Tree Canopy Analysis Report: Cherokee Park (2022-2024)

Jonathan Moreira, Chris Myhre, Kadin McWilliams

Introduction and Research Question

Urban parks play a critical role in maintaining ecological balance, supporting biodiversity, and providing recreational and environmental benefits to communities. Olmsted Park, known for its intricate landscaping and rich biodiversity, serves as a vital green space in its region. Over the years, environmental changes, urban pressures, and invasive species have posed challenges to its ecological health. Understanding these dynamics, particularly through the lens of tree canopy cover, is crucial for effective park management and conservation efforts.

Tree canopy cover is a key ecological indicator, influencing factors such as temperature regulation, habitat availability, and soil health. Analyzing changes in canopy cover over time provides insights into the health of the park's ecosystem and the impact of both natural and anthropogenic factors. This project focuses on comparing tree canopy data from 2022 and 2024 to identify regions where the canopy has expanded or reduced and explore potential causes for these changes.

To guide this analysis, we aim to address the following research questions: Is there a relationship between tree canopy and tree species? This question explores whether specific tree species contribute more significantly to canopy coverage and how species composition influences canopy dynamics. What is the relationship between invasive species and canopy changes? This question examines how the presence and activity of invasive species may influence or be influenced by changes in tree canopy, shedding light on potential threats to the park's ecological balance.

By conducting Exploratory Data Analysis (EDA) on the Olmsted Park dataset, this project seeks to uncover trends and patterns related to canopy change. Integrating insights from a recent field trip and visual evidence will further enrich the analysis, offering a holistic understanding of the park's evolving landscape.

EDA Part 1

We found that the top 10 most common canopy species were as follows: Maple, Hackberry, Walnut, Tulip Poplar, Sycamore, Cherry, Beech, Box Elder, Ash and Sugar Maple. However, when looking at the correlations between the top 10 trees and average canopy coverage, there were no strong correlations found between the two.

When looking closer at the correlation matrix you will see that sycamore trees and elder box trees have a moderate correlation. Elder box trees canopies tend to be more vertical while sycamores tend to be more horizontal. These canopy shapes tend to complement each other

meaning that they would be able to more efficiently fill in the canopy space. If there was an effort to plant trees to make up for canopy space, then these two trees might be a good combination.

Similar results were found when looking at the undergrowth as well. When looking at the invasive species however, we found that having a tree downed and Ampelopsis Brevipedunculata, commonly known as the porcelain vine, have moderate negative correlations with average canopy coverage. Considering the English ivy and the wintercreeper too, the invasive species that tend to be together could potentially be easier to treat. Another possibility is that the treatment for the invasive species that tend to be found together could be different meaning only one is being treated. From this EDA we think that the type of invasive species present will have the most significant factor in canopy coverage. We do not think that undergrowth will play a significant factor and the type of tree in the canopy will also not have a significant effect.

A large majority of the data did not change between 2022 and 2024. The only notable difference was that the invasive species correlations were higher and that there were less box elder saplings. I would expect that our machine learning models are going to keep most of their accuracy between years.

EDA Part 2

While the order of the average canopy coverage trees changes, the same trees remain above the 80% mark. However, some trees did make an increase compared to 2022 such as the red oak and the pawpaw tree. Looking at the data between the two groups, I would believe that there is some connection between canopy coverage and type of tree.

Compared to 2022, 2024 had overall better canopy coverage when looking at areas with an increased presence of invasive species in figures 3 and 4. Another notable increase was that ailanthus went from having one of the lower coverages to having a significantly higher coverage. These values are unexpected and are noteworthy information to keep in mind when interpreting the data.

From looking at the graphs during our more advanced EDA, I believe that we can answer our two questions. I think that the average canopy coverage for a type of tree being somewhat consistent shows that the type of tree does play a role in average coverage. Looking at the presence of invasive species, the overall increase of average canopy coverage between groups and having ailanthus increase by a significant margin, makes answering the question of if invasive species type and average canopy coverage more difficult. I think that, considering the heat map correlations, there is not a significant impact on the type of invasive species and canopy coverage.

Machine Learning Insights

The two machine learning models we used for this project were the Random Forest Classifier and the Support Vector Machine model. The Random Forest Classifier achieved 92.72% accuracy on the 2022 dataset and 75.35% when tested on 2024 data, identifying Maple and Hackberry species as key positive contributors to coverage, while invasive species like Ligustrum had a negative impact. The Support Vector Machine (SVM) model, though slightly less accurate with 89.09% training accuracy and 71.83% testing accuracy, highlighted similar predictors but struggled with the dataset's complexity, resulting in comparatively lower performance.

The Random Forest Classifier (RFC) and Support Vector Machine (SVM) were both evaluated for their predictive performance on the 2024 dataset, showcasing distinct strengths. RFC achieved high accuracy on both the test and 2024 datasets with values of 92.72% and 75.35%, particularly after hyperparameter tuning. The model's ROC curve and AUC demonstrated the models accuracy and ways it tended to fail, while the confusion matrix highlighted models performance towards the negative outcome. the top 20 features had an almost even mixture of invasive species, canopy and understory columns.

SVM, on the other hand, performed better after tuning parameters but not as good as RFC with accuracy scores of 89.09% and 71.83%, with the RBF kernel emerging as optimal. While SVM lacked inherent feature importance metrics, Recursive Feature Elimination (RFE) was used to rank features. The ROC curve and AUC for SVM showed how the dataset was imbalanced and how well each version of the model performed. The confusion matrix provided a detailed breakdown of inaccurate predictions. Overall, both models exhibited strong performance, but ultimately had lower accuracy scores when compared to RFC, making it the more optimal choice.

Graphs and Visuals

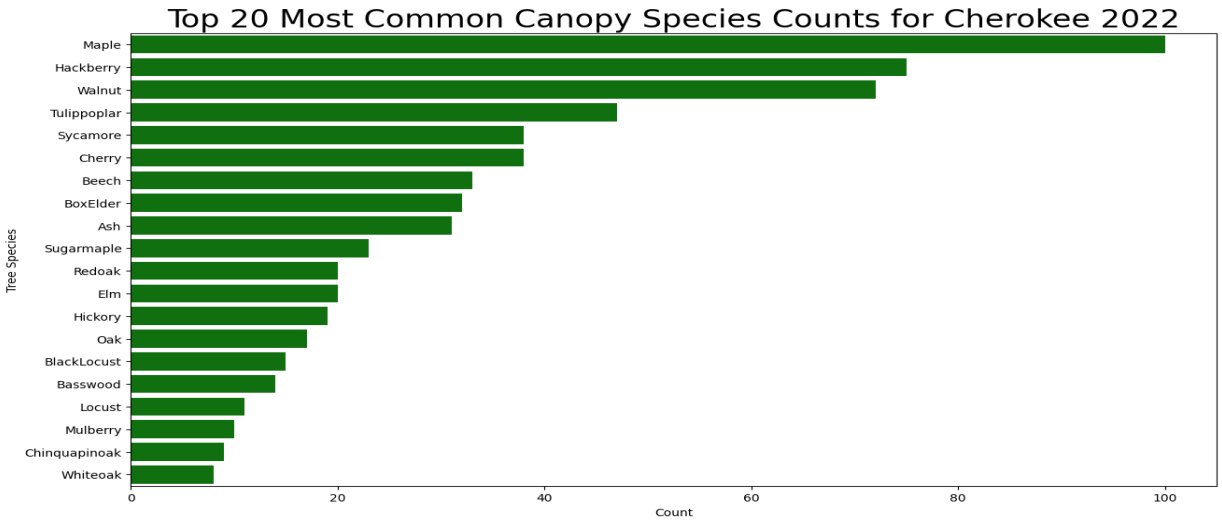

Figure 1.

In figure 1, Maple is the most common type of canopy species with them being one of the most common species at 100 out of 164 trimble stations. Hackberry and Walnut are the second and third for most common canopy species.

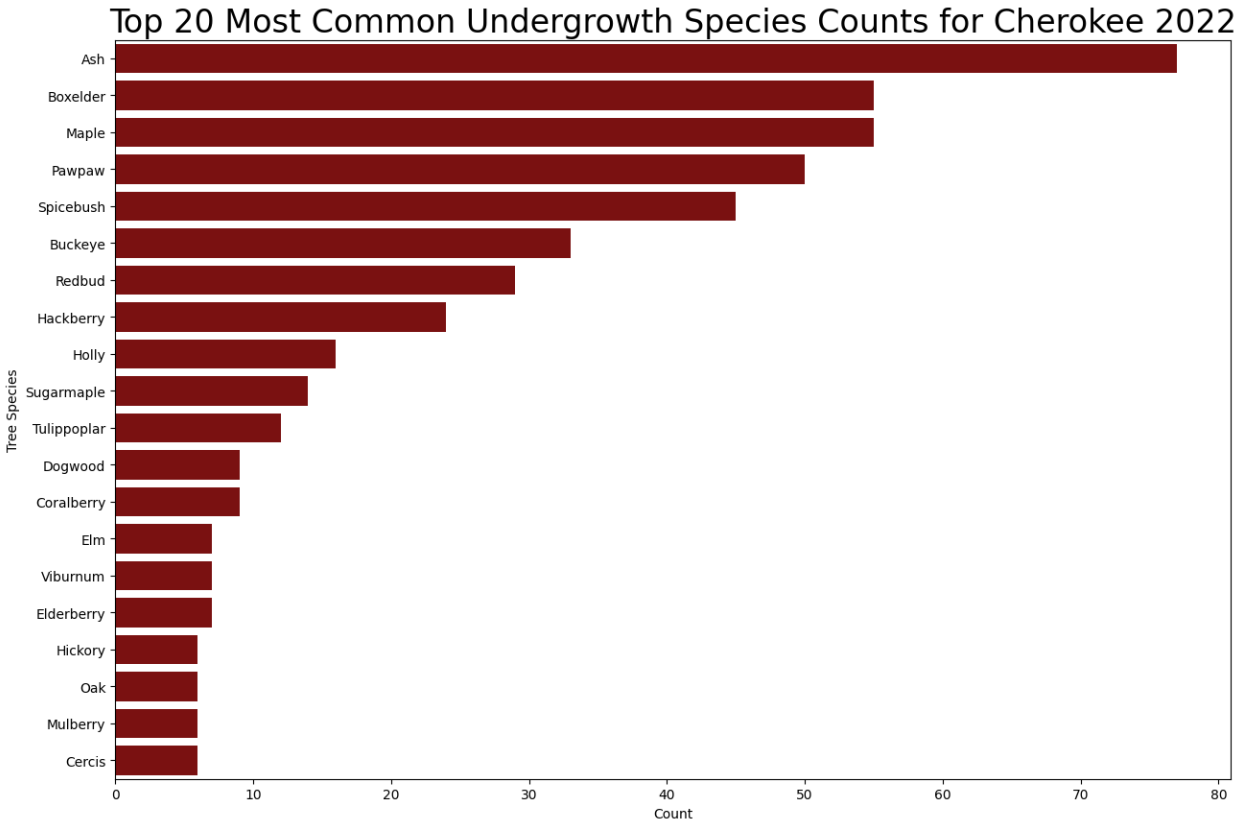

Figure 2.

In figure 2, Ash is the most common type of undergrowth species with them being one of the most common species at 75 out of 164 trimble stations. Boxelder and Maple are the second and third for most common canopy species.

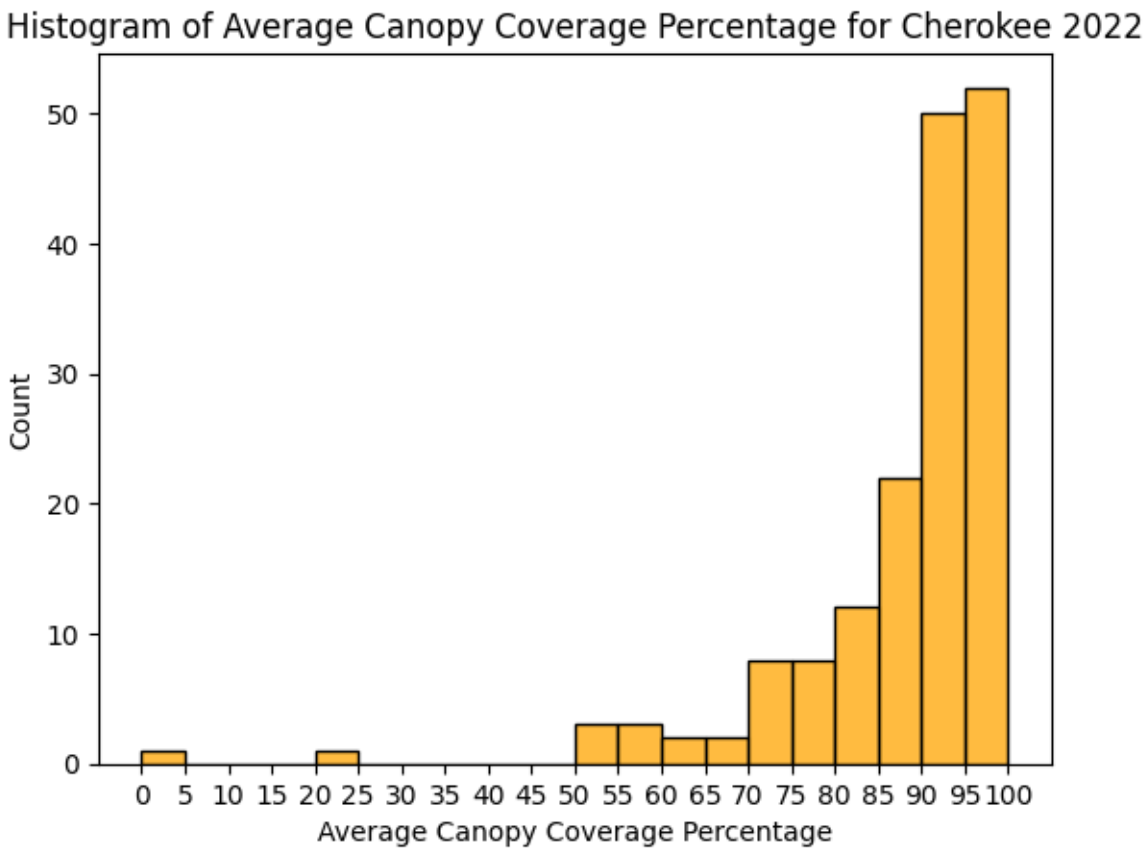

Figure 3.

In figure 3, most trimble stations have at least 85% average canopy coverage. The data set is left skewed.

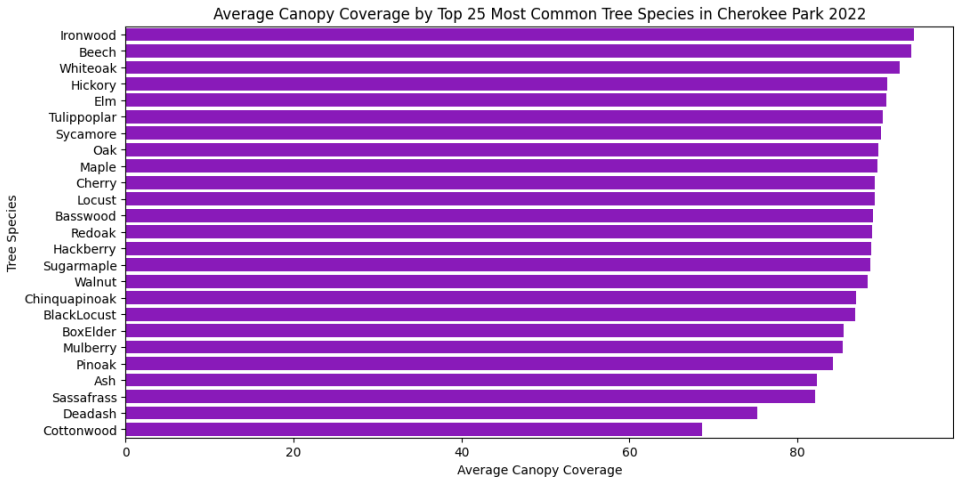

Figure 4.

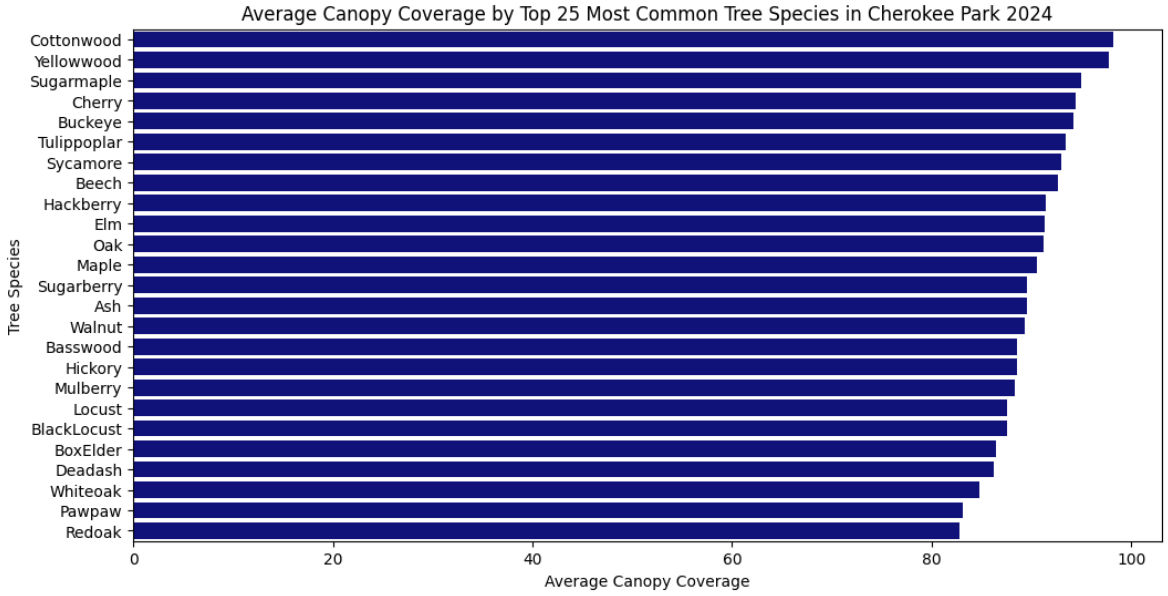

Figure 5.

Looking at Figures 4 and 5, while the order of the average canopy coverage trees changes, the same trees remain above the 80% mark. However, some trees did make an increase compared to 2022 such as the red oak and the pawpaw tree. Looking at the data between the two groups, I would believe that there is some connection between canopy coverage and type of tree

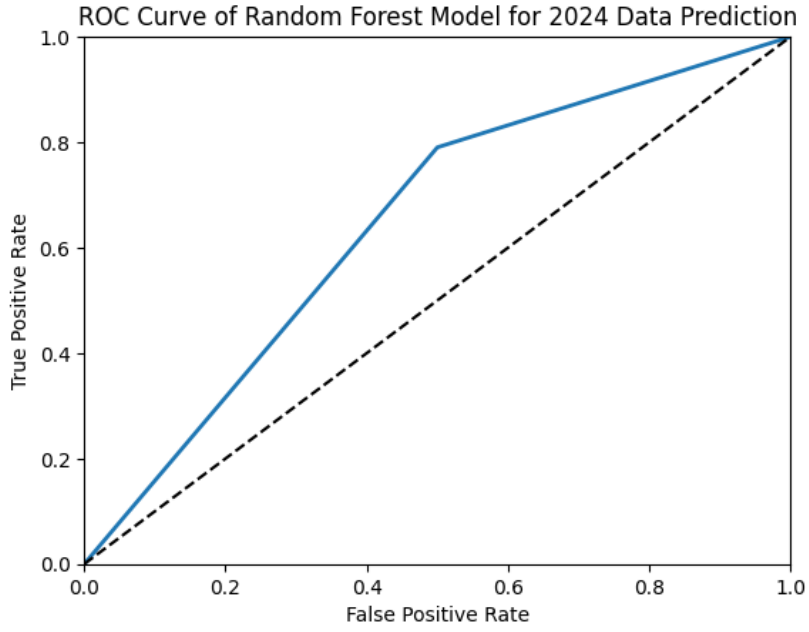

Figure 6.

In figure 6, it seems that the model had significantly more false negatives in the model causing the accuracy on the graph to decrease.

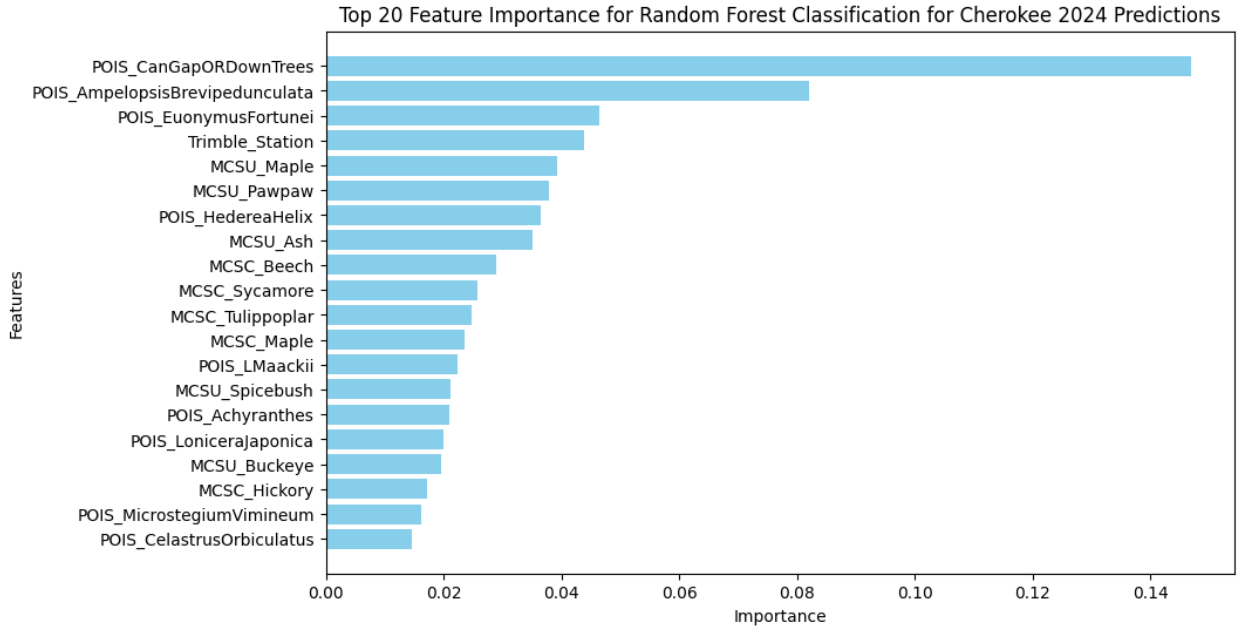

Figure 7.

In figure 7, there was a significant amount of understory columns in the important features which we did not expect. There seems to be a good mixture of canopy, invasive species and understory data points in the top 20 features. The 3 most important features were downed trees, Ampelopsis Brevipedunculata and Euonymus Fortunei

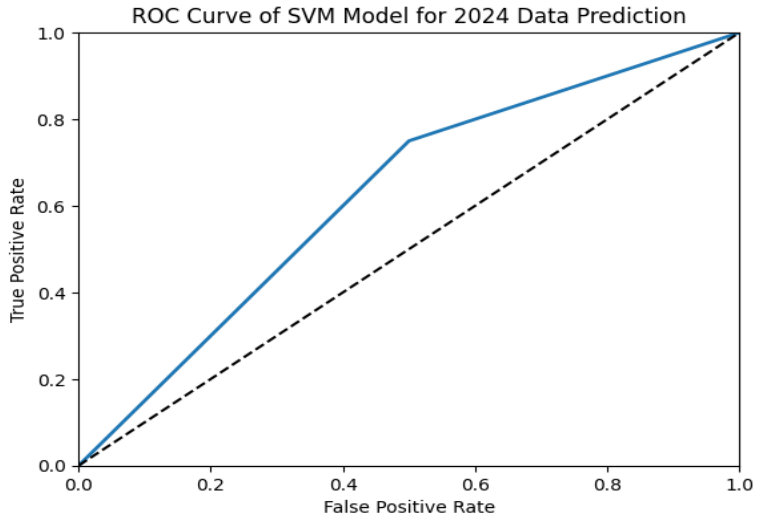

Figure 8.

In figure 8, this model took a larger hit on the prediction on the 2024 data set and went down nearly a whole 0.2. This graph shows how this model leaned into making more errors in its prediction.

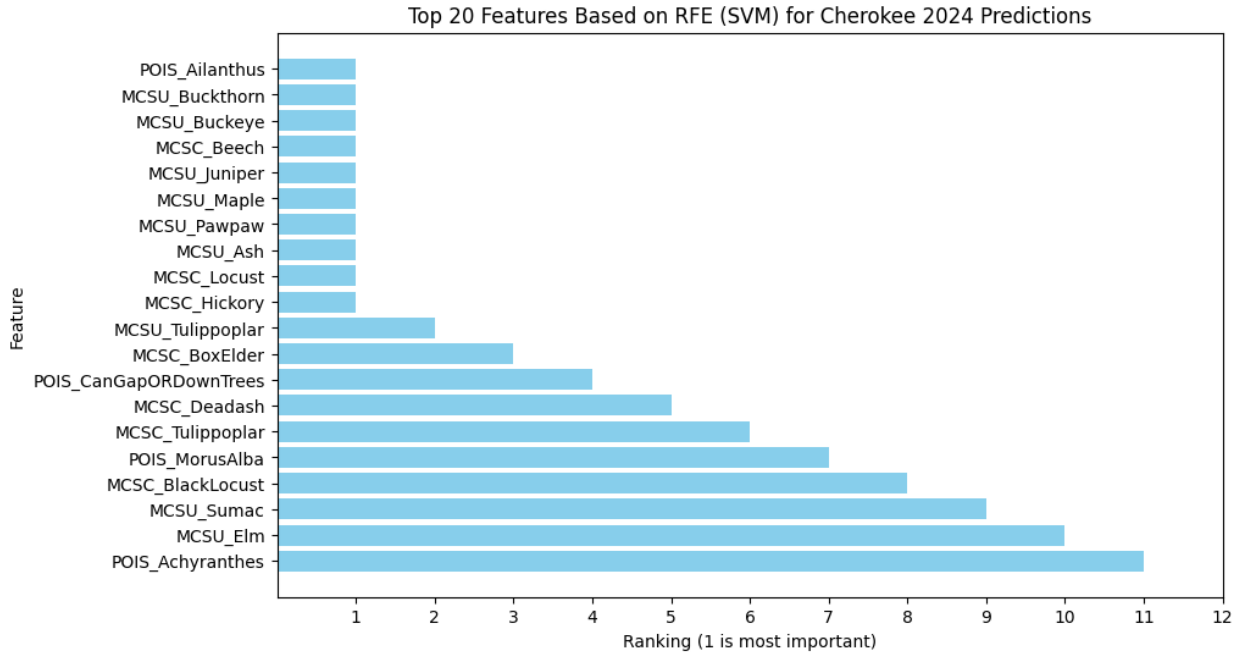

Figure 9.

In figure 9, six of the ten features ranked as one for most important have to do with the understory. This could suggest that the understory plays a bigger role in the canopy coverage of Cherokee park. There is more understory prediction for canopy coverage in this model than in the previous model.

Findings

The exploratory data analysis (EDA) revealed minimal differences between the 2022 and 2024 Cherokee Park datasets, except for stronger correlations between invasive species in 2024. When comparing machine learning models, the Random Forest model demonstrated superior performance with an accuracy of 0.92 on testing data and 0.75 for predictions on the 2024 dataset, while the SVM model achieved slightly lower scores of 0.89 and 0.71, respectively. Feature importance analysis showed that the Random Forest model provided a balanced emphasis on invasive species, canopy, and understory features, whereas the SVM model relied primarily on understory data. Additionally, comparisons of average canopy coverage for tree types between 2022 and 2024 suggested a potential correlation between the two years. Notably, areas with dense invasive species coverage exhibited increased canopy coverage, which appeared to diminish the influence of invasive species on canopy dynamics. Overall, the EDA and

machine learning models highlighted a stronger connection between tree canopy coverage and understory species, offering valuable insights into Cherokee Park's ecological patterns

Recommendations

We would recommend focusing more on what is happening with the undergrowth in the canopy regions. We believe that the lack of undergrowth would suggest that there would be less canopy space to provide protection. Potentially providing ways to improve undergrowth competition for native plants could have a significant impact in the future. Prioritizing treating Ampelopsis Brevipedunculata and Euonymus Fortunei would also seem to have a significant impact on the tree canopy coverage seen in the park.

Conclusion & Reflection

Looking back on our project, we realize that paying more attention to the understory—the plants and smaller trees growing under the canopy—could have improved our ability to make accurate predictions for 2024. While our analysis worked well with the 2022 data, it didn't perform as strongly when we applied it to 2024. One reason for this could be that our method focused too much on certain patterns in the 2022 data, making it less effective when looking at new information. Choosing a different approach, such as a method inspired by how the human brain processes information (called a neural network), might have worked better.

We also faced challenges in preparing the data for analysis. Ensuring the numbers were accurate and that the data was set up correctly was more complicated than we expected. At first, we considered a method that predicts specific values (like percentages), but based on feedback, we switched to a simpler method that groups areas into categories, like "high coverage" and "low coverage." We were also advised to make our explanations and notes easier to understand, so we worked hard to make the project clear and accessible, even for people who don't specialize in this kind of work.

In the end, we learned that projects like this require a lot of planning and organization. Keeping notes and labels clear and writing things in an easy-to-follow way are essential to avoid confusion and frustration. This experience taught us how important it is to stay organized and focused when working on a big project.

Group Contributions

Kadin took the lead on several critical tasks, including writing code for exploratory data analysis (EDA) and machine learning algorithms, producing visuals for analysis, drafting conclusions and interpretations, completing the preprocessing of the 2024 dataset, and contributing to the report. Jonathan supported coding for the random forest model, created the

PowerPoint presentation, worked on its design and interpretations, contributed brainstorming ideas, and wrote the introductory informational statements. Chris assisted with coding for the SVM model, contributed to the PowerPoint, participated in brainstorming sessions, helped prepare the 2024 dataset, and provided support for the report.