**Final Project Report**

Analysing the Impact of Tree Species Diversity on Foliar Fungal Disease Incidence

in European Forests

University of Maryland, Baltimore County

Rahul Kaushik Mallela

Professor Devin Fensterheim

**Report**

**Business Problem: -**

Forests in Europe, housing a diverse array of tree species, face an escalating threat from fungal pathogens. The implications of this threat extend to the broader ecosystem and its functioning. Experimental studies propose a potential correlation between reduced fungal disease incidence and increased tree species diversity, yet evidence from natural forests remains scarce. The primary business problem is to investigate whether foliar fungal disease incidence is negatively influenced by tree species diversity across various European forest types.

This involves understanding the impact of tree species richness, functional type (conifer vs. broadleaved), and phylogenetic diversity on overall fungal disease incidence. The dataset, derived from field assessments of leaves, comprises information on 16 tree species across 209 plots in six countries, spanning a forest-type gradient from the Mediterranean to boreal forests.

The business problem centers on understanding and predicting the total foliage of trees. Accurate predictions can assist in various domains, such as environmental studies, urban planning, and landscaping. The goal is to develop a model that generalizes well to unseen data, providing reliable estimates of total foliage based on relevant features.

**Approach: -**

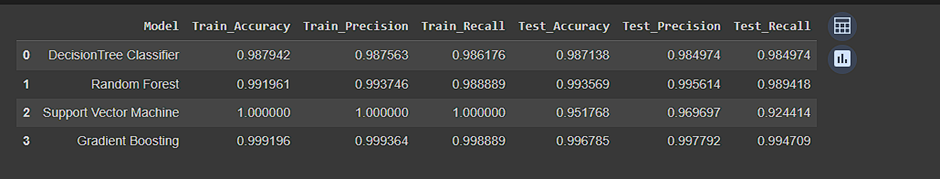
* **Data Collection**
* Data has been collected from the following Source link: - <https://zenodo.org/records/5022387>
* The dataset encompassed both numerical and categorical features, with the target variable being "total.foliage."
* **Data Exploration and Preprocessing**
* Executed meticulous data cleaning procedures, eliminating duplicates, and addressing missing values.
* Applied the Interquartile Range (IQR) method for outlier detection and subsequent removal, ensuring data integrity.
* Conducted Exploratory Data Analysis (EDA) visualizations to understand feature distributions and dataset characteristics.
* **Dimensionality Reduction**
* Employed Principal Component Analysis (PCA) to reduce the dataset's dimensionality.
* Utilized PCA to transform features, contributing to a refined dataset for improved model training.

**Model Selection: -**

**four classification models were chosen:**

* DecisionTree Classifier
* Random Forest
* Support Vector Machine (SVM)
* Gradient Boosting

Models were evaluated based on test accuracy, and the results were as follows:

****

* Adopted four well-established classifiers: Decision Tree Classifier, Random Forest, Support Vector Machine (SVM), and Gradient Boosting.
* Each model was chosen based on its suitability for the task and potential to handle the dataset's complexity.
* **Hyperparameter Optimization**
* Utilized GridSearchCV for hyperparameter optimization of each model.
* Customized hyperparameters to meet the specific requirements of the dataset, enhancing model performance.
* **Evaluation**
* Focused on evaluating models primarily based on train & test accuracy.
* Decision Tree Classifier and Random Forest emerged as top-performing models, exhibiting high accuracy, precision, and recall metrics.
* Considered Support Vector Machine (SVM) and Gradient Boosting models, acknowledging their potential that may be relevant for specific tasks.

**Conclusion: -**

* The comprehensive analysis of fungal disease incidence along tree diversity gradients across European forests provides valuable insights into the dynamics of forest ecosystems.
* The findings reveal relationships between tree species diversity, latitude, and functional type, influencing foliar fungal disease incidence. The variation observed across different forest types underscores the complexity of these ecological interactions.
* The results highlight the need for tailored forest management strategies, considering factors such as latitude and tree functional types, to effectively mitigate fungal diseases and promote overall forest health.

**Recommendations and Future Steps: -**

* Customize forest management based on identified fungal disease patterns, with a focus on species-specific strategies.
* Prioritize tree diversity, especially in regions where it significantly influences disease frequency, to enhance overall resilience.
* Conduct more detailed research on tree age, soil conditions, and weather patterns to improve the ability to predict fungal diseases.
* Consider specific disease symptoms and their impact on different tree species to refine and strengthen forest management practices.

**References: -**

* Sklearn.tree.DecisionTreeClassifier. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html "An Introduction to Statistical Learning" by James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Springer: New York, NY.
* "Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy (2012). The MIT Press: Cambridge, MA.
* sklearn.ensemble.RandomForestClassifier. (n.d.). Scikit-learn. [https://sciki](https://sciki/)-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html "An introduction to support vector machines" by Nello Cristianini and John Shawe-Taylor. (2000). Cambridge University Press.
* Support vector machines. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/svm.html>
* sklearn.ensemble.GradientBoostingClassifier. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>
* *Sklearn.decomposition.PCA*. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
* *sklearn.model\_selection.GridSearchCV*. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>
* Data from: Fungal disease incidence along tree diversity gradients depends on latitude in European forests. (2022). Zenodo. <https://doi.org/10.5061/dryad.389mt>