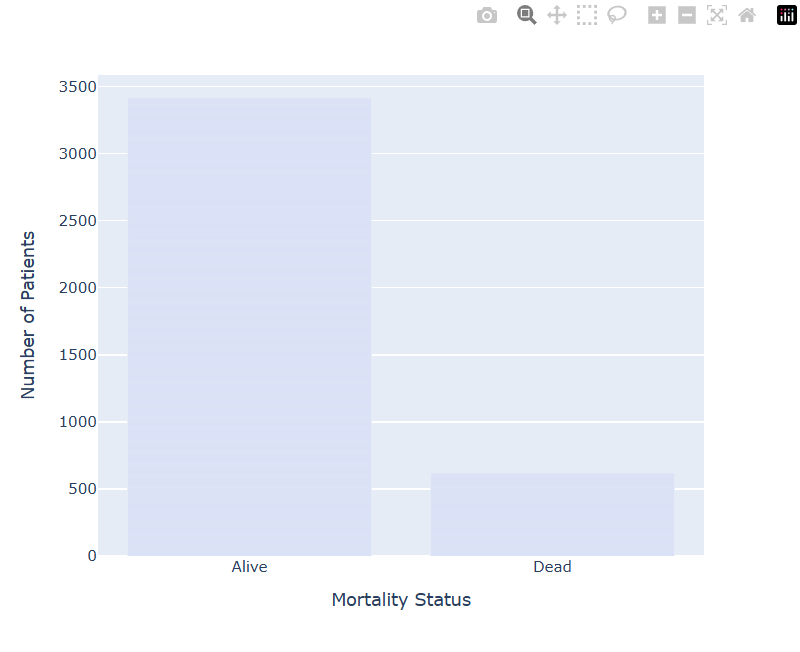
**Case Study (A) Analyses Report for Predicting Mortality Status Tasks**

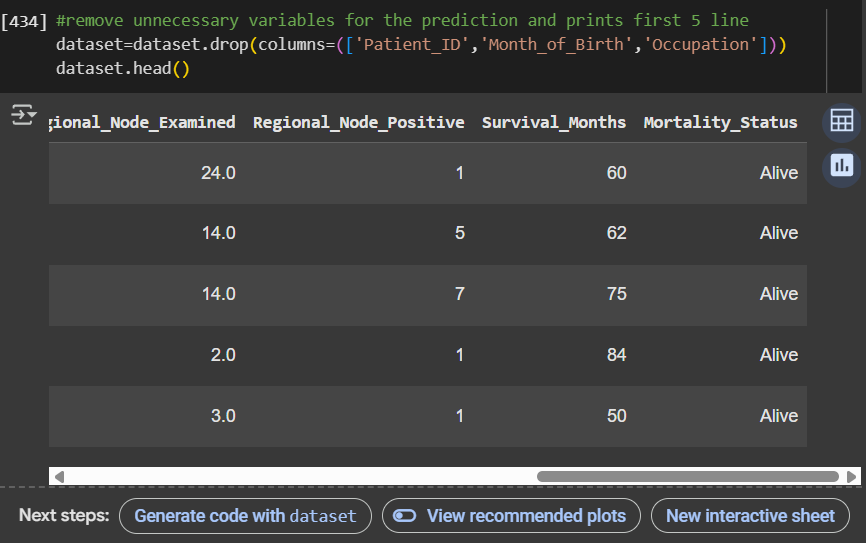
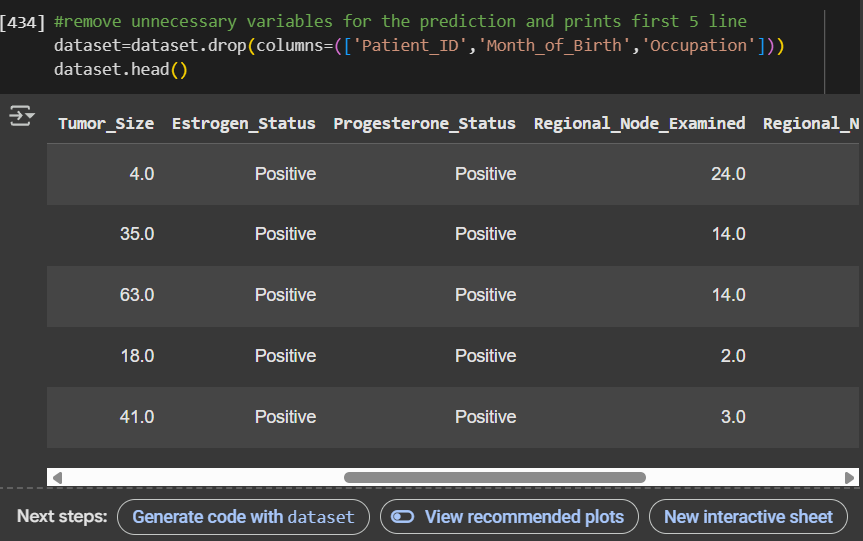
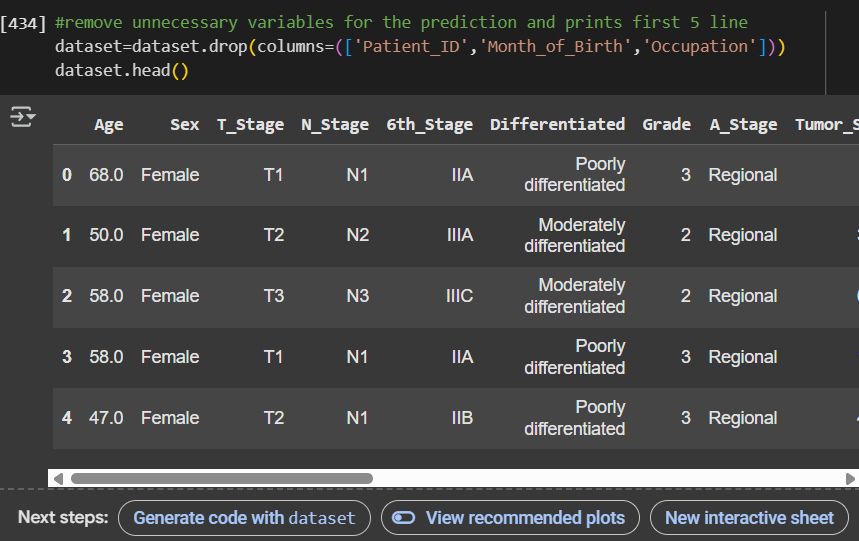
**Task (1) – Domain Understanding: Classification**

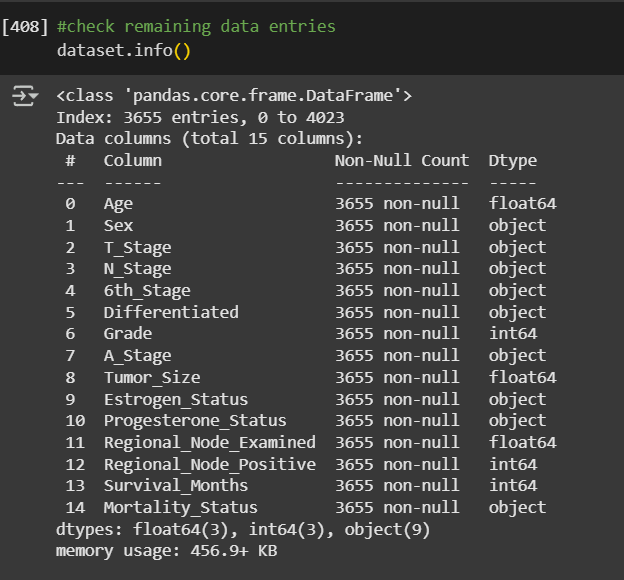
|  |  |  |
| --- | --- | --- |
| **Variable Name** | **RETAIN or DROP** | **Brief justification for retention or dropping** |
| Patient ID | DROP | Unique identifier for patients, These are providing non-predictive values |
| Month of Birth | DROP | Risk of overfitting, does not relevant to the mortality 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. |
| Tumour 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 1: Domain Understanding*

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







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

|  |  |  |  |
| --- | --- | --- | --- |
| 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 2: Data manipulation table*

1. 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.

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 program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

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.

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.

A screenshot of a computer

AI-generated content may be incorrect.

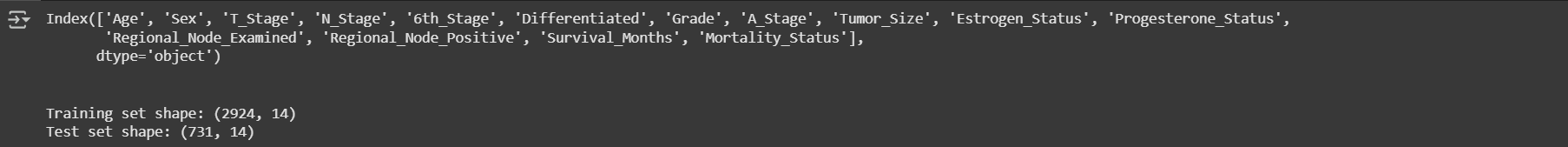
**Task (4) – Classification Modelling of Cancer Patients Mortality Status**



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

*Table 3: Hyperparameters for specific model*

*B)*

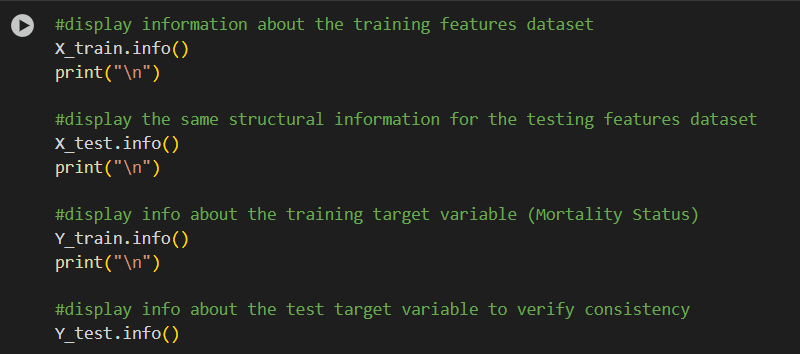
* 1. 

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.

A screenshot of a computer program

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.

A screenshot of a computer

AI-generated content may be incorrect.

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**

* + 1. A screenshot of a computer screen

       AI-generated content may be incorrect.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.

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.  
  
A chart with different colored squares

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

AI-generated content may be incorrect.

b)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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* | 0.90 |
| *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 |

1. 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)

i. A screenshot of a computer program

AI-generated content may be incorrect.

ii.

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.

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

1. 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.
2. i.
   * 1. A screenshot of a computer program

        AI-generated content may be incorrect.

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.

A screenshot of a computer screen

AI-generated content may be incorrect.

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

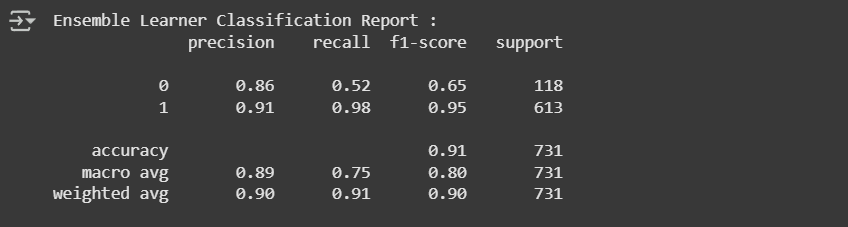
AI-generated content may be incorrect.

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 chart with a number of colored squares

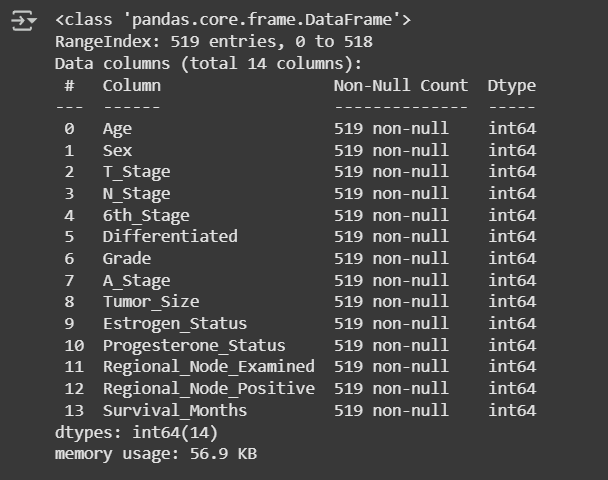
AI-generated content may be incorrect.

Recommendation

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) Analyses Report for Predicting Survival Months Tasks**

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



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



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).

* + 1. A screenshot of a computer program

       AI-generated content may be incorrect.

A screenshot of a computer program

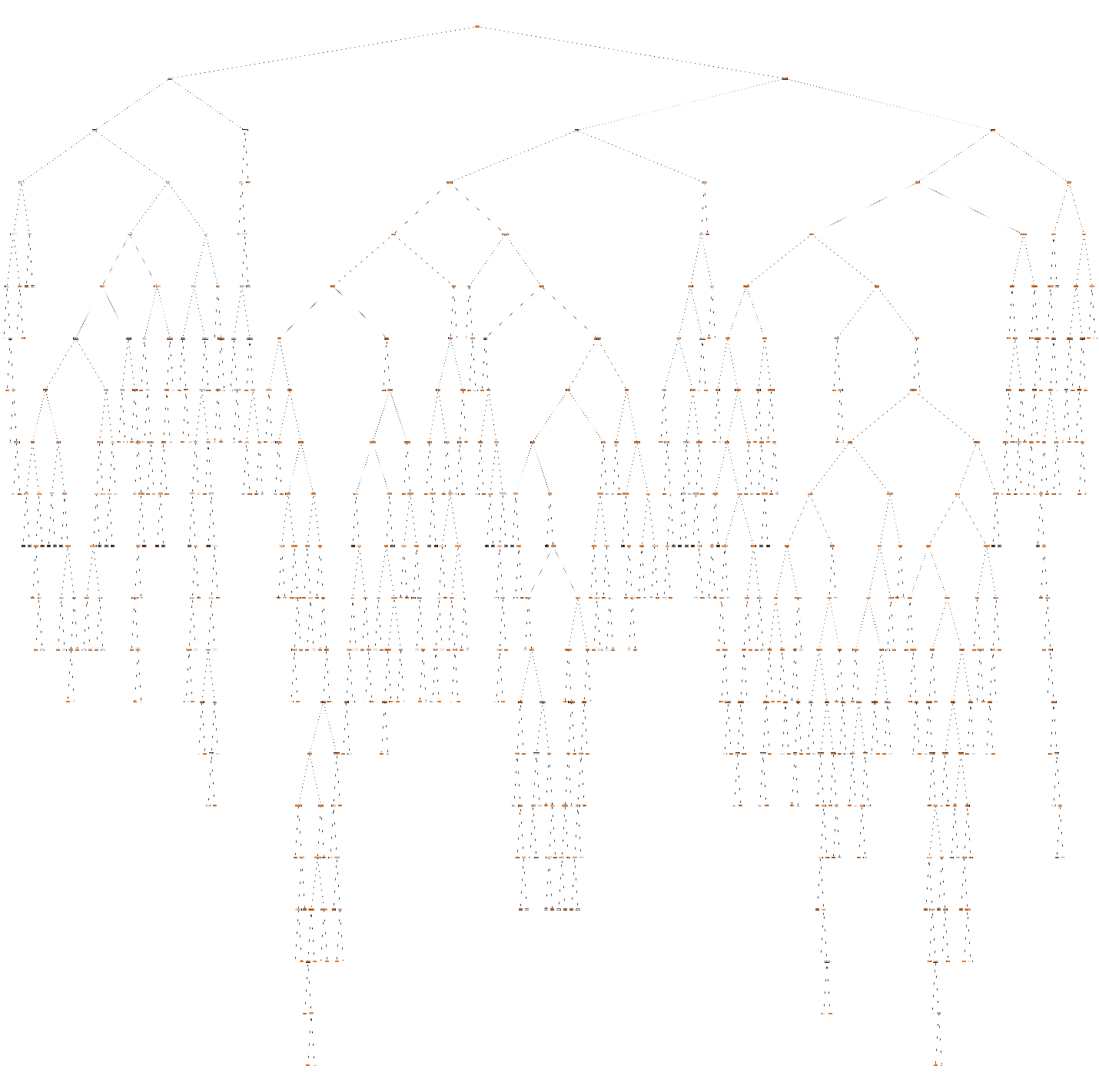
AI-generated content may be incorrect.

ii.

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 the suitable level needs to be done to maintain the model accurate and interpretable.

* 1. 

A group of orange and black objects

AI-generated content may be incorrect.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |
| 0.4341 | 0.4341 |

*Table 5:Regression Evaluation Table*

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 in patient prognosis. Pruning also avoids overfitting, leading to better generalization for new, unseen patients, fulfilling the success criteria to a larger degree.

c)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: Age 29, Sex coded as 2 (Female), T Stage coded as 3 (T3), N Stage coded as 1 (N1), 6th Stage coded as 3 (IIIC), Differentiation coded as 0 (Moderately differentiated), Grade 2, A Stage coded as 1 (Regional), Tumor Size 41, Estrogen Status coded as 0 (Negative), Progesterone Status coded as 1 (Positive), Regional Node Examined 5, and Regional Node Positive 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.