Inspiring Excellence

CSE422

Project report

Project Title : **Tree Seedling Survival Prediction under Varying Light and Soil Conditions**

Student Name : Srijon Basak - 22101359

Md. Sakibur Rahman - 22101063

Course Name : Artificial Intelligence

Date : 11-05-25

Introduction:

This project aims to explore the survival of tree seedlings under varying light conditions and soil types, focusing on how plant-soil feedback mechanisms mediate survival responses. By examining the interactions between different tree species, soil sources, and light availability, the project seeks to understand the ecological factors that influence seedling survival in forest ecosystems. The motivation behind this project stems from the need to better understand how environmental factors, such as light and soil composition, impact forest regeneration and the long-term success of tree species in varying ecological conditions. This knowledge can inform forest management strategies, especially in the face of changing climate conditions and habitat fragmentation.

Dataset Description:

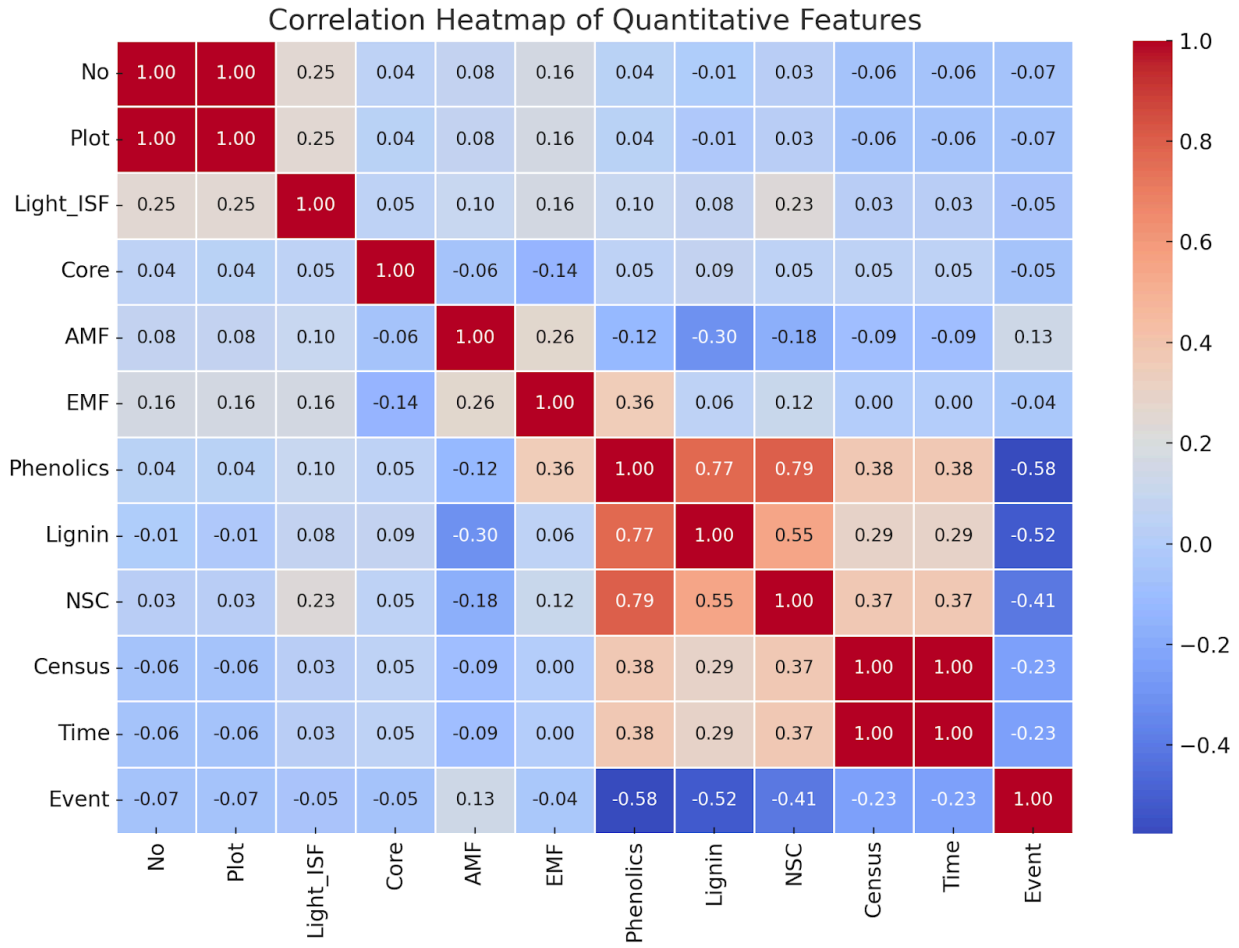
- **Number of Features**: The dataset contains **24 features** (columns).
- **Classification or Regression Problem?**: This is a **classification** problem, as the target variable, "Alive" is binary, indicating whether the seedling survived (1) or died (0).
- **Number of Data Points**: The dataset consists of **3,024 data points** (observations).

Types of Features:

- **Quantitative Features**: These are continuous variables, including Light_ISF, AMF, EMF, Phenolics, Lignin, NSC, Census, and Time.
- **Categorical Features**: These include variables like Species, Light_Cat, Soil, Myco, SoilMyco, Sterile, Conspecific, and Alive.

Correlation Analysis:

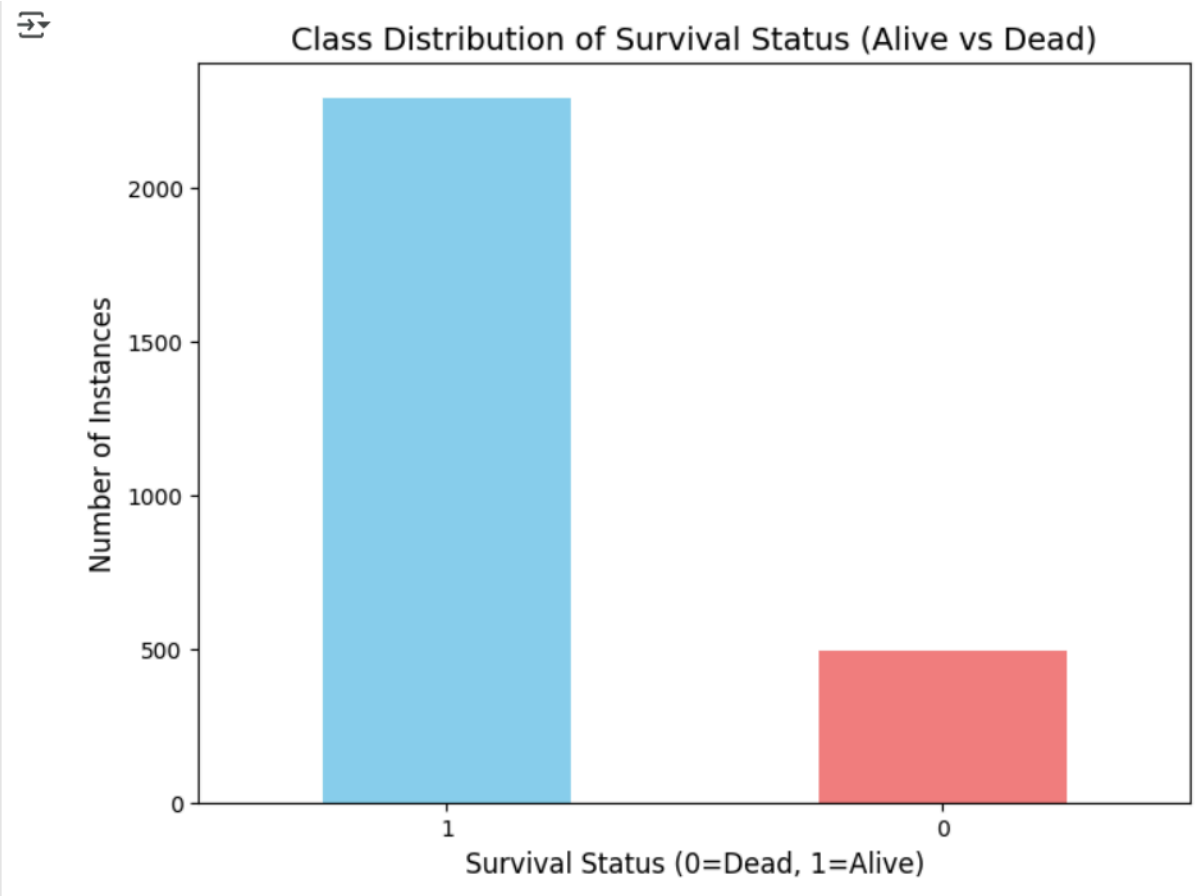
- The correlation heatmap of numerical features shows relationships between variables like mycorrhizal colonization (AMF and EMF), light levels (Light_ISF), and survival outcomes. High correlations (above 0.75) indicate strong relationships between features, while lower correlations suggest less dependency.



Imbalanced Dataset

- **For the output feature, do all unique classes have an equal number of instances or not?**
  No, the dataset is imbalanced. The class representing seedlings that are **alive** (1) has a significantly higher number of instances compared to the class representing seedlings that are **dead** (0). This indicates that there are far more instances of alive seedlings in the dataset than dead ones.

- **Bar Chart Representation**:
  The bar chart below illustrates the distribution of the two classes in the Alive feature. The "Alive" class (1) contains a much larger number of instances than the "Dead" class (0), highlighting the imbalance between the two classes.



Dataset Pre-processing

1. **Null / Missing Values**:
    - **Solution**: Rows with missing values in the Event column were removed to ensure the dataset only contains complete data.
2. **Categorical Values**:
    - **Solution**: Categorical features were converted into numerical values using **Label Encoding**, making them compatible with machine learning models.
3. **Feature Scaling**:
    - **Solution**: Although feature scaling was initially applied, it was found to decrease model accuracy and was subsequently removed for better performance.

Model Selection/Comparison Analysis:

Based on the evaluation of various classification models, **Random Forest** performed the best, achieving the highest accuracy (91.5%), precision, and recall, especially for the positive class (class 1). It significantly outperformed the other models in terms of balanced performance across both classes. **Neural Network** also performed well, but it had slightly lower recall for class 1 compared to Random Forest. **Logistic Regression** showed good results, particularly for class 1, but its performance was still lower than Random Forest. **KNN** had the weakest performance, especially for class 0, with low recall and precision.

Key Insights:

- **Random Forest** is the most effective model, with high accuracy and a good balance of precision and recall.

- **Neural Network** performed well but slightly underperformed compared to Random Forest.

- **Logistic Regression** performed reasonably well but is less effective for classifying the negative class (class 0).

- **KNN** showed the least accuracy and recall, indicating it's not the best fit for this problem.

Reasons for Results:

- **Random Forest** excels due to its ensemble approach, which handles both class imbalances and complex data patterns well.

- **Neural Networks** require more data to fully capture complex patterns and might not have been fully optimized for this dataset.

- **Logistic Regression** is a simpler linear model, which can struggle with non-linear data.

- **KNN** suffers due to the curse of dimensionality and its sensitivity to irrelevant features.

Challenges:

- **Class Imbalance**: Class imbalance made it harder for some models (e.g., Logistic Regression and KNN) to predict class 0 effectively.

- **Feature Selection**: Some models struggled due to improper handling of features or noise.

- **Model Complexity**: More complex models like Neural Networks may not have provided the best results for this dataset.

- **Overfitting and Underfitting**: KNN appeared to overfit the data, leading to poor performance.

Improvements:

- **Addressing Class Imbalance**: Techniques like oversampling the minority class or using class weights could improve model performance.

- **Hyperparameter Tuning**: Fine-tuning models like Neural Networks and Random Forest could lead to better results.

- **Feature Engineering**: Improving feature selection and transformation techniques could help optimize model performance.

Conclusion:

Random Forest emerged as the most effective model, delivering the highest accuracy and a strong balance of precision and recall. This suggests that it is well-suited for this classification task, likely due to its ensemble approach, which helps manage class imbalance and complex data patterns. The **Neural Network** performed well but did not quite match Random Forest in terms of accuracy, likely due to the dataset's characteristics and the complexity of tuning the neural network. **Logistic Regression** performed reasonably well but struggled with classifying the "Dead" class (0). **KNN** was the least effective model, especially for classifying "Dead" seedlings, possibly due to its sensitivity to irrelevant features and overfitting.