Machine Learning Assignment 2

Report

**Name:** Sistla Keerthana

**Date:** 5th October 2025

**Topic:** Using Random Forest to Learn Imbalanced Data

**Reference Paper:** <https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf>

**Github Repository:** [github.com/KeerthanaSistla/ML/blob/main/Assignment2/Report.doc](https://github.com/KeerthanaSistla/ML/blob/main/Assignment2/Report.doc)

## **1. Project Overview**

Imbalanced datasets are a major challenge in machine learning, where the minority class (e.g., malignant tumors) is underrepresented, leading to biased model performance and poor detection of critical cases.

This project focuses on improving classification performance on the Mammographic Masses dataset by using Weighted Random Forest (WRF) models with class weighting and decision threshold tuning. The goal is to maximize minority class detection (malignant tumors) while maintaining a reasonable majority class performance (benign tumors).

Unlike previous studies that used fixed thresholds and unweighted Random Forest (or balanced Random Forest), this work demonstrates hyperparameter tuning and threshold optimization to enhance model performance on imbalanced data.

## **2. Technologies and Tools Used**

* **Programming Language:** Python
* **Dataset:** Mammographic Masses Dataset (UCI Repository)
* **Libraries:** scikit-learn, pandas, numpy, matplotlib, seaborn
* **Methods and Model:** Weighted Random Forest, Label Encoding, Threshold Tuning

## **3. System Architecture**

**Raw Dataset (CSV):** Load mammographic mass data from CSV.

**Data Preprocessing:** Handle missing values, encode categorical variables, and normalize numerical features.

***Weighted Random Forest*:** Train a Random Forest classifier with higher weights assigned to the minority class, improving model focus on underrepresented classes.

**Hyperparameter Tuning:** Optimize parameters like class\_weight to improve model performance.

**Threshold Optimization:** Adjust decision threshold by maximizing G-mean to balance minority and majority class recall.

**Evaluate Metrics:** Assess performance using Acc+ (recall for malignant), Acc- (recall for benign), Precision, F1-score, G-mean, and Weighted Accuracy.

The feedback loop between the Weighted Random Forest and Hyperparameter Tuning blocks represents iterative optimization, ensuring the best performing model configuration is achieved before final evaluation.

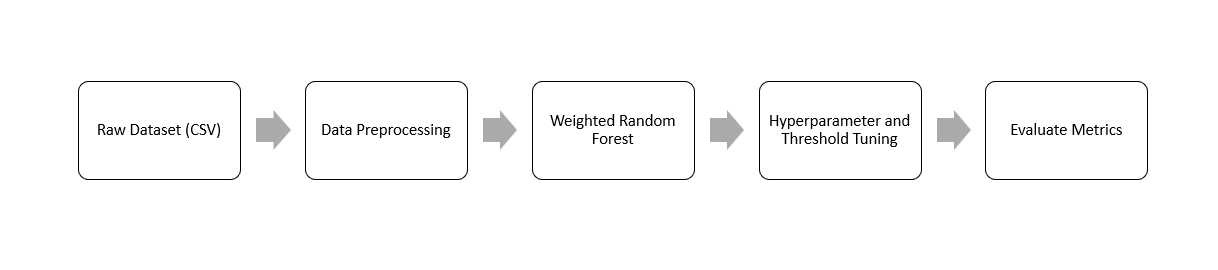


Fig 3.1: Workflow Diagram

## **4. Methodology**

### **Step 1: Data Loading & Preprocessing**

**Dataset:** /content/mammographic\_masses.data

**Features:** BI-RADS, Age, Shape, Margin, Density, Severity

**Encoding:** Converting categorical values into numeric.

### **Step 2: Handle Class Imbalance**

**Class distribution:** Benign (0): 516, Malignant (1): 445

WRF assigns higher weight to the minority class (class\_weight={0:1, 1:2} and {0:1, 1:3})

### **Step 3: Model Building**

**Model Used:** Weighted Random Forest (WRF)

**Hyperparameters:**

* n\_estimators = 300
* max\_depth = None (full depth)
* class\_weight = 1:2 and 1:3

**Train-test split:** 70% training, 30% testing

### **Step 4: Threshold Optimization**

Predicted probabilities and varied thresholds from 0.1 - 0.9

Optimal threshold selected by maximizing G-mean, balancing minority and majority recall.

## **5. Screenshots**



Fig. 5.1: Loading Data

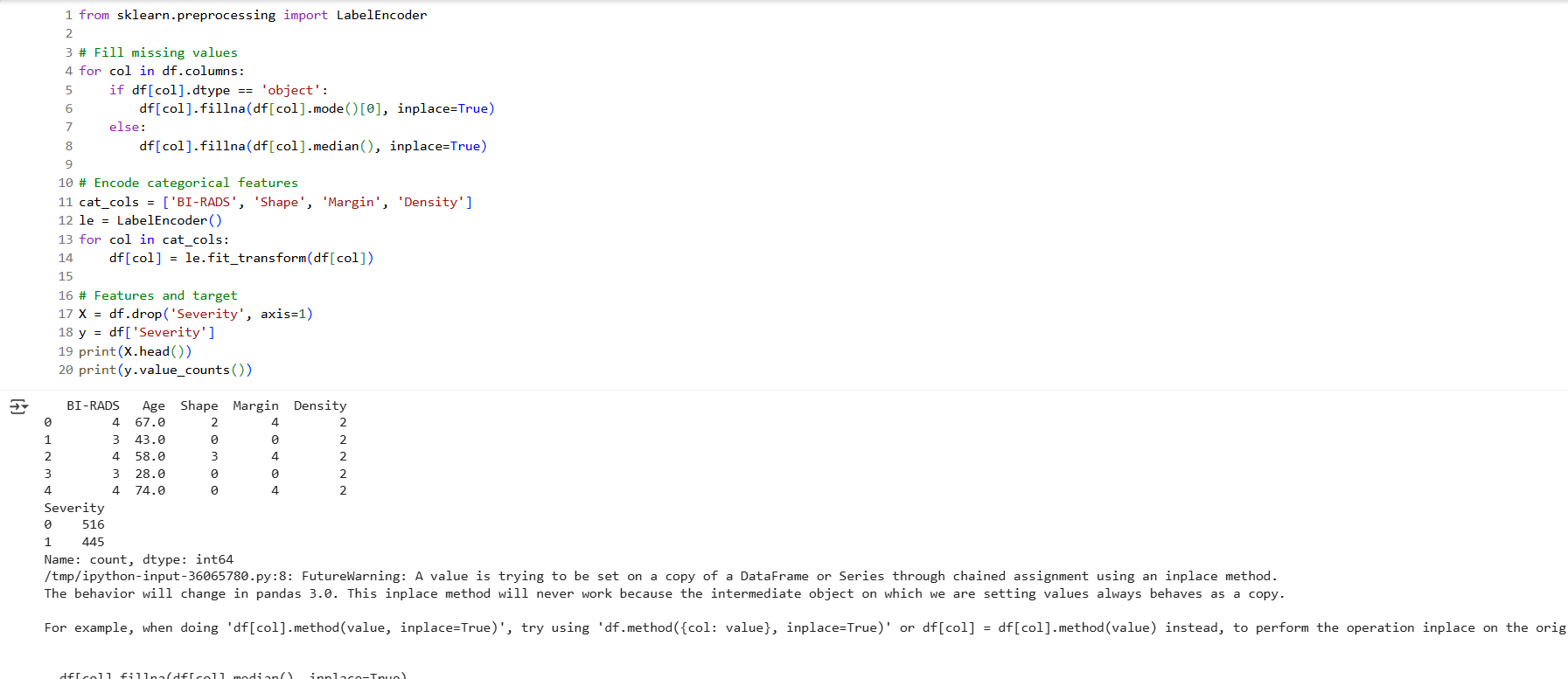


Fig 5.2: Data Preprocessing

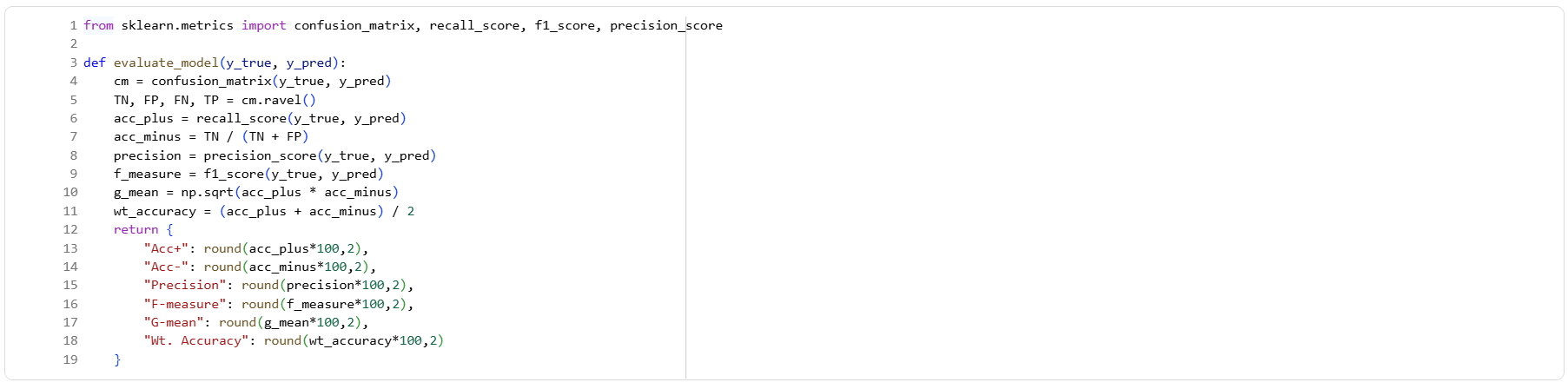


Fig 5.3: Model Evaluation Function

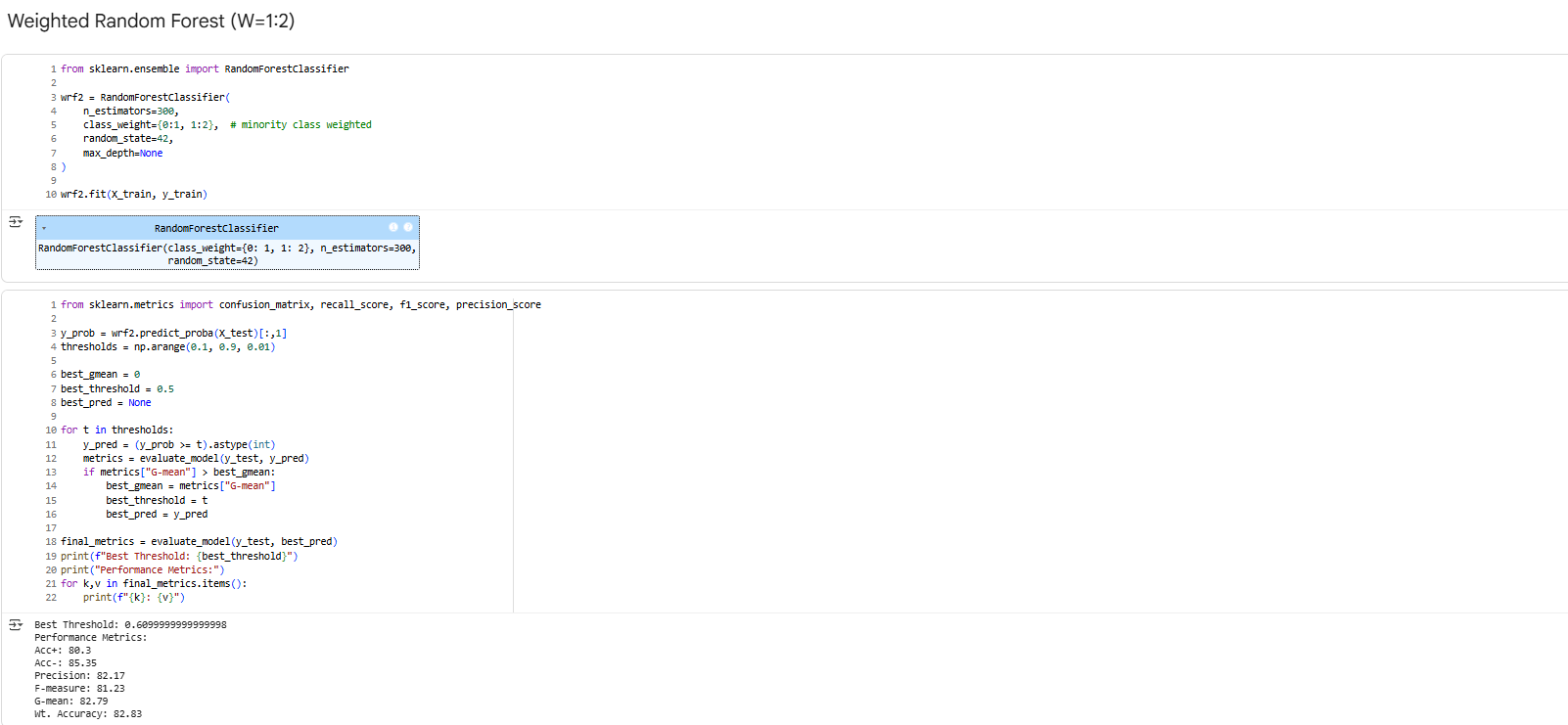


Fig 5.4: Weighted Random Forest (W=1:2)

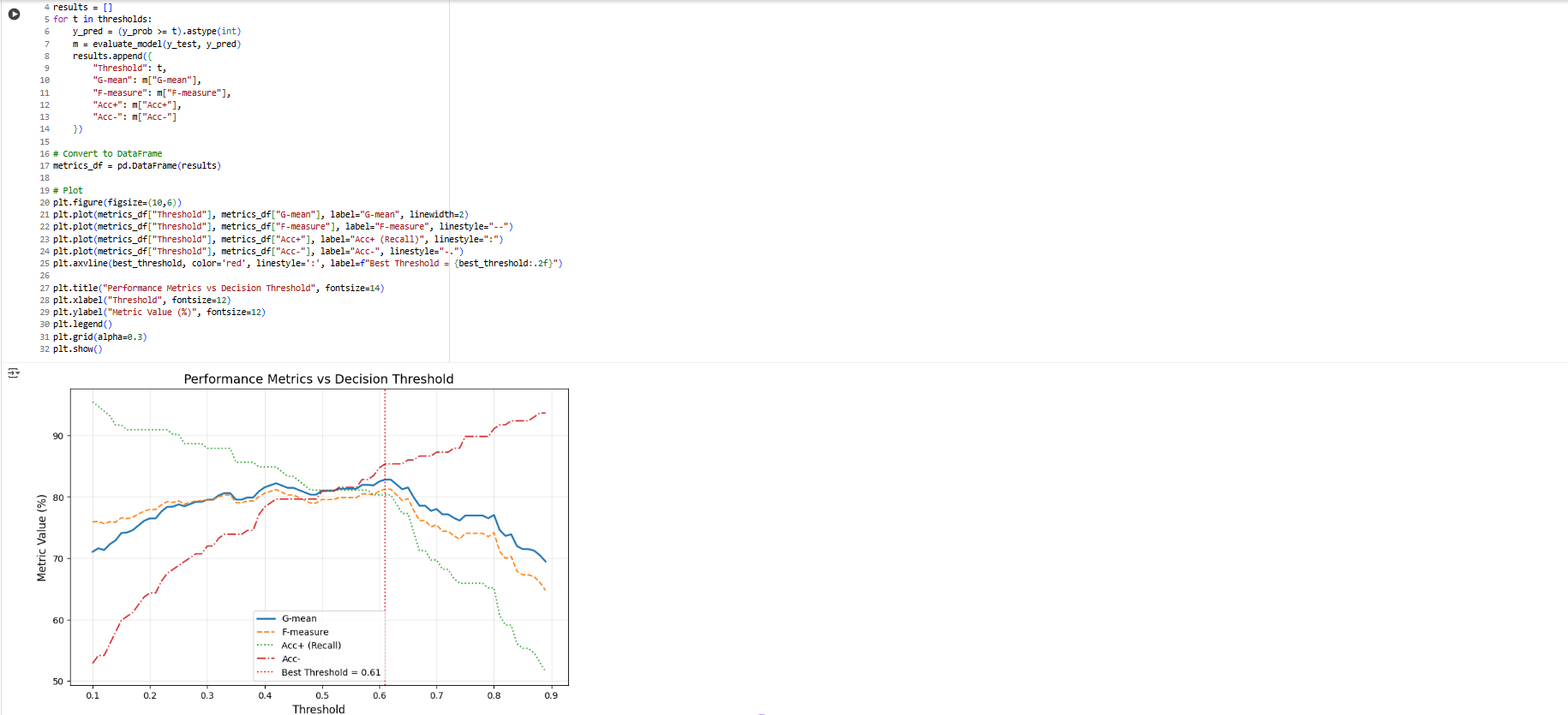


Fig 5.5: Weighted Random Forest Plot

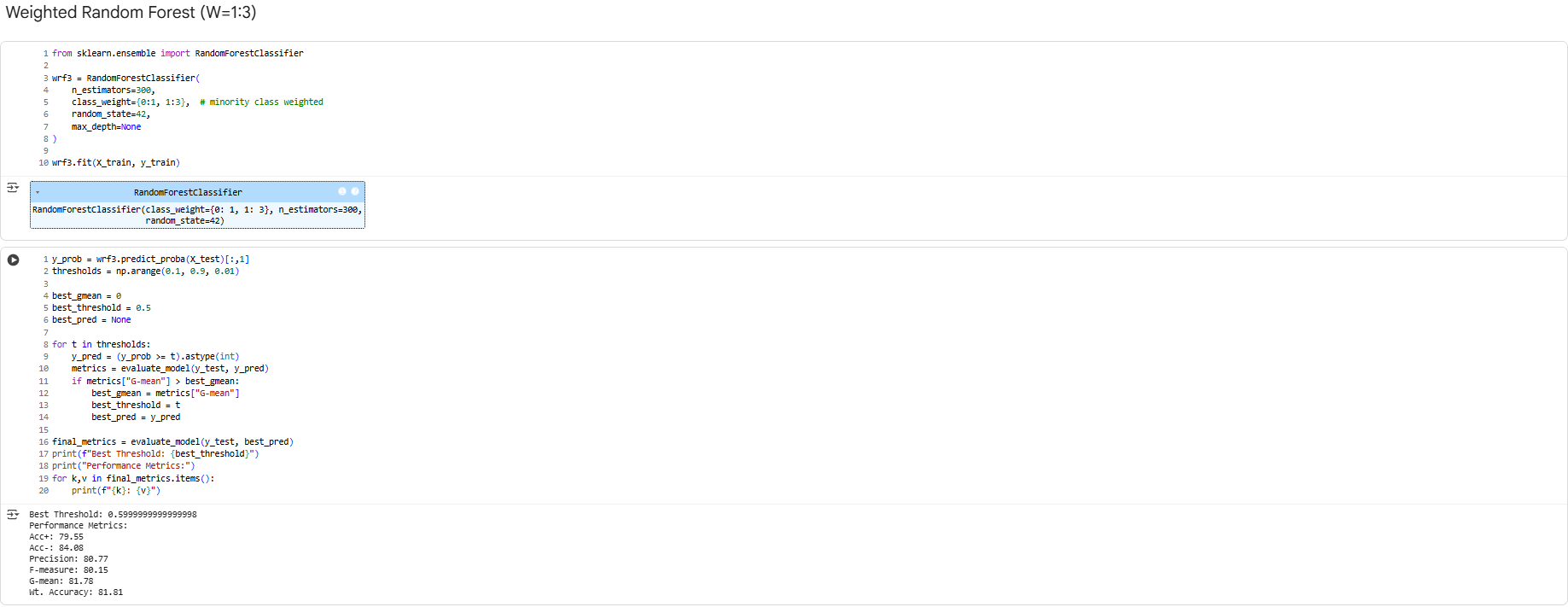


Fig 5.6: Weighted Random Forest (W=1:3)

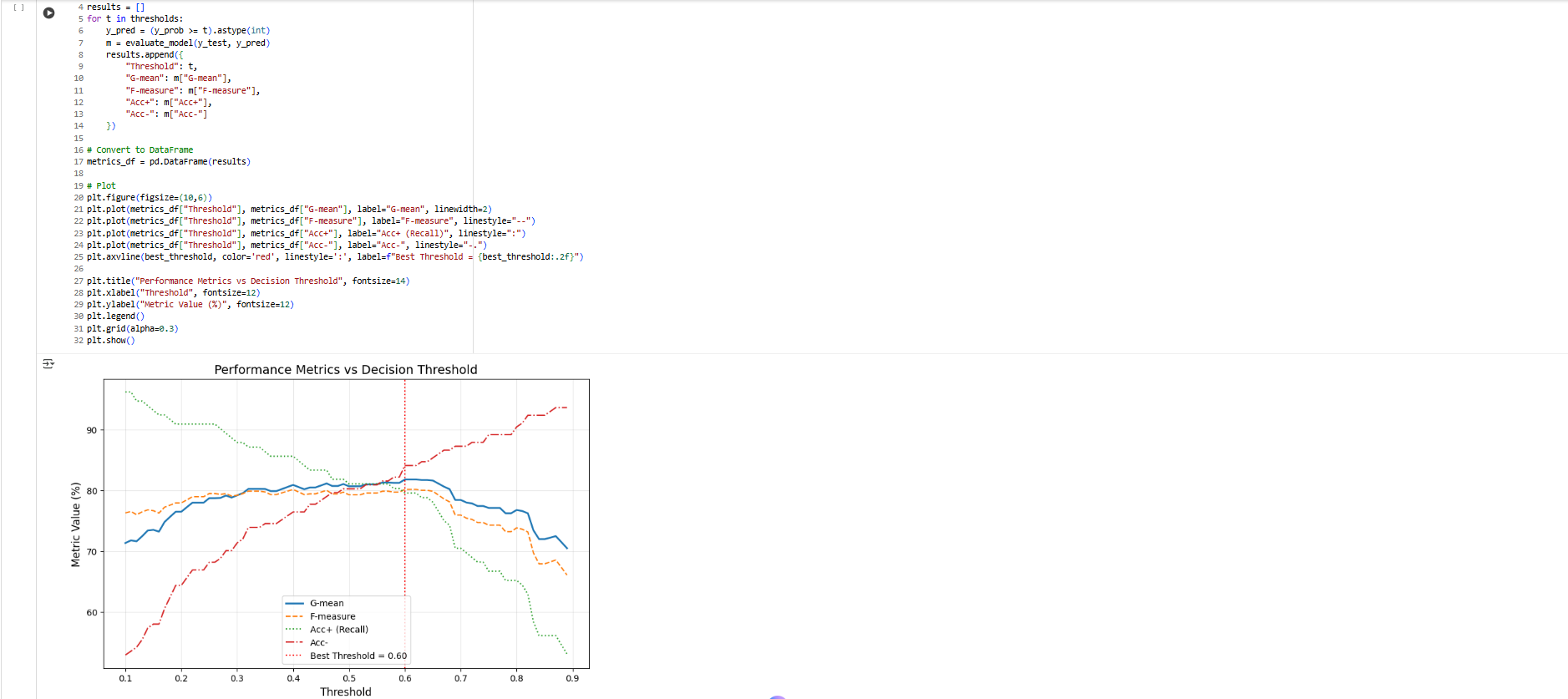


Fig 5.7: Weighted Random Forest Plot

## **6. Testing and Results**

Table 6.1: Comparison of Performance Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **WRF 2:1 (Paper)** | **WRF 3:1 (Paper)** | **WRF 2:1 (My Model)** | **WRF 3:1 (My Model)** |
| Acc+ (Recall +) | 65.38 | 72.69 | 80.30 | 79.55 |
| Acc- (Recall -) | 99.57 | 99.25 | 85.35 | 84.08 |
| Precision | 78.34 | 69.74 | 82.17 | 80.77 |
| F-measure (F1) | 71.28 | 71.18 | 81.23 | 80.15 |
| G-mean | 80.68 | 84.94 | 82.79 | 81.78 |
| Weighted Accuracy | 82.48 | 85.97 | 82.83 | 81.81 |
| Best Threshold | N/A | N/A | 0.6098 | 0.5998 |

**Observation:**

**Minority Class Detection (Acc+):** Custom WRF models improve malignant tumor recall, demonstrating better detection.

**Majority Class Detection (Acc-):** Slightly lower than paper’s WRF, showing the trade-off in improving minority class detection.

**Precision & F1-score:** Both metrics are higher in custom models, indicating better balance between precision and recall.

**G-mean & Weighted Accuracy:** Comparable to paper’s results, confirming balanced performance.

**Threshold Tuning:** Optimized thresholds (~0.6) outperform the paper’s fixed cutoff (0.3), emphasizing minority class detection.

## **7. Learning Outcomes**

* Learned to handle imbalanced datasets using class weighting.
* Understood threshold tuning to optimize G-mean.
* Improved understanding of evaluation metrics for imbalanced classification.
* Developed skills in hyperparameter tuning and model evaluation using Python.

## **8. Conclusion**

The Weighted Random Forest with threshold optimization demonstrates significant improvement in minority class detection and F1-score on the Mammographic Masses dataset. This approach effectively balances the trade-off between minority and majority class performance, confirming the importance of class weighting and threshold tuning in handling imbalanced datasets.