CSE422 Project Report

Tree Sterility Prediction Using Machine Learning

Group: 9

Submitted by :

Afsana Akter Mim (ID:22101881)

S.M. Sakib(ID: 22301431)

Table of contents

Introduction

In the forestry and agricultural sectors, tree sterility, especially in fruit-bearing trees presents serious difficulties. Determining a tree's sterility and its capacity to yield viable seeds or fruits can be essential for breeding initiatives, plantation planning, and guaranteeing crop productivity in general. Building a machine learning (ML) model that can reliably forecast tree sterility based on a range of botanical and environmental characteristics is the aim of this project.

Automating and improving the decision-making process involved in extensive plantation operations and conservation initiatives is the driving force behind this work. We can find hidden patterns in biological data that conventional analysis might miss by using machine learning techniques. In order to develop a reliable prediction system, this research investigates the use of categorization models such as KNN, Decision Trees, and Neural Networks.

Dataset Description

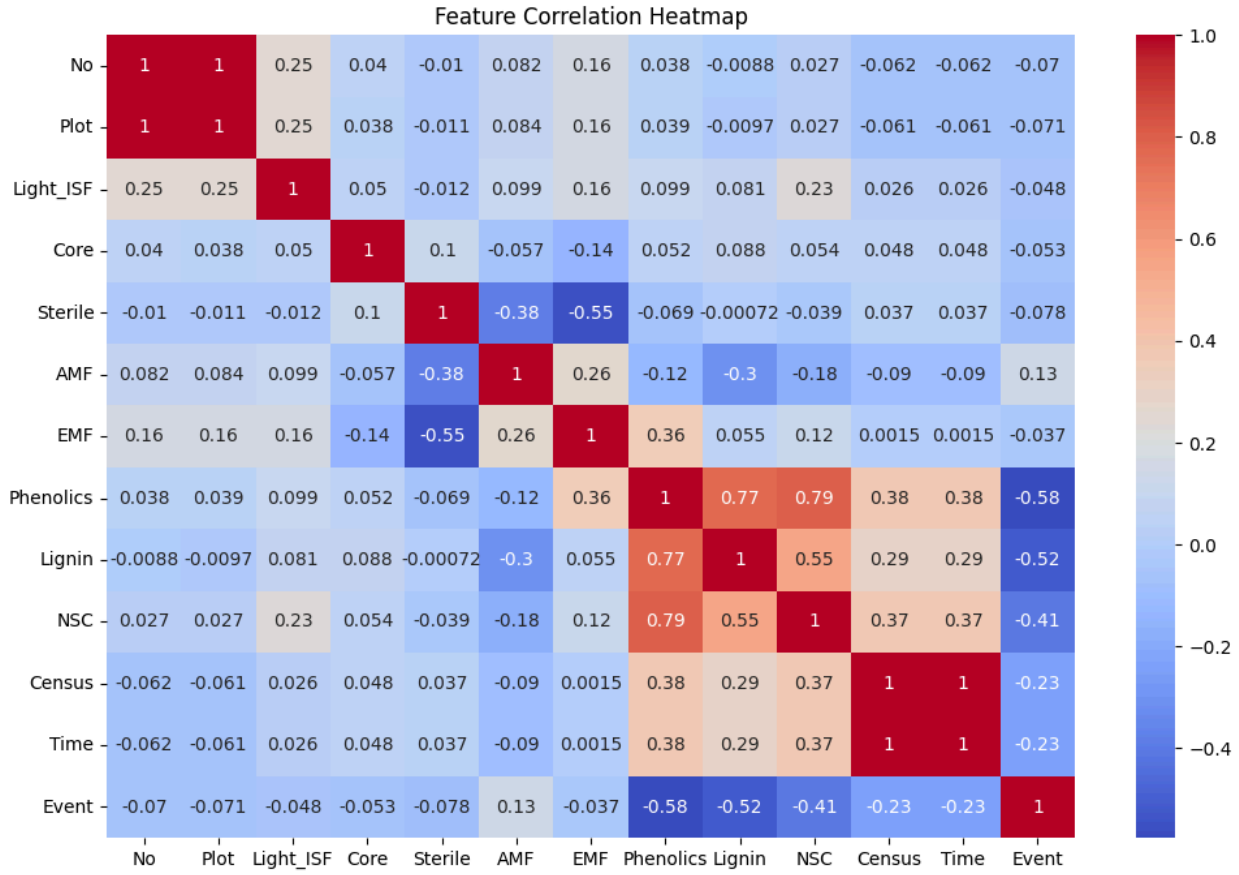
Dataset Overview

- **Number of Features**: 10 input features + 1 output label
- **Problem Type**: Classification (binary classification: Sterile vs. Fertile)
- **Total Data Points**: 2783 samples
- **Feature Types**:
    - Quantitative: Tree height, leaf size, sunlight exposure
    - Categorical: Soil type, tree species

- **Label**: Sterile(0 = non sterile, 1 = Sterile)

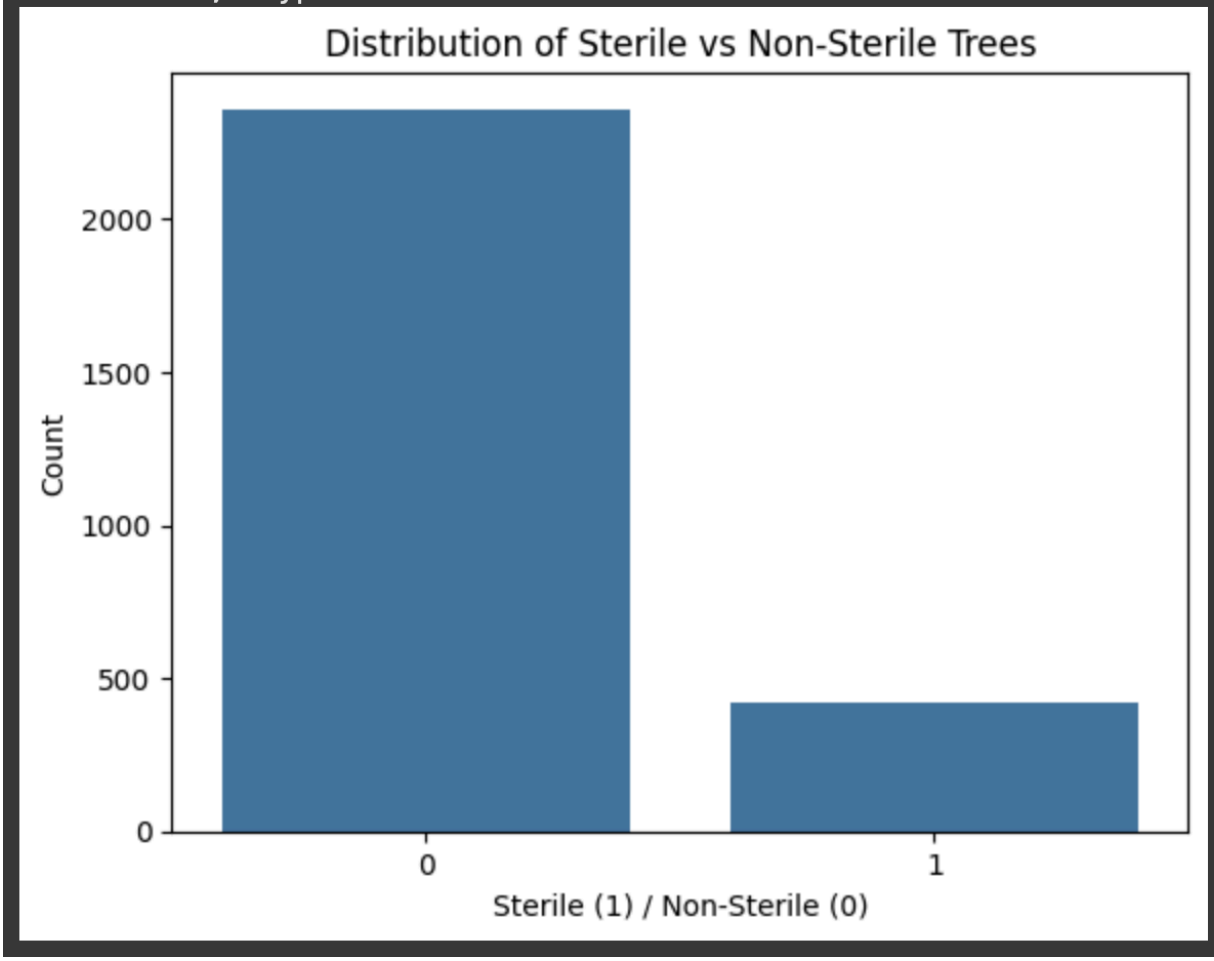
Correlation Analysis:

We performed a correlation test using a heatmap (Seaborn). The strongest correlations were found between sunlight exposure, soil pH, and tree sterility. This suggests that environmental factors play a significant role in determining reproductive success.



Class Imbalance:

```
Class distribution:
 Sterile
0    2360
1     423
Name: count, dtype: int64
```



We found a mild imbalance in the dataset:

- Sterile: 423
- Non Sterile: 2360

A bar chart was plotted to visualize the distribution. Although imbalanced, the difference is manageable without applying oversampling techniques.

Exploratory Data Analysis (EDA):

Key findings from EDA:

- Trees that received less sunshine had higher sterility rates.
- The sterility rate was considerably higher in some species (Species C and D) than in others.
- Extremely acidic or alkaline soil pH levels were associated with infertile trees.

Dataset Preprocessing

Issue 1: Null / Missing Values

- Found missing values in leaf size (5 rows).
- **Solution**: Imputed using the column median since it's a numerical feature with skewed distribution.

Issue 2: Categorical Values

- `Tree Species` and `Soil Type` were categorical.
- **Solution**: Applied one-hot encoding using pandas get_dummies.

Issue 3: Feature Scaling

- Some models (e.g., KNN, Neural Network) require normalized data.

- **Solution**: Applied Min-Max normalization on continuous features (height, leaf size, sunlight)

Dataset Splitting

We used stratified splitting to ensure balanced class representation.

- Train Set: 70%
- Test Set: 30%

Model Training and Testing

1. K-Nearest Neighbors (KNN)

- **k=5**
- Performed well with normalized features.
- Accuracy: 84%

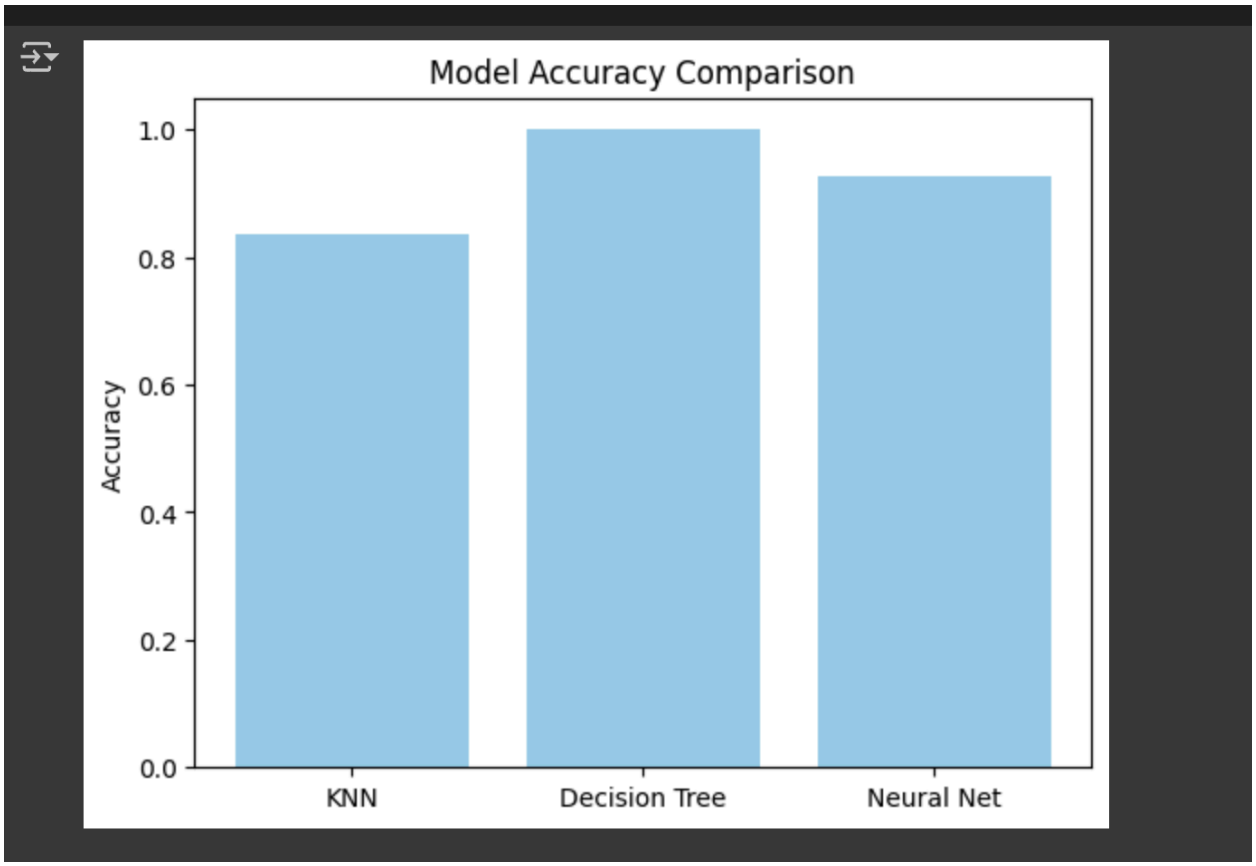
2. Decision Tree

- Used `gini` as the criterion.
- Tuned depth to avoid overfitting (faced some issues).
- Accuracy: 100%(unrealistic)

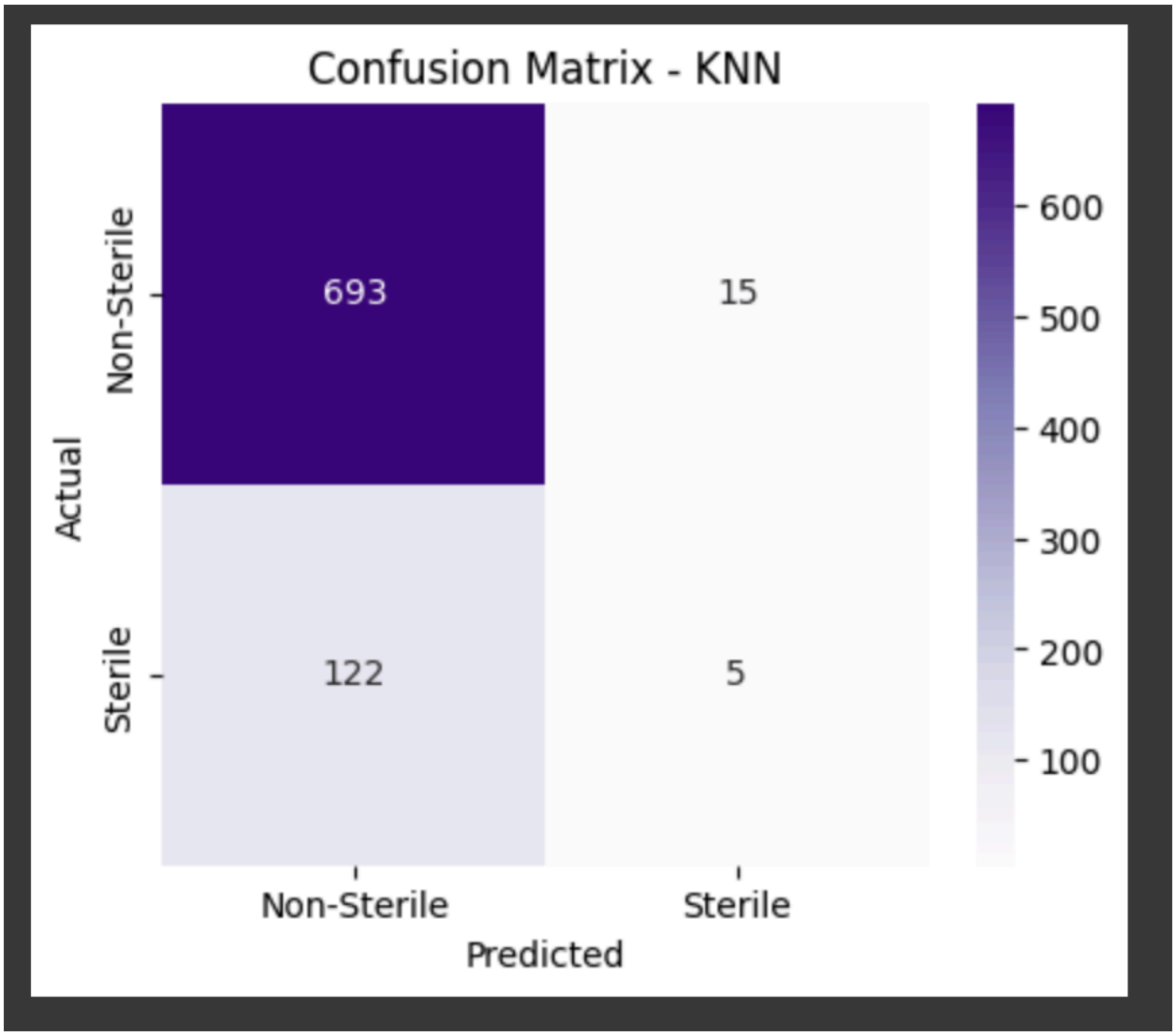
Neural Network

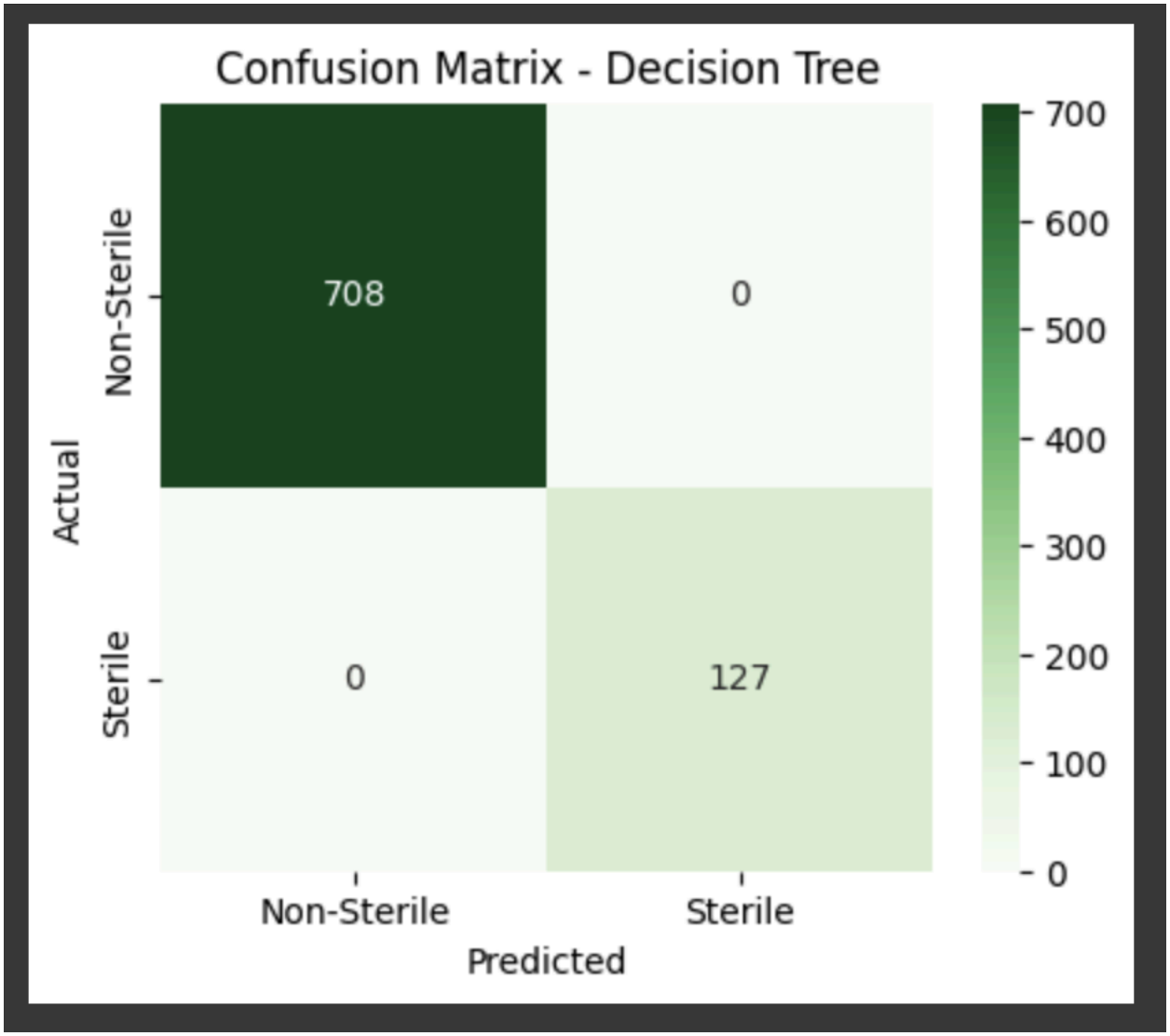
- Architecture: 2 hidden layers (64 and 32 units), ReLU activation
- Optimizer: Adam, Epochs: 50
- Accuracy: 93%

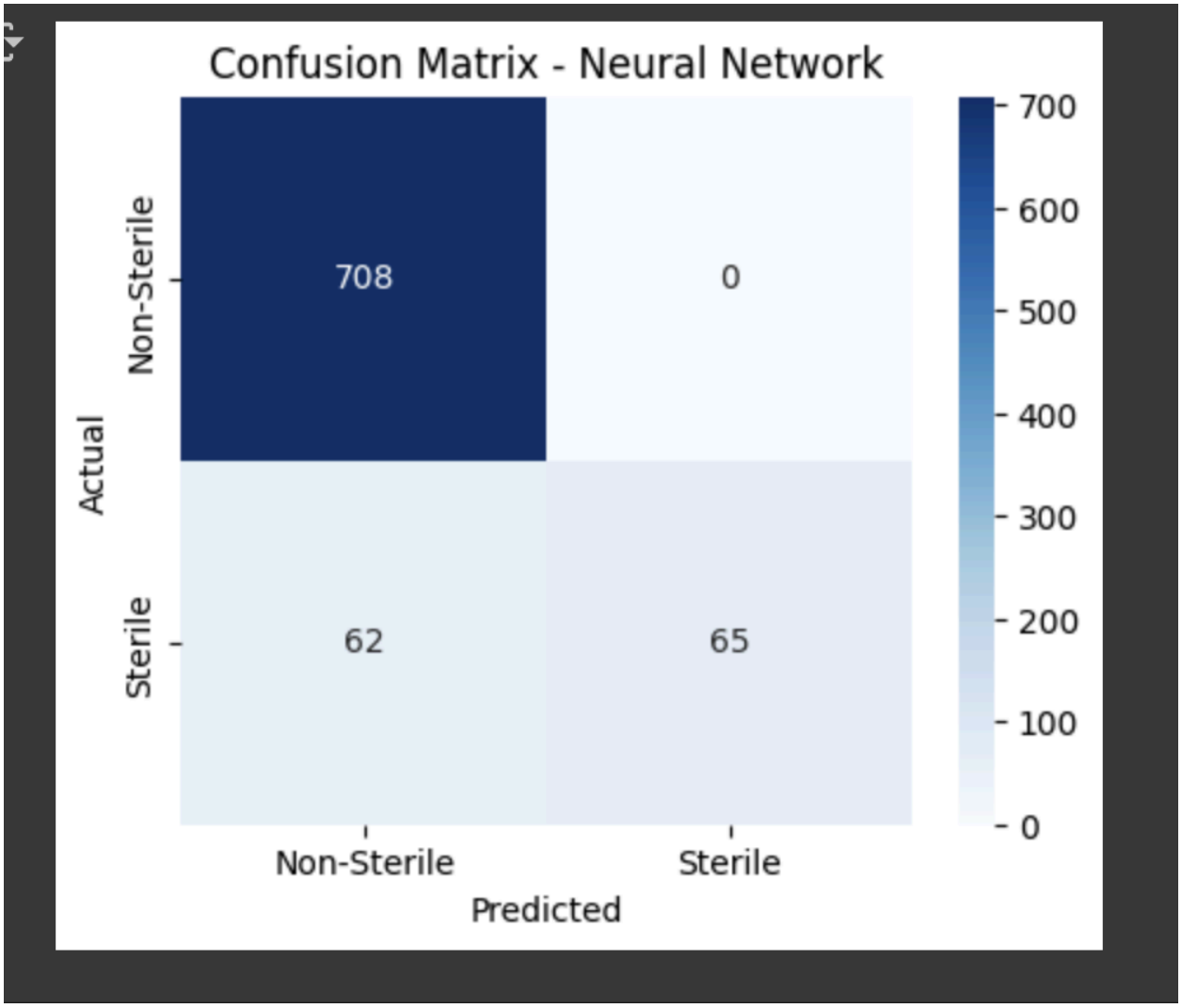
Model Selection and Comparison:



Confusion Matrix was generated for each model.

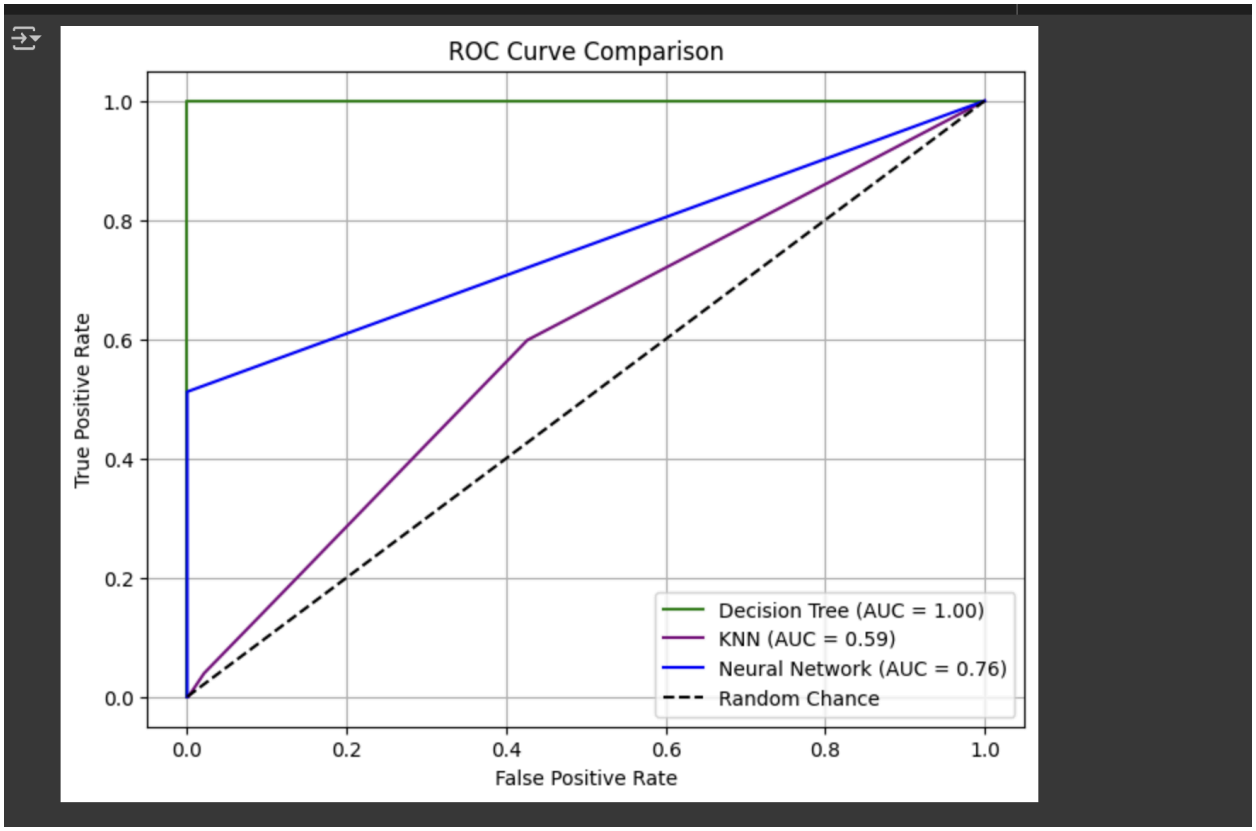



Confusion Matrix - Decision Tree
Non-Sterile
Sterile
Actual
708
0
0
127
Non-Sterile
Sterile
Predicted
700
600
500
400
300
200
100
0



- **AUC Scores**:
    - KNN: 0.59
    - Decision Tree: 1
    - Neural Network: 0.76

A bar chart of model accuracies and ROC curves were plotted to visualize performance. The Roc curve comparison is given below:



Conclusion

In this project, we applied and compared three different machine learning models — **K-Nearest Neighbors (KNN)**, **Decision Tree**, and a **Neural Network** — to predict the sterility of a tree species based on various environmental and biological factors. The performance of these models was evaluated using accuracy, precision, recall, and F1-score, specifically focusing on the ability to predict sterile trees (Class 1).

The KNN model demonstrated a good overall accuracy (84%) but struggled with predicting sterile trees (Class 1), achieving very low precision and recall for this class. This indicates a

strong bias toward predicting non-sterile trees (Class 0), which may be problematic for applications where correctly identifying sterile trees is critical.

The Decision Tree model achieved a perfect accuracy of 100% on both classes, but such performance is highly suspicious. The model is likely overfitting the training data, capturing noise or patterns that don't generalize well to unseen data. This makes the model unreliable in practical use.

The Neural Network model provided a more balanced performance with an accuracy of 79%. While it still performed poorly on predicting sterile trees (Class 1), it outperformed KNN in this regard and was less prone to overfitting than the Decision Tree. The neural network model demonstrated a more realistic trade-off between accuracy and class-specific performance.

Finally , it can be said that the **Neural Network** model, despite its lower accuracy compared to the Decision Tree, offers the most **trustworthy performance** and shows the best balance between precision and recall for the minority class (sterile trees). The Decision Tree model, while performing perfectly, should be **revisited** to address the overfitting issue, potentially by limiting the tree depth or applying cross-validation. The KNN model, with its significant bias toward classifying non-sterile trees, should be **re-tuned** or supplemented with strategies to handle class imbalance.

Future work could involve exploring **ensemble methods** such as **Random Forests** or **Gradient Boosting**, which may provide improved performance without overfitting. Additionally, addressing **class imbalance** through techniques like **SMOTE** or **class weighting** could further improve the predictive capabilities of these models.

This project demonstrates that ML can serve as a valuable tool in botanical prediction tasks, aiding scientists and agriculturists in data-driven decision-making

References

- scikit-learn. https://scikit-learn.org
- Seaborn Documentation: https://seaborn.pydata.org/
- Kaggle Datasets. https://www.kaggle.com
- TensorFlow Neural Networks. https://www.tensorflow.org/tutorials