Species Analysis

S2689073 **S2635437** **S1922129** **S2748996**

Current Progress

- Collected and preprocessed data from iNaturalist and WorldClim.
- Visualize species distribution and identify main environmental features.
- Trained initial models using machine learning techniques (RF, SVM, GB, SVC, MLP, KNN), with preliminary results indicating that RF, SVM, and GB perform the best.

Plans for Completion

- Conduct further data analysis through dimensionality reduction techniques.
- Apply Principal Component Analysis (PCA) for feature analysis and selection.
- In EDA part, maybe do some analyse about data that including extra features.

Abstract

Biodiversity is essential for ecological balance, and understanding species distributions is crucial for conservation efforts. In this study, machine learning techniques were used to predict species distributions in different geographical regions using large-scale datasets from iNaturalist and WorldClim. Machine learning models were introduced in the Species Distribution Model (SDM) and Random Forest (RF), Support Vector Machine (SVM) and Gradient Boosting (GB) algorithms were used to construct a classification model that predicts the presence of species based on environmental and geographical factors. The study aimed to identify key environmental variables that influence habitat suitability and provide insights into species-location associations on a global scale. Results show that these machine learning models significantly improve prediction accuracy compared to traditional methods, providing a valuable tool for conservation planning. This report provides an overview of the dataset and pre-processing methods, details the model selection and evaluation process, and concludes with recommendations for future research.

1 Introduction

Biodiversity plays an important role in ecological balance, and species distribution is one of the critical indicators that is necessary for conservation. The monitoring of species across various geographical locations can provide insights into habitat preferences of species, interspecies interactions, and possibly environmental threats. This project will develop classification models using the *iNaturalist* and *WorldClim* datasets to predict the presence of species within certain areas and examine associations between species and locations. Through machine learning methods, this study will explore applications that involve large, multi-dimensional biodiversity datasets, aiming to predict species distribution with geographical and climatic variables to contribute to conservation strategies and environmental decision-making.

Species distribution modelling (SDM) is a frequently used method in ecology for estimating the geographic ranges of species based on environmental suitability. Traditional models such as Maximum Entropy (MaxEnt) use known presence locations and environmental factors and then predict species presence with limited data [1]. However, MaxEnt and similar models are limited by their reliance on linear relationships, limiting their ability to capture complex ecological interactions. Recent advances in machine learning have introduced more complex models to capture these relationships, such as Random Forests (RF), Support Vector Machines (SVMs), and Gradient Boosting (GB) are increasingly used in SDM research because of their ability to deal with large multi-dimensional datasets and model non-linear relationships[2] [3]. These techniques broaden the scope of SDM and allow ecologists to achieve higher prediction accuracy in different environments[4] [5].

The aim of this project is to construct models that predict species presence within specific regions while also examining species-location associations across multiple geographic and bioclimatic dimensions. Unlike previous studies that are often restricted to single-species analysis or focused on limited areas, this project adopts a multi-species, global perspective, allowing for a broader exploration of biodiversity patterns. These insights can be instrumental for conservationists, helping to identify critical habitats, anticipate range shifts due to climate change, and develop proactive conservation measures.

This report is structured as follows: Section 2 provides an overview of the datasets and preprocessing steps. Section 3 presents exploratory data analysis, including visualizations of species distribution. Section 4 describes the machine learning models and feature selection methods employed in this study. Section 5 discusses model evaluation and performance metrics, and finally, Section 6 concludes with potential future directions.

2 Data preparation

The basic data set consists of 272073 samples indicating the location of different species that have been observed with a 2 dimensional geographical coordinates – Latitude and Longitude. The data was offered by *iNaturalist - www.inaturalist.org*[6]. Moreover, supported by *inaturalist taxa*[7], the type of species are recorded in a list with 500 integer taxon IDs and the corresponding names of each species. The final test for our models was under 288122 more data points obtained from *IUCN*'s database[8]. In addition, we prepared 1067592 more samples of 1918 species in total as backup for carrying out some potential exploration.

Furthermore, in our later experiment, based on data from WorldClim Bioclimatic, we add 19 more bio-climatic features which is introduced in more detail in 'Learning Methods - Feature Selection' section.

3 Exploratory Data Analysis

We utilise some exploratory data analysis methods to help us better understanding the data. Only training data is analysed in this section.

Figure1 plots all sample points in the form of scatter plots. As the basic data set are in the form of a simple 2-dimensional geographic feature, we can thus intuitively feel the geographical distribution of the species sample in general this way. Comparing it with the layout of the world continents, we can found that almost all of the samples were collected on the mainland or on small islands, with only a small number coming from the ocean. Few samples came from Antarctica or Greenland. This conclusion is also supported by the figure 2. Furthermore, as a heat-map, figure2 also demonstrate the density of the data in different area and its location. Especially for the northwestern part of North America and the southeastern part of Australia, they are the most intense area on the heat-map.

The histogram in Figure3 illustrates the distribution of the count of samples for each species. Among the 500 data points in the training set, there are more than 60 of them, each of which has about 2000 data points instances that were observed. For most of the remaining species, there are only fewer than 800 samples in the dataset for each species.

Convex Hull Algorithm is a commonly used method to measure sparsity, spread, or distribution trend of data in Geographic Information System according to Alkathiri et al. (2016)[9]. Given the set of locations on a two-dim plane, the convex hull is the smallest convex polygon that encloses these

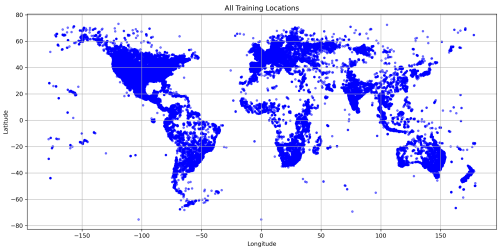

Figure 1: Data point distribution map

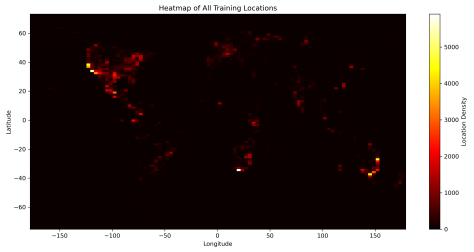

Figure 2: Data point distribution heat-map

points. We apply this algorithm to each species and generate box plot, where each point indicates the convex hull area of one particular species, to compare the sample distribution between various species which is presented in figure 4. The two black bars represent the 'minimum' and 'maximum' value of the whole dataset respectively. And the left and right sides of the box indicates the data of the top quarter and the bottom quarter of the data. As for the orange bar in the middle, it is the median value. To analysis, for most of the species, their geographical distributions are not large zones with the area smaller than 2500. When it comes to the remaining points on the right hand side of the box, the largest convex hull area even approaches 25000, which is more than 10 times the maximum value of ordinary data in statistics. We can infer that there is a small portion of all listed species widely distributed across the globe.

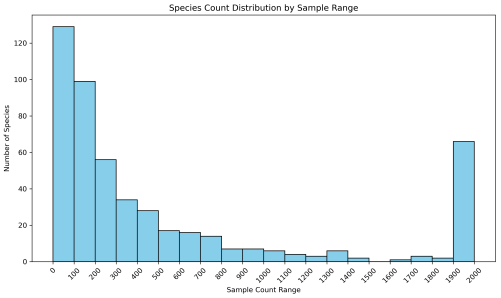

Figure 3: species count distribution

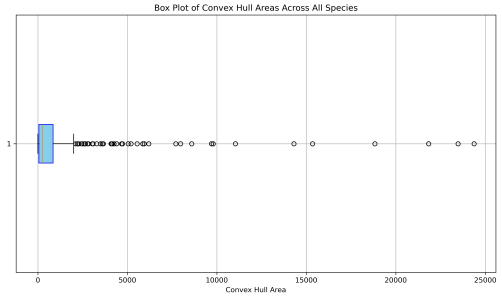

Figure 4: species convex hull area box plot

4 Learning methods

4.1 Feature Selection

The table in appendix (Table 1) shows the features used in our classification models.

In addition to basic geographic features, the **Latitude** and **Longitude**, we also utilized a clustering feature, **Cluster ID**, generated from K-means clustering based on these coordinates. This feature allows the model to capture regional distribution patterns.

Furthermore, we incorporated climate data from WorldClim Bioclimatic variables, providing additional features for each coordinate. These bioclimatic variables (BIO1 to BIO19) capture annual, seasonal, and monthly trends in temperature and precipitation, enhancing the model's ability to understand species distribution across various environmental conditions.

4.2 Model Selection

As shown in the section of Exploratory Data Analysis, the training data is rather sparse and may be complex for a linear model to generalise. And the conditional independence assumption of naive bayes made it a poor choice as well. Because intuitively the location of a point associates with the precipitation and temperature etc at that point.

For this report, we trained our data with Random Forest (RF), Gradient Boosting (GB) and Support Vector Classifier (SVC) for their potentials when working on multi-dimensional data. Gradient Boosting has the ability to minimize bias and variance. And for some species has much fewer samples comparing to the others. The robustness of SVC with small training set made it promising for this job.

4.3 Data Cleaning and Standardization

To improve model performance and reduce noise, we conducted outlier detection on the training data. Using the z-score method, we identified and removed data points with z-scores greater than 3 as outliers.

Next, all features were standardized by applying the StandardScaler method to both training and testing data. This standardization ensured the model's robustness across features with varying scales.

4.4 Model Training and Evaluation

Our approach treats the task as multiple binary classification problems, with a separate model for each species. For each species-specific model, positive samples (presence of the species) are significantly fewer than negative samples (absence of the species), creating a class imbalance. To address this, we employed random undersampling before training each model, ensuring an equal number of positive and negative samples to facilitate balanced model training.

For the Random Forest and Gradient Boosting models, we also computed feature importance scores to understand which features contributed most to classification decisions.

During testing, each model produced probability outputs for all species. To ensure higher confidence in the predictions, we applied a threshold of 0.7, meaning that only samples with a predicted probability of 0.7 or above were classified as positive for a given species. Given that there are 500 species in total, but any single location is unlikely to contain many species, a higher threshold helps prevent over-prediction and provides a more accurate reflection of species presence at each location.

To comprehensively evaluate model performance, we converted the model outputs into a multi-label classification format, where the predicted labels for each location corresponded to the species classified as positive by the respective species-specific models. We calculated Precision, Recall, and F1-Score using both Micro and Macro settings, providing insights into the model's effectiveness on the test set. Both the evaluation results and feature importance analysis are presented in the Results section. Note since SVM transfers the feature matrix into coefficients, we will not able to analysis its feature importance directly.

4.5 Model tuning

This report utilised the Elbow method when finding an optimal k value for K-Means clustering, by plotting the inertia of each on 5. With this method, we determined the optimal k value to be 3.

Also, we determines a better set of hyperparameters for each machine learning model by grid searching over possible combinations. We used `sklearn.model_selection.GridSearchCV` with f1 score as the performance matrix to work out the best hyperparameters for each species, and taking the parameters with the most occurrences as the final setting for all of the species for a simpler pipeline.

5 Results

Appendix Figure 6 illustrates the micro and macro f1 score obtained for each machine learning model on the test set. Micro and macro are two different weighting parameters used in computing an f1 score. Micro weighting emphasis on the overall accuracy while macro weighting emphasis on per-class accuracy in multi-class classification scenarios.

Our best macro f1 score of 0.609 is obtained with random forest classifier, while the best micro f1 score of 0.590 is obtained from SVC. Gradient boosting on the other hand, produced the worst

accuracy among three classifiers. Although the recall of gradient boosting is the highest achieved, its precision is much lower than the other classifiers (see Appendix Figure 7 and Appendix Figure 8).

As seen in Appendix Figure 9, the coordinates is one of the most important feature, especially the longitude. The other significant features in the order of dominance includes K-Means clusters, BIO3, BIO1, BIO4 and BIO6. We observed a similar trend in Grandient Boosting classifier (see Appendix Figure 10), where the top 4 dominent features are BIO1, longitude, K-Means clusters and BIO10. In both classifiers, longitude dominants latitude as a feature and BIO1 is much more importance than the other BIO variables. Although the longitude can take a range wider than latitude (-180 to 180 comparing to -90 to 90), we employed `sklearn.preprocessing.StandardScaler` to normalise features, so the difference in importance should be irrelavent to their scales. Furthermore, BIO1 denotes the Annual Mean Temperature of a given region as detailed in Appendix Table 1, suggesting the presence of a species might depending more on the mean temperature of that region than precipitation.

References

[1] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

[2] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[3] Qinghua Guo, Maggi Kelly, and Catherine H Graham. Support vector machines for predicting distribution of sudden oak death in california. *Ecological modelling*, 182(1):75–90, 2005.

[4] Robert J Hijmans, Susan E Cameron, Juan L Parra, Peter G Jones, and Andy Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 25(15):1965–1978, 2005.

[5] Wilfried Thuiller, Sandra Lavorel, Miguel B Araújo, Martin T Sykes, and I Colin Prentice. Climate change threats to plant diversity in europe. *Proceedings of the National Academy of Sciences*, 102(23):8245–8250, 2005.

[6] inaturalist, 2024. Accessed: 2024-10-26.

[7] iNaturalist. inaturalist taxa directory, 2024. Accessed: 2024-10-26.

[8] International Union for Conservation of Nature. Iucn red list spatial data download, 2024. Accessed: 2024-10-26.

[9] Mazin Alkathiri, Jhummarwala Abdul, and MB Potdar. Geo-spatial big data mining techniques. *International Journal of Computer Applications*, 135(11):28–36, 2016.

Appendices

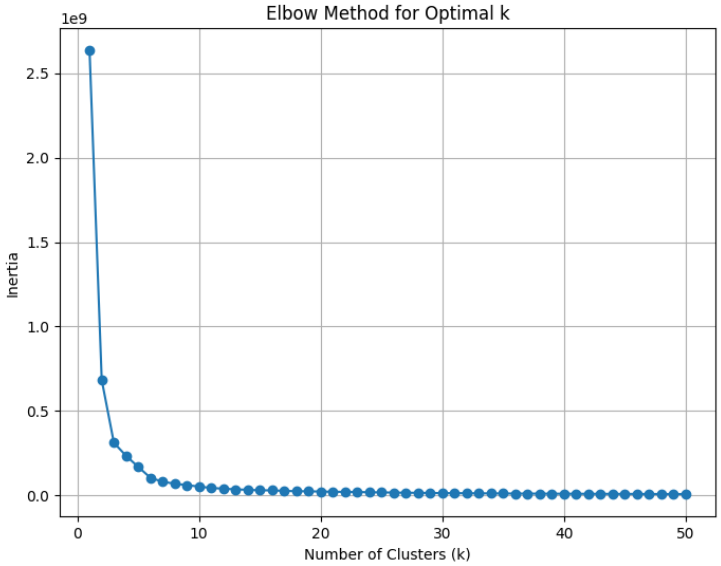

Figure 5: Optimal k value for K-Means clustering.

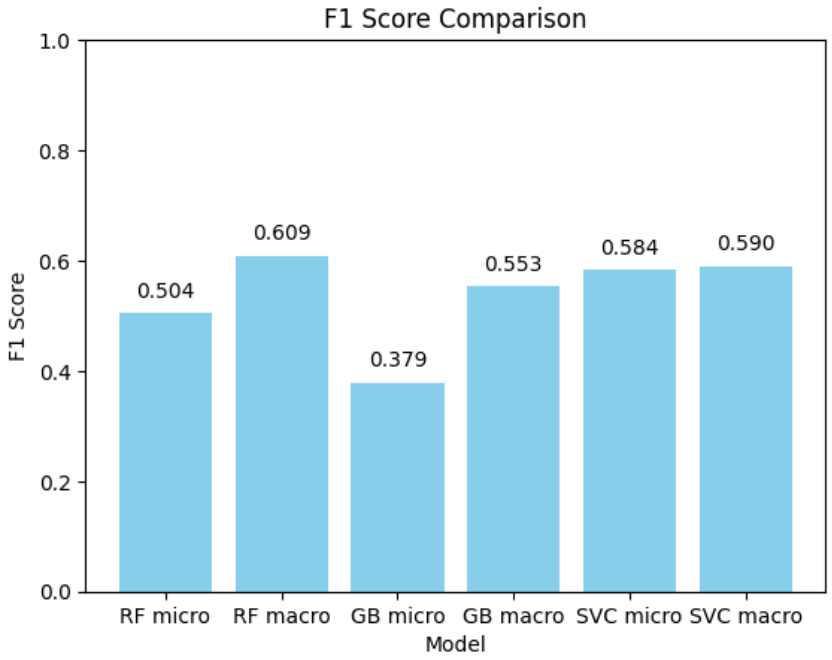

Figure 6: F1 score of ML models. RF = Random Forest, GB = Gradient Boosting, SVC = Support Vector Classifier.

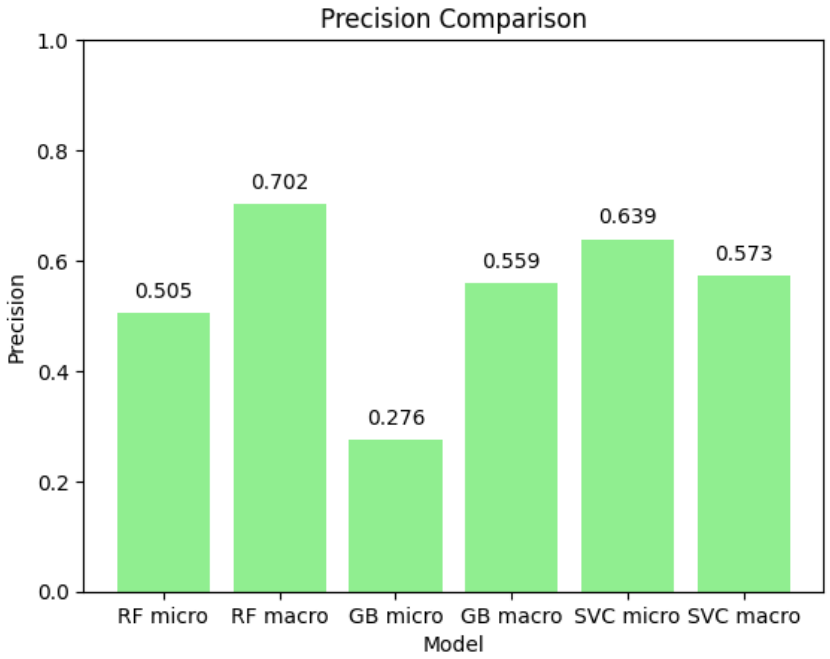

Figure 7: Precision of ML models. RF = Random Forest, GB = Gradient Boosting, SVC = Support Vector Classifier.

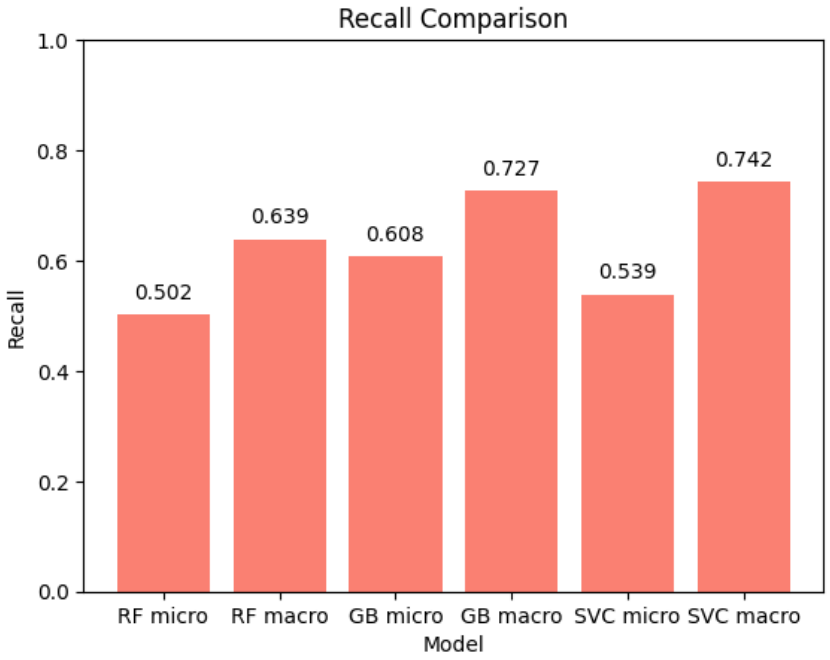

Figure 8: Recall of ML models. RF = Random Forest, GB = Gradient Boosting, SVC = Support Vector Classifier.

Table 1: Features and Descriptions

Feature	**Description**
Latitude	Geographic latitude
Longitude	Geographic longitude
Cluster ID	Cluster ID from K-means clustering based on Latitude and Longitude
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) ($\times$100)
BIO4	Temperature Seasonality (standard deviation $\times$100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

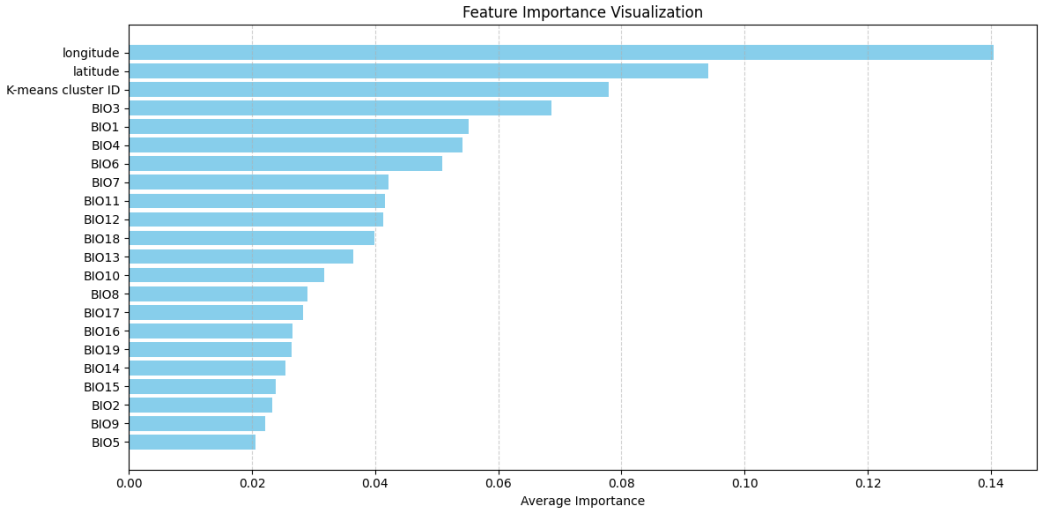

Figure 9: Random Forest feature importance.

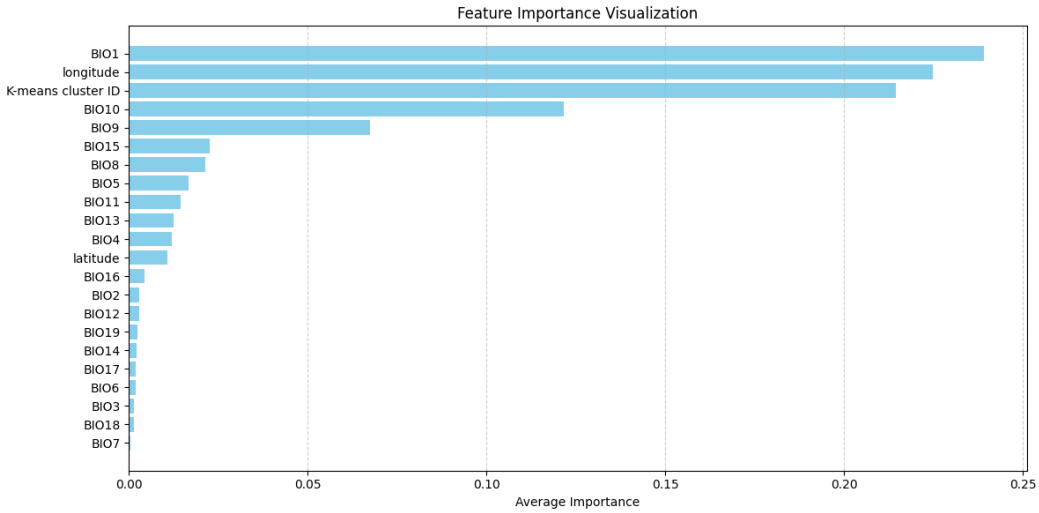

Figure 10: Gradient Boosting feature importance.