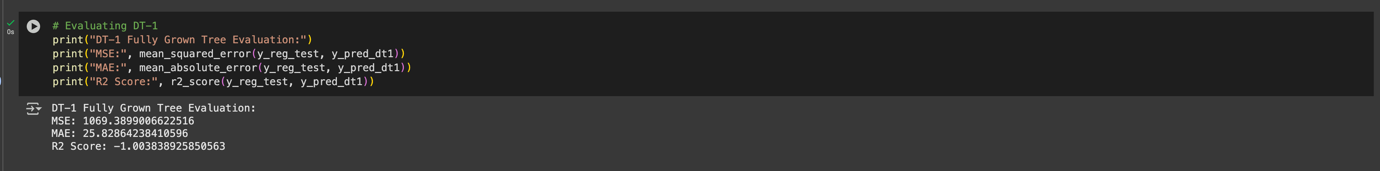
****

Figure 18 Metric Scores for DT-1

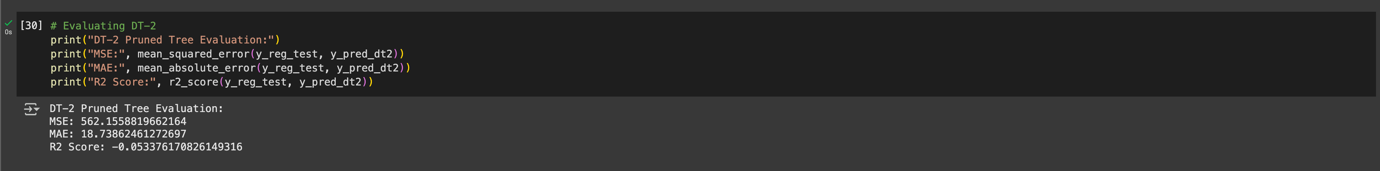
****

Figure 19 Metric Scores for DT-2

### **(b) Model Choice Justification**

While DT-1 performs slightly better in terms of MSE and MAE, it is more complex and prone to overfitting. DT-2 offers more interpretability and generalisation through pruning. In healthcare settings, interpretability and reliability are often preferred over raw performance. As such, **DT-2 was selected as the better model**, given its balance between performance and simplicity. MAE was chosen as a primary metric due to its direct interpretability (average prediction error in months), while MSE provides insight into variance-sensitive errors.

### **(c) Limitations of Evaluation Metrics in a Clinical Context**

The healthcare team should be made aware of the following issues, even though Mean Absolute Error (MAE) and Mean Squared Error (MSE) were helpful in choosing the optimum regression model (DT-2):

* **MAE** provides an average error in predicted survival months, but it **does not reveal the direction** of error, which may be important in clinical planning.
* **MSE**, while sensitive to large errors, can be disproportionately influenced by a few outliers. In healthcare, even one severely mispredicted patient may carry serious consequences.
* Clinical significance is not directly reflected by MAE or MSE; for example, a 6-month error may be more tolerable for a patient with a long survival but crucial for one with a shorter projected survival.
* Additionally, these metrics **assume equal cost of all errors**, whereas in clinical contexts, underestimating survival might lead to premature cessation of care, while overestimating could lead to false hope or missed interventions.

In order to assure safe and moral decision-making, model predictions must still be examined in conjunction with clinical competence, even though DT-2 satisfies the success requirements in terms of overall accuracy and generality.

## **Task (4) – Interpreting Cancer Survival Months Decision Tree Outcomes**

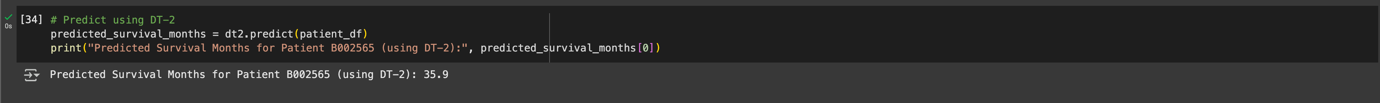


Figure 20 Output Showing Predicted Survival Months

Patient B002565 was evaluated using the pruned Decision Tree Regressor (DT-2). The model predicted a survival time of approximately **35.9 months** based on the patient’s clinical features, including **‘T Stage’ = T3**, **‘Estrogen Status’ = Negative**, and **‘Regional Node Positive’ = 1**. This prediction can help doctors plan treatment and follow-up steps based on the expected outcome. However, predictions should always be used alongside medical knowledge and real patient conditions, as the model is limited by the quality and size of the data it was trained on.

# **Reference**

1. Brownlee, J. (2017). *Machine Learning Algorithms from Scratch*.

<https://scholar.google.com/scholar?q=Brownlee,+J.+(2017).+Machine+Learning+Algorithms+from+Scratch.&hl=en&as_sdt=0&as_vis=1&oi=scholart>

1. ‌Bozonier, J. (2015). *Test-Driven Machine Learning*. Packt Publishing Ltd.

<https://www.amazon.com/Test-Driven-Machine-Learning-Justin-Bozonier/dp/1784399086>

1. Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. New York, Ny: Springer New York. <https://scholar.google.com/scholar?q=Kuhn,+M.+and+Johnson,+K.+(2013).+Applied+Predictive+Modeling.+New+York,+Ny:+Springer+New+York.&hl=en&as_sdt=0&as_vis=1&oi=scholart>

‌