BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members:
(Surname, first name)
1. Donado Agudelo, Valentina
2. Aliieva, Niiara
3. Farina Alcedo, Blanca
4. Kiefer, Marius
5.
6.
7.

GROUP NUMBER: **6**

MSc in: Business Analytics

Module Code: SMM636

Module Title: Machine Learning

Lecturer: Dr Rui Zhu

Submission Date: 28/02/24

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: %

Decision Tree & Random Forest

Breast Cancer Prediction

SMM636 Machine Learning
Group 6
Words: 1096

Niiara Aliieva, Valentina Donada Agudelo, Blanca Farina Alcedo, Marius Kiefer

February 2025

The Dataset's Task

The Breast Cancer Wisconsin dataset is a widely used resource in machine learning and medical research for classifying breast tumors. It comprises 569 instances, each representing a patient's biopsy, with 30 numerical features computed from digitized images of fine needle aspirate samples of breast masses. These features describe characteristics of the cell nuclei present in the images. The target variable, labeled as '1' for malignant and '0' for benign, indicates whether the tumor is cancerous. The dataset was adjusted to 500 observations while preserving 60/40 class imbalance. This allows us to examine how models handle skewed data and uncertainty.

App Design and How It Works

This app is designed to be accessible to clients with limited knowledge of machine learning. Users can explore decision trees and random forests models, to understand how to classify breast cancer cases based on cell measurements. The app provides an interactive experience with explanations and visualizations that guide users through the process. It features a clean and user-friendly interface, with tabs at the top of the screen that allow users to switch between the two models. Each model has its own dedicated sections that explain and visualize different aspects of the models.

A Decision Tree Classifier is a model that helps classify data by following a series of decision rules. In a decision tree, each node represents a decision based on a feature, the branches show possible outcomes, and the leaves give the final classification or prediction. When setting up a decision tree, there are several ways to control its functionality. In our Shiny app, users can adjust key parameters to influence how the tree is built. One of these is **complexity**, which determines how much the tree will be pruned. Another is **maximum depth**, which limits how many decision levels the tree can have. Lastly, there's **minimum split**, which defines the smallest group of data points that can be split further. By adjusting these settings, users can control whether the tree is general and simple or detailed and complex.

Random Forest creates multiple decision trees using different subsets of data and features. Instead of relying on a single tree, it combines the predictions of all trees, leading to a more reliable and accurate diagnosis. Our app includes adjustable hyperparameters to see how changes affect the model's predictions and performance. The **Number of Trees (ntrees)** parameter controls how many trees are in the forest. A higher number of trees generally improves the model's accuracy but can increase computation time. Users can adjust this option to observe how the prediction results change as more trees are added. The **Number of Variables at Each Split (mtry)** parameter controls how many features are considered when splitting a node in a tree. A smaller value leads to greater diversity among trees but may reduce accuracy. The **Minimum Node Size (nodesize)** parameter controls the smallest number of data points required in a group before the tree stops splitting. A smaller value allows the tree to grow deeper, capturing more detailed patterns but increasing the risk of overfitting. A larger value simplifies the model, making it less prone to overfitting but potentially missing finer patterns. Adjusting this parameter helps balance model complexity and performance. Lastly, the **Bootstrap Sample (replace)** parameter controls whether each tree in the forest is trained on a random sample of the data with replacement (meaning some data points can be selected more than once). Users can switch this setting on and off for comparison.

Both models contain an **Exploration** section that provides an overview of how the model works and how it's applied to classify breast cancer cases. The **Performance** section presents key metrics to evaluate the model's effectiveness. The **ROC Curve** helps assess how well the model distinguishes between benign and malignant cases, where a higher AUC indicates better performance. A perfect model has an AUC of 1, while a model that randomly guesses would have an AUC of 0.5. The **Confusion Matrix** helps users understand where the model makes errors by displaying the counts of true positives, true negatives, false positives, and false negatives. In this context, false negatives are particularly critical, as they mean the model fails to identify cancer when it is actually present, potentially leading to delayed diagnosis and treatment. The **Performance Metrics** include accuracy (percentage of correct predictions), precision (correctly predicted positives out of all predicted positives), recall (ability to identify all positives, with higher recall reducing false negatives, which can be critical in certain applications), specificity (correctly identified negatives, with fewer false positives), and the F1 score (balance between precision and recall).

The **Variable Importance** tab highlights the importance of different input variables in making predictions, helping users understand which factors contribute the most to classifying cancer cases. The **Prediction Tool** tab allows users to train and adjust both the hyperparameters and values of the features to predict the likelihood of cancer.

The Decision Tree model is visualized in **Tree Plot**. Each node in the tree has three rows. The top row represents the predicted class, middle row shows the probability of the majority class in that node and bottom row indicates the percentage of total samples that reach this node.

Analysis of Results

For most parameter choices, the Random Forest Classifier demonstrates better accuracy by combining multiple decision trees. This ensemble approach reduces overfitting (and bias) and creates more robust predictions across variations in training data. The Decision Tree classifier shows high sensitivity to complexity parameters and splitting criteria, with performance fluctuating significantly as these parameters change. This sensitivity often creates performance cliffs where small parameter adjustments lead to dramatic accuracy changes. Random Forest handles imbalanced data better by combining trees built on different data samples, reducing overfitting to the majority class. This explains why predictions show less extreme probabilities compared to Decision Tree results under the same settings. The ROC curves consistently demonstrate that Random Forest achieves higher AUC values across most configurations, indicating superior discriminative power.

On the other hand, confusion matrices reveal that Decision Trees tend toward higher variance in predictions, sometimes excelling with one class while performing poorly with another. However, increased accuracy comes with less interpretability. Unlike Decision Trees, which show clear rules for classification, Random Forest combines many trees, making it harder to follow the path of each decision. Users can visually inspect a Decision Tree to understand exactly how classifications are made, while Random Forest functions more as a ”black box” ensemble.

Ultimately, this comparison allows users to determine whether Random Forest’s performance improvements justify the trade-off against interpretability in their specific use case, while visualizing these differences across various metrics including accuracy, recall and F1 score.