A black background with a black square

Description automatically generated with medium confidence

A black background with a black square

Description automatically generated with medium confidence

**Informatics Institute of Technology**

**University of Westminster**

**5DATA001C. Machine Learning and Data Mining**

**Coursework Analysis Report**

**Student Name : Sundaralingam Gokula Nandhan**

**UOW ID : W2082091**

**IIT ID : 20233027**

**Tutorial Group : L5 SE G9**

# **Table of content**

[**Table of content** 1](#_Toc198081423)

[**Case Study (A): Predicting Cancer Patients Mortality Status** 1](#_Toc198081424)

[**Case Study (A) Analyses Report for Predicting Mortality Status Tasks** 1](#_Toc198081425)

[**Task (2) – Exploring and Understanding the Dataset** 2](#_Toc198081426)

[**Task (3) – Data Preparation: Cleaning and Transforming your data** 2](#_Toc198081427)

[**a) Data Quality Issues in the Cancer Dataset** 2](#_Toc198081428)

[**b) Fixing Data Quality Issues Using Python (with Evidence Screenshots)** 3](#_Toc198081429)

[**Task (4) – Classification Modelling of Cancer Patients Mortality Status** 5](#_Toc198081430)

[**a) Parametric vs Non-Parametric Algorithms and Their Parameters** 5](#_Toc198081431)

[**b) Building Classification Models Using Train–Test Split** 5](#_Toc198081432)

[i. Screenshot Requirement 5](#_Toc198081433)

[ii. Justification of the Training–Test Split Ratio 6](#_Toc198081434)

[iii. Ensuring Consistency and Label Balance in Train–Test Split 6](#_Toc198081435)

[**Task (5) – Evaluating your Cancer Mortality Status Classification Models** 8](#_Toc198081436)

[**a) Evaluation Outputs for Classification Models** 8](#_Toc198081437)

[1.Logistic Regression 8](#_Toc198081438)

[2. K- Nearest Neighbors 8](#_Toc198081439)

[3. Naïve Bayes 8](#_Toc198081440)

[**b) Justification of Evaluation Metrics Based on Success Criteria** 9](#_Toc198081441)

[**c) Best Classification Model Recommendation for Mortality Status** 9](#_Toc198081442)

[**d) Hyperparameter Tuning of the Best Mortality Prediction Model using GridSearchCV** 10](#_Toc198081443)

[i. GridSearchCV Hyperparameter Tuning for Optimized Model 10](#_Toc198081444)

[ii. Evaluation of Hyperparameter Tuning Impact on Best Model 10](#_Toc198081445)

[**e) Critical Evaluation and Ethical Concerns of the Best Mortality Prediction Model** 10](#_Toc198081446)

[**f) Building a Probability-Based Voting Ensemble Classifier** 11](#_Toc198081447)

[i. Ensemble Classifier Declaration and Fitting 11](#_Toc198081448)

[ii. Justification for Base Learner Selection 11](#_Toc198081449)

[1. Logistic Regression 11](#_Toc198081450)

[2.K-Nearest Neighbors 12](#_Toc198081451)

[3. Ensemble Learner Classification Model 12](#_Toc198081452)

[**Case Study (B): Predicting Cancer Patients Survival Months** 12](#_Toc198081453)

[**Case Study (B) Analyses Report for Predicting Survival Months Tasks** 12](#_Toc198081454)

[**Task (1) – Domain Understanding and Designing Your Regression Experiments** 12](#_Toc198081455)

[**Task (2) – Modelling: Build Predictive Regression Models** **a) Benefits of Using a Decision Tree Regressor for Survival Month Prediction** 13](#_Toc198081456)

[**b) Building and Testing Decision Tree Regression Models (DT-1 and DT-2)** 13](#_Toc198081457)

[i. Python Code Blocks to Import, Declare, and Fit DT-1 and DT-2 13](#_Toc198081458)

[ii. Explanation of Pruning Method and Its Impact on Cancer Patient Modelling 13](#_Toc198081459)

[**c) Visualisation of Regression Decision Trees DT-1 and DT-2** 14](#_Toc198081460)

[i.Descision Tree-1(Fully Grown DT) 14](#_Toc198081461)

[ii. Decision Tree -2(Pruned DT) 14](#_Toc198081462)

[**Task (3) – Evaluating your Cancer Survival Months DT Regression Models** 15](#_Toc198081463)

[**a) Regression Evaluation Metrics – Use or Not Use** 15](#_Toc198081464)

[**b) Suggesting the Best Regression Model Based on Performance Metrics** 15](#_Toc198081465)

[**c) Concerns About Selected Performance Metrics** 15](#_Toc198081466)

[**Task (4) – Interpreting Cancer Survival Months Decision Tree Outcomes** 16](#_Toc198081467)

[**References** 17](#_Toc198081468)

**List of Figures**

[Figure 1:Bar graph of mortality status 2](#_Toc198088169)

[Figure 2:Information of the dataset 2](#_Toc198088170)

[Figure 3:Instances of Dataset 2](#_Toc198088171)

[Figure 4:Misspelled variable before and after (Regional\_Node\_Positive) 3](#_Toc198088172)

[Figure 5:Outlier imputation (before and after) 4](#_Toc198088173)

[Figure 6:Numerical value imputation for null values(Before and After) 4](#_Toc198088174)

[Figure 7:Unnecessary label reduction of mortality status (Before and after) 4](#_Toc198088175)

[Figure 8:Labialization for differentiated column (Before and after) 4](#_Toc198088176)

[Figure 9:Converting datatypes of columns (Before and After) 5](#_Toc198088177)

[Figure 10:List of Classification data columns 5](#_Toc198088178)

[Figure 11:Code block of split train and test data 6](#_Toc198088179)

[Figure 12:Code block for getting information about train and test dataset 6](#_Toc198088180)

[Figure 13:Variable distribution of train and test dataset 7](#_Toc198088181)

[Figure 14:Outputs of the above code block 7](#_Toc198088182)

[Figure 15:Classification report, Confusion matrix and ROC-Curve of Logistic Regression 8](#_Toc198088183)

[Figure 16:Classification report, Confusion matrix and ROC-Curve of KNN 8](#_Toc198088184)

[Figure 17:Classification report, Confusion matrix and ROC-Curve of Naive Bayes 8](#_Toc198088185)

[Figure 18:Code block for implementation for GridSearchCV 10](#_Toc198088186)

[Figure 19:Confusion metrics of KNN (Pre and Post tuned) 10](#_Toc198088187)

[Figure 20:Code block for ensemble learner implementation 11](#_Toc198088188)

[Figure 21:Classification report, Confusion matrix and ROC-Curve of Logistic Regression 11](#_Toc198088189)

[Figure 22:Classification report, Confusion matrix and ROC-Curve of KNN 12](#_Toc198088190)

[Figure 23:Classification report, Confusion matrix of Ensemble learner 12](#_Toc198088191)

[Figure 24:Column distribution of regression dataset 12](#_Toc198088192)

[Figure 25:Decision tree regressor declaration of fully grown tree and pruned 13](#_Toc198088193)

[Figure 26:Fully grown decision tree 14](#_Toc198088194)

[Figure 27:Pruned decision tree 14](#_Toc198088195)

# **List of Tables**

[Table 1:Justification table of variable drop and retain for the dataset 1](#_Toc198088591)

[Table 2:Justification table of variable issues 3](#_Toc198088592)

[Table 3:Details of classification algorithms 5](#_Toc198088593)

[Table 4:Evaluation metrics of classification algorithms 9](#_Toc198088594)

[Table 5:KNN report of pre and post tuning 10](#_Toc198088595)

[Table 6:Evaluation metrics of regression tree 15](#_Toc198088596)

# 

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

### **Case Study (A) Analyses Report for Predicting Mortality Status Tasks**

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

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Retain or Drop** | **Brief justification for retention or dropping** |
| Patient ID | Drop | Unique identifier for patients. These are providing non-predictive values |
| Month of Birth | Drop | Risk of overfitting. Does not relevant to the mortality status prediction |
| Age | Retain | Important variable because directly affects cancer survival risk |
| Sex | Retain | Gender – based variable for the breast cancer analysis |
| Occupation | Drop | It lacks direct clinical relevance to cancer survival and may introduce unnecessary bias into the model |
| T Stage | Retain | Reflects the tumor size and severity, which are critical factors in predicting breast cancer mortality status |
| N Stage | Retain | Provides information about lymph node involvement, Signals cancer progression and mortality |
| 6th Stage | Retain | Providing a definitive baseline of established cancer critical for predicting patient mortality |
| Differentiated | Retain | Shows how abnormal cancer cells are compared to normal cells, which means more aggressive cancer and a worse outlook for survival |
| Grade | Retain | Measures cancer aggressiveness and strongly predicts breast cancer patient outcomes |
| A Stage | Retain | Essential as it reveals cancer spread (regional vs. distant), a critical factor for predicting patient survival |
| Tumor Size | Retain | Directly impacts the patient's likely outcome, and smaller tumors often mean better survival |
| Estrogen Status | Retain | Reveals if estrogen drives the cancer. Impacting treatment and survival |
| Progesterone Status | Retain | Often indicates a better outcome and influences treatment in breast cancer |
| Regional Node Examined | Retain | Essential for accurate cancer staging, which directly influences patient mortality prediction |
| Regional Node Positive | Retain | Independent predictor of patient survival and mortality risk |
| Survival Months | Retain | Strong predictors for mortality status. Enhance the accuracy of the model |
| Mortality Status | Retain | Target variable for the mortality status prediction |

Table :Justification table of variable drop and retain for the dataset

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

A screenshot of a computer

AI-generated content may be incorrect.

Figure :Bar graph of mortality status

A screenshot of a computer

AI-generated content may be incorrect.

Figure :Information of the dataset

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure :Instances of Dataset

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

##### **a) Data Quality Issues in the Cancer Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Issue found** | **Proposed fix** | **Justification for used fix method** |
| Regional\_Node\_Positive | Misspelled as “Reginol\_Node\_Positive” | Rename it with the correct spelling | Typos in variable names can create basic errors that become confusing to resolve |
| Age | Null values were found in every variable | All null values were filled by using the mean of each respective variable.(Mode used for Sex variable) | It allows models to work correctly and provide reliable results without losing too much information |
| Sex |
| Tumor\_Size |
| Regional\_Node\_Examined |
| Age | Outliers found in every variable | Replaced maximum amount of outlier values by using mean of each respective variable | It makes the results and predictions more sensible by stopping weird data from messing them up |
| Tumor\_Size |
| Regional\_Node\_Examined |
| Regional\_Node\_Positive |
| Survival\_Months |
| Mortality\_Status | Found output labels in many spellings (E.g : For Alive – Alive, ALIVE, Alive, alive) | Converted all values to lowercase and mapped as “Alive” and “Dead” | Like Alive' and 'alive' as different things, which would make counts and predictions inaccurate |
| Sex | Found all values are assigned with the categorical values | Mapped All the variables to numeric value | It allows machine learning algorithms to optimize based on numeric values. Ensuring the model can learn from the input features effectively |
| T\_Stage |
| N\_Stage |
| 6th\_Stage |
| A\_Stage |
| Eastrogen\_Status |
| Projesterone\_Status |
| Mortality\_Status |
| Age | Found variables with float and object types | Converted all the variables to integer type. For differentiated it was encoded with numerical values | To avoid data type mismatches or errors during model training |
| Regional\_Node\_Examined |
| Tumor\_Size |
| Differentiated |

Table :Justification table of variable issues

##### 

##### **b) Fixing Data Quality Issues Using Python (with Evidence Screenshots)**

A computer screen shot of a computer code

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure :Misspelled variable before and after (Regional\_Node\_Positive)

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Figure :Numerical value imputation for null values(Before and After)

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.

Figure :Outlier imputation (before and after)

A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure :Unnecessary label reduction of mortality status (Before and after)

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure :Labialization for differentiated column (Before and after)

A screenshot of a computer

AI-generated content may be incorrect.A computer screen shot of a computer program

AI-generated content may be incorrect.

Figure :Converting datatypes of columns (Before and After)

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

##### **a) Parametric vs Non-Parametric Algorithms and Their Parameters**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm Name** | **Algorithm Type** | **Learnable Parameters** | **Some Strategic Hyperparameters** |
| NB | Parametric | Class prior probabilities, Feature likelihoods | Smoothing parameter |
|  |
| LR | Parametric | Feature weights, | Regularization strength, penalty type, solver |  |
|  |
| KNN (N=5) | Non-Parametric | None | Number of neighbors, distance metric, weighting |  |
|  |

Table :Details of classification algorithms

#### **b) Building Classification Models Using Train–Test Split**

##### i. Screenshot Requirement

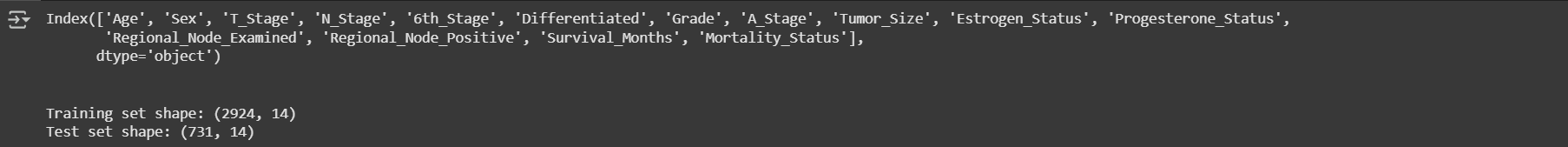


Figure :List of Classification data columns

##### ii. Justification of the Training–Test Split Ratio

An 80:20 training-test split was utilized to ensure a sufficient dataset for model learning while leaving sufficient data intact for proper evaluation. There were 3,655 cleaned records, out of which 2,924 were allocated for training and 731 for testing, such that models could learn meaningful patterns without compromising validation integrity. 80:20 ratio is a commonly proposed strategy in machine learning because it provides a balanced trade-off between bias and variance in performance assessment (Zhang et al., 2019). The approach also helps decrease overfitting and allows the realistic estimation of the model's ability to generalize to new instances. With the moderate size of the dataset, this ratio seems suitable and adequate.

##### iii. Ensuring Consistency and Label Balance in Train–Test Split

A screenshot of a computer program

AI-generated content may be incorrect.

Figure :Code block of split train and test data

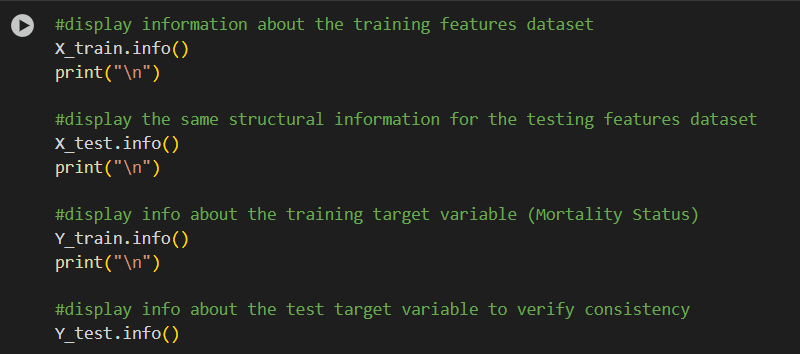


Figure :Code block for getting information about train and test dataset

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure :Variable distribution of train and test dataset

A screenshot of a computer

AI-generated content may be incorrect.

Figure :Outputs of the above code block

To ensure unbiased judgment and similar outcomes between models, a fixed random\_state was used in train\_test\_split(), meaning all models would be trained on and tested with the same set of data subsets. This eliminates randomness as the source of variance in performance. The stratify parameter was also specified as the target variable (Mortality Status: "Alive" vs "Dead") in order to preserve the original distribution of classes. Stratified sampling is necessary for class-balanced classification problems to guarantee that the model learns from a representative sample (Kohavi, 1995).

The info() function is used to verify the integrity and structure of training and test datasets. It checks for the number of rows, columns, data types, and missing values in X\_train, X\_test, Y\_train, and Y\_test. This confirms features are in correct format and missing data is none, ensuring the data is divided correctly and available for model evaluation and training.

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

##### **a) Evaluation Outputs for Classification Models**

##### 1.Logistic Regression

A chart with numbers and a few colored squares

AI-generated content may be incorrect.A graph of a logistic rate

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

Figure :Classification report, Confusion matrix and ROC-Curve of Logistic Regression

##### 2. K- Nearest Neighbors

A chart with numbers and a few colored squares

AI-generated content may be incorrect.A graph of a positive rate

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

Figure :Classification report, Confusion matrix and ROC-Curve of KNN

##### 3. Naïve Bayes

A chart with different colored squares

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

A graph of a positive label

AI-generated content may be incorrect.

Figure :Classification report, Confusion matrix and ROC-Curve of Naive Bayes

##### 

##### **b) Justification of Evaluation Metrics Based on Success Criteria**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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** | Do not use | Can be misleading in imbalanced datasets; may favor the majority class ("Alive") | NB | 0.84 |  |
| LR | 0.9 |  |
| KNN (K=5) | 0.89 |  |
| **Recall** | Use | Measure ability to correctly identify "Dead" cases (true positives) | NB | 0.52 |  |
| LR | 0.47 |  |
| KNN (K=5) | 0.45 |  |
| **Precision** | Do not use | Precision is less important when false negatives (missed dead patients) are more harmful | NB | 0.45 |  |
|  |
| LR | 0.75 |  |
| KNN (K=5) | 0.64 |  |
| **F1-Score** | Use | Balances recall and precision. Useful when there's class imbalance | NB | 0.48 |  |
| LR | 0.58 |  |
| KNN (K=5) | 0.53 |  |
| **AUC-ROC** | Use | Shows model's ability to separate classes at various thresholds. Useful for imbalanced classification | NB | 0.82 |  |
| LR | 0.86 |  |
| KNN (K=5) | 0.78 |  |

Table :Evaluation metrics of classification algorithms

##### 

##### **c) Best Classification Model Recommendation for Mortality Status**

Based on the performance metrics considered - Recall, F1-Score, and AUC-ROC - the KNN (K=5) classifier is chosen as the best model. Though Logistic Regression is marginally better than KNN on all three metrics, the differences are small (within 0.1), and KNN is an acceptable alternative. KNN achieves a good balance between predicting "Dead" cases correctly and not having too many false positives. It also allows for easier interpretation and more flexibility in dealing with nonlinear data. Hence, KNN satisfies the needs of healthcare professionals by delivering stable and balanced performance in identifying high-risk patients with steady overall classification accuracy.

##### 

##### **d) Hyperparameter Tuning of the Best Mortality Prediction Model using GridSearchCV**

##### i. GridSearchCV Hyperparameter Tuning for Optimized Model

A screenshot of a computer program

AI-generated content may be incorrect.

Figure :Code block for implementation for GridSearchCV

##### ii. Evaluation of Hyperparameter Tuning Impact on Best Model

A chart with numbers and a few colored squares

AI-generated content may be incorrect.

A chart with different colored squares

AI-generated content may be incorrect.

Figure :Confusion metrics of KNN (Pre and Post tuned)

##### 

|  |  |  |
| --- | --- | --- |
| KNN Report | Before Tuning | After Tuning |
| Recall | 0.45 | 0.48 |
| F1 – Score | 0.53 | 0.57 |
| AUC - ROC | 0.78 | 0.91 |

Table :KNN report of pre and post tuning

##### **e) Critical Evaluation and Ethical Concerns of the Best Mortality Prediction Model**

Although the performance of the KNN model improved with hyperparameter tuning, it remains far from perfect. It is very sensitive to scaling and distribution of the data, and its performance decreases in the presence of irrelevant features. Further, KNN may not generalize well to new unseen data, especially with class imbalance, a problem with this dataset considering the fewer "Dead" cases. Ethically, KNN can introduce bias against underrepresented patient groups in the training data. Also, false negatives can be costly in healthcare, potentially overlooking patients at risk.

##### 

##### **f) Building a Probability-Based Voting Ensemble Classifier**

##### i. Ensemble Classifier Declaration and Fitting

A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.

Figure :Code block for ensemble learner implementation

##### ii. Justification for Base Learner Selection

Logistic Regression and KNN were chosen as base learners since they complement one another. Logistic Regression is good at  linear relationships, whereas KNN can model complex patterns. Their differences make their combination a well-balanced ensemble that can take advantage of the strengths of both.

###### 1. Logistic Regression

A graph of a logistic rate

AI-generated content may be incorrect.A chart with numbers and a few colored squares

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

Figure :Classification report, Confusion matrix and ROC-Curve of Logistic Regression

###### 2.K-Nearest Neighbors

###### A graph of a positive rate AI-generated content may be incorrect.A chart with a number of colored squares AI-generated content may be incorrect.A screenshot of a computer AI-generated content may be incorrect.

Figure :Classification report, Confusion matrix and ROC-Curve of KNN

###### . Ensemble Learner Classification Model

A chart with a number of colored squares

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Figure :Classification report, Confusion matrix of Ensemble learner

The Voting Ensemble Learner shows somewhat better overall accuracy and better stability in classifying both classes, particularly enhancing recall and F1-score for the minority "Dead" class. These improvements are essential to accurately identify high-risk cases in mortality prediction. It achieves a balanced and stable classification by reducing KNN's risk of overfitting and Logistic Regression's majority class bias. Therefore, the ensemble learner is a better reliable choice to help healthcare professionals attain accurate mortality estimates.

## **Case Study (B): Predicting Cancer Patients Survival Months**

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

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

A screenshot of a computer

AI-generated content may be incorrect.

Figure :Column distribution of regression dataset

##### **Task (2) – Modelling: Build Predictive Regression Models** **a) Benefits of Using a Decision Tree Regressor for Survival Month Prediction**

A Decision Tree Regressor has several benefits for healthcare survival month prediction. It is easy to interpret and visualize, which is beneficial for medical professionals to see the impact of input features (e.g., age, diagnosis, treatment) on survival outcomes. It can handle both numerical and categorical data without requiring normalization or scaling. Decision trees are capable of learning non-linear dependencies and feature interactions, which are common in medical data. They also perform well with missing values and provide feature importance, which can be utilized to identify the most influential factors in survival prediction. These advantages make decision trees an interpretable and valuable tool in healthcare analytics (Loh, 2011).

##### **b) Building and Testing Decision Tree Regression Models (DT-1 and DT-2)**

##### i. Python Code Blocks to Import, Declare, and Fit DT-1 and DT-2

A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.

Figure :Decision tree regressor declaration of fully grown tree and pruned

##### ii. Explanation of Pruning Method and Its Impact on Cancer Patient Modelling

In DT-2, less-complex pruning was applied with max\_depth = 4, which limits the decision tree's growth before its full expansion. This will avoid overfitting as the model is made simpler and less complicated, especially in healthcare data that is susceptible to noise and anomalies. A more shallow tree is preferable for generalization and interpretation in this scenario of cancer patient survival month prediction, where medical professionals can interpret important decision factors.

The key advantage of this technique is that it focuses the model on the most relevant patterns, leading to more stable predictions on new patient data. Additionally, the reduced model complexity speeds up computation and results interpretation in the clinical setting.

However, it has one major drawback: underfitting can result, where meaningful relationships in the information are missed because of the depth limitation. Therefore, pruning at a suitable level needs to be done to maintain the model accurate and interpretable

##### **c) Visualisation of Regression Decision Trees DT-1 and DT-2**

##### i.Descision Tree-1(Fully Grown DT)

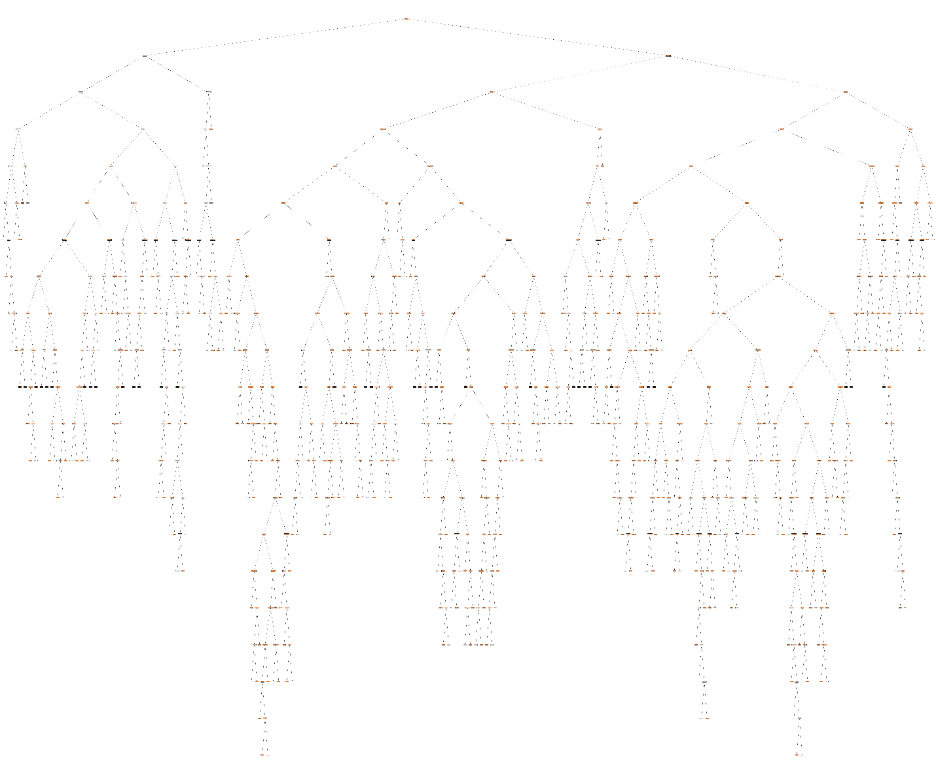


Figure :Fully grown decision tree

##### ii. Decision Tree -2(Pruned DT)

A group of orange and black objects

AI-generated content may be incorrect.

Figure :Pruned decision tree

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

#### **a) Regression Evaluation Metrics – Use or Not Use**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **USE or DO NOT USE** | **Justification in relation to the success criteria** | **Model Name** | **Test Score** |
|  |
| MSE | Do not use | Squares errors, making it sensitive to outliers, which can distort performance in survival predictions with extreme values. | DT-1 (Fully Grown DT) | 0.1477 |  |
| DT-2 (Pruned DT) | 0.0766 |  |
| MAE | Use | Gives average error, is more robust to outliers, and better reflects real-world survival month deviations for cancer patients. | DT-1 (Fully Grown DT) | 0.1477 |  |
| DT-2 (Pruned DT) | 0.1473 |  |
| R-Square | Use | Indicates how well the model explains variation in survival months; higher R² means better prediction power and reliability. | DT-1 (Fully Grown DT) | -0.0914 |  |
| DT-2 (Pruned DT) | 0.4341 |  |

Table :Evaluation metrics of regression tree

##### **b) Suggesting the Best Regression Model Based on Performance Metrics**

Based on the performance metrics used-MAE and R²-DT-2 is the better regression model. DT-2 achieves a lower MAE (0.1473), indicating that it makes lower average prediction errors than DT-1. DT-2 also achieves a positive R² value (0.4341), indicating that it explains about 43% of the variability in survival months, while DT-1 results in a negative R², indicating poor model fit. These results show that DT-2 provides more accurate and reliable survival month predictions, which is very important in helping healthcare professionals with patient prognosis. Pruning also avoids overfitting, leading to better generalization for new, unseen patients, fulfilling the success criteria to a larger degree.

##### **c) Concerns About Selected Performance Metrics**

MSE is not appropriate for medical forecasting because it's outlier-sensitive, which could make model performance biased. MAE is more reliable because it's outlier-robust and provides a clearer reflection of average errors. DT-2 is lower than DT-1 in MAE, hence it's more accurate. R² describes the effectiveness of a model to account for variance, and DT-2 have a positive R² (0.4341) as opposed to the negative R² of DT-1, showing better performance than DT-1. The better model is DT-2, but replication according to clinical advice is recommended.

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

To estimate the predicted survival months of patient B002565, the DT-2 model (Decision Tree Regressor trimmed down to a maximum depth of 4) was employed. It was employed instead of the fully trained DT-1 because it outperformed the latter, as indicated by its lower Mean Absolute Error (MAE) and Mean Squared Error (MSE), along with a positive R² score that indicates a better generalization.

With the aid of the high-resolution graphical representation of DT-2, the B002565 forecast was executed in a series of decision rules as per her clinical features. Features of the patient are:

1. Age coded as 29
2. Sex coded as 2 (Female)
3. T\_Stage coded as 3 (T3)
4. N\_Stage coded as 1 (N1)
5. 6th\_Stage coded as 3 (IIIC)
6. Differentiation coded as 0 (Moderately differentiated)
7. Grade coded as 2
8. A\_Stage coded as 1 (Regional)
9. Tumor\_Size 41
10. Estrogen\_Status coded as 0 (Negative)
11. Progesterone\_Status coded as 1 (Positive)
12. Regional\_Node\_Examined as 5
13. Regional\_Node\_Positive as 1

Decision Path (Rules from the Tree):

1. **6th Stage ≤ 2.5?** → **No** → move right
2. **T Stage ≤ 2.5?** → **No** → move right
3. **Tumor Size ≤ 43.5?** → **Yes** → move left
4. **Estrogen Status ≤ 0.5?** → **Yes** → move left → **Predicted survival months: 53.06**

Using the **pruned Decision Tree Regressor (DT-2)**, patient **B002565** is predicted to survive approximately **53.06 months**. This estimate is based on a series of decision rules derived from the patient’s tumor stage, size, and hormone receptor status.

### **References**

Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), pp.14–23. https://doi.org/10.1002/widm.8 [Accessed 13 May 2025]

Simplilearn, 2021. Model Evaluation in Machine Learning | Machine Learning Tutorial For Beginners | Simplilearn. [video] YouTube. Available at: https://youtu.be/hDKCxebp88A?si=6XQiYppPK5gpfzYi [Accessed 13 May 2025]

DataHeroes.ai, 2024. Unleashing the Power of ML: The Art of Training Models and Its Vital Significance. [online] Available at: https://dataheroes.ai/blog/unleashing-the-power-of-ml-the-art-of-training-models-and-its-vital-significance/ [Accessed 13 May 2025]

Zhang, K., Ye, B., Wu, L., Ni, S., Li, Y., Wang, Q., Zhang, P. and Wang, D., 2023. *Machine learning-based prediction of survival prognosis in esophageal squamous cell carcinoma*. Scientific Reports, 13, Article number: 13532. Available at: <https://www.nature.com/articles/s41598-023-40780-8> [Accessed 14 May 2025].

Kohavi, R., 1995. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Volume 2. Montreal, Canada, 20–25 August 1995. San Francisco: Morgan Kaufmann, pp.1137–1143 [Accessed 13 May 2025].