Random Forest Classification Tutorial Report

Breast Cancer Wisconsin (Diagnostic) Dataset

GitHub link: https://github.com/Anirudh2302/breast-cancer-wisconsin-rf-analysis

1. Introduction

The use of a Random Forest Classifier on the Breast Cancer Wisconsin (Diagnostic) Dataset is explained in this report's tutorial. In addition to developing an efficient classification model, the tutorial aims to explain the procedures that were employed in its creation, including the rationale behind each procedure, how it operated in this setting, and the implications for comprehending the dataset and classification model. The process includes training a baseline Random Forest and conducting exploratory data analysis (EDA).
Dimensionality reduction with PCA and NCA.

- feature importance analysis (Gini and permutation importance)
  Hyperparameter behaviour research.
- GridSearchCV-based hyperparameter tuning.
- 2D decision boundary visualization in various feature spaces
  The final assessment of the model

2. Dataset Overview and Exploratory Data Analysis (EDA)

The Breast Cancer Wisconsin (Diagnostic) collection contains 569 samples, each with 30 numerical features taken from digital images of breast mass fine needle aspiration. A two-class classification challenge arises when each sample is categorized as either benign or malignant. The characteristics reveal details about the form and texture of cell nuclei, such as:

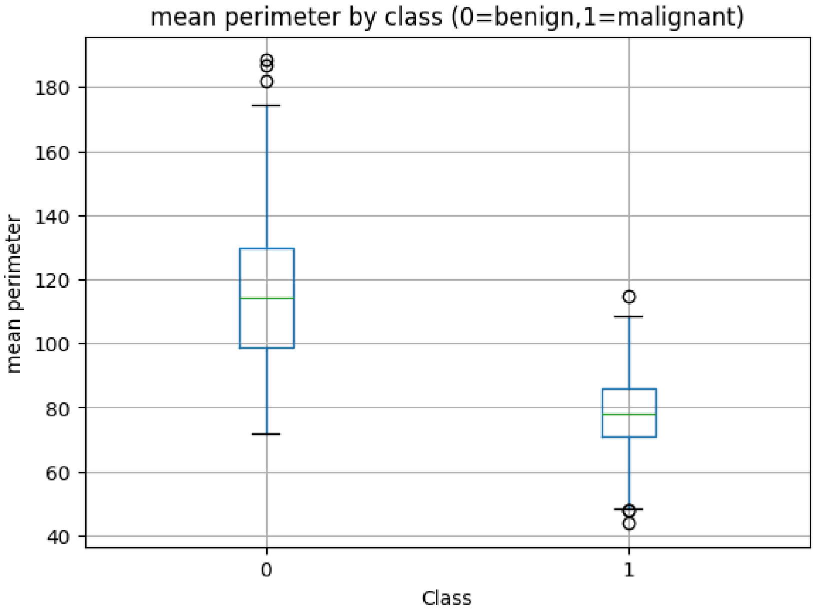
- Mean, Standard Error and Worst (Maximum) for Radius, Texture, Perimeter, Area
- Geometry Features Including: Smoothness, compactness, concavity and symmetry.

As reported by the exploratory data analysis (EDA), EDA is a great starting point for developing an understanding of

- The differences between classes based on the features
- The features which may have predictive capabilities.
- The existence of a high level of correlation (or redundancy) among variables.

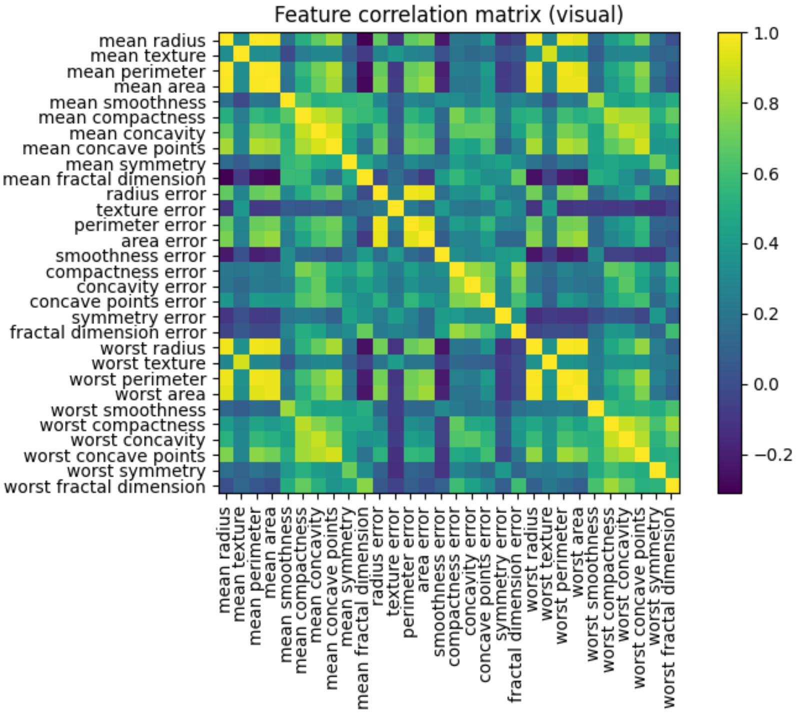
Boxplots of the samples "by class" show that, in comparison to benign samples, malignant tumours typically exhibit larger values of the geometry-based attributes (or mean radius and mean perimeter). This suggests that the two groups of breast samples might be distinguished using the geometry-based metrics.

Fig 1 Boxplot – mean perimeter by class (benign vs malignant)



A correlation matrix revealed that most geometric feature categories, such as radius, perimeter, and area, have strong connections with one another. Many geometry-related metrics have excellent correlations with one another. Some concentricity and texture measurements also indicated strong connections. The correlation structure will have a big impact on PCA since Random Forest will provide more insight into the link between these features, and PCA uses the greatest variance and highest correlated features, which will correlate with these PCA-related characteristics.

Fig:2 Correlation matrix heatmap



3. Random Forest Model: Generic Explanation and Rationale

Random Forest is an ensemble learning technique that makes use of different decision trees. Every decision tree uses a randomly chosen subset of the input variables at each split and is constructed using a bootstrap sample of the dataset. Each decision tree votes on the final class label when used for classification, and the Random Forest model returns the class with the majority vote. Random Forest models have a few built-in benefits:

- Random Forests provide fewer wrong predictions than individual decision trees because they combine forecasts from multiple trees.
- Because the output of a Random Forest is a compilation of forecasts from numerous trees, they are far less likely to generate inaccurate predictions than the output from any one decision tree.
- RF can naturally represent intricate, previously unseen interactions between predictor variables as well as nonlinear correlations between the predictor's properties and the dependent variable.
- They offer an automatic evaluation of feature relevance for any prediction model built with them; they can effectively handle tabular data that contains both tiny and large values.

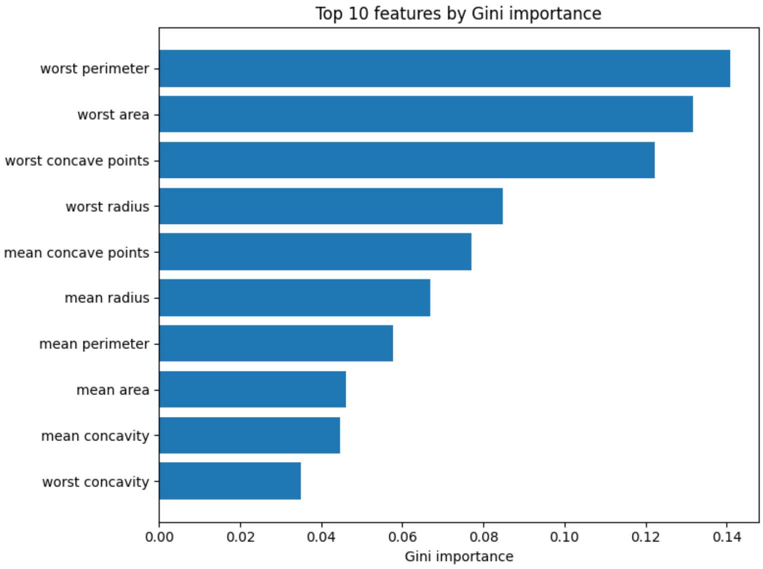
Since the dataset's structure is tabular and all its components are numerical, RF seems suitable for this dataset

- The dataset has a tabular structure, and all its components are numerical.
- There will probably be nonlinear correlations between each feature's values and the target class.
- Feature importances are helpful in interpreting the models because the model will be employed in a medical context.

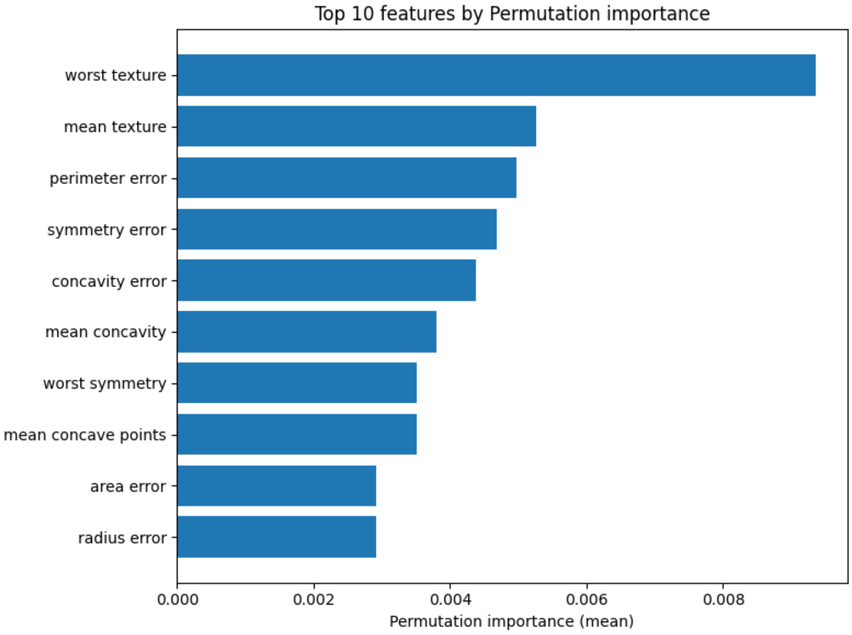
4. Feature Importance Methods and Results

Understanding the features Random Forest employs for prediction is critical to improving model interpretability and mechanism. As a result, the Random Forest employs two distinct methods to assess feature relevance.

Initially, the construction of trees is used to determine Gini significance, which quantifies the amount of impurity (Gini) lowered by each characteristic across all splits and trees. Gini importance is usually easier to compute and requires fewer resources than permutation importance because it is measured as trees are constructed. Gini significance, however, might Favor some feature types over others.



Furthermore, Permutation Importance can be computed after training because it is model-agnostic and only needs the Random Forest model to be learned. This method entails figuring out how much accuracy is lost when a single feature's values are randomly rearranged. Permutation Importance, in contrast to Gini Importance, is a direct indicator of how a feature is expected to impact the overall performance of the model.



Similar feature categories are indicated by both Gini and Permutation Importance, although attributes pertaining to average or maximum perimeter, concavity/concave points, and texture/area are especially well-represented. Because Permutation Importance gives a clear indicator of how much a feature contributes to overall model performance, it is utilized to choose the top two features to include in upcoming raw 2D visualizations.

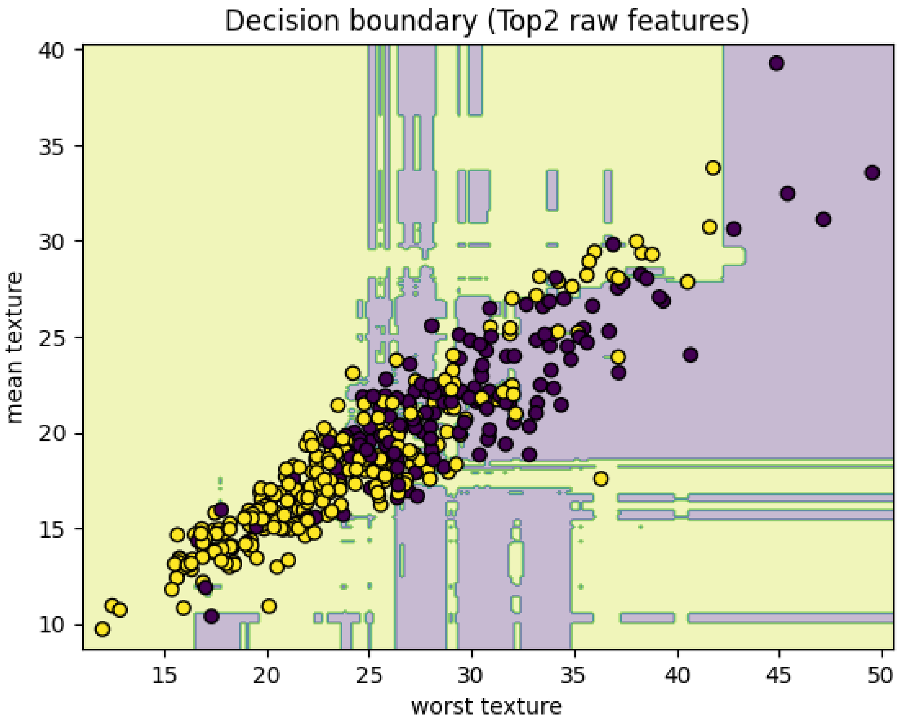
5. 2D Decision Boundaries Using Top Two Raw Features

The study's goal was to construct two-dimensional visual representations of the models based on the two aspects determined by permutation importance. To produce a representative visualization of its prediction bounds, a 2D Random Forest was constructed. This made it possible to visualize where predictions fall within 2D coordinates while streamlining the entire dataset.

The result is a scatter plot where:

- Benign and malignant observations are represented by points in the 2D scatter plot.
- The unusual shape of the decision boundary, which has numerous edges and denotes where the splits were made along each axis because of the nature of decision trees; the background is coloured according to the Random Forest model prediction classes translated to the feature space.

Raw Decision Boundary top 2 raw features



Observations:

- The classes are reasonably separated but still overlap in some regions.
- Two features alone cannot perfectly capture the full class structure, but they are informative enough to illustrate how the Random Forest partitions the space.

This provides a baseline for comparison with PCA and NCA projections.

6. PCA (Principal Component Analysis) – Generic and Applied

Generic explanation:

PCA is an unsupervised dimensionality-reduction approach. It identifies new orthogonal axes (principal components) that capture the most variance in the data. The first few components can frequently explain much of the variance, allowing for lower-dimensional representations.

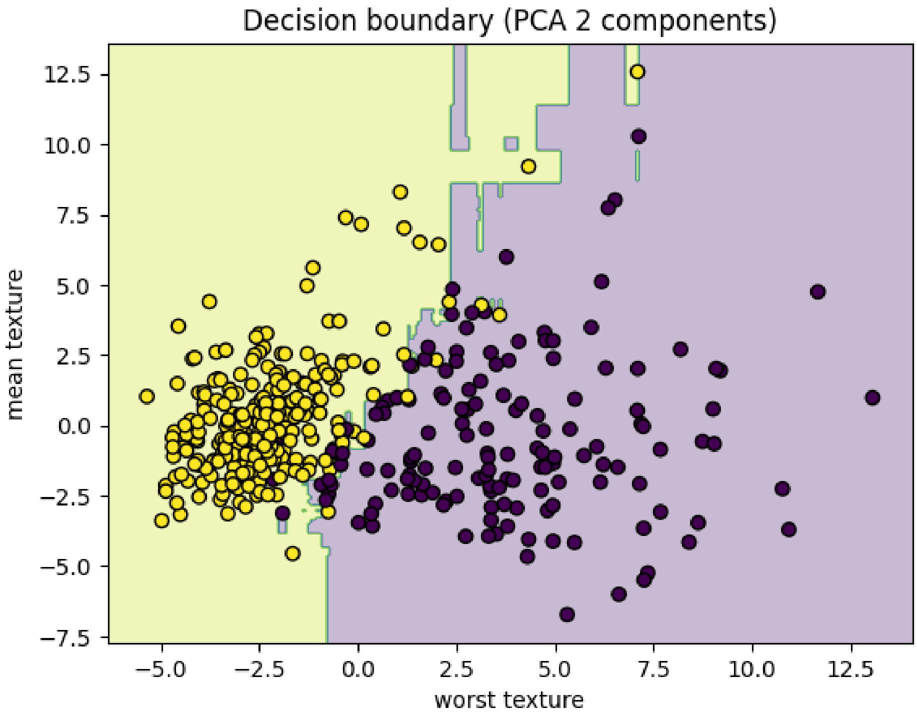
Key points:

- PCA is variance-driven, not separation-driven, and does not require class labels. • Features with high correlation, such as radius, perimeter, and area, frequently dominate the initial components.

In this project:

- The features are standardised before applying PCA.
- The data are projected onto the first two principal components.
- Training a Random Forest on 2D PCA features and visualizing the decision boundary.

PCA Decision Boundary plot



Interpretation:

- PCA distributes data based on variance, revealing some structure. • However, binary classes still overlap due to the lack of classification optimization.
- The decision border is smoother than the raw top two feature boundary, but it is still not perfectly separated.

7. NCA (Neighbourhood Components Analysis) – Generic and Applied

Generic explanation:

NCA is a supervised linear dimensionality reduction method. Instead of maximizing variance, NCA learns a projection that enhances the classification performance of a nearest-neighbour classifier. It effectively seeks to bring same-class points closer together while pushing different-class points farther away in the projected space.

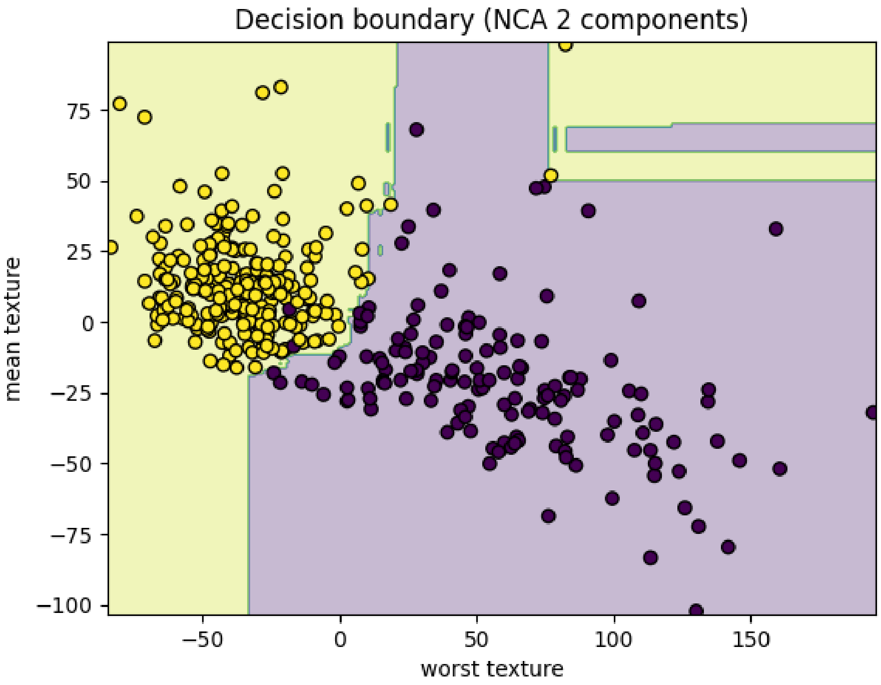
Key points:

- NCA utilizes label information directly.
- It focuses on enhancing class separability rather than PCA.

In this project, features are scaled before NCA.

- Features are scaled before NCA.
- NCA is applied with 2 output components.
- The 2D NCA space is used to train a Random Forest and define its decision limit.

NCA Decision Boundary plot]



Interpretation:

- The NCA projection produces much clearer class separation than both raw top two features and PCA.
- Decision regions are more compact and intuitive, and the Random Forest achieves higher accuracy in this 2D space.

This comparison highlights the difference between unsupervised (PCA) and supervised (NCA) projections and shows why NCA is often better for classification-focused visualisation.

8. Hyperparameter Behaviour Study

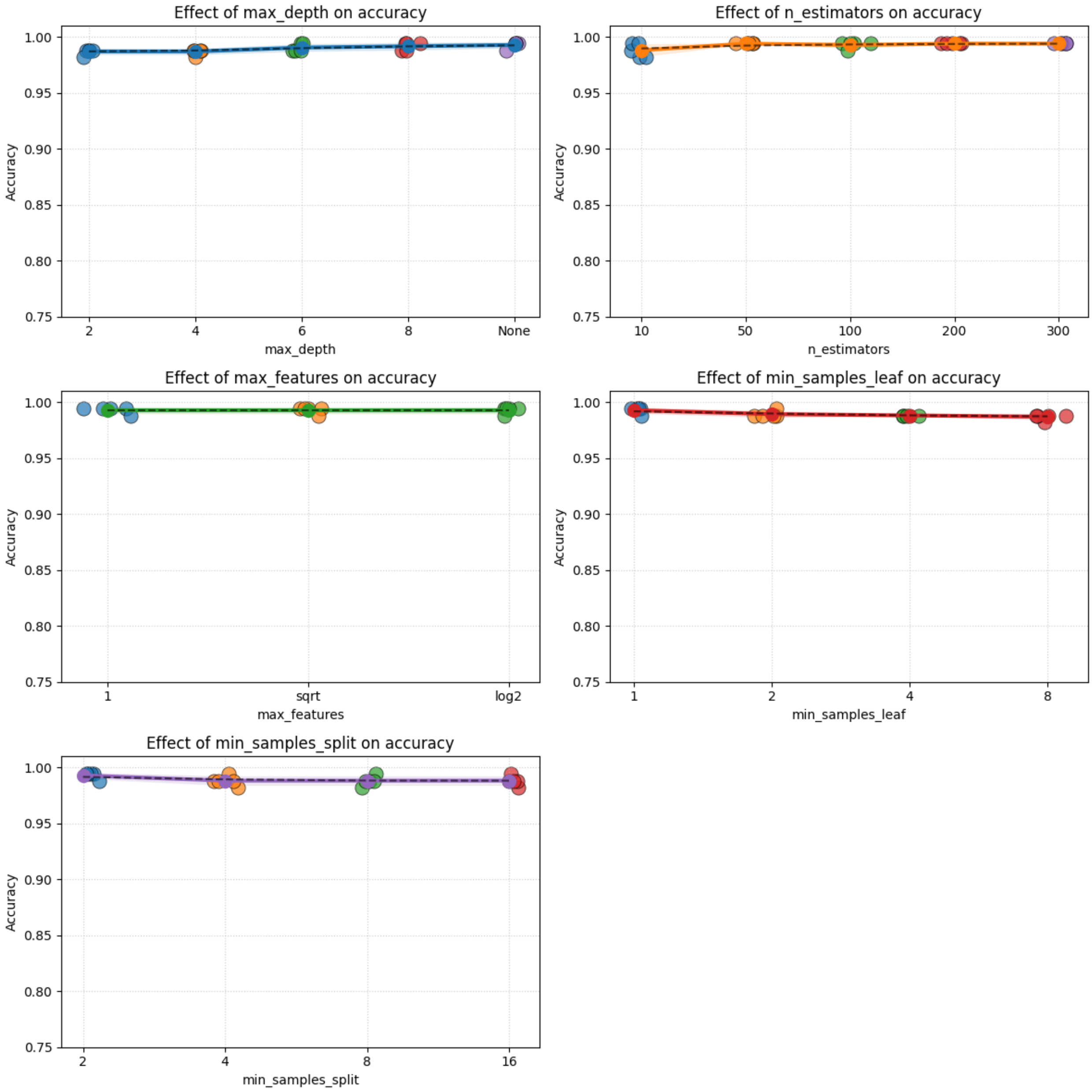
Hyperparameters control model complexity and behaviour. For Random Forests, key hyperparameters include:

- max_depth – maximum depth of each tree
- n_estimators – number of trees in the forest
- max_features – number of features considered at each split

- min_samples_leaf – minimum samples required in a leaf node
- min_samples_split – minimum samples required to split an internal node

Each hyperparameter is changed throughout a range of values in this project, and the accuracy of the model on the test set is noted. Each setup can be assessed several times using various seeds to capture stability.

Effects of HyperParameters plot (combined)



Typical observations:

- max_depth: very small values underfit; moderately deep trees perform best; unrestricted depth can sometimes overfit.
- n_estimators: accuracy improves as trees are added, then plateaus around a certain number (e.g. 100–200 trees).
- max_features: using 'sqrt' often balances tree strength and diversity; using 1 feature per split can make the model unstable.
- min_samples_leaf / min_samples_split: larger values regularise the model, smoothing decision boundaries and reducing overfitting, but excessively large values hurt accuracy.

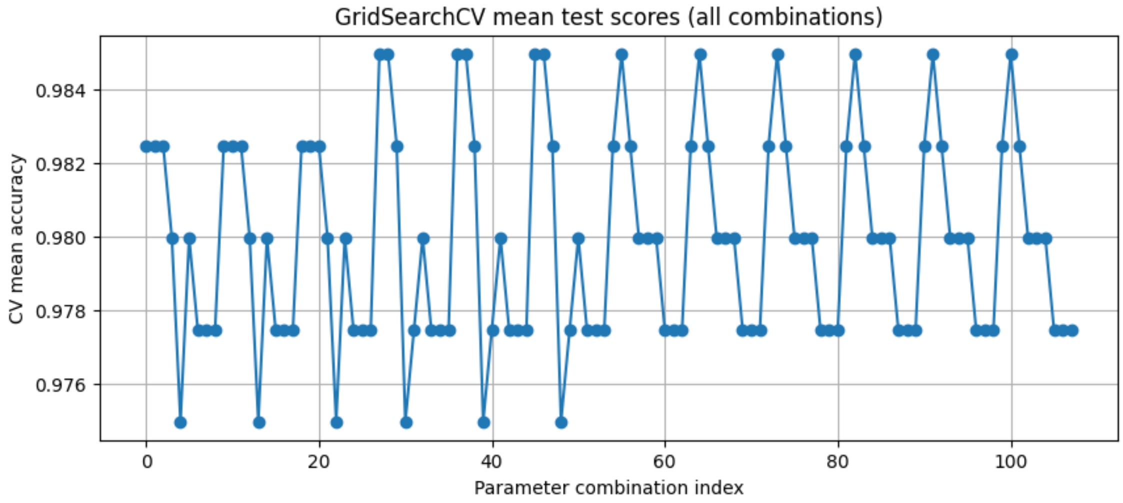
These experiments help build intuition for the bias–variance trade-off.

9. Hyperparameter Tuning with GridSearchCV

After analysing the effects of specific hyperparameters, GridSearchCV is used to conduct a more systematic search.
Generic explanation:
GridSearchCV utilizes cross-validation to exhaustively analyse a preset grid of hyperparameter combinations. It returns the optimum combination based on a specific parameter

In this project, the grid includes combinations of:

- n_estimators
- max_depth
- max_features
- min_samples_leaf



GridSearch mean scores plot

Observations:

- Several parameter combinations produce similar high scores, suggesting the model's robustness in a specific hyperparameter range.
- GridSearchCV picks the combination with the highest mean cross-validation accuracy, providing a principled final choice.

10. Final Model Evaluation

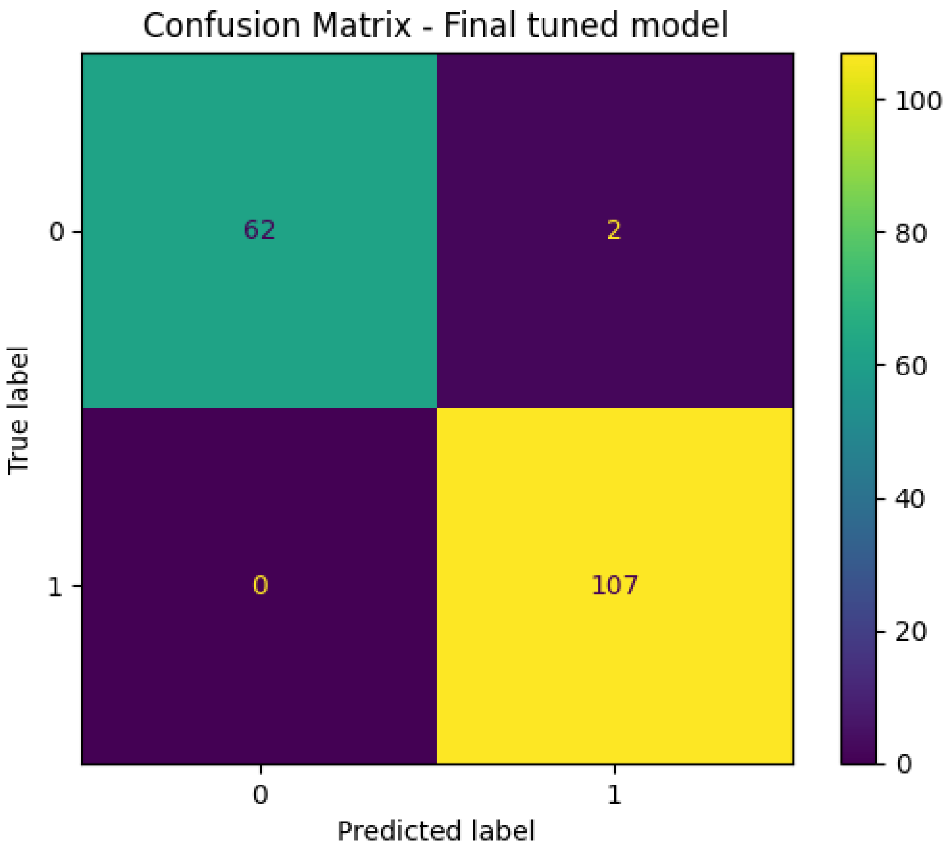
The best Random Forest configuration from GridSearchCV is then evaluated on the test set (data not seen during training or cross-validation).

Metrics considered:

- Accuracy
- Precision, Recall, F1-score

- Confusion matrix

Final confusion matrix plot



Interpretation:

- The tuned Random Forest achieves high accuracy on the test set.
- The model performs well in malignant cases, with a low false-negative rate.
- Misclassifications are rare and mostly occur in borderline cases, as shown by the confusion matrix.

This confirms that the combination of good feature representations, sensible hyperparameter settings, and assembling leads to a robust diagnostic classifier.

11. Reflections on Techniques and Model Choices

- Each technique made a unique contribution to the tutorial.
- EDA provided an initial grasp of feature distribution and correlation.
- Random Forests supplied reliable baseline and natural feature significance measures. Feature significance analysis identified clinically relevant features (shape and concavity-related).
- Visualizations comparing PCA and NCA revealed differences in unsupervised and supervised dimensionality reduction.
- Decision boundaries provided insight into how the classifier distinguishes classes.
- Hyperparameter studies reveal how model complexity and regularization impact performance and stability.

- GridSearchCV offers a systematic and reproducible method for selecting a final model. Overall, the procedure was made more efficient by combining interpretability, visualisation, and adjustment.

12. Conclusion and References

This study presents an end-to-end Random Forest workflow for the Breast Cancer Wisconsin dataset, with a particular emphasis on explanation and understanding. The course seeks to improve both practical skills and intellectual knowledge by integrating generic machine learning concepts (ensembles, feature significance, dimensionality reduction, hyperparameters) to real experiments and visualizations.

The final tuned Random Forest exhibited outstanding classification performance, with good recall for malignant cases and clear interpretability thanks to feature importances and 2D projections. While no model should be employed in isolation for clinical decisions, this study demonstrates how classical machine learning can aid diagnostic thinking in a straightforward manner.

References:

Breast Cancer Wisconsin (Diagnostic) Dataset, UCI Machine Learning Repository. Reiman, L. (2001). Random Forests. Machine Learning. Pedregosa et al. (2011). Scikit-Learn: Machine Learning in Python. Jolliffe, I. T. (2002). Principal Component Analysis. OpenAI. (2025). *ChatGPT (GPT-5.1)* [Large language model]. https://chat.openai.com/