**Report Title**

Donald Philp

01/11/2024

**Table of Contents**

1. **Domain Understanding**
   * Task 1: Classification of Variables
2. **Data Understanding**
   * Task 2: Statistical Summary and Target Distribution
3. **Data Preparation**
   * Task 3: Data Issues and Solutions
4. **Modelling**
   * Task 4: Classification Algorithms and Model Construction
5. **Evaluation**
   * Task 5: Model Performance Analysis
6. **Breast Cancer Survival Rate Prediction**
   * Tasks 1–5: Regression Model Development and Evaluation
7. **References**

**PART A:** **Breast Cancer Mortality Prediction**

1. **Domain Understanding**

**Task 1: Classification of Variables**

* ***Introduction to Variables****: Briefly explain the classification objective, e.g., understanding variables for breast cancer mortality prediction.*
* ***Retained and Dropped Variables****:*
  + ***Table****: Include a table of retained and dropped variables with justification.*

{Insert Table 1}

**Summary:** Summarize your choices for retaining or dropping variables, using concise explanations and in-text citations where necessary.

1. **Data Understanding**

**Task 2: Statistical Summary and Target Distribution**

* ***Statistical Summary****: Provide a summary of the retained variables, including central tendencies and distribution details.*

{Insert Table 2}

1. ***Target Variable Distribution****: Include a screenshot or description of the target variable distribution.*

{Insert Figure 1}

* Describe any initial observations and their relevance for modelling.

1. **Data Preparation**

**Task 3: Data Issues and Solutions**

1. ***Issues Identification****: Summarize data quality issues discovered in the retained dataset.*

{Insert Table 3}

*~~Missing Data Imputation with mean Consistent with dataset structure (Brown, 2021).~~*

*~~Outliers Winsorization Minimizes outlier impact (Taylor, 2020).~~*

1. **Resolution Implementation**: Provide brief explanations of the transformations applied and include screenshots of key results if required.
2. **Modelling**

**Task 4: Classification Algorithms and Model Construction**

1. **Algorithm Summary**: Summarize algorithms, parameters, and hyperparameters used.

{Insert Table 4}

1. **Model Construction**: Describe the training-test split ratio with justification. Include screenshots of feature names and shape outputs.
   1. Take a screenshot of the feature names and the data shape output used in your classification models.
   2. In less than 100 words, research and justify your choice of the training-test split ratio and provide an in-text citation.
   3. In less than 100 words, discuss the overall purpose of using a training-test approach in contrast to the use of validation sets in K-fold cross-validation and describe the case/s when to apply each of those approach is used.
   4. Provide as evidence the code line from your source code that ensures that all models were tested on the same test dataset, also ensure that the labels ratio of Mortality Status “Alive” to Mortality Status “Dead” is the same in the training and test sets.

**Purpose of Training-Test Split: Briefly discuss its purpose and the reason for choosing this method over K-fold cross-validation.**

1. **Evaluation**

**Task 5: Model Performance Analysis**

1. **Confusion Matrix and Evaluation Metrics**: Present confusion matrices for each model and explain metric choices (USE or NOT USE).

{Insert Figure 1-4; one for each model}

1. **Paste the performance Table**

{Insert Table 5}

Accuracy USE High sensitivity aligns with criteria (Doe, 2022). LR 0.85

Recall USE Supports early detection of mortality risk.

1. Best Model Recommendation, based on the ‘USED’ performance metrics scores you identified in Task (5b): Describe the selected model based on the success criteria, with justification for its suitability.
2. Investigate with evidence to establish whether your selected model is a good fit, underfit or overfit.
3. **Retune some models with hyperparameters indicated in task 4a, retrain the algorithm again with GridSearchCV.**
   1. Indicate the number of cross-validation K folds used.
   2. For the newly tuned model, document the estimated best hyperparameters
   3. Present the test confusion matrix for the best models before and after tuning
   4. Calculate the document the new score/s of the “USED” Performance metric/s of your choice to interpret the success criteria identified in task 5.b before and after tuning
   5. Use your observations to **comment on whether the tuning of hyperparameters of your best model improved its positive predictive ability** in line with the success criteria
4. Based on your best model, draft an answer for the research question, criticise your best-performing model, and state any limitations you may have identified. Research and try to explain why your selected algorithm overtook all other models in no more than 100 words. State any ethical issues your model may raise if used to screen for breast cancer mortality.

**PART B: Breast Cancer Survival Rate Prediction**

**Task 1: Domain Understanding: Regression**

**Task 2: Data Understanding: Producing Your Experimental Designing**

**Task 3: Data Preprocessing: Transforming your data**

**Task 4: Modelling: Build Predictive Regression Models**

**Task 5: Evaluation: How good are your Models**

**AI Usage**

Add a table to explain how you used AI

**References**

*(List all references alphabetically in Harvard style)*

Example:

* Doe, J. (2022) *Understanding Data Science*. 2nd ed. London: Sage.
* ~~Smith, A. and Jones, B. (2023) ‘Classification Methods in Healthcare’,~~ *~~Journal of Medical Data~~*~~, 10(3), pp. 205-213.~~