

Urban Treehuggers Project Final Report

Introduction: Trees are crucial in cities, lowering air temperatures, providing shade, filtering pollutants, and absorbing UV radiation (Nowak and Heisler, 2010), as well as managing stormwater and providing eco-friendly community spaces (Nitoslawski et al., 2019). Naturally, the maintenance and design of urban forests are vital for smart city planning. We seek to investigate the factors that influence the health of trees in US cities, and how it relates to environmental quality and public health, particularly in Los Angeles.

Problem Definition: Urban forests are vital for maintaining environmental quality and community well-being. However, there is limited integrated research that connects tree health, environmental pollution, and public health outcomes. Our project explores the health of urban forests across US cities and evaluates how environmental conditions affect both tree health and public health, with a specific focus on Los Angeles.

Literature Survey: Current studies on urban forests remotely monitor tree dimensions, species diversity, and moisture levels via LiDAR and computer vision (Estrada et al., 2023), along with image segmentation (Francis et al., 2023). Nowak et al. (2014) study the impact of healthy trees using datasets such as the National Land Cover Database and the U.S. EPA's BenMAP to get estimates for tree cover, hourly pollution reduction, and other factors. Other studies exist that use longitudinal maps to analyze urban forest growth potential (Liu et al., 2024). Previous research has shown that trees are crucial in cities for lowering air temperatures, filtering pollutants, and absorbing UV radiation (Nowak and Heisler, 2010). Urban forests also help with energy savings, increase biodiversity, and provide community spaces, so their growth, maintenance, and design are critical for smart urban planning (Nitoslawski et al., 2019). Irga et al. (2015) find that particulate matter fractions are negatively correlated with urban forest density. With over 90% of the world's population in regions where air pollution exceeds safe limits (Oliveira e Almeida et al., 2020), urban forests are key to improving public health. Previous studies have shown the impact of quantifying individual and organizational use of carbon in promoting individual contributions to urban forests (Rowntree and Nowak, 1991). Optimally placing green spaces can help reduce health risks associated with poor air quality, improving urban living conditions (Lai et al., 2017).

However, very few studies predict the future of urban forests, but focus on past trends and current conditions (Araújo et al., 2021). Few studies account for variables such as tree height and volatile organic compound emissions, which could unintentionally contribute to pollution further (Nowak, 2002). Environmental impacts of urban forests such as temperature differences (Nowak, 2002) and the effects of species diversity, have also been neglected to be studied in-depth (Chen and Jim, 2008). Attempts to apply the latest AI/ML methods such as explainable artificial intelligence are limited in scope and do not emphasize air quality (Feng et al., 2024). Random forest classification methods have been used to accurately detect forest expansion activities in China, but its lack of generalizability and inefficiency make it hard to apply the method elsewhere (Liu et al., 2024). Beyond quantifying the benefits, there is room for novel contributions in quantifying the costs of maintenance and resilience of urban forests (Ordóñez and Duinker, 2012). Taking into account that urban forest management practices vary across different cities may limit the generalizability of models (Saheer et al., 2022). Few studies integrate urban forests, air quality, and public health (Arantes et al., 2019), highlighting the need for comprehensive projects comparing model predictions with actual assessments across cities, supported by interactive visualizations (Han et al., 2018).

Proposed Method:

Intuition: Our proposed method is better than current approaches because it considers urban forestry, air quality, and human health holistically, which most studies do not. We combined those three topics and used machine learning and deep learning to identify predictive relationships. Our analytics at the city level provide a novel way to assess and compare urban forests and urban forestry management practices.

By determining the variables that influence the health of trees and urban air quality and identifying how other cities can learn from Los Angeles (and vice-versa), along with showing interactive visualizations at both the city and national level, our study goes beyond the state of the art. Our work will help bridge the gap between data analysis and practical decision-making, leading to more informed urban planning.

Detailed Description: We downloaded our datasets from Kaggle ([5M Trees](#), [Urban Air Quality and Health](#)), and first processed them (using pandas and glob) by removing unnecessary columns and combining parts of the datasets. We transformed the datasets so that we can more easily and directly work on them. For visualizations, some outlier trees were filtered out if the location of the tree, as determined by the longitude and latitude coordinate, did not even remotely correspond to the tree's designated city.

In classifying the health of trees, we considered trees from all major US cities. We handled missing data (particularly for tree dimensions) by replacing missing measurements with the average of its city, and computed the age of each tree. Then, using sklearn's LabelEncoder, we encoded categorical variables such as the city a tree is in and its species. Using the variables of interest (species, city, native or not, height, diameter, age), we trained a random forest classifier of initially 25 trees (with sklearn's RandomForestClassifier) to classify the health of a tree (excellent, good, fair, poor, dying, dead). We also visualized each decision tree with matplotlib.

To try to improve the accuracy of our model, we explored and attempted many methods. First, we tuned the hyperparameters (with both sklearn's GridSearchCV and RandomizedSearchCV). We also introduced new features, such as an individual tree's difference in age and height from the averages of its city. To prevent overfitting, we limited the classifier to variables with feature importance of at least 0.01. These approaches did not significantly improve our accuracy, so we implemented an XGBoost model (with xgboost's XGBClassifier) of 100 estimators, and combined the XGBoost with our random forest classifier in an ensemble model (i.e. sklearn's VotingClassifier). Unfortunately, our accuracies never exceeded 76%.

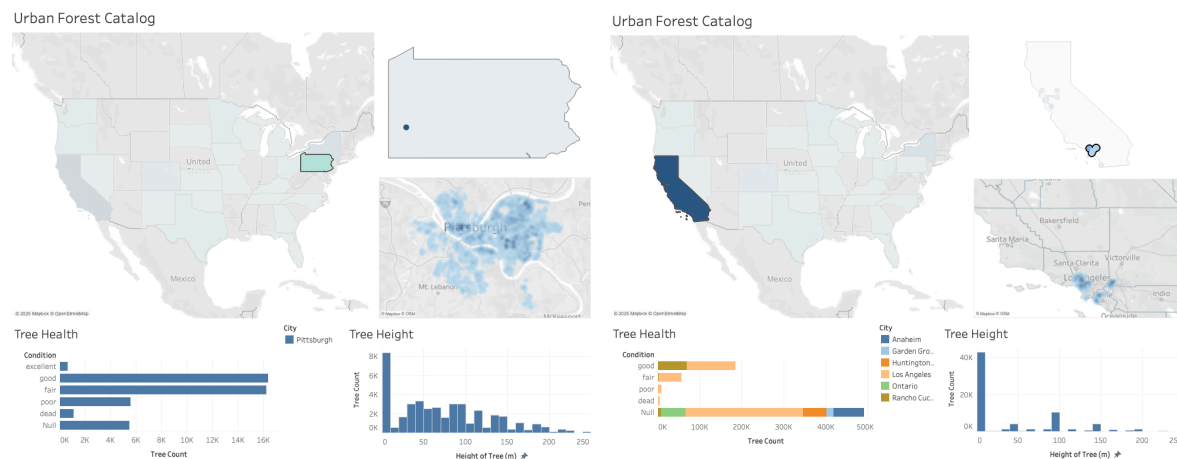
We compared the health status distribution of trees in Los Angeles to those in other major US cities, using seaborn and matplotlib. We computed the percentage of trees in each city with each type of status (i.e. dead, dying, poor, fair, good, excellent), and produced a bar chart visualization.

For analyzing the relation between air quality and health, we focused on air quality in Los Angeles. First, using sklearn's LabelEncoder, we encoded categorical variables such as condition (e.g. clear, cloudy). Then, out of all the possible explanatory variables (e.g. chemical concentrations, air quality index, temperature, humidity, windspeed, UV index), we identified the eight primary variables (using sklearn's RFECV and LinearRegression) that correlate with health risk score. Using those eight features, we implemented a random forest regressor of 10 decision trees (using sklearn's RandomForestRegressor), a gradient boosting regressor (using sklearn's GradientBoostingRegressor), and a deep neural network

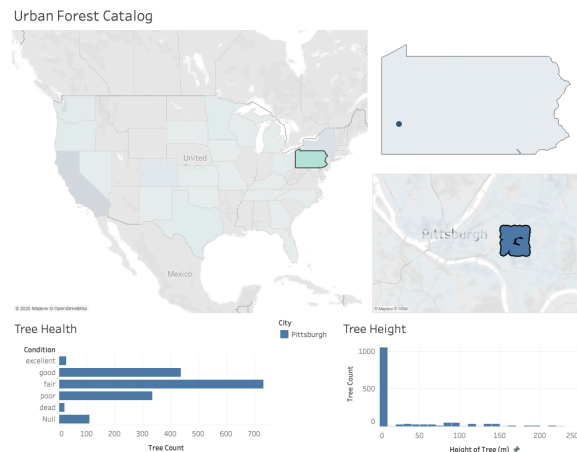
(DNN) (using TensorFlow) and Keras to predict health risk score. For the random forest regressor, we visualized each tree using matplotlib.

Our DNN consists of three hidden layers with 128, 64, and 32 neurons, respectively, each of which leverages the ReLU activation function. The final layer of the DNN outputs a number that represents the health of the tree. Health is usually described in words (such as good or fair), but we wanted to use a numerical encoding so that the model could better learn the gradual differences between different health levels. We compiled the DNN using the Adam optimizer and used mean squared error (MSE) as our loss function. We also included mean absolute error (MAE) as an additional evaluation metric. We trained the model over 50 epochs and tested the model on a separate test set that consisted of 20% of the original data. We evaluated the model's performance using MSE, RMSE, MAE, and the R^2 score and created a visualization that compares the actual vs. predicted health scores using matplotlib.

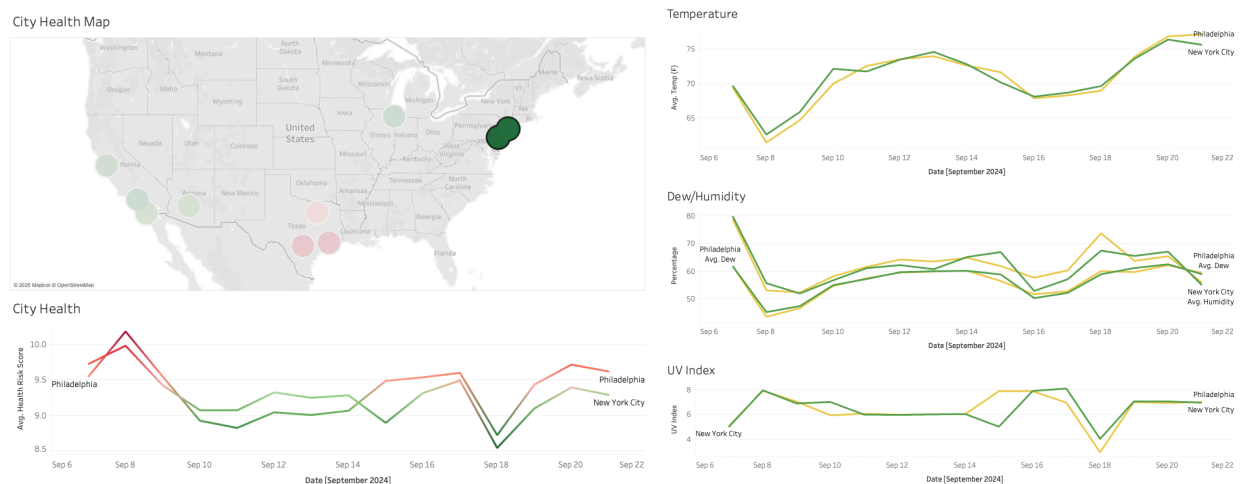
For visualizations, we created two Tableau dashboards to provide interactive catalogs of the urban trees and health quality of select US cities, which complement our models described above. Drill-down features of the dashboards let users easily search through our urban forest catalog geospatially, by state, city, and exact location. Once a specific city or a group of cities are reached, a density heat map visualizes all trees, and aggregate health and height statistics are displayed through stacked bar charts and histograms. For example, the figures below demonstrate the first dashboard with Pittsburgh selected and Los Angeles selected:



Our urban forest dashboard allows urban planners to inspect and compare the urban trees located in major US cities. Bird-eye city views of urban forests can help city planners visualize the relationship between tree density and other urban features, maintain the health of trees, and identify optimal locations to plant trees. The dispersion and density of the trees in the cities can be easily determined, and health statistics can be displayed for any specific subset of urban trees. For example, health and height statistics of trees near Carnegie Mellon University in Pittsburgh can be viewed if that subset of trees is specifically selected by dragging the mouse through Carnegie Mellon University's campus on the density map of Pittsburgh.



Our second Tableau visualization facilitates health and air quality comparisons between select US cities. Cities included in our air quality models are displayed on a US map, color-coded by their average health score. When cities are selected by dragging the mouse over the US map, health, temperature, humidity, dew, and UV index time series trends are displayed over the two weeks of data collection. The figure below displays the dashboard with the East Coast cities selected:



The city health dashboard visualizes correlations between health risk score and environment variables. Correlations are quantified explicitly with the air quality models described above.

Evaluation:

Description of Testbed: There are a few key questions that our experiments, including computational models and visualizations, are designed to answer.

- What are the relative importances of the features that influence air quality in Los Angeles, or tree health in general?
- Given data about a tree, can someone predict its health? Similarly, given a weather forecast, can someone predict the health risk score?

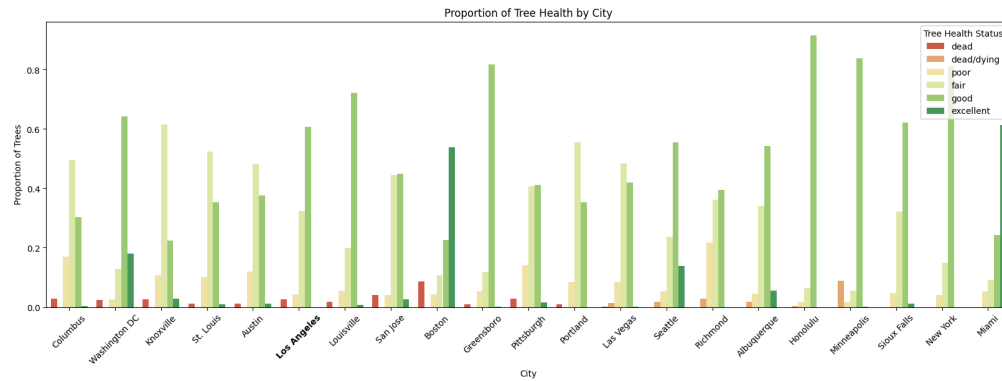
- How intuitive is it to access information about urban forests using our visualizations? How insightful is the information displayed?
- How is the environmental quality and public health in Los Angeles compared to other major US cities? How can Los Angeles learn from other cities' forestry practices to improve its own air quality and public health, or vice-versa?

Description of Experiments and Observations: For our computational models, we aimed to achieve optimal performance measures (e.g. accuracy score for classification and R^2 for regression) when comparing our models' predictions with ground truth data.

For tree health classification, we have an accuracy of no more than 76%, even after extensive hyperparameter tuning with GridSearchCV and combining XGBoost with the random forest in an ensemble model. Our random forest classifier accurately identifies trees with good health status, but seems to struggle with identifying trees that are dead or of poor health status. These results show that the health of trees could be inherently difficult to accurately classify, probably because so many different factors influence their health. Also, there could be other factors, not included in our datasets, that impact tree health; such factors could include presence of bugs, soil quality, and sunlight and rain conditions. Nonetheless, we can conclude that based on the individual decision trees, a tree's diameter at breast height, city, species, and age are the most important factors in classifying its health.

For analysis of air quality and public health in Los Angeles, we achieved an R^2 score of about 0.99 with both random forest and gradient boosting, with predicted results being very close to actual results. In the random forest regressor, minimum daily temperature, humidity, and UV index are the top factors in determining health risk score. Higher daily temperatures and higher humidities tend to increase health risk score, and influence health risk score more than UV index does. Our deep neural network (DNN) achieved an R^2 score of approximately 0.99, which indicates that the predicted and actual health risk scores are very close. The DNN also has a very low mean squared error (0.0012) and mean absolute error (0.026), which again reinforces the close fit between the actual and predicted scores. The value of these metrics suggests that our DNN was able to capture the relationships between different environmental factors and the health risk scores. Furthermore, like in the tree-based models, variables such as daily temperature, humidity, and UV index played a key role in influencing the health risk score, with higher temperatures and humidity levels generally associated with greater health risks.

Comparing the proportions of tree health in US cities, the visualization below shows that Los Angeles has a non-trivial percentage of dead trees and that while most of its trees are in good condition, it does not have any trees in excellent condition (compared to cities like Boston or Miami which have a large proportion of trees with excellent health). These comparisons indicate that Los Angeles could improve its environmental quality and lower its overall health risk score by removing dead trees, taking care of its trees in good condition, and planting new trees in the most optimal locations to achieve excellent health.



We evaluated our visualization dashboards by conducting functional experiments and qualitative interviews. Each participant was asked to drill-down to determine the health of the urban trees near Georgia Tech. The time it takes to quantify the distribution of tree health, as well as qualitative comments on the efficiency, ease-of-use, and quality of the visualizations, were recorded.

The average time it took participants in our study to quantify the distribution of tree health of trees near Georgia Tech using our dashboard was 76.5 seconds. Most participants did not have trouble mentally locating Georgia Tech from the set of maps provided in the dashboard, and many commented that the design of the dashboard made it intuitive to geospatially locate locations of interest. However, many participants struggled with manipulating the state and city map to focus and zoom into Atlanta, which took about 75% of the total time. Tableau requires users to switch viewing modes to switch between moving the map and selecting trees of interest. We've received comments that it was not clear how to switch between modes in Tableau, and more guidance on map manipulation would have been useful. However, once the map manipulation learning curve was overcome, participants had no trouble generating statistics on any subset of trees as they wished.

Conclusion and Discussion:


Our primary goal is to determine the factors that influence the health of urban trees, and based on the air quality and public health status in Los Angeles, determine if Los Angeles could improve by learning from other cities' urban forestry practices, or vice-versa. We achieved outstanding (99%) accuracy for analyzing the relation between environmental quality and health risk in Los Angeles, but only around 75% accuracy for tree health classification. This suggests that tree health could be a lot more nuanced or unpredictable compared to public health risk. Our findings and visualizations will benefit those seeking to improve the environment and public health by planting trees. However, our project focuses on the environmental quality and health in Los Angeles, so it should generalize well to similar large cities like Chicago or New York, but maybe not all cities. Another limitation is that we only considered a subset of all the available features, especially when analyzing the impacts of environmental quality on public health. While it made our computation faster, it also inevitably resulted in some loss of information.

Potential future extensions could be taking into account trends over time to study how quality of urban forestry, air quality, and public health have evolved. It would also be beneficial to focus more on the species of trees and their relative locations and densities in cities to better optimize urban forestry design.

All team members have contributed an approximately equal amount of effort.

References

- Arantes, B. L., Mauad, T., & Da Silva Filho, D. F. (2019). Urban forests, air quality and health: a systematic review. *The International Forestry Review*, 21(2), 167–181.
<https://www.jstor.org/stable/27101467>
- Chen, W. Y., & Jim, C. Y. (2008). Assessment and valuation of the ecosystem services provided by Urban Forests. In *Springer eBooks* (pp. 53–83). https://doi.org/10.1007/978-0-387-71425-7_5
- de Lima Araujo, H. C., Martins, F. S., Cortese, T. T. P., & Locosselli, G. M. (2021). Artificial intelligence in urban forestry—A systematic review. *Urban Forestry & Urban Greening*, 66, 127410.
- e Almeida, L. D. O., Favaro, A., Raimundo-Costa, W., Anhô, A. C. B. M., Ferreira, D. C., Blanes-Vidal, V., & dos Santos Senhuk, A. P. M. (2020). Influence of urban forest on traffic air pollution and children respiratory health. *Environmental monitoring and assessment*, 192, 1-9.
- Estrada, J. S., Fuentes, A., Reszka, P., & Auat Cheein, F. (2023). Machine learning assisted remote forestry health assessment: a comprehensive state of the art review. *Frontiers in plant science*, 14, 1139232.
- Feng, F., Ren, Y., Xu, C., Jia, B., Wu, S., & Laforteza, R. (2024). Exploring the non-linear impacts of urban features on land surface temperature using explainable artificial intelligence. *Urban Climate*, 56, 102045-. <https://doi.org/10.1016/j.uclim.2024.102045>
- Francis, J., Disney, M., & Law, S. (2023). Monitoring canopy quality and improving equitable outcomes of urban tree planting using LiDAR and machine learning. *Urban Forestry & Urban Greening*, 89, 128115.
- Han, J., Zhang, W., Liu, H., & Xiong, H. (2018). Machine Learning for Urban Air Quality Analytics: A Survey. arXiv. <https://doi.org/10.48550/arXiv.2310.09620>
- Irga, P. J., Burchett, M. D., & Torpy, F. R. (2015). Does urban forestry have a quantitative effect on ambient air quality in an urban environment?. *Atmospheric Environment*, 120, 173-181.
- Lai, Y., & Kontokosta, C. E. (2017). Measuring the Impact of Urban Street Trees on Air Quality and Respiratory Illness. arXiv. <https://doi.org/10.48550/arXiv.1710.11046>

- Liu, Z., Zhang, Y., & Zheng, X.. (2024). *Improving urban forest expansion detection with Landtrendr and machine learning*. MDPI. <https://www.mdpi.com/1999-4907/15/8/1452>
- Nitoslawski, S. A., Galle, N. J., Van Den Bosch, C. K., & Steenberg, J. W. (2019). Smarter ecosystems for smarter cities? A review of trends, technologies, and turning points for smart urban forestry. *Sustainable Cities and Society*, 51, 101770.
- Nowak, D. J. (2002). The effects of urban trees on air quality. *USDA forest service*, 96-102.
- Nowak, D., & Heisler, G. (2010). Air quality effects of urban trees and parks. *Research Series Monograph. Ashburn, VA: National Recreation and Parks Association Research Series Monograph. 44 p.*, 1-44.
- Nowak, D. J., Hirabayashi, S., Bodine, A., & Greenfield, E. (2014). Tree and forest effects on air quality and human health in the United States. *Environmental pollution*, 193, 119-129.
- Ordóñez, C., & Duinker, P. N. (2012). Ecological integrity in urban forests. *Urban Ecosystems*, 15(4), 863–877. <https://doi.org/10.1007/s11252-012-0235-6>
- Rowntree, R., & Nowak, D. (1991). Quantifying the role of urban forests in removing atmospheric carbon dioxide. *Arboriculture & Urban Forestry*, 17(10), 269–275. <https://doi.org/10.48044/jauf.1991.061>
- Saheer, L. B., Bhasy, A., Maktabdar, M., & Zarrin, J. (2022). Data-Driven Framework for Understanding and Predicting Air Quality in Urban Areas. *Frontiers in Big Data*. <https://doi.org/10.3389/fdata.2022.822573>
- Urban Air Quality and Health Impact Analysis*. (2024, September 7). Kaggle. <https://www.kaggle.com/datasets/abdullah0a/urban-air-quality-and-health-impact-dataset>
-  5M Trees Dataset. (2023, August 11). Kaggle. <https://www.kaggle.com/datasets/mexwell/5m-trees-dataset>