University of Texas at Arlington
Applied Statistics and Data Science

6302-MACHINE LEARNING

Street Tree Health Prediction Using Machine Learning

NYC 2015 STREET TREE CENSUS

Group-03(DATA SPATANS)

Presented By:

Steve Adjorlolo
Pranav Karthik Chinya Umesha
Mahender Bandi
Girish Sunkadakatte Chandrappa

INDEX

I. Abstract

Urban forests play a critical role in enhancing the quality of life in cities. Street trees health in urban environments like New York City is increasingly threatened by environmental factors and infrastructure challenges. This study has leveraged the data from the 2015 NYC Street Tree Census to analyse health patterns across different geographic locations, tree species and environmental factors mainly associated with the tree health. By applying machine learning models, we aimed to accurately classify the tree based on their health condition and highlighting the trees whose health are at risk. Among the models trained, XGBoost emerged as the top-performing model, demonstrating strong predictive power in identifying and classifying the trees whose health are at risk. The findings provide actionable insights for urban planners and arborists to prioritize interventions and safeguard the city green infrastructure.

II. Introduction

Our project focuses on applying machine learning classification models to accurately classify the tree based on their health condition and highlighting the trees whose health are at risk. The primary goal is to evaluate the effectiveness of machine learning in identifying and classifying the trees whose health are at risk which provide actionable insights based on the factors affecting the trees health.

Exploratory data analysis (EDA) was conducted to visually examine patterns and trends within the data. Classification models were chosen for its suitability with the binary nature of the target variable, "health". The model's performance was assessed based on the parameters like accuracy, sensitivity and specificity.

Class imbalance in the dataset is addressed with under sampling the majority class approach, and outliers were handled with the by capping method. These preprocessing steps ensured the robustness and reliability of the model. The analysis shows that environmental factors and tree species were highly influencing the target variable "health". Leveraging these insights can help urban planners and arborists to prioritize interventions and safeguard the city green infrastructure.

III. Problem Statement

Urban trees play a crucial role in enhancing city ecosystems and their health is often threatened by physical damage, root constraints, insufficient care and unsuitable planting environments. Traditional inspection approaches are time consuming and limit timely intervention. Our project addresses the need for a scalable, predictive system to assess street tree health across New York City location and species. Using machine learning classification models, we aim to predict the binary health status of trees based on environmental attributes, location and tree species. Mainly focusing on minimizing false negatives to enable early identification of Trees health which are at risk. Furthermore, the analysis explores borough specific vulnerabilities, species resilience, role of stewardship and infrastructure in influencing urban tree vitality.

IV. Summary of Dataset

Data for this study is related to NYC 2015 Street Tree Census, a comprehensive survey of 683,788 trees across New York City conducted by the Department of Parks & Recreation with volunteer support between May 2015 and October 2016.

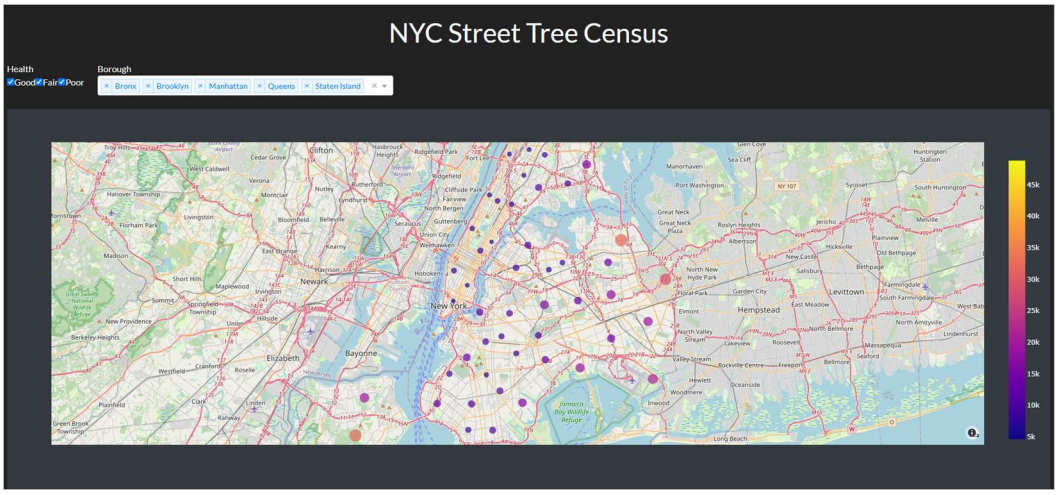
This dataset comprises 683,788 observations with 45 features (7 numerical, 38 categorical) covering Geospatial, Physical Characteristics and Environment characteristics of the trees. For the purpose of predictive modeling, the multi-class feature "health" variable was converted into a binary outcome (Good or Poor).

Dataset was then cleaned and preprocessed through outlier handling, missing value imputation, categorical encoding, feature scaling and under sampling to manage class imbalance, creating a well-prepared dataset suitable for model training.

Geospatial Dashboard https://www.data-spartans.com/

In this project, we set out to transform the 2015 New York City Street Tree Census into an interactive web-based dashboard that makes it easy to explore tree health distribution and demographics across the five boroughs. Leveraging the power of Plotly Dash, a Python framework that wraps React components as simple Python functions, we have built a two page application styled with Bootstrap's Darkly theme. Behind the scenes, the dashboard unites HTML and CSS design elements (through Dash's html and dash bootstrap components) with rich data driven visualizations powered by Plotly Express.

The first page, Map View, invites users to filter the trees by health status (Good, Fair, Poor) via a horizontal checklist and by borough through a multiselect dropdown. Behind the scenes, a callback recomputes district level aggregates, total tree counts, and average coordinates then renders them on an OpenStreetMap background with marker sizes reflecting count, and hover tooltips revealing the exact numbers. A linked data table below lists individual tree ID, addresses, species, health, status, and council district, all styled to match the dark theme.

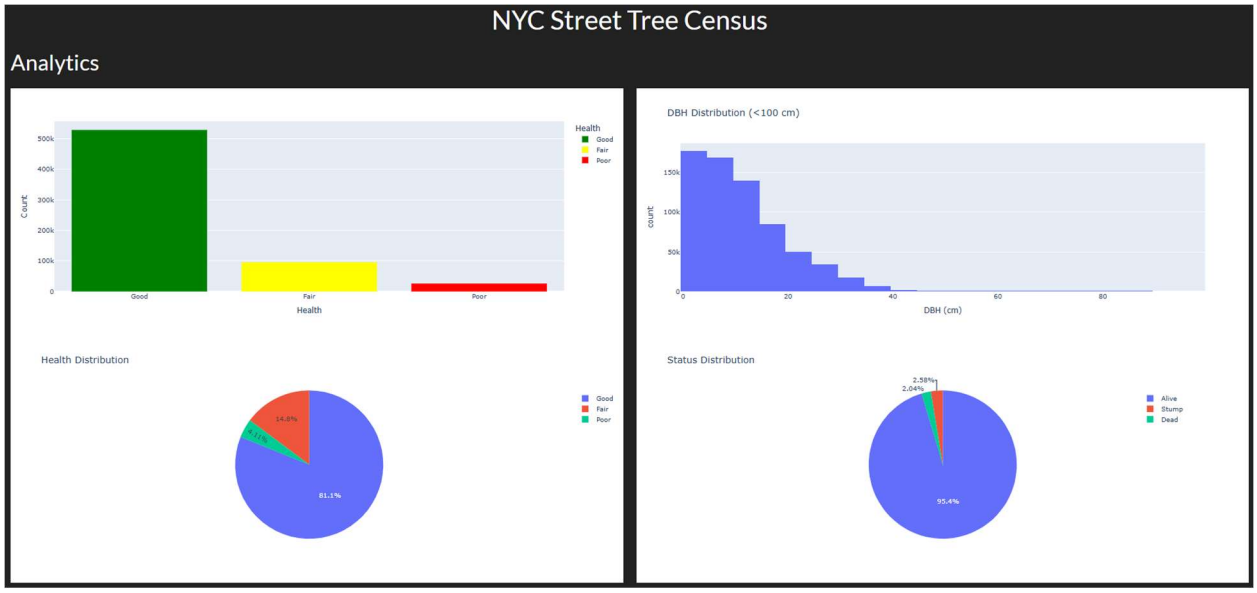


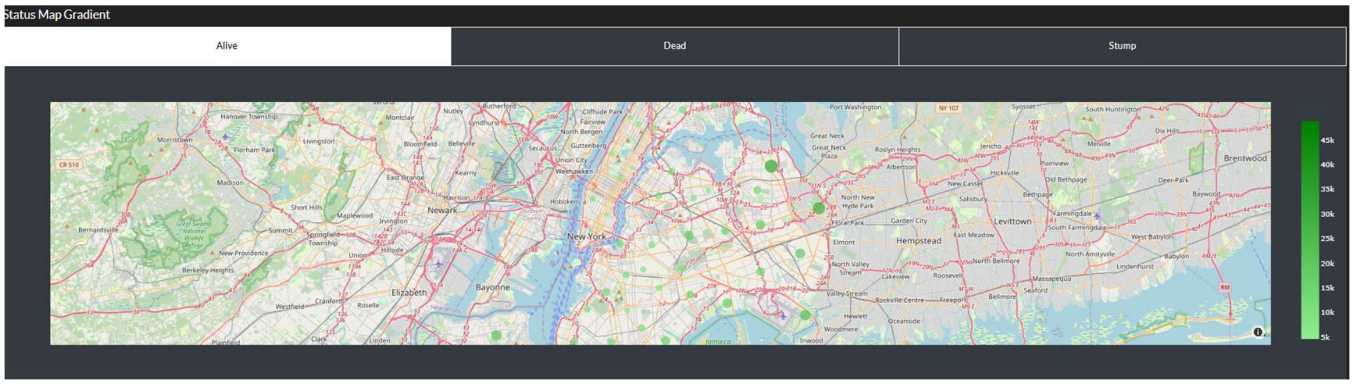
Map Data Table

Tree Id	Address	Spc Common	Health	Status	Council District
180683	108-005 70 AVENUE	red maple	Fair	Alive	29
200540	147-074 7 AVENUE	pin oak	Fair	Alive	19
204026	390 MORGAN AVENUE	honeylocust	Good	Alive	34
204337	1027 GRAND STREET	honeylocust	Good	Alive	34
189565	603 6 STREET	American linden	Good	Alive	39
190422	8 COLUMBUS AVENUE	honeylocust	Good	Alive	3
190426	120 WEST 60 STREET	honeylocust	Good	Alive	3
208649	311 WEST 50 STREET	American linden	Good	Alive	3
209610	65 JEROME AVENUE	honeylocust	Good	Alive	
192755	638 AVENUE Z	London planetree	Fair	Alive	47

1 / 65218

Switching to the Analytics page, the dashboard showcase four static charts a bar chart of overall health counts, a histogram of tree diameters under 100 cm and two pie charts illustrating the distributions of health and status. Below these, a gradient map uses Dash Tabs to let the user select “Alive,” “Fair,” or “Dead” status the map color scales from light to intense green, yellow, or red based on district-level tree counts. This arrangement tells a story at both the macro and micro levels: you can see broad patterns in the summary charts, then drill into spatial clustering on the map.





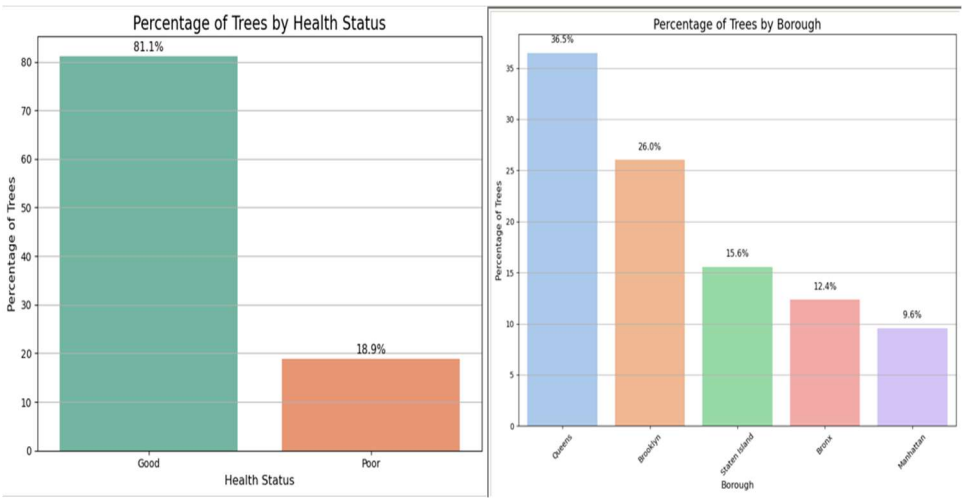
To make this app publicly accessible, we connected the GitHub repository to Render.com. Finally, we bound the custom domain, “www.data-spartans.com” to the Render service by adding a CNAME record for www and A-records for the root domain using Namecheap DNS settings. Explore the dashboard at https://www.data-spartans.com, where every GitHub push triggers an automatic redeploy so the live site always reflects the latest code

Feature details:

Category	Features
Tree Characteristics	tree_id, block_id, tree_dbh, stump_diam, spc_latin, spc_common, status, health, problems, steward, guards, sidewalk
Damage Indicators	curb_loc, root_stone, root_grate, root_other, trunk_wire, trnk_light, trnk_other, brch_light, brch_shoe, brch_other
Location & Geography	borough, borocode, boro_ct, nta, nta_name, postcode, zip_city, address, state, latitude, longitude, x_sp, y_sp, , bin, bbl
Civic Boundaries	community board, cncldist, council district, st_assem, st_senate, census tract
Metadata	created_at, user_type

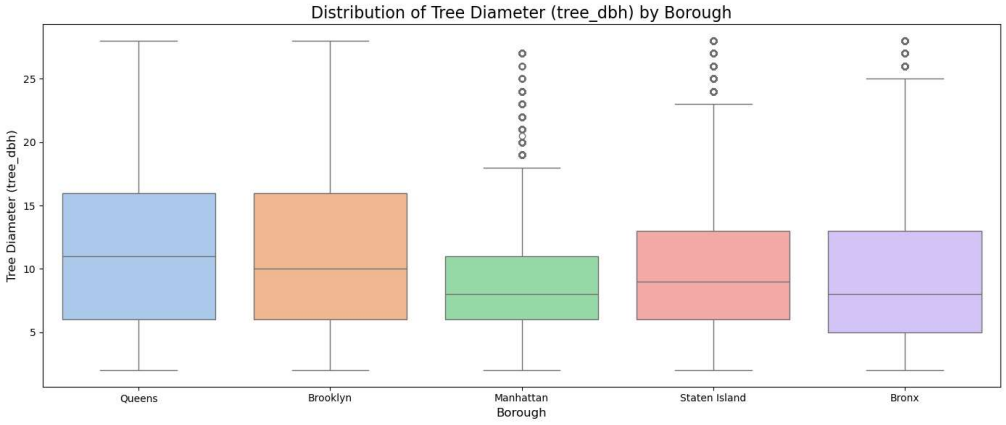
V. Exploratory Data Analysis

1) Borough and Target Feature Distribution



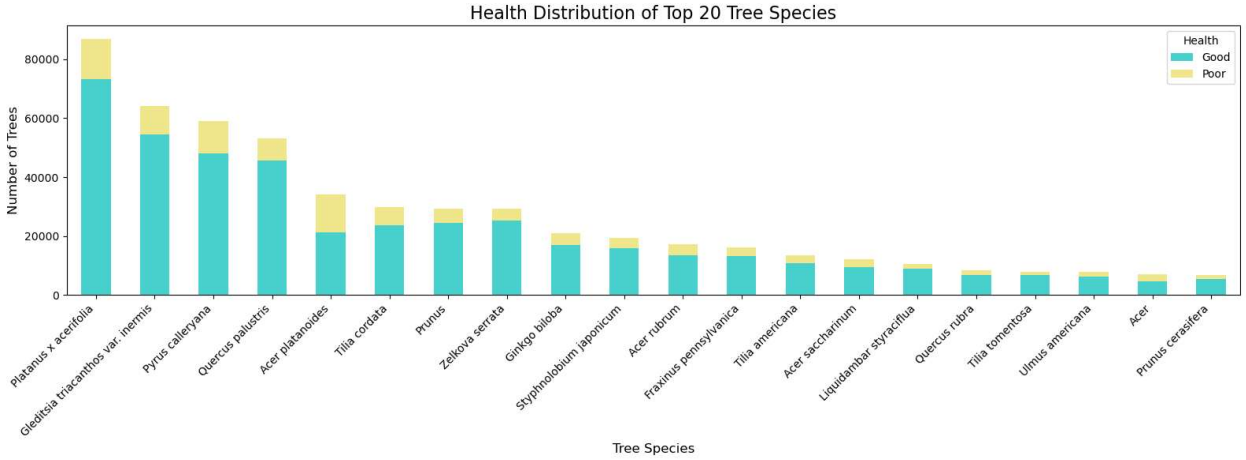
- The bar graph illustrates the percentage distribution of street trees by health status, categorized as "Good" or "Poor." A significant majority of the trees about 81.1%, classified as being in good health, while only 18.9% are categorized as having poor health. This indicates a notable imbalance in the dataset which will be addressed prior model training.
- The bar graph displays the percentage of street trees distributed across the five boroughs of New York City. Queens accounts for the largest proportion of trees at 36.5%, followed by Brooklyn at 26.0%. The remaining boroughs, Staten Island, Bronx, and Manhattan, have progressively smaller percentages of the city's street tree population.

2) Tree diameter distribution across all Boroughs.



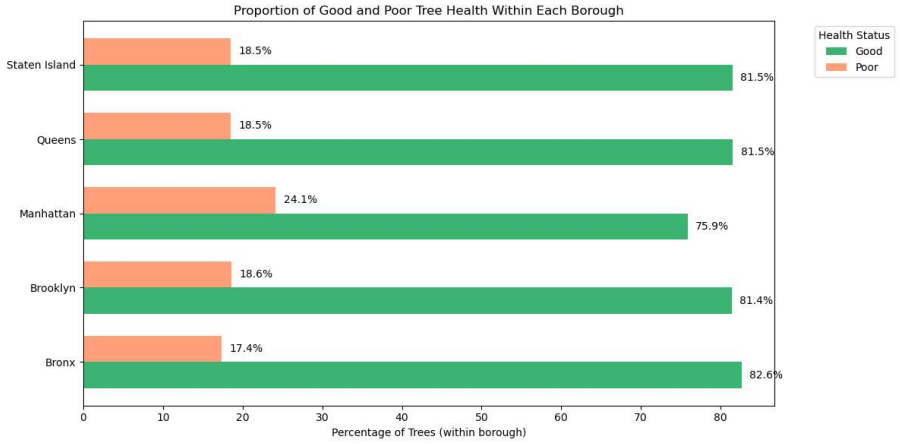
- Above boxplot compares the distribution of tree diameters across the five boroughs of New York City.
- Queens and Brooklyn exhibit a wider range of tree diameters and generally larger median diameters compared to Manhattan, Staten Island, and the Bronx.
- Manhattan shows a more compressed distribution with a lower median diameter and several outliers indicating exceptionally large trees.

3) Health Distribution of Top 20 Tree Species



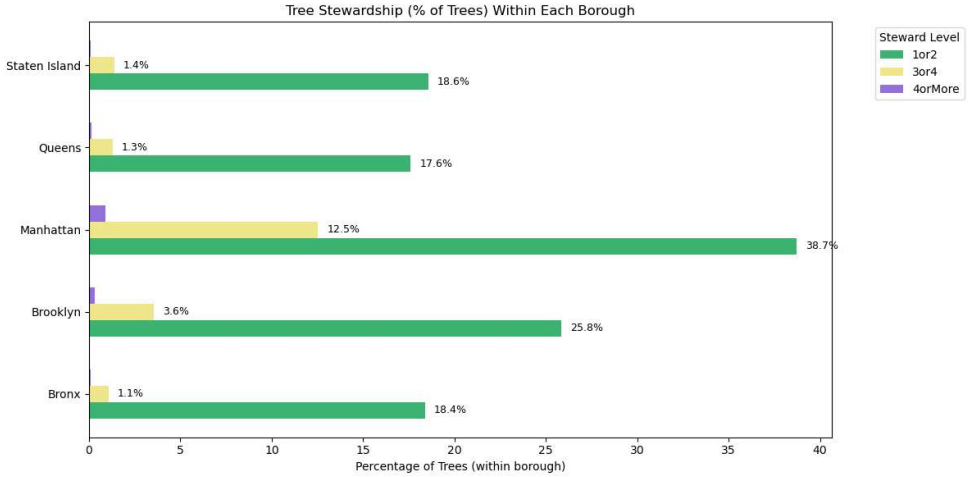
- Platanus x acerifolia is the most abundant species and has a large portion of trees in poor health.
- Gleditsia triacanthos var. inermis and Pyrus calleryana also rank high in total count and show noticeable poor health counts.
- Quercus palustris has high abundance but relatively fewer poor-health trees.
- Ginkgo biloba and Prunus species have moderate counts, but with a high proportion of trees in good health and small poor-health segments, suggesting they are more resilient and better suited for urban environments.

4) Proportion of Good and Poor trees in each borough



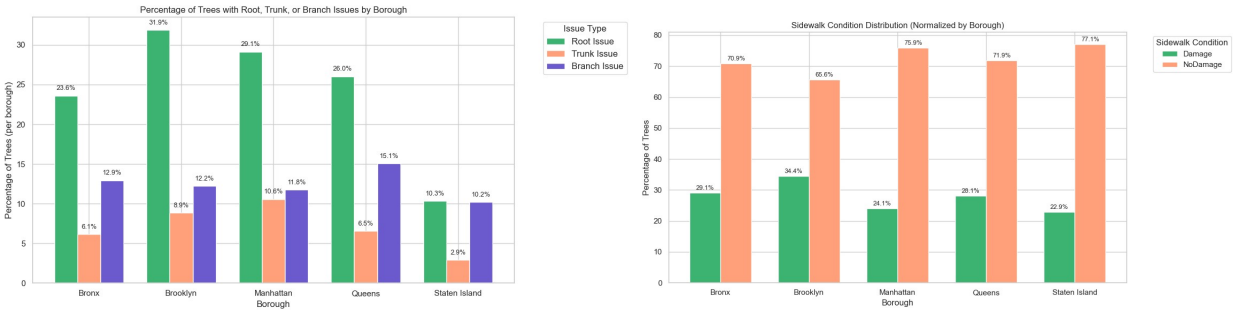
- Above, chart shows that majority of trees in all five boroughs are in good health, with Queens and Staten Island having the highest percentages (81.5%).
- Manhattan has the lowest, with only 75.9% of trees in good condition and the highest proportion of poor health (24.1%).
- This suggests borough level differences in environmental conditions or tree care practices.

5) Tree Stewardship within each borough



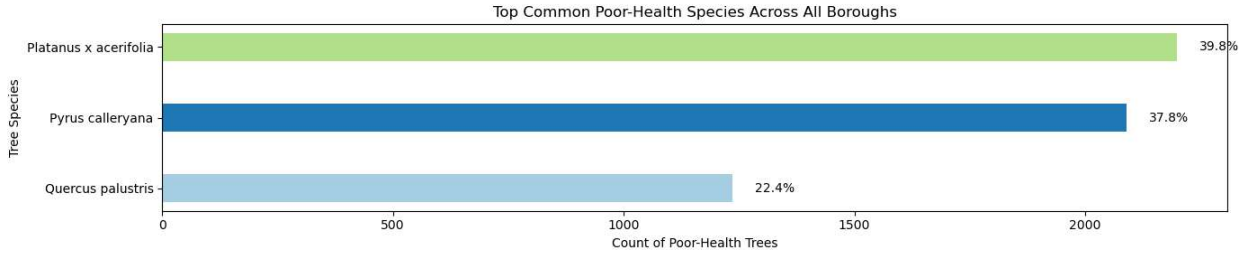
- Brooklyn also has a relatively strong stewardship presence (38.7%), while the Bronx, Queens, and Staten Island exhibit lower levels of community tree stewardship.
- Manhattan has the highest level of active stewardship, with 12.5% of trees under 3or4 stewardship indicating strong community involvement.
- Bronx, Queens and Staten Island have over 90% of trees with minimal stewardship.

6) How do different types of problems vary by borough?



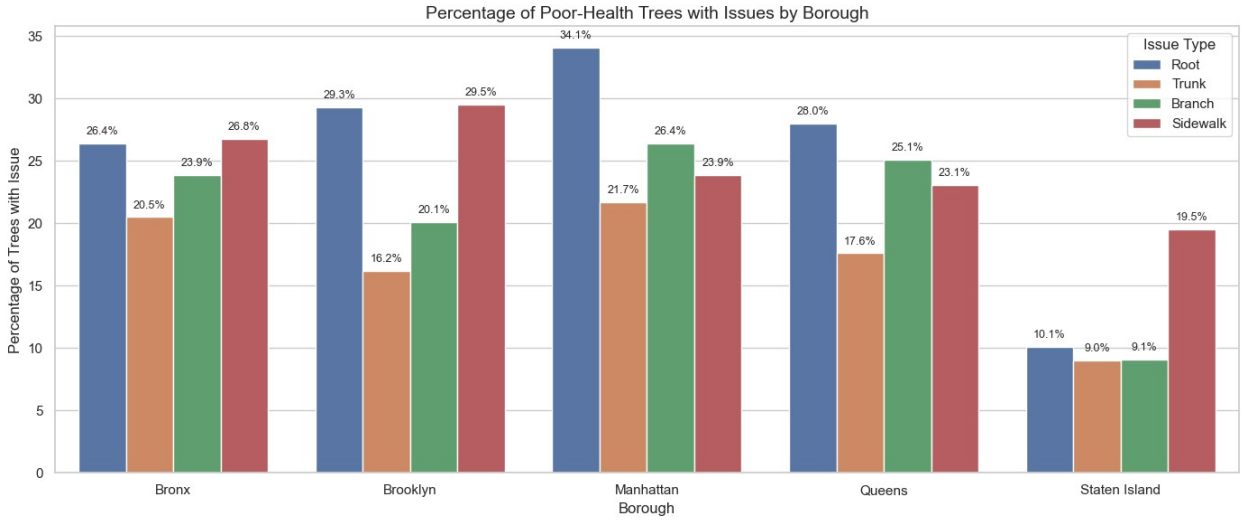
- Brooklyn and Manhattan have the highest percentage of trees with root issues suggesting potential underground or planting site problems.
- Branch issues are most common in Queens (15.1%) and Bronx (12.9%), possibly indicating aging.
- Staten Island consistently reports the lowest percentages across all issue types, implying relatively healthier tree structures.
- Brooklyn has the highest percentage of trees associated with sidewalk damage (34.4%), indicating potential root-related uplift or structural impact.
- Staten Island and Manhattan report the lowest sidewalk damage rates, under 25%, suggesting more stable sidewalk tree interactions in those boroughs.

7) Most Vulnerable Species in NYC



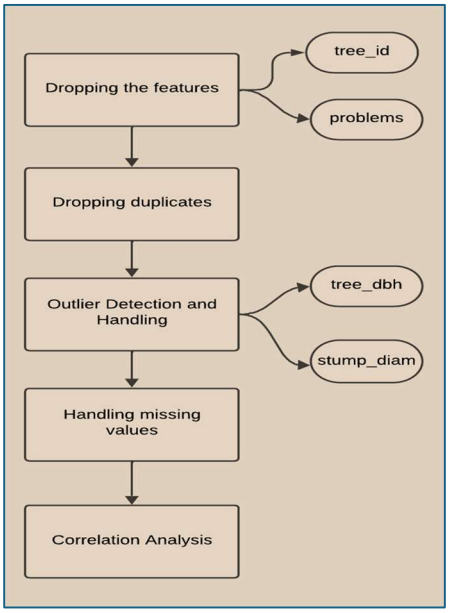
- Platanus x acerifolia is the most frequently found species in poor health across all boroughs indicating a highly vulnerable species despite its popularity.
- Other notable species showing significant vulnerability include Pyrus calleryana and Quercus palustris, suggesting these species may require closer monitoring.

8) The most common issues in the Poor Health tree



- Manhattan has the highest rate of root-related issues (34.1%), while Brooklyn shows the highest sidewalk-related concerns (29.5%).
- Staten Island consistently reports the lowest percentage of issues across all categories, suggesting comparatively better conditions.
- Sidewalk problems are prominent in most boroughs, especially in Brooklyn, Bronx, and Manhattan indicates the infrastructure might not be good.

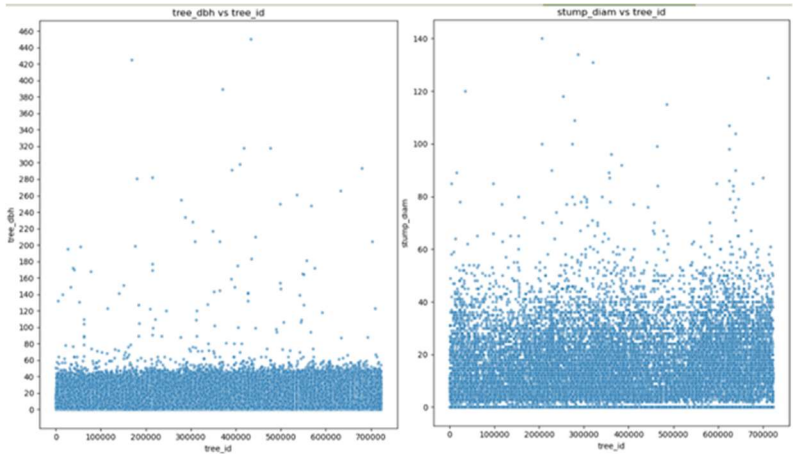
VI. Data Preprocessing



- Features with unique values (tree_id), redundant feature like “problems” and feature with single class “status” are dropped.

Handling Outliers

- Outliers in the numerical features like tree_dbh and stump_diam are detected and handled via capping method.



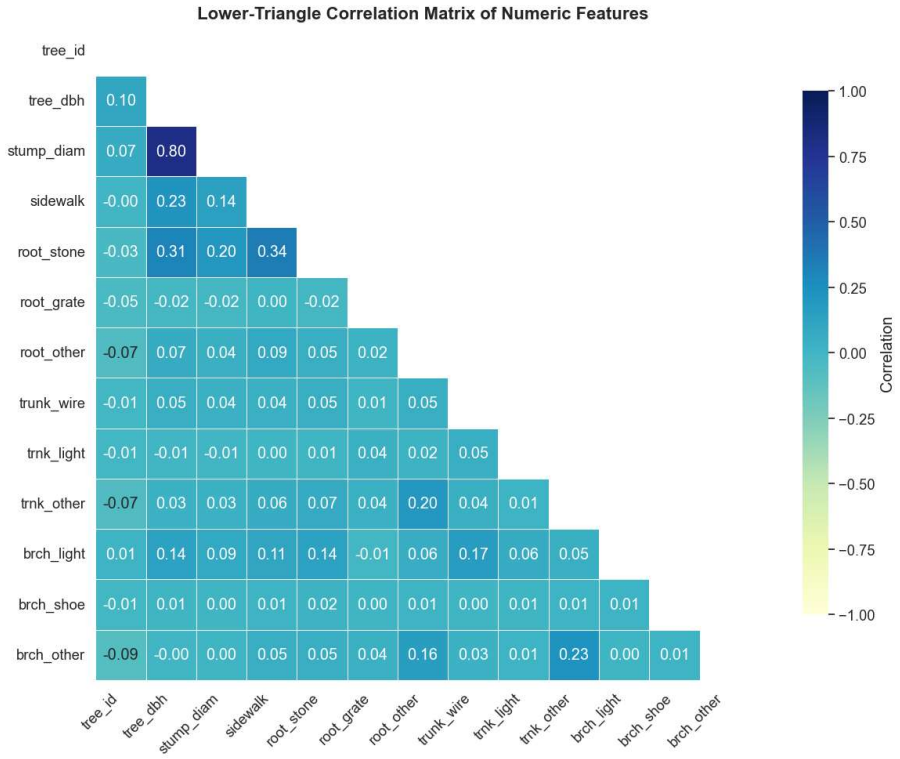
Handling missing values

- For the tree status with Stump and Dead was not having details related to their species and factors affecting tree health hence we have updated as No observation.
- Binary features were converted into factors.
- Target “health” was a multi class feature was converted into binary class by combing Fair and Poor as Poor. Also converted them into factor with “1” indicating Good health and “0” with poor health.

Handling data imbalance

- Target variable is highly imbalanced, and imbalance issue is handled by Under sampling the majority class (Good).
- Encoded the categorical columns.

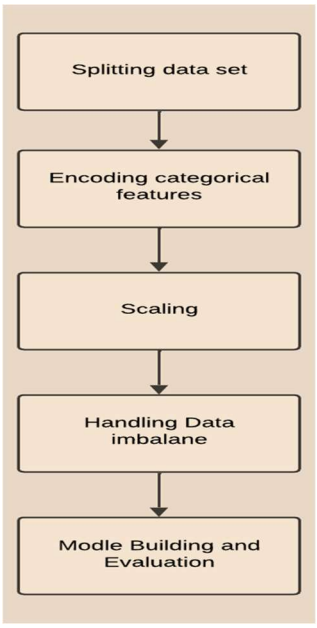
Correlation Analysis.



- Correlation analysis was done on the numerical features and found that feature tree_dbh and stump_diam are highly correlated.

VII. Model Building and Evaluation

Work flow



Due binary nature of the Target Feature “Revenue” Classification models are best suited for prediction. We have performed 4 different Classification models with and without hyper parameter tuning. Cross validation is performed to analyze the performance consistency.

Models without hyper parameter tuning.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Recall (Class 0)
Logistic Regression	0.67	0.7508	0.67	0.6991	0.6519	0.5006
Decision Tree	0.7087	0.7513	0.7087	0.7263	0.6498	0.4359
Random Forest	0.6974	0.7582	0.6974	0.7205	0.6714	0.4908
XGBoost	0.6795	0.754	0.6795	0.7067	0.6604	0.501

- Decision Tree model achieved the highest overall accuracy of 0.7087 and F1-score of0.7263 when compared with other models.
- Random Forest had the highest AUC value of 0.6714, suggesting slightly better discrimination between classes.
- XGBoost showed balanced performance with decent precision and recall, and the highest recall for Class 0 of 0.5010, making it more effective in identifying the minority class compared to others.
- Cross validation analysis confirmed that Random Forest and XGBoost maintained the most consistent and highest mean F1 scores, indicating reliable generalization performance.

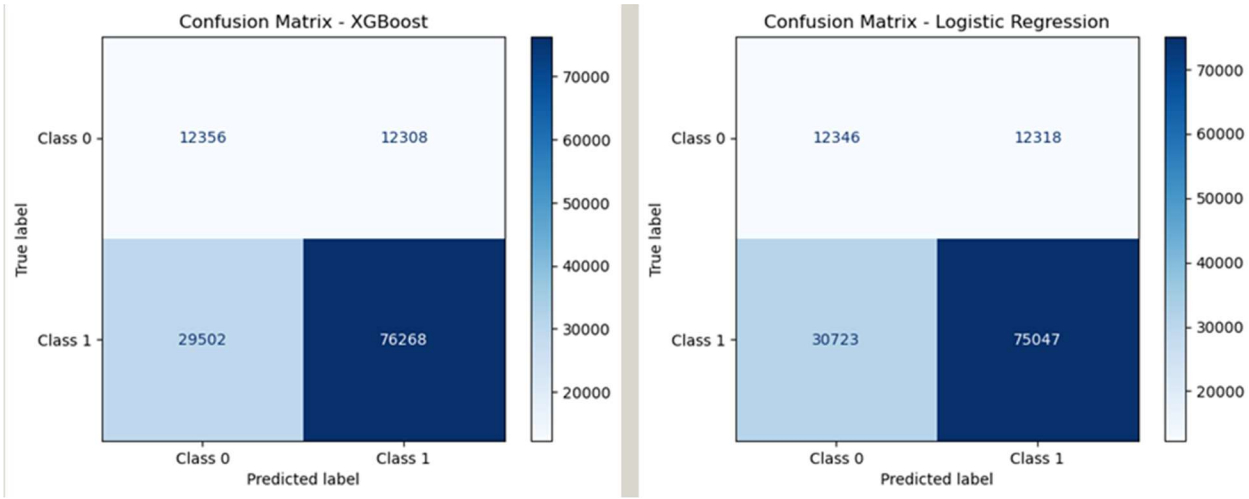
Models with hyper parameter tuning.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Recall (Class 0)
Logistic Regression (HP)	0.6686	0.7504	0.6686	0.698	0.6519	0.5011
Decision Tree (HP)	0.7473	0.7482	0.7473	0.7478	0.6249	0.3366
Random Forest (HP)	0.7228	0.7544	0.7228	0.7363	0.6608	0.4237
XGBoost (HP)	0.7099	0.7497	0.7099	0.7265	0.6471	0.4256

- Logistic Regression hyper parameter tuning model achieved moderate overall performance with an F1-score of 0.6980 and accuracy of 0.6686. Model had the highest Class 0 recall (0.5011), making it most effective in detecting minority cases.
- Decision Tree model with hyper parameter tuning model produced the highest accuracy (0.7473) and F1-score (0.7478), showing strong overall predictions. However, it had the lowest Class 0 recall (0.3366), indicating poor sensitivity to the minority class.
- Random Forest model hyper parameter tuning demonstrated balanced and robust performance, with an F1-score of 0.7363 and the highest AUC (0.6608). It maintained good precision and decent Class 0 recall (0.4237), making it the most reliable overall.
- XGBoost hyper parameter tuning model offered competitive results with an F1-score of 0.7265 and accuracy of 0.7099.
- Its Class 0 recall (0.4256) slightly outperformed Random Forest, making it strong for minority class detection.
- Post comparing the performance metrics of the models with hyper parameter tuning
- Random Forest with hyperparameter tuning is the most balanced and robust model with high accuracy, precision, and recall. Also, it has the better AUC and reasonable sensitivity to Class 0.
- It offers a good trade-off between detecting both classes and overall performance quality.

Confusion matrices for Top performing models.

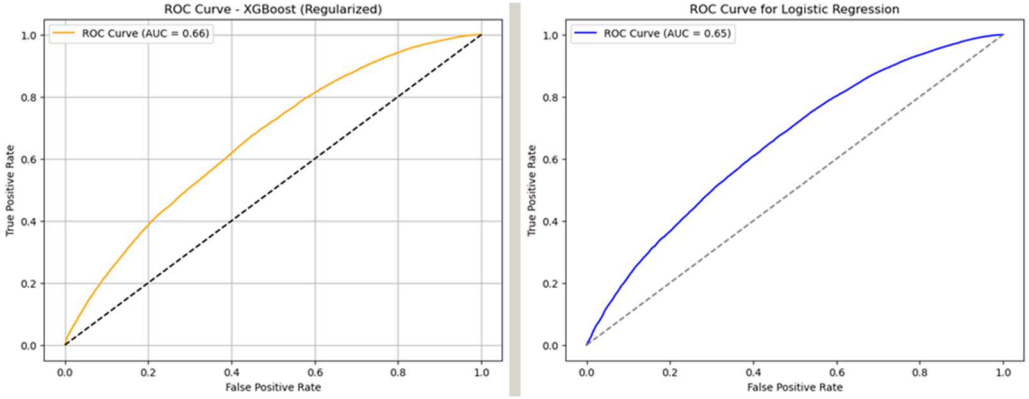
- **XGBoost and Logistic Regression**



- Class 0 Detection: XGBoost correctly identified 12,356 Class 0 instances, while Logistic Regression identified 12,346—indicating near-equal performance in minority class recognition.
- Class 1 Detection: XGBoost had fewer false negatives for Class 1 (29,502) compared to Logistic Regression (30,723), showing stronger sensitivity to detecting at-risk trees.

- Overall Balance: XGBoost maintained a better balance between true positives and false positives, reducing the likelihood of missclassification cases.
- Model Choice Justification: The confusion matrix reinforces XGBoost suitable for this classification task, especially when minimizing false negatives is a key objective.

ROC Curves for both models

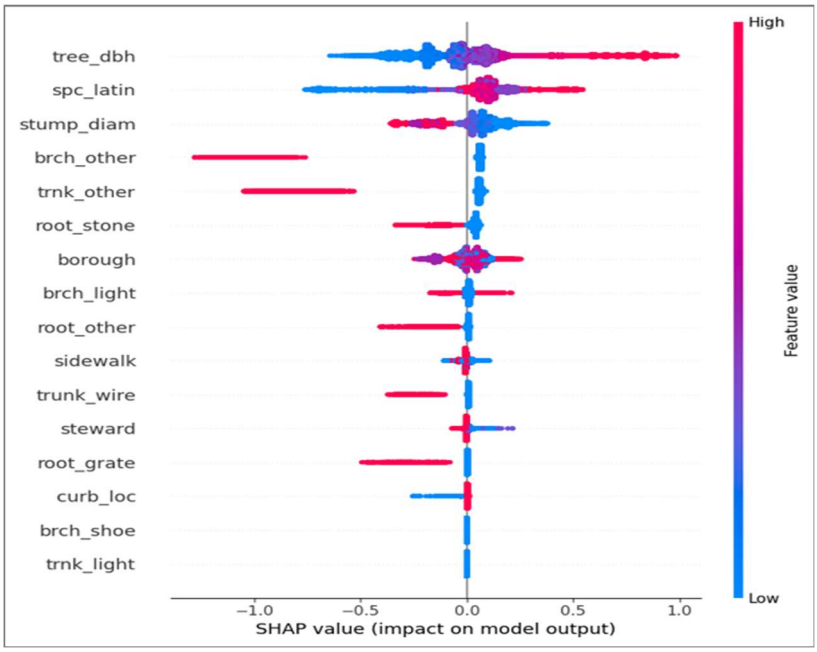


- The ROC curve shows the trade-off between true positive rate and false positive rate across thresholds. XGBoost achieved a marginally higher AUC (0.66) compared to Logistic Regression (0.65), indicating slightly better discriminatory power. Both curves rise above the diagonal, confirming that the models perform better than random guessing.
- XGBoost achieved the highest overall F1-score and AUC, making it slightly better suited for this classification problem, particularly when working with complex patterns across tree species, locations, and condition variables. However, Logistic Regression offers greater model interpretability, which can be valuable in policy or community-facing applications.

Key insights from model analysis

- Both the Logistic Regression and XGBoost models performed about the same in figuring out which trees were in poor health.
- XGBoost model was slightly better overall. It had a little higher scores in things like precision, F1 score, and AUC.
- Logistic Regression is simpler to understand while XGBoost can handle more complicated patterns in the data.
- XGBoost model is better at finding poor health trees and handles complexity hence from our analysis this is best model.

Feature Importance - XGBoost



- spc_latin and tree_dbh have the highest impact with both high and low values have significant influence on target.
- brch_other and trnk_other are also strong contributors, with consistently negative SHAP values suggesting their presence tends to reduce the model's prediction score.
- Other features like stump_diam, root_stone, and borough have moderate influence, while variables such as trnk_light, brch_shoe, and root_grate show very minimal impact.

VIII. Future Scope

Our project presents a strong foundation for data-driven urban forestry analysis and can be extended in the following ways:

- **Advanced Feature Engineering and Class Imbalance Handling**
  Future work can focus on incorporating more relevant features (e.g., soil type, historical maintenance) and applying advanced imbalance-handling techniques to improve prediction accuracy, especially for underrepresented health categories.
- **Integration of Future Tree Census Data**
  Using updated census datasets in future years will enable longitudinal analysis and help evaluate model generalization, improving robustness over time.

- **Species Trends and Urban Forest Planning**
  Analyzing long-term species performance across neighborhoods can support biodiversity goals and guide data-informed tree planting strategies.
- **Environmental and Socioeconomic Data Fusion**
  Combining tree data with external datasets such as air quality, heat island maps, and socioeconomic indicators can enhance insights into environmental equity and resilience.
- **Development of Interactive Tools**
  Building interactive dashboards and incorporating real-time feedback from field experts and the public can support continuous monitoring and iterative model improvement.

IX. Conclusion

In conclusion, our work proves that we can successfully use machine learning to figure out the health of street trees. The XGBoost model is good at this because it can understand complex details in the tree data. By knowing which trees are at risk and why, we can help the city take care of them more effectively.

X. References

- NYC Open Data. (2016). 2015 Street Tree Census: Tree Data. NYC Department of Parks and Recreation. *https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh*
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. *https://doi.org/10.1023/A:1010933404324*
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). *https://doi.org/10.1145/2939672.2939785*
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1705.07874
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
- Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021. *https://doi.org/10.21105/joss.03021*