Data 603 MSML
Principles of Machine Learning

Balamurugan Manickaraj
UID: 121291417

14 October 2024

Project Title: Predictive Modeling for Urban Tree Health

1 Research questions: Describe your research question. What were you trying to find in the dataset? Describe why is this problem and important? How did you formulate the problem? Which ML task you used? Is it classification, regression, object detection, segmentation, or unsupervised data mining?

This project's main research question is: How can machine learning be used to predict the health status of urban trees based on various environmental and physical factors? This involves developing predictive models that classify trees into health categories such as healthy, sub-healthy, or poor health. The focus is on analyzing features like tree diameter, species, location, and environmental conditions to determine their impact on tree health. The machine learning task employed here is classification, as the goal is to assign discrete labels to each tree's health status based on these features.

This problem is important because trees play a critical role in maintaining ecological balance, improving air quality, and enhancing environments. By accurately predicting tree health, urban planners and environmentalists can make informed decisions about resource allocation for tree maintenance and care. This leads to more efficient urban forestry management, ensuring the longevity and vitality of city trees. Ultimately, the project aims to contribute to sustainable urban development by leveraging data-driven insights to support healthier urban ecosystems.

Using this method, the study hopes to determine if patients with persistent plantar heel pain can experience a placebo effect by feeling they are receiving an active medication. Only the group who believes they are receiving an active treatment, according to the researchers' hypothesis, is going to significantly reduce their feeling of pain due to placebo-induced anesthesia. This approach contributes to a better understanding of the psychological processes behind the placebo effect by not only trying to determine the amount of the benefit but also offering insights into how patient expectations can affect clinical outcomes.

2 Dataset: Describe the dataset that was used. What was the modality, size, sample size, features and labels (if any), how the data is collected (if known)? Describe the importance of this dataset (if any).

2.1 Dataset: The dataset used for this project is the "2015 Street Tree Census" from New York City Open Data.

Modality: The dataset consists of structured tabular data, capturing detailed information about urban trees, including their health status and physical characteristics.

Size and Features: The dataset includes information on over 683,788 trees with 45 features. These features encompass various aspects such as tree species, diameter, health status, and precise geospatial coordinates.

Sample Size: The full dataset is utilized without downsampling, providing a comprehensive view of the urban forest ecosystem in New York City.

Labels: The dataset is used for classification tasks, with the health status of trees serving as the target variable. Labels include categories like healthy, sub-healthy, and poor health.

Data Collection: The data was collected through the 2015 Street Tree Census conducted by New York City Open Data. It involved volunteers and city staff, ensuring a wide-ranging and detailed collection of tree-related data across the city.

By analyzing this data, researchers and urban planners can gain a better understanding of the factors affecting tree health and make informed decisions about resource allocation for maintenance and care.

3 Reason that the dataset is chosen: Describe the reason of you selecting this dataset. Is it job related, personal interest, or just curiosity?

I chose the "2015 Street Tree Census" dataset out of curiosity and a desire to explore how machine learning can impact environmental benefits, particularly in urban settings. The dataset offers a comprehensive view of New York City's urban forestry, which piqued my interest in understanding the factors affecting tree health. By analyzing this rich dataset, I aim to gain insights into how data-driven approaches can support sustainable urban development and contribute to greener city environments. This project aligns with my personal interest in environmental sustainability and urban planning, providing an opportunity to apply machine learning techniques to real-world ecological challenges.

4 Methods that will be used: List the algorithm(s) that you are planning to test. Why did you choose these algorithms? This is a tentative list that can change as you learn to use more algorithms.

For the project "Predictive Modeling for Urban Tree Health," I plan to experiment with several machine learning algorithms to determine the most effective model for predicting tree health. The initial list of algorithms includes:

Logistic Regression: This algorithm will serve as a baseline model due to its simplicity and effective-

ness in binary classification tasks. It helps in understanding the influence of individual features on tree health and provides a straightforward interpretation of results.

Random Forest: Chosen for its ability to handle large datasets and capture complex interactions between features, Random Forests are particularly effective in dealing with class imbalance and providing robust predictions. Their ensemble nature helps reduce overfitting, making them suitable for this dataset.

Decision Tree: Selected for its interpretability and capability to model non-linear relationships, Decision Trees allow for easy visualization of how different factors contribute to tree health. This makes them valuable for understanding the decision-making process within the model.

XGBoost: Known for its high performance on structured data, XGBoost provides robust predictions and can handle missing data effectively. Its gradient boosting framework enhances accuracy and efficiency, making it a strong candidate for this project.

Also, I may explore additional algorithms like Neural Networks or Support Vector Machines (SVM) to further improve model performance and capture more complex patterns within the data. Comparing these models will help identify which one offers the best predictive accuracy and reliability for urban tree health classification.

5 Preprocessing: If this applies to your problem, describe any preprocessing steps you will be using and why? Are there any missing data? Do you need to clean the data? If so, describe how are you planning to handle those?

Firstly, Handling Missing Data is a critical step, as the dataset contains missing values across multiple features. Categorical missing values will be addressed by replacing them with the mode of their respective columns, ensuring that the categorical distribution is preserved. For numerical columns, a strategy based on the presence of outliers will be employed: median imputation will be used where outliers are present, while mean imputation will be applied otherwise. This approach ensures that the dataset remains robust and suitable for subsequent analysis, minimizing the impact of missing values on model performance.

Feature Encoding is necessary for categorical variables such as tree species and health status, which will be transformed using label encoding to make them compatible with machine learning algorithms. Additionally, Data Scaling will be performed on numerical features using StandardScaler to ensure that all features are on a comparable scale, which can improve the convergence of machine learning models. To address potential class imbalance in the dataset, resampling techniques such as oversampling of minority classes or undersampling of majority classes may be employed to ensure balanced representation in training data.

6 Hyperparameters: [List the initial set of hyperparameters that will be used. The hyperparameters may change in the final report.]

The initial set of hyperparameters I plan to use includes:

Random Forest:
- Number of estimators: 100
- Maximum depth: None
- Minimum samples split: 2

Logistic Regression:
- Regularization (C): 1.0
- Maximum iterations: 100

XGBoost:
- Learning rate: 0.1

- Number of estimators: 100
- Maximum depth: 6
- Subsample: 0.8

Support Vector Machine (SVM):
- Kernel: 'RBF'
- Regularization parameter (C): 1.0
- Gamma: 'scale'

These values will serve as the starting point, and I will use grid search and cross-validation to optimize these hyperparameters as I proceed with model development and evaluation.

7 Performance Metrics: List and describe the metrics that will be used to evaluate the performance. Accuracy, mean squared error, mean average precision (mAP),intersection over union (IOU)?

I plan to use standard performance metrics to evaluate and compare the effectiveness of various machine learning models.

Accuracy: Measures the overall correctness of predictions, showing the proportion of correctly classified instances.

Precision and Recall: Evaluates how good the model predicts positive instances (e.g., healthy trees). Assesses the model's ability to capture all actual positive instances, crucial for identifying unhealthy trees.

F1 Score: Balances precision and recall, useful for handling class imbalance.

Confusion Matrix: Provides breakdown of true positives, negatives, false positives, and negatives.

AUC-ROC Curve: Measures the model's ability to classify between classes across various thresholds.