

MSML Final Report

FINAL PROJECT

***Predictive Machine Learning Models for Tree Health Prognosis***

*Submitted by*

**Balamurugan**

**121291417**

**Executive Summary:**

The project aimed to predict the health of trees in New York City using machine learning techniques. Leveraging the NYC 2015 Street Tree Census dataset, I explored various predictive models to classify tree health based on factors such as tree diameter, location, and environmental conditions. The dataset contained information on over 600,000 trees, including their health status, location, and physical characteristics. Through exploratory data analysis, I identified key variables and potential challenges, such as class imbalance and missing data. Various machine learning algorithms were employed: Logistic Regression, Random Forest, and Decision Tree. To improve model accuracy, future efforts could focus on feature engineering, advanced ensemble techniques, and collaboration with domain experts to better understand the underlying factors influencing tree health. Additionally, exploring alternative datasets or incorporating additional environmental variables may provide valuable insights for more accurate predictions.

Top of Form

**Problem Statement:**

The aim of this project is to develop machine learning models capable of accurately predicting the health status of trees in New York City. By leveraging predictive techniques, the project seeks to empower urban planners and environmentalists with actionable insights to optimize resource allocation and streamline tree management processes. By effectively predicting tree health, the project aims to contribute to the sustainable development and resilience of urban ecosystems, ultimately fostering a greener, healthier environment for city residents. Through the creation of robust predictive models, this endeavour aims to harness the power of data-driven decision-making to enhance urban forestry practices and promote the longevity and vitality of city trees.

**Dataset Used:**

The "2015 Street Tree Census," obtained from New York City Open Data, offers a comprehensive and expansive glimpse into the urban forest ecosystem of New York City. With detailed information on over 683,788 trees spanning the city's landscape, the dataset provides a wealth of insights into various aspects of tree life, including species diversity, physical dimensions, health status, and geographical distribution. With 45 variables capturing a wide range of tree attributes, such as species, size, health, and precise geospatial coordinates, this dataset serves as a valuable resource for researchers, urban planners, and environmentalists seeking to understand and manage the intricate dynamics of urban forestry in one of the world's largest metropolitan areas.

**Pre-process of Data:**

Data cleaning is a crucial step in preparing a dataset for analysis, particularly when dealing with missing values and outliers. In this dataset, a significant number of missing values were identified across 11 features, necessitating correction to ensure the integrity of the data. Categorical missing values were addressed by replacing them with the mode of their respective columns, preserving the categorical distribution of the data. Moreover, the presence of outliers in numerical columns was assessed using z-scores, allowing for the computation of the percentage of outliers in each column. To handle null values in numerical columns, a strategy was devised based on the presence of outliers: median imputation was employed in cases where outliers were present, while mean imputation was utilized otherwise. This systematic approach to data cleaning ensures that the dataset is robust and suitable for subsequent analysis, minimizing the impact of missing values and outliers on the validity of results.

**Data Visualization:**

The visualization component of the analysis offers compelling insights into the health and management of trees in New York City. The bar chart depicting tree health showcases a predominantly positive outlook, with the majority of trees being categorized as healthy. However, it also highlights a concerning proportion of trees classified as sub-healthy, representing approximately one in seven trees, alongside a smaller fraction identified as in poor health. This visual representation underscores the importance of monitoring and addressing tree health issues to maintain the overall vitality of the urban forest ecosystem.

Furthermore, the visualization illustrating the presence of guards around trees reveals an encouraging trend, with most trees benefitting from the presence of helpful guards. However, it also acknowledges the existence of harmful guards, signalling a need for further investigation and potential intervention to mitigate any adverse effects on tree health.

Lastly, the bar chart depicting volunteer data collection efforts across different boroughs provides valuable insights into the distribution of data contributors. It highlights the substantial involvement of volunteer teams in Brooklyn and Queens, with notable contributions from NYC Park’s staff in Staten Island. However, it also sheds light on disparities in data collection efforts, with Trees Count staff in certain areas lagging behind their counterparts in terms of data collection. This visualization underscores the importance of equitable data collection practices and resource allocation to ensure comprehensive and accurate data coverage across all boroughs.

**Predictive Models Used:**

In model prediction, accuracy, and optimization, the primary objective is to develop a predictive model that accurately forecasts the target variable based on input features. Initially, categorical columns were encoded using label encoding, ensuring compatibility with machine learning algorithms. The 'health' column was designated as the target variable, while the remaining columns served as input features for prediction.

To evaluate model performance, the dataset was split into 80% training and 20% testing subsets, enabling the assessment of model accuracy on unseen data. Addressing class imbalance is crucial to prevent biased predictions, and this was achieved through data resampling techniques.

Various machine learning algorithms were implemented, including decision trees, Naive Bayes, logistic regression, random forest, XG Boost, and stochastic gradient descent (SGD). Each algorithm was evaluated based on its accuracy in predicting tree health, with random forest achieving the highest accuracy of 91.9%.

Further optimization was pursued through hyperparameter tuning, where grid search techniques were employed to determine the optimal values of hyperparameters for each model. Additionally, k-fold cross-validation was implemented with Support Vector Machine (SVM) using different kernel functions (linear, polynomial, and RBF) to assess model robustness and generalization performance.

To understand model behaviour and prevent overfitting or underfitting, a graph depicting different complexity levels was analysed. This visualization aids in selecting the optimal model complexity that balances bias and variance, ensuring the model's ability to generalize well to unseen data while capturing the underlying patterns in the dataset. Overall, these steps contribute to the development of accurate and reliable predictive models for tree health prediction in New York City.

**Metrics Used To Compare:**

In my project, I evaluated several machine learning models to predict tree health in New York City. After thorough analysis, the Random Forest model emerged as the top performer in terms of accuracy, achieving an impressive accuracy score of 92.2%. This indicates that the Random Forest model correctly classified the health status of trees in the dataset with high precision. Additionally, when considering other metrics such as recall, precision, and F1 score, the Random Forest model also demonstrated strong performance, indicating its ability to effectively capture positive instances while minimizing false positives and false negatives. These metrics collectively suggest that the Random Forest model is well-suited for predicting tree health in urban environments. While other models such as Decision Trees, Naive Bayes, Logistic Regression, XG Boost, and Stochastic Gradient Descent were also evaluated, none surpassed the performance of the Random Forest model in terms of accuracy. Therefore, the Random Forest model stands out as the most effective and reliable option for predicting tree health in my project.

**Conclusion:**

In conclusion, the analysis I aimed to develop accurate predictive models for assessing tree health in New York City using machine learning techniques. Through extensive pre-processing steps, including label encoding, data splitting, and class imbalance handling, I prepared the dataset for modelling. Various algorithms, such as decision trees, Naive Bayes, logistic regression, random forest, XG Boost, and stochastic gradient descent, were implemented and evaluated based on their predictive performance.

Random forest emerged as the top-performing algorithm, achieving an impressive accuracy of 92.2%. Further optimization was pursued through hyperparameter tuning and k-fold cross-validation, ensuring robustness and generalization capability of the models. Additionally, visualizations depicting model complexity helped us understand and mitigate issues of overfitting and underfitting.

Overall, my study provides valuable insights into the factors influencing tree health in urban environments and lays the groundwork for future research and interventions in urban forestry management. By leveraging machine learning techniques and domain expertise, I can enhance my understanding of urban ecosystems and contribute to the sustainable development of green spaces in metropolitan areas.

**Future Research:**

Moving forward, conducting additional research and analysis is essential to comprehensively understand the multifaceted factors influencing the health of various tree species in the urban environment. Exploring the intricate relationships between environmental variables, urban infrastructure, and tree health could provide valuable insights for developing more accurate predictive models. Additionally, considering more advanced machine learning techniques, such as neural networks, may enable the capture of complex, non-linear relationships within the dataset, potentially improving model performance. Collaboration with domain experts, including arborists, urban planners, and environmental scientists, is paramount to integrating domain knowledge into the modelling process effectively. By leveraging the expertise of professionals with practical experience in urban forestry, the accuracy and relevance of predictive models can be enhanced, facilitating informed decision-making and sustainable management of urban tree populations.