BAYES
BUSINESS SCHOOL
CITY ST GEORGE'S
UNIVERSITY OF LONDON

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Alessandro Rinaldi 2. Lorenzo Rossi 3. Stephane Leboyer	4. 5. 6. 7. **GROUP NUMBER:** 3
MSc in: Mathematical Trading and Finance	
Module Code: SMM748	
Module Title: Machine Learning for Quantitative Professionals	
Lecturer: Rui Zhu	**Submission Date: 4.3.2025**

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: %

Machine Learning for Quantitative Professionals

Alessandro Rinaldi, Lorenzo Rossi, Stéphane Leboyer

March 4, 2025

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Word count: 956

Contents

1 Dataset

This report aims to provide the analysis of the Brest Cancer Wisconsin (Diagnostic) Dataset[1], using a shiny app designed to visualize decision trees and random forests.
The dataset originated from the University of Wisconsin Hospitals, can be used as a machine learning tool to aid in the diagnosis of breast cancer patients; containing 569 observations categorized into 30 features each corresponding to a patient who underwent a fine needle aspiration (FNA) procedure.This dataset provides quantitative measurements of various characteristics of cell nuclei, which are the basis in distinguishing between malignant (M-cancerous) and benign (B-non-cancerous).

The use of a binary classification to predict whether a tumor is cancerous or non-cancerous is crucial in medical diagnostics, as early and accurate detection of breast cancer can significantly improve patient outcomes. The 30 features in the dataset are derived from digitized images of cell nuclei and include three statistical measures for each of the 10 key characteristics of the cell nuclei: the mean, standard error, and worst (or largest) value observed. The 10 characteristics include: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These features are computed for each cell nucleus, and their mean, standard error, and worst values are recorded, resulting in a total of 30 features per observation. The target variable is the diagnosis, which is binary: 1 for malignant (cancerous) and 0 for benign (non-cancerous).

2 App.py

This Shiny app allows users to explore the classification of breast cancer diagnoses using Decision Trees and Random Forest models. The app provides a user-friendly interface where users can select their preferred model type (Decision Tree or Random Forest), criterion (Gini or Entropy), and various visualizations to analyze the model's performance. The app first preprocesses the data by selecting relevant features and encoding the target variable numerically from M/B to 1/0. The dataset is then split into training and testing sets. Based on the user's selection, the app trains either a Decision Tree or a Random Forest classifier and makes predictions on the test data. The app dynamically updates interface components, allowing users to visualize different aspects of the model, including trees, confusion matrices, ROC curves, feature importance, and error rates function of number of trees (for Random Forests only). The plots are generated using Matplotlib and as well as Seaborn for the confusion matrix, offering intuitive insights into the model's behavior and effectiveness. This design allows for interactive exploration of tree-based classification models while providing valuable insights in their performances.

3 Results

3.1 Approach and Model Training

In this study, we analyzed the performance of two classification models—Decision Tree and Random Forest—applied to a dataset with features extracted from medical imaging. Our approach included training both models using two different impurity criteria: Gini impurity and entropy. Additionally, we incorporated cost-complexity pruning (alpha parameter set to 0.01) to optimize the Decision Tree model and prevent overfitting. The analysis was conducted using a Shiny application that visualized model structures, confusion matrices, ROC curves, and feature importance rankings to provide an intuitive understanding of classification performance.

[1] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

3.2 Decision Tree Analysis

The Decision Tree model was evaluated using both Gini impurity and entropy as splitting criteria. The tree structure generated under each criterion exhibited distinct patterns of depth and complexity. With Gini impurity, the tree was relatively shallow, indicating that the splitting decisions were based on minimizing class impurity efficiently. On the other hand, the entropy-based tree exhibited a deeper structure with more splits, reflecting its tendency to account for information gain rather than impurity reduction alone.

The confusion matrices for both approaches showed strong classification performance. Using Gini, the model correctly classified 47 benign cases and 29 malignant cases, with minor misclassifications (7 false positives and 3 false negatives). The entropy criterion slightly improved accuracy, correctly identifying 48 benign and 30 malignant cases, while reducing false positives to 6. The ROC curves demonstrated high AUC values (~ 0.89–0.92), indicating robust predictive ability.

Feature importance rankings revealed that concave_points3 was the most significant variable in both cases, strongly influencing classification. Other key features included area3, concavity2 texture3 and fractal dimension1.

3.3 Random Forest Analysis

Moving to the Random Forest model, which aggregates multiple decision trees, the performance improved significantly in both versions (Gini and entropy). The added randomness and averaging effect of the ensemble approach reduced overfitting while increasing robustness. The confusion matrices showed further refinement in classification, particularly with entropy, where false positive and false negative rates decreased further. The entropy-based Random Forest correctly classified 52 benign and 29 malignant cases, making only 2 false positive predictions.

The ROC curves exhibited AUC values of approximately 0.99, surpassing those of the Decision Tree, demonstrating the increased stability and accuracy of ensemble learning. Feature importance distributions were more balanced compared to the Decision Tree, where a few dominant features had disproportionate influence. The most critical features included concave_points3, perimeter3 radius3, concave_-points1, and area3, with importance values ranging from 0.11 to 0.16. These features encapsulate the contour characteristics of the observed medical structures, supporting their relevance in classification tasks.

3.4 Final Comparison and Interpretation

	Accuracy	Malignant Cases	Precision	AUC	False Pos	False Neg
Decision Tree (Gini)	85.71%	90.63%	80.56%	0.89	7	3
Decision Tree (Entropy)	88.57%	93.75%	83.33%	0.92	6	2
Random Forest (Gini)	90.00%	90.63%	87.88%	0.99	4	3
Random Forest (Entropy)	92.86%	90.63%	93.55%	0.99	2	3

Table 1: Results

Comparing the two models, the Random Forest outperformed the Decision Tree in both accuracy and stability, reducing classification error and demonstrating superior AUC values. The impact of feature importance also differed, with Random Forest distributing importance across more variables, making it more resilient to noise and overfitting. However, while Random Forest generally delivers better predictive reliability, its increased complexity and reliance on multiple trees can lead to overfitting, especially if not properly regularized. The Decision Tree, while more interpretable, exhibited a reliance on a smaller subset of key features, which could lead to higher variance. These findings confirm that while Decision Trees provide an intuitive and explainable model, Random Forest offers enhanced predictive performance, though care must be taken to prevent excessive complexity from negatively impacting generalization.