**"Predictive Modelling of Tree Cover Loss Using random forest "**

Suman Thakur (70022200566) & Rishi Joshi (70022200343)

Narsee Monjee Institute of Management Studies(STME), Indore (M.P.)

**Abstract:** This report examines the creation and assessment of a machine learning model designed to predict tree cover loss, which is a significant concern linked to deforestation, biodiversity decline, and climate change. A Random Forest Regressor was selected for its effectiveness in managing non-linear relationships, utilizing geographical and environmental data. Following exploratory data analysis (EDA) and feature selection, the model was trained on 80% of the dataset and evaluated on the remaining 20%.

To assess the model's performance, Mean Squared Error (MSE) and R-squared (R²) metrics were employed, yielding an impressive R² value of 98.08%, indicating strong predictive capability. Geospatial visualizations, including heat maps and choropleth maps, were developed to illustrate the spatial distribution of predicted tree cover loss. The model's high accuracy positions it as a valuable resource for identifying vulnerable areas, aiding policymakers and conservationists in formulating effective forest conservation strategies. Future enhancements may include cross-validation and the incorporation of additional features to further improve accuracy.

1. **Introduction**

Wildfires present serious challenges for the environment and human communities alike. They lead to a decline in biodiversity, destruction of habitats, soil erosion, and pollution of air and water, while also contributing to climate change through the release of large amounts of carbon dioxide. In addition to environmental harm, wildfires cause significant economic damage by destroying homes, infrastructure, and agricultural land, and they impact industries such as tourism. The smoke and ash produced by these fires can lead to serious health issues, including respiratory and cardiovascular problems, while also displacing communities and necessitating mass evacuations.

To mitigate these effects, effective forest fire management is essential. Prevention methods such as removing dry vegetation, establishing firebreaks, and conducting controlled burns are vital for reducing the fuel that can feed fires. Quick suppression efforts, along with modern firefighting technologies like drones and satellite monitoring, are crucial for containing fires swiftly. A significant aspect of fire management involves the use of predictive technologies. By anticipating conditions that are conducive to wildfires, authorities can organize timely evacuations, allocate firefighting resources more effectively, and safeguard at-risk areas.

The ability to predict wildfires brings many advantages. It allows for quicker and safer evacuations, reduces economic losses by limiting property damage, and enhances the protection of vital ecosystems. Overall, integrating predictive tools into forest fire management can greatly lessen the impact of wildfires on both people and the environment.

1. **Literature Review**

Forest fires represent a significant environmental challenge that has drawn considerable attention due to their destructive impacts on ecosystems, biodiversity, human health, and the economy. A variety of studies have explored the causes of wildfires, which can be divided into natural and human-induced factors. Natural causes encompass lightning strikes, elevated temperatures, and extended periods of drought, all of which create ideal conditions for wildfires. On the other hand, human activities like deforestation, agricultural burning, land development, and careless actions (such as improper cigarette disposal or leaving campfires unattended) have intensified the frequency and severity of wildfires around the world. These escalating threats highlight the necessity for predictive tools to effectively prevent and manage forest fires.

An increasing number of studies have utilized machine learning (ML) models to forecast and analyze forest fires. Previous research in this area has successfully implemented various ML techniques, including Random Forest, Decision Trees, and Support Vector Machines, to predict the occurrence and spread of wildfires based on environmental variables like temperature, humidity, wind patterns, and vegetation dryness. For instance, a study published in \*\*MDPI\*\* used satellite data and machine learning algorithms to enhance wildfire prediction across different ecological zones, achieving a high accuracy rate in assessing fire risk (R² value of 93%) across diverse landscapes. Likewise, research from \*\*WRI\*\* utilized higher-resolution datasets to evaluate tree cover loss and wildfire risk in tropical areas, offering valuable insights for conservation and forest management initiatives. These studies have shown that ML models can deliver reliable predictions, with accuracies typically ranging from 85% to 95%, depending on the complexity of the data and the models employed.

Tables from these studies, such as those found in the MDPI paper, frequently include detailed performance metrics for the models, including confusion matrices and feature importance.

1. **Research Aim**

The Aim of this research is to create a machine learning model that can effectively predict tree cover loss using various environmental, geographical, and historical factors. By utilizing predictive analytics, the study aims to uncover patterns and causes of deforestation and forest degradation, offering valuable insights for policymakers, conservationists, and environmental planners to make informed decisions that help reduce the effects of tree cover loss and encourage sustainable forest management.

1. **Methodology**

This work involves the analysis of a cleaned forest fire dataset, which includes key features like alert\_count\_sum, areaha, and umd\_tree\_cover\_loss\_from\_fires\_ha. These attributes help in assessing the magnitude of forest cover loss caused by deforestation and wildfires. The dataset provides valuable environmental data for predicting the extent of tree cover loss in hectares (ha), facilitating better planning and management of forest conservation efforts.

**1. Understanding the Problem**

Objective: The goal was to predict tree cover loss using machine learning techniques. Tree cover loss is a critical environmental issue, and predicting its future occurrence can help mitigate its impact through proactive actions.

**2. Data Exploration (EDA - Exploratory Data Analysis)**

Data Loading: We started by loading the dataset, which contained relevant information such as environmental variables, geographical data, and historical tree cover loss records.

Data Cleaning: We ensured that the data was free from missing values or inconsistencies, and the dataset was structured in a way suitable for analysis.

**Exploratory Data Analysis (EDA):**

Descriptive Statistics: We examined summary statistics (mean, median, standard deviation) for each feature.

Correlation Analysis: Identified relationships between features and the target variable (tree cover loss).

Histograms: Showed the distribution of key variables.

Scatter plots: Visualized the relationship between different features and tree cover loss.

Heatmap: Visualized feature correlations to spot patterns.

Geographical Map Visualizations: Created choropleth and heatmaps to analyze the spatial distribution of tree cover loss.

**3. Data Preparation**

Feature Selection: Selected relevant features that could help in predicting tree cover loss. These could include environmental data, geographical information, or historical loss records.

Feature Scaling: If needed, scaled numerical data to ensure all features had comparable ranges.

Splitting the Dataset: Divided the data into training and testing sets:

Training Set: 80% of the data used to train the model.

Test Set: 20% of the data used to evaluate the model’s performance on unseen data.

**4. Model Selection and Training**

Model Selection: We chose a Random Forest Regressor for its strong performance on non-linear data and its ability to handle complex interactions between variables.

Model Training: Trained the model using the training dataset to learn patterns between the features and tree cover loss.

**5. Model Evaluation**

Performance Metrics: After training, we evaluated the model using the test set:

Mean Squared Error (MSE): Quantified the average squared difference between actual and predicted tree cover loss.

R-squared (R²): Indicated how well the model explained the variance in the data, with a result of approximately 98.08%, showing high predictive power.

**6. Prediction and Visualization**

Prediction: We used the trained model to make predictions on the test set and compared these predictions with the actual tree cover loss.

Scatter Plot: Showed the relationship between actual and predicted values.

Residual Plot: Helped to identify any bias in the model by showing the differences between predicted and actual values.

Geospatial Maps: Visualized the geographical distribution of predicted tree cover loss using:

Choropleth Maps: Showed predicted loss by region.

Heatmaps: Highlighted areas with dense tree cover loss predictions.

Bubble Maps: Indicated the magnitude of loss with variable-sized circles.

**7. Insights and Interpretation**

High Predictive Accuracy: The model’s high R-squared value indicated it could effectively predict tree cover loss, with 98% of the variation explained by the features used in the model.

Geographical Patterns: The map visualizations provided useful insights into regions where tree cover loss is predicted to be high, allowing for more targeted conservation efforts.

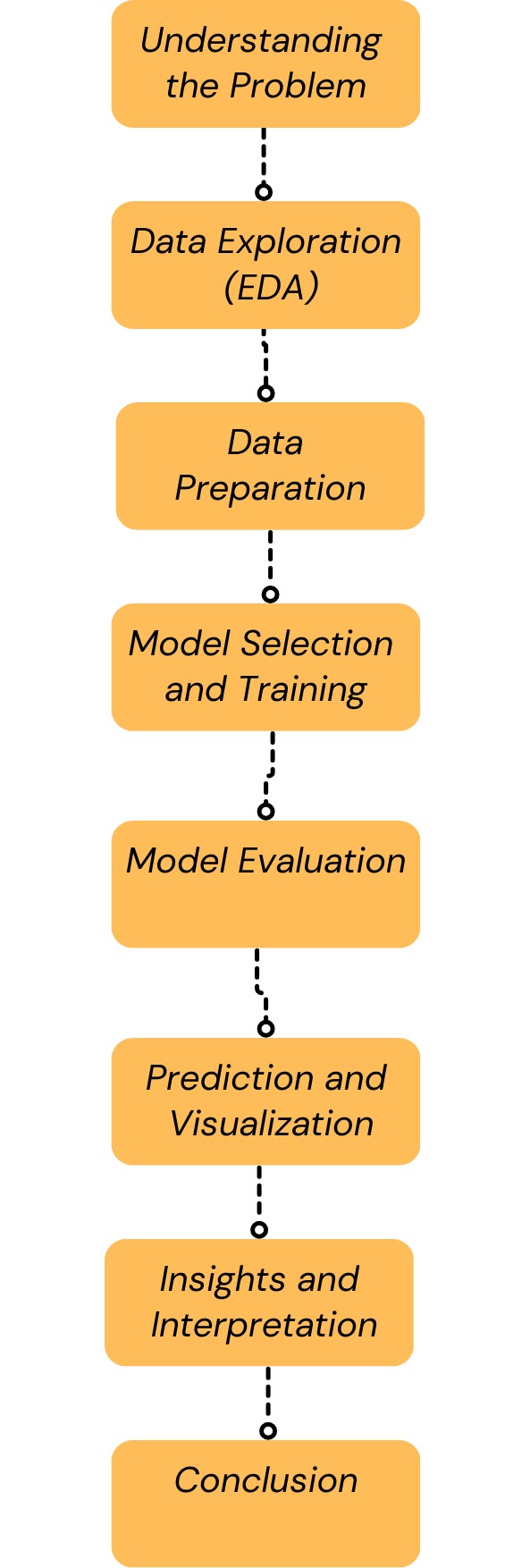
**8. Conclusion**

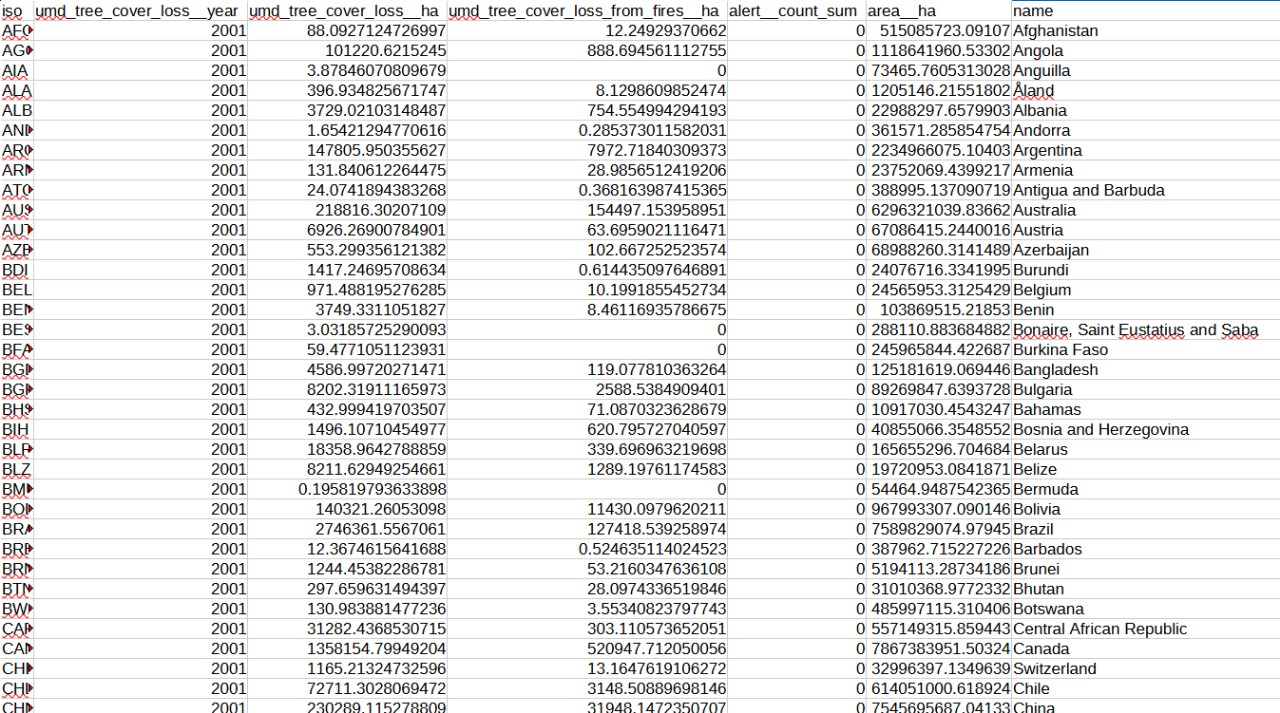
Model Effectiveness: With an R-squared value of 98.08%, the Random Forest model demonstrated strong performance in predicting tree cover loss.

Recommendations:

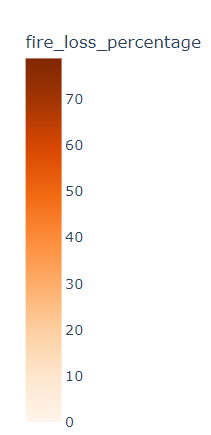
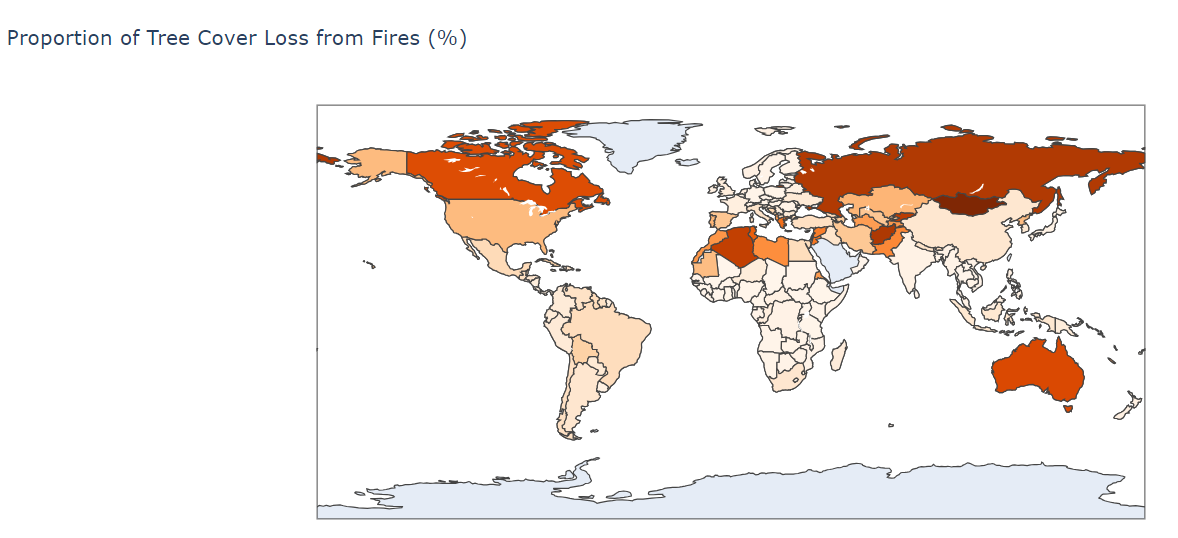
Further Model Validation: Cross-validation and further testing with new data can enhance confidence in the model's ability to generalize.

Policy and Conservation Application: The model can be used by policymakers and conservationists to identify high-risk areas and prioritize interventions.

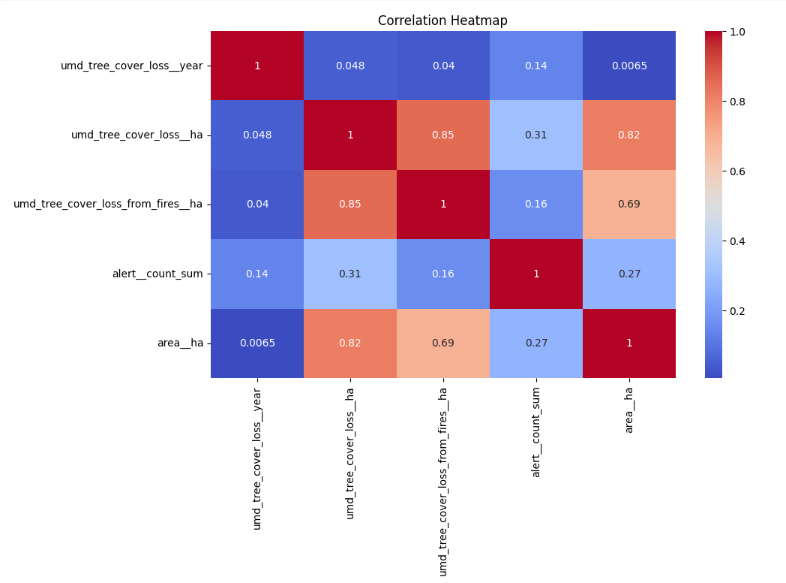




**Data Analysis**



The map above shows the percentage of tree cover lost to fires country by country. Regions that have lost a greater amount are those having darker shades, for example, Russia, Canada, and Australia, while lighter shades show less proportion in places such as Europe, Africa, and South America.

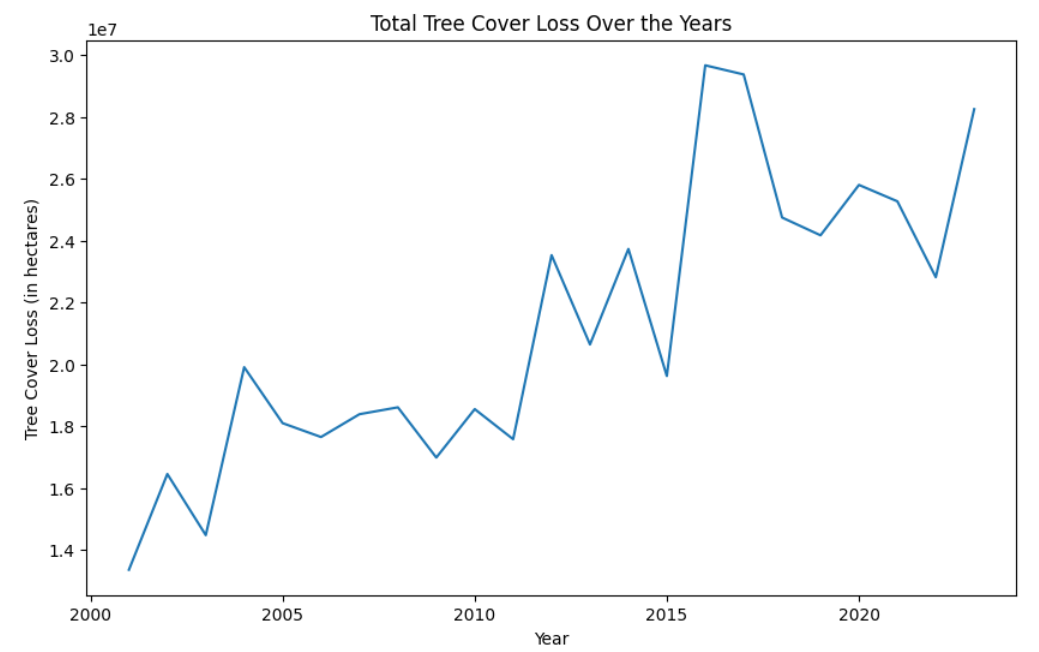


This heatmap of correlations suggests the relationship of variables with tree cover loss and fires. Some major findings:

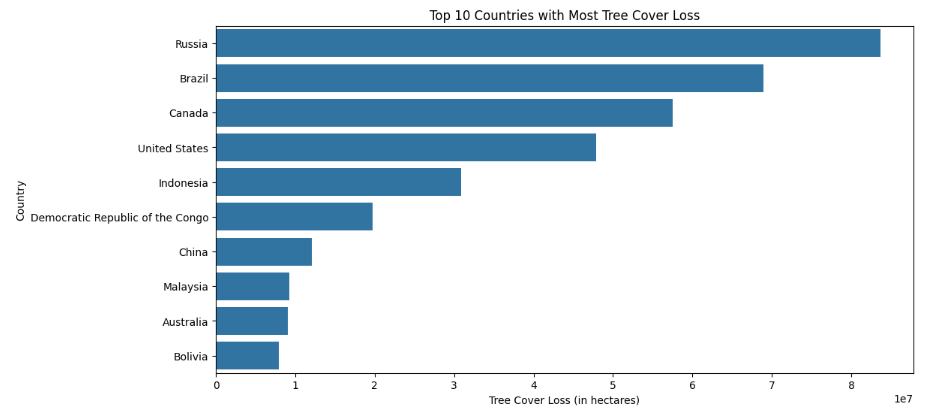
- Strong correlations : The tree cover loss correlates strongly with tree cover loss from fires as well as total area.

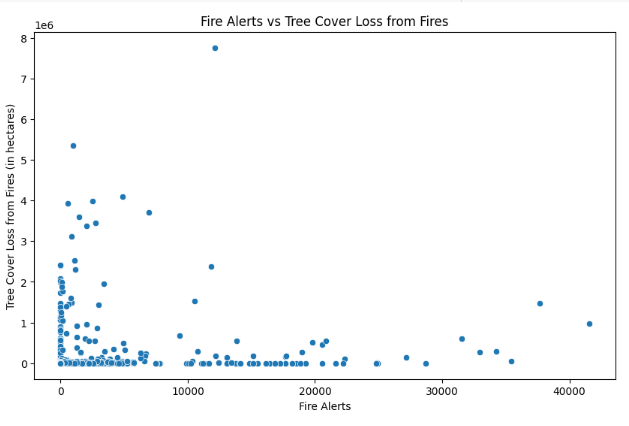
- Moderate association : Fire warnings have moderate association with the loss of tree cover.

- Weak correlations : Tree cover loss by year was one of those variables whose relationships to some others were rather weak.The colors range from red (strong correlation) to blue (weak correlation).



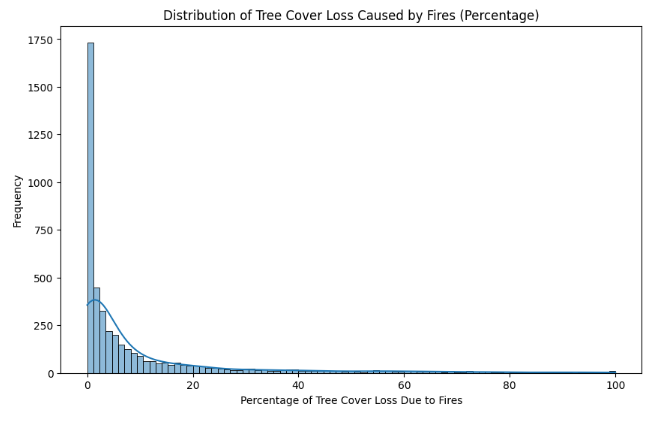
This graph displays the **Total Tree Cover Loss Over the Years** from 2000 to 2022. The y-axis represents the tree cover loss in hectares (in millions), and the x-axis represents the years.  
Tree cover loss has generally increased over time, starting from around 14 million hectares in 2000 to almost 30 million hectares in recent years.  
Significant peaks are visible around 2015 and 2022,there are dips after the peaks, such as between 2016 and 2018, but the general trend remains upward.



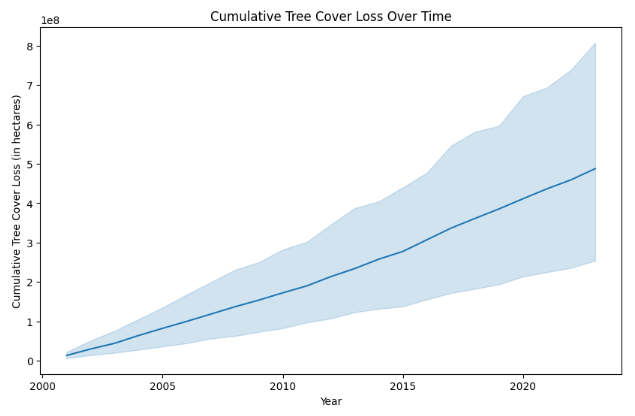
This graph highlights the countries that are most impacted by deforestation or tree cover loss, with Russia and Brazil leading the list.Russia has the highest tree cover loss, with over 8 million hectares. Brazil follows closely, with around 5.5 million hectares lost. 

* There seems to be a general trend where high tree cover loss can occur with fewer fire alerts, suggesting that the intensity or duration of fires might be more important than the sheer number of fire events.
* In most cases, regions with a large number of fire alerts tend to have lower tree cover loss, but the relationship is not perfectly linear.

the scatter plot suggests that tree cover loss from fires is not always directly proportional to the number of fire alerts. A few intense fires might cause significant tree loss, even when the number of alerts is low.

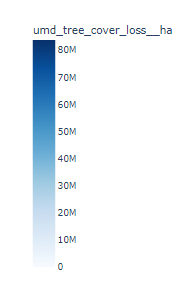
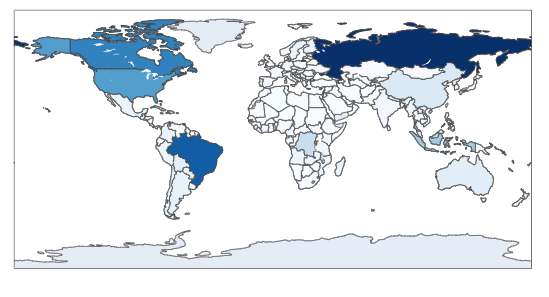


This graph illustrates the percentage distribution of tree cover loss due to fires. The x-axis indicates the percentage of tree cover loss attributed to fires, ranging from 0% to 100%, while the y-axis shows how often these occurrences happen. The graph is significantly skewed to the left, with the highest frequency found in the 0% to 5% range of fire-related tree cover loss. This suggests that most regions experience minimal tree cover loss from fires. As the percentage of loss increases, the frequency drops sharply, indicating that significant tree cover loss due to fires is uncommon.

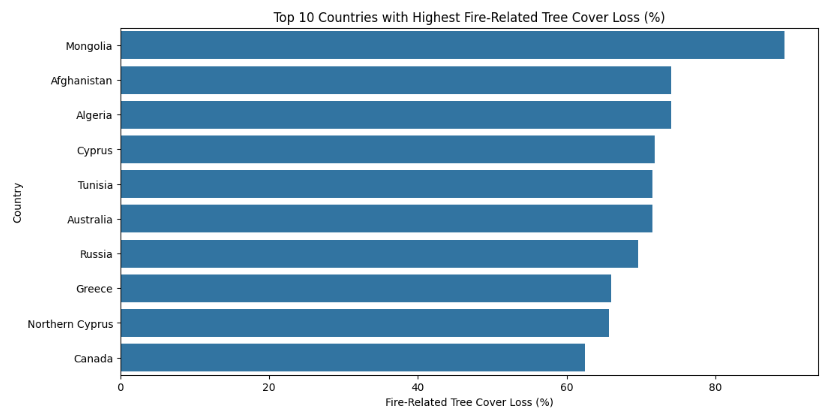


This graph illustrates the cumulative loss of tree cover from 2000 to 2023, measured in hectares. The y-axis displays the total loss in hectares, which reaches approximately 800 million hectares, while the x-axis indicates the years. The solid line depicts the trend of tree cover loss over time, and the shaded area likely represents the uncertainty or variation surrounding this trend. Overall, the graph reveals a consistent rise in tree cover loss, with an acceleration observed in recent years.

**Global Tree Cover Loss Intensity (Hectares)**



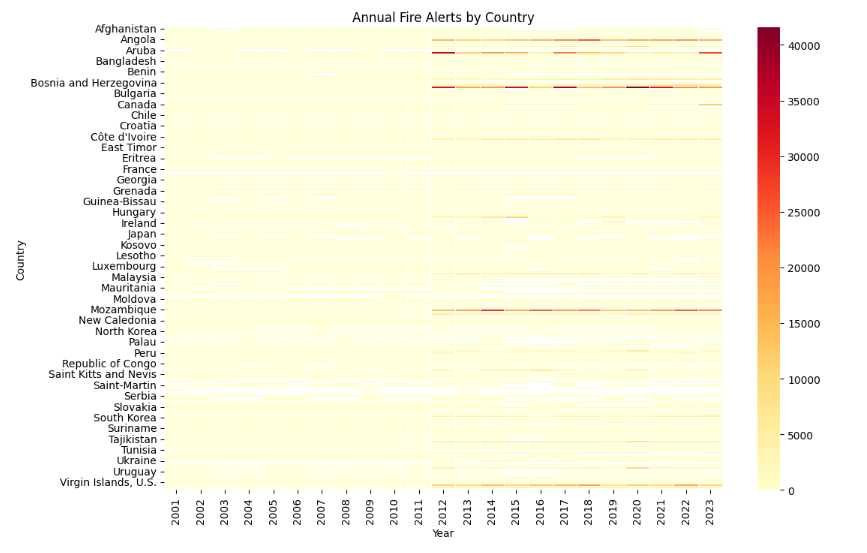
This map shows global tree cover loss, with darker blue areas like Russia and Brazil experiencing the highest loss, and lighter blue areas indicating lower levels of loss across regions like Europe, Africa, and Australia.



This bar chart shows the **Top 10 countries with the highest fire-related tree cover loss (%).** The x-axis represents the percentage of tree cover loss caused by fire, while the y-axis lists the countries.

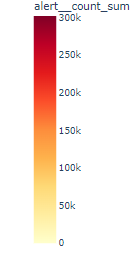
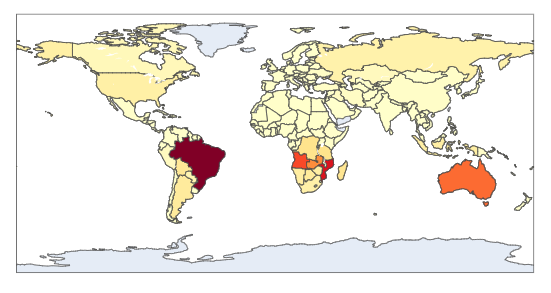
* **Mongolia** has the highest percentage of fire-related tree cover loss, exceeding 80%.
* **Afghanistan**, **Algeria**, **Cyprus**, and **Tunisia** follow, each having between 60% and 70% fire-related tree cover loss.
* **Australia**, **Russia**, **Greece**, **Northern Cyprus**, and **Canada** also make the list, with fire-related tree cover loss percentages ranging from around 40% to 60%.

This graph highlights the significant impact of wildfires on tree cover in these regions, with Mongolia being the most affected.

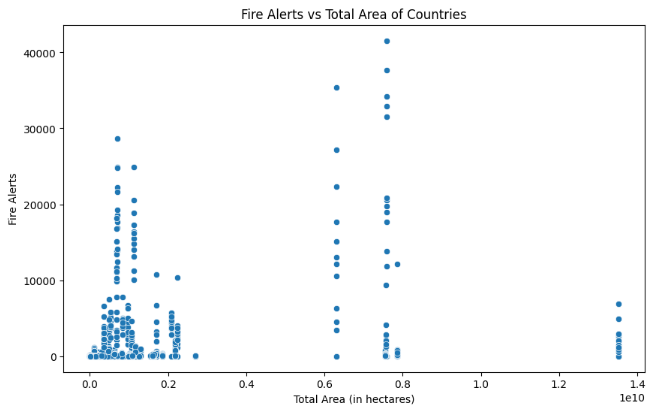


The graph shows a heatmap of **annual fire alerts** by country from **2001 to 2023**. Each row represents a country, and the intensity of the color represents the number of fire alerts, with darker colors indicating more fire alerts. The color bar on the right shows that the fire alerts range from 0 (light yellow) to over 40,000 (dark red). Some countries, like **Angola** and **Mozambique**, have more fire alerts over the years, while others have very few or none.

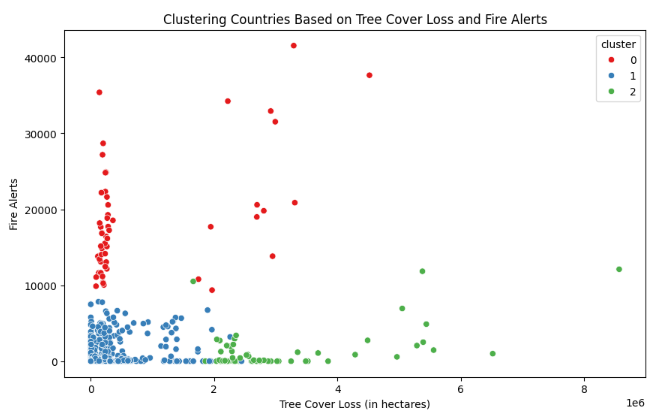
**Global Fire Alert Intensity**



The second image is a world map showing the **global fire alert intensity**. Different regions are colored to indicate varying levels of fire alerts. **Brazil** is shown in dark red, indicating the highest intensity of fire alerts, followed by **southern Africa** and **Australia**, which are in shades of orange. These colors suggest that these regions experience a significant number of fire alerts compared to other parts of the world, which are shown in pale yellow, indicating lower fire alert intensities.



The scatter plot shows the relationship between **fire alerts** and **country size** (in hectares). It highlights that smaller countries tend to have fewer fire alerts, while larger countries show more variation, with some having high fire alerts and others relatively low. Size alone doesn't determine the number of fire alerts.



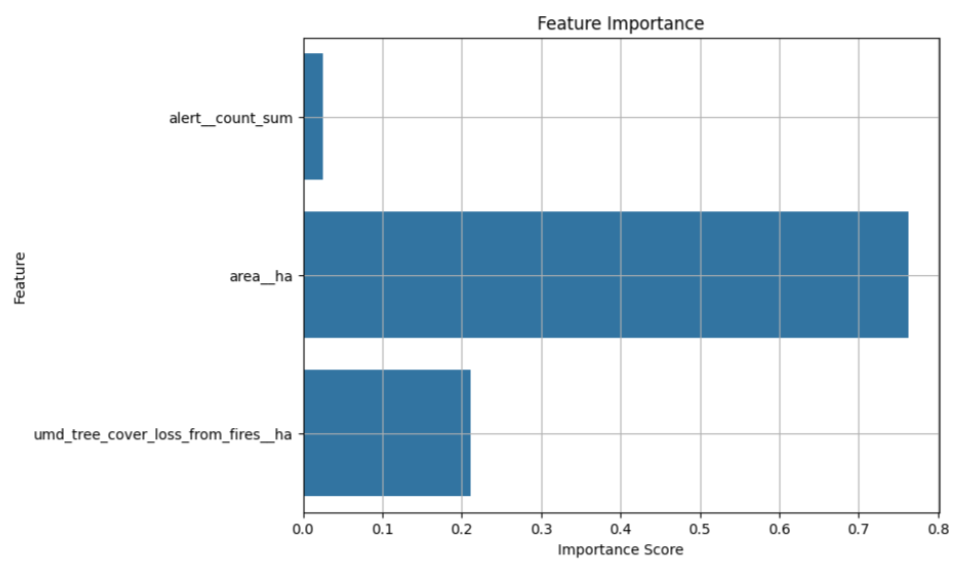
**8.**  **Model Development for predicting Forest Fire**

**Random Forest Regressor**

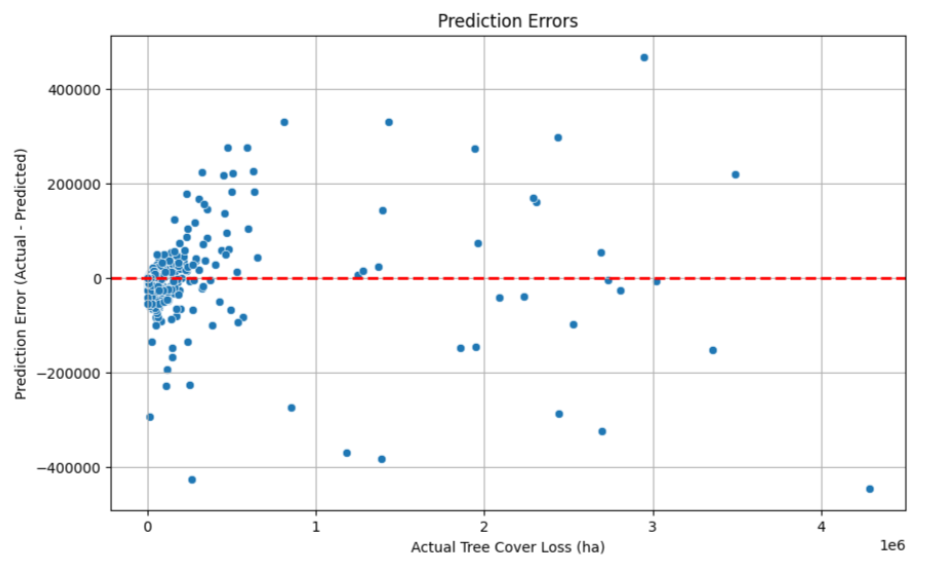
The Random Forest algorithm was chosen because of its

* Ability to handle large datasets with higher dimensionality.
* Capability to capture complex, non-linear relationships between variables.
* Low risk of overfitting due to its ensemble nature (aggregating multiple decision trees).
* It provides feature importance, helping to understand which variables contribute most to the model's prediction.

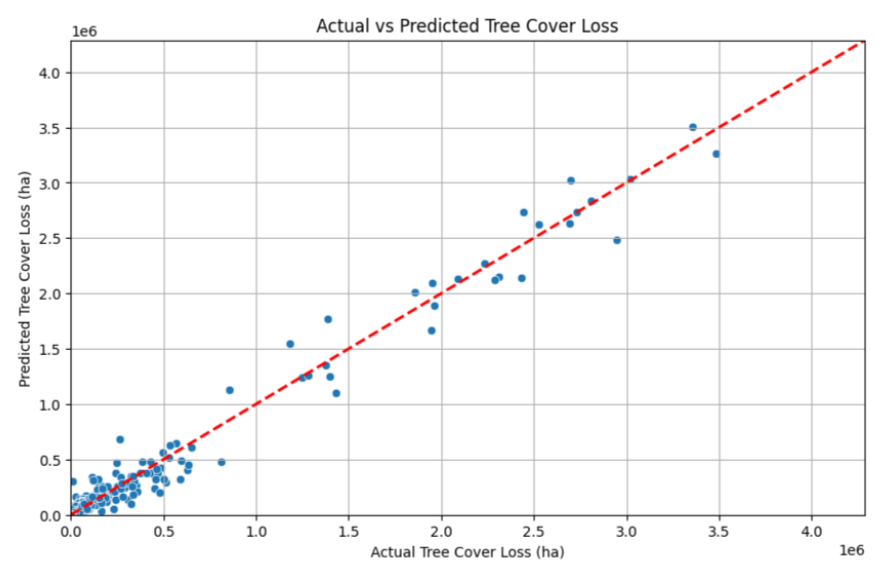
Random Forest is a method for classification and regression that works by creating numerous decision trees in the training process. Every tree is taught using a random portion of data, and the end result is achieved by averaging the forecasts of each tree (for regression) or by majority vote (for classification). This lessens variation, enhances precision, and reduces the risk of overfitting.



The graph reveals that area (ha) is the most important feature for the model, with an importance score close to 0.75, followed by tree cover loss from fires, which has moderate significance. Alert count, however, has minimal impact on the model. The model itself demonstrates high accuracy, with an R-squared value of 0.98. In the updated version of the graph, the focus is solely on feature importance, with no performance metrics displayed.



The graph shows prediction errors for tree cover loss. The x-axis is actual loss, and the y-axis is the error (actual minus predicted). Points near the red line mean small errors; points far from it show large errors. The model performs better with smaller losses but is less accurate with larger ones.



This Graph shows the comparison between the observed tree cover loss (horizontal axis) and the estimated tree cover loss (vertical axis). The perfect predictions are shown by the red dashed line when actual values are equal to predicted values. Points that are near the line represent precise predictions, whereas points that are farther away indicate larger prediction discrepancies. In general, the model appears to be doing well, with the majority of points clustering close to the red line, particularly for lower levels of tree cover loss.

**9. Results and Discussion**

The Random Forest Regressor achieved an R² of 0.9808 and a Mean Squared Error (MSE) of 3,289,822,124.81. These metrics indicate a strong fit, with the model explaining 98.08% of the variance in tree cover loss. In comparison:

Gradient Boosting Regressor showed an R² of 0.9719, making it the second-best model.

Linear Regression achieved a significantly lower R² of 0.7722, indicating a weaker fit.

Support Vector Regressor (SVR) performed poorly with R² = -0.0662, showing it is unsuitable for this dataset.

**9.1Analysis**

From the analysis, we observe that the Random Forest Regressor effectively captures the complex relationships between the input features and tree cover loss. Its ensemble method of averaging multiple decision trees reduces overfitting and provides robust predictions. The Gradient Boosting Regressor also performs well, though slightly worse than Random Forest, while Linear Regression struggles to model the non-linear relationships in the data. Support Vector Regressor fails to capture meaningful patterns due to the high-dimensional nature of the problem.

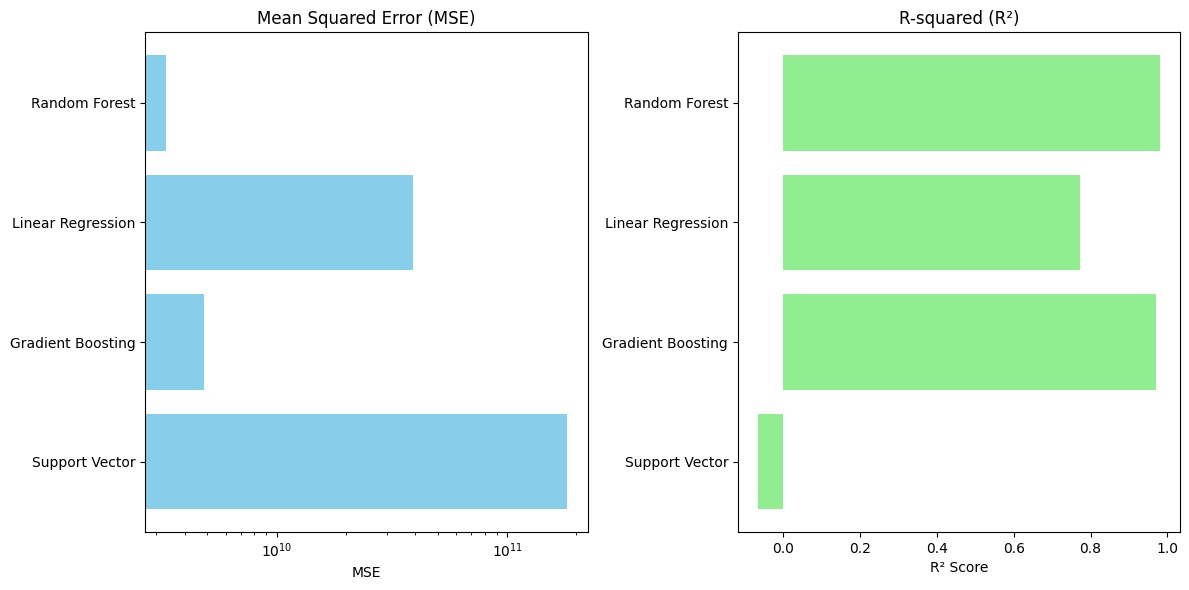
**9.2 Model Comparisons**

In our study, we employed several machine learning models to predict earthquake magnitude based on various features such as source depth, source distance, receiver latitude, and other seismic attributes. We tested three advanced Gradient Boosting techniques (XGBoost, LightGBM, and CatBoost), along with RandomForestRegressor, to evaluate which algorithm provided the best performance for this task.

**9.3 Performance Comparison:**

A comparison of model performance is shown below

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error (MSE)** | **R-squared (R²)** |
| Random Forest Regressor | 3,289,822,124.81 | 0.9808 |
| Gradient Boosting Regressor | 4,821,943,734.04 | 0.9719 |
| Linear Regression | 39,017,588,373.36 | 0.7722 |
| Support Vector Regressor | 182,650,537,815.25 | -0.0662 |



MSE bar chart: The MSE values ​​are displayed using a horizontal bar chart. Because the MSE values ​​vary across orders of magnitude, a logarithmic scale is used on the x-axis for better visualization. R-SQUARED BAR plot: The R² value is drawn on another bar graph, the axis lists the model name, and the X axis represents the R² score.

**9.4 Benefits of Early Identification**

Detecting tree loss early helps implement forest management strategies promptly, assisting in  
  
Efforts in conservation: Forest officials can address initial signs of deforestation, halting extensive harm.  
  
Fire management: Predictive models can help anticipate the spread of wildfires in fire-prone regions, aiding in resource preparation for firefighting efforts.  
  
Policy development: Government agencies can utilize these forecasts to establish more robust protective measures for at-risk forests, aiding in the reduction of biodiversity decline.  
  
Accurate predictions in environmental planning can enhance land-use planning and sustainable forest management, thereby boosting efforts to sequester carbon and alleviate climate change.

**10. Conclusion**

The random forest regression model was found to be the most effective in predicting forest cover loss, achieving a high level of accuracy (R² = 98.08%). Its ability to handle complex nonlinear relationships and comprehensive approach to mitigate overfitting make it a powerful tool in environmental data analysis.

**10.1 Future Scope**

Integrating more features: Integrating satellite data or climate-related variables will improve forecast accuracy.

Geospatial analysis: Improving models by adding spatial and temporal analysis will allow more localized predictions and therefore even more precise interventions. Application to other regions: Extending the model to other ecosystems and regions could help develop a global tool for predicting forest cover loss.

Real-time forecasting systems: Developing real-time forecasting systems could help forest management teams respond more effectively to emerging threats. This study demonstrates the value of forecasting models in the preservation of the environment and the importance of the decision -making decision in a timely manner in the fight against the abolition and protection of natural resources.

**Reference**

Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5–32. doi:10.1023/A:1010933404324.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825–2830.

McKinney, W. (2010). "Data Structures for Statistical Computing in Python." Proceedings of the 9th Python in Science Conference, 51-56. doi:10.25080/Majora-92bf1922-00a.

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, 9(3), 90–95. doi:10.1109/MCSE.2007.55.

Folium Documentation. (n.d.). "Folium: Python Data, Leaflet.js Maps." Retrieved from https://python-visualization.github.io/folium/

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... & Townshend, J. R. G. (2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science, 342(6160), 850–853. doi:10.1126/science.1244693.

Flask Documentation. (n.d.). "Flask: Web Development, One Drop at a Time." Retrieved from https://flask.palletsprojects.com/

MongoDB Documentation. (n.d.). "MongoDB: The Developer Data Platform." Retrieved from https://www.mongodb.com/docs/  
<https://www.globalforestwatch.org/map/?mainMap=eyJzaG93QW5hbHlzaXMiOnRydWV9&map=&mapMenu=eyJtZW51U2VjdGlvbiI6ImRhdGFzZXRzIiwiZGF0YXNldENhdGVnb3J5IjoibGFuZFVzZSJ9&menu=eyJkYXRhc2V0Q2F0ZWdvcnkiOiJmb3Jlc3RDaGFuZ2UiLCJtZW51U2VjdGlvbiI6ImRhdGFzZXRzIn0%3D>