

**Group Coursework Submission Form**

Specialist Masters Programme

|  |  |  |
| --- | --- | --- |
| **Please list all names of group members:**  1. 1. Al-Gaaod, Sadeem: 200013927  2. Lai, Hoang Duong: 240034730  3. Alshaikh, Budour: 240038687  4. Linkevich, Victor: 210028599 | **GROUP NUMBER:**  **2** | |
| **MSc in: Business Analytics** | | |
| **Module Code: SMM636** | | |
| **Module Title: Machine Learning** | | |
| **Lecturer: Dr Rui Zhu** | | **Submission Date: 28/2/2025** |
| **Declaration:**  By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.  We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded. | | |
| **Marker’s Comments (if not being marked on-line):** | | |

**%**

**Deduction for Late Submission: Final Mark:**

## 1. Introduction

The dataset used for this analysis is the Breast Cancer Dataset from Kaggle (Wolberg et al., 1995), comprising 569 observations and 32 features. For illustrative purposes, we have down sampled the dataset to include 500 observations while preserving its predictive integrity.

This dataset is designed for binary classification, aiming to predict whether a tumor is malignant or benign based on various diagnostic measurements. The features encompass mean values, standard errors, and worst-case values of attributes such as radius, texture, perimeter, and smoothness.

This report focuses on leveraging Decision Tree and Random Forest models to develop a predictive framework for tumor classification. Emphasis is placed on optimizing the model’s accuracy in identifying malignant tumors, as early and accurate detection is crucial for effective medical intervention.

**2. Shiny App**

Shiny is an application in Python/R that allows models and graphs to display real time changes to parameters display an active and interactable dashboard. The code is split into 2 main sections, the first deals with the user interface of the model. The second part deals with the rendering and output of the models and graphs.

**Layout and User Interface of the App**

Our Shiny App is designed to be user friendly and minimalistic to avoid overwhelming the user. A clear title for the app ‘Breast Cancer Diagnosis’, and 5 tabs:

1. **Introduction:** gives context to the data and case and provides the objective of the models and more detailed information on what the features of the data represent.
2. **Decision Tree:** displays the visualization of the Decision tree with a scroll to change the alpha parameter (pruning), and the impact on the GINI/ impurity.
3. **Features of Importance:** display what the Random Forest relied on the most for predictions. The number of trees and features used by the random forest are available to change.
4. **ROC:** this curve focuses on the comparison of AUC and model prediction across thresholds, displaying the improved performance of the random forest. The number of trees, features and alpha are all available for adjustment.
5. **Confusion Matrix:** provides an in-depth look at predictions and building on the previous tab parameters, allows you to change the threshold of the model to push the random forest prediction models bias.

## 3. Model Analysis

A Decision Tree makes predictions by recursively splitting data based on features, while a Random Forest combines multiple decision trees to improve accuracy, reduce overfitting, and enhance generalization through ensemble learning.

After cleaning and processing, the data was split into training (80%) and test (20%). The split utilized stratify to improve the balance between classification, creating a final balance of 58:42 as compared to 64:36.

**Creating the Decision tree and Random Forest with AUC maximization.**

To find the best model parameters, we have iterated loops to maximize the Area Under the ROC Curve. The AUC measures the model’s ability to rank positive instances higher than negative ones across all probability thresholds. This is computed in Figure 2; by calculating the Positive and Negative recalls at each threshold for both models. Given that ROC relatively robust to class imbalance compared to other metrics. the AUC provides a more comprehensive overview of each model’s predicative performance when compared to accurate calculations (Mantas et al., 2018).

For the Decision tree to maximize the AUC, we found that optimal parameters were an alpha of 0 (no pruning), and a tree size of 5 as illustrated in Figure 1. The Gini-index for this decision tree was 0 at the final branch splits meaning complete purity in the final classification. The tree size of 5 minimized the test error, and a tree size of six and greater would lead to overfitting.

A graph of a tree size

AI-generated content may be incorrect.

Figure 1. Decision Tree Error Testing

The Random Forest model that maximizes AUC had tuners set to 2, meaning two features would be considered at each split. This increases randomness and reduces overfitting of the model. For this exercise, the number of trees has been set to 500. Although this increases computation time, it helps reduce variance and produces more stable predictions. The model utilizes both Bootstrap sampling and out-of-bag samples to create diverse trees and give an internal validation estimate.

**Comparison between Decision Tree & Random Forest**

Comparing the performance of both models, the ROC curves and AUC of the Decision tree and Random Forest are plotted against each other. The Random Forest is superior across all thresholds for the same randomly sampled test data. The Random Forest and Decision tree returned AUC scores of 0.9918 and 0.9060 respectively.

A graph of a positive rate

AI-generated content may be incorrect.

Figure 2. ROC Curve Comparison of RF and DT

**Interpretation of results using confusion matrices**

The confusion matrices and the respective calculations of recall, precision and accuracy provide a detailed look into how the model performs against the test data.

|  |  |
| --- | --- |
| Confusion Matrix 1: Decision Tree | Confusion Matrix 2: Random Forest |
| |  |  |  |  | | --- | --- | --- | --- | | Truth:  Predicted: | 0 | 1 | Precision | | 0 | 54 | 5 | 92% | | 1 | 4 | 37 | 90% | | Recall | 93% | 88% | Accuracy= 91% | | |  |  |  |  | | --- | --- | --- | --- | | Truth:  Predicted: | 0 | 1 | Precision | | 0 | 55 | 4 | 93% | | 1 | 3 | 38 | 93% | | Recall | 95% | 90% | Accuracy = 93% | |

The 2 Confusion matrixes above illustrate the exact areas of improvements of the Random Forest over the Decision Tree: precision, recall and accuracy all increased by approximately 2%.

However, in the context of this case where 1 represents that a tumor is malignant and 0 represents Benign, we recognized the importance of creating a model that ensures minimizes False Negatives. If a malignant tumor is misclassified as benign (FN), the patient may not receive treatment in time. On account for this we decreased the threshold for malignant classification to 0.1, increasing the sensitivity recall for malignant classification to 100% as shown in confusion matrix below:

|  |
| --- |
| Confusion Matrix 3: Random Forest (0.1 threshold) |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Truth:  Predicted: | 0 | 1 | Precision |  | | 0 | 47 | 0 | 100% |  | | 1 | 11 | 42 | 79% |  | | Recall | 81% | 100% | Accuracy = 89% |  | |

Although Malignant cases are now always recognized, this led to a dramatic decrease in precision and recall in Benign cases (More FP). This leads to a decrease in Accuracy to 89%.

### **A graph of blue and black text AI-generated content may be incorrect.Feature Importance**

Figure 3. Feature Importance by Mean Decrease in Impurity

The most important feature used by the trees was ‘Perimeter worst’, ‘area worst’, and “the mean and worst of ‘concave points’ as displayed in Figure 3. Perimeter Worst measures the largest boundary length of the tumor’s nuclei, whereas Area Worst measures the largest cross-sectional area of the tumor’s nuclei. The Concave points measure the average and most pronounced concave shapes and edges in the tumor’s nuclei.

All these features are highly essential in the classification of a malignant tumor nuclei, as the malignant tumors tend to have large, irregular, and jagged shapes. Moreover, the ‘Worst’ represents the most extreme values in these categories which help identify anomalies.

### **Limitations**

* High feature correlation complicates interpretation by making feature importance rankings less reliable. While tree-based models tolerate correlation, redundant features add computational overhead without improving accuracy. Removing them could enhance interpretability and efficiency.
* Additionally, computational constraints limit model scalability. Excessive features increase processing time without significant predictive gains, making optimization crucial for balancing accuracy and efficiency.

## 4. References

* Abhinav Mangalore (2015). Breast Cancer Dataset [Wisconsin Diagnostic UCI]. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/abhinavmangalore/breast-cancer-dataset-wisconsin-diagnostic-uci?resource=download [Accessed 28 Feb. 2025].
* Mantas, C.J., Castellano, J.G., Moral-García, S. and Abellán, J. (2018). A comparison of random forest based algorithms: random credal random forest versus oblique random forest. Soft Computing. doi:https://doi.org/10.1007/s00500-018-3628-5.
* Shiny for Python. (n.d.). Shiny for Python. [online] Available at: <https://shiny.posit.co/py/>.
* Wolberg, W., Mangasarian, O., Street, N. and Street, W. (1995). UCI Machine Learning Repository. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.