Coral Reef Fish Species

predicts the diversity index of fish species

Shalutha Perera

TABLE OF CONTENT

CONTENT

LIST OF FIGURES

LIST OF TABLES

1. Introduction

The diversity index of fish species is a crucial ecological metric used to assess the biodiversity within aquatic ecosystems. It reflects not only the variety of species present but also their relative abundance, providing insights into the health and stability of an ecosystem. A high diversity index indicates a balanced ecosystem with a wide range of species, while a lower index may suggest environmental stress, habitat degradation, or dominance by a few species. Understanding and predicting the diversity index is vital for conservation efforts, resource management, and maintaining the ecological balance of water bodies. This project aims to predict the diversity index based on factors such as morphological traits and environmental conditions, which influence the distribution and abundance of fish species.

2. DESCRIPTION OF THE QUESTION

In this project, we aim to develop a machine learning model that predicts the diversity index of fish species based on various factors, including morphological and environmental characteristics. Using the provided dataset, the workflow will follow a structured approach, starting with data preprocessing, exploratory data analysis to uncover patterns and relationships within the data. This will be followed by feature engineering and selection to enhance the predictive power of the model.

3. DESCRIBE THE DATASET

The dataset includes density data for 109 fish species included in the analysis, environmental data used for the density analysis

Variable	Description	Data Type
site	numeric site descriptor matching NOAA Reef Visual Census sites	Categorical
model	factor used to subset data for two separate models	Categorical
Year	year of RVC fish survey	Quantitative
Month	month of RVC fish survey	Categorical
Latitude	latitude of RVC fish survey	Quantitative
Longitude	longitude of RVC fish survey	Quantitative

Depth	depth of RVC fish survey averaged for each surveyor	Quantitative
Region	jurisdiction of RVC fish survey	Categorical
Coral_cover	percentage of benthos made up of living hard coral visually estimated by RVC surveyors	Quantitative
Reef_complexity	maximum hard relief measured by averaging the height of the highest rigid point above the lowest point in 8 segments of the cylinder for RVC surveys	Quantitative
SST	minimum monthly average sea surface temperature in Celsius derived from CoRTAD database from 2012-2016	Quantitative
NPP	net primary productivity derived from remotely sensed chlorophyll-a from the OSU VGPM model	Quantitative
Wave_exposure	exposure calculated using linear wave theory	Quantitative
Habitat_type_classLV0	habitat classification of each site according to the FWC Unified Reef Map level 0	Categorical
Habitat_type_classLV2	habitat classification of each site according to the FWC Unified Reef Map level 2	Categorical
Coral_area_UFRTM_20km	area classified as reef by Unified Reef Map level 0 within 20 km of each site	Quantitative
Coral_area_UFRTM_200km	area classified as reef by Unified Reef Map level 0 within 200 km of each site	Quantitative
Depth_Sbrocco	remotely sensed depth of survey sites	Quantitative
Deepwater	euclidean distance in meters over water to the 30-meter bathymetric line	Quantitative
FSA	number of marina slips over 45 ft within 10 km of each site	Quantitative
Marina_slips_10km	number of marina slips over 45 ft within 10 km of each site	Quantitative
Marina_slips_25km	number of marina slips over 45 ft within 25 km of each site	Quantitative
Marine_reserve	protected status of site; whether fishing was allowed or not	Categorical
Population_20km	human population living within 20 km of reef sites derived from LandScan dataset	Quantitative
Population_50km	human population living within 50 km of reef sites derived from LandScan dataset	Quantitative
Recreational_fishermen_50km	number of recreational fishing licenses within 50 km of reef sites derived by zip code	Quantitative

Tourist_fishing	statistics from Johns et al. 2001 and publicly available dataset of hotel units in Florida	Quantitative
Artificial_reefs_1km	number of artificial reefs within 1 km	Quantitative
SG_permits_50km	number of commercial snapper-grouper fishing permits within 50 km	Quantitative
SG_charter_permits_25km	SG_permits_50km: number of recreational snapper-grouper fishing permits within 25 km	Quantitative
Total_gravity_intercept	number of people in population centers within 500 km divided by the square of travel time	Quantitative
Total_gravity	number of people in population centers within 500 km divided by the square of travel time	Quantitative
Keys_Divisions	sub-jurisdictions of Florida Keys including Upper, Middle, Lower Keys and Marquesas; NAs for non Florida Keys sites	Categorical
FKNMS	Florida Keys National Marine Sanctuary sites; NAs for non Florida Keys sites	Categorical
DryTortugas	Dry Tortugas sites; NAs for non-Dry Tortugas sites	Categorical
BNP	Biscayne National Park sites; NAs for non-BNP sites	Categorical
CoralECA	Coral Ecological Conservation Area sites; NAs for non-ECA sites (also known as SEFCRI)	Categorical
Nursery_seagrass	connectivity of reef sites to continuous seagrass patches within 10 km	Quantitative
Nursery_mangroves	connectivity of reef sites to mangrove stands within 12 km	Quantitative
connectivity	number of larva from upstream modeled to a connectivity matrix	Quantitative
Comm_engagement	metrics of commercial engagement based on landings and permits provided by NOAA	Quantitative
Comm_reliance	metrics of commercial engagement based on landings and permits relative to size of fishing community provided by NOAA	Quantitative
Rec_engagement	metrics of recreational engagement based on landings and permits provided by NOAA	Quantitative
Rec_reliance	metrics of recreational engagement based on landings and permits relative to size of fishing community provided by NOAA	Quantitative

Commercial_pounds_landed	annual number of pounds of fish reported by commercial anglers	Quantitative
Pop_per_area_reef_20km	human population divided by area of reef within 20km	Quantitative
Random	random number assigned to each column	Quantitative
impact	fishing impact variable	Quantitative
YEAR	year of RVC surveys	Quantitative
HABITAT_CD	habitat code used by NOAA RVC surveys to stratify sites	Categorical
REGION	jurisdiction of RVC survey sites	Categorical
PCT_CORAL	percent coral cover	Quantitative
MAX_HARD_RELIEF	maximum hard relief	Quantitative
no.divers	number of divers for RVC survey	Quantitative
Diversity_index	score that varies between 0 and 1	Quantitative

Table 1 - Variable Description

4. DATA PREPROCESSING

At the start of the project, our main objective was to develop a model that predicts the diversity index of fish species.

We began the Data cleaning process by removing unnecessary columns in data set. In the beginning our data set have 3999 rows and 56 columns.

According to the Domain Knowledge of Fishers diversity index we select key Variables, To select the most appropriate variables for predicting the diversity index of fish species, We will focus on variables that are directly or indirectly related to environmental, ecological, and habitat conditions, as well as those related to the species diversity itself. Here's the reasoning:

1. **Removing Unnecessary Columns**: Variables not contributing to the diversity index prediction were dropped.

Key Variables to Include:

1. Latitude and Longitude:

Reason: Geographic coordinates are essential for location-based variations in species diversity, environmental conditions, and fish habitat availability.

2. Depth:

Reason: Fish diversity is strongly influenced by the depth of the habitat, as species adapt to different depth levels and related environmental factors (e.g., light, pressure).

3. Coral_cover:

Reason: Coral cover is a significant determinant of biodiversity in reef ecosystems. A higher coral cover generally correlates with more diverse species due to the complex habitats it provides.

4. Reef_complexity:

Reason: Structural complexity of reefs provides hiding spots and resources for fish, directly influencing species richness and abundance.

5. SST (Sea Surface Temperature):

Reason: Sea surface temperature affects the metabolic rates and habitat ranges of fish species. It is a critical driver of marine biodiversity and can influence the diversity index.

6. NPP (Net Primary Productivity):

Reason: Higher productivity generally supports a richer food web, which can support a higher diversity of species.

7. Wave_exposure:

Reason: Exposure to waves affects reef health and the species that can thrive in a particular area. It impacts habitat stability, which in turn affects biodiversity.

8. Habitat_type_classLV0:

Reason: Habitat classification at a high level (LV0) helps in understanding the type of ecosystem present at the site, which influences species diversity.

9. Coral_area_UFRTM_20km:

Reason: The area of coral reef within 20 km of the site is relevant for assessing the regional diversity of fish species that can inhabit the coral areas.

10. Marine_reserve:

Reason: Protected areas often have higher biodiversity due to reduced fishing pressure, which allows fish populations to grow and diversify.

11. Nursery_seagrass and Nursery_mangroves:

Reason: Connectivity to seagrass and mangrove nurseries can enhance fish species diversity due to their role as breeding grounds and juvenile habitats for many species.

12. Population_20km or Population_50km:

Reason: Human population density near reefs can indicate the degree of anthropogenic pressure (pollution, overfishing), which impacts the health of marine ecosystems and fish diversity.

13. Region:

Reason: Geographic regions influence ecological conditions and species composition. Including this variable can help capture regional differences in fish diversity.

14. MAX_HARD_RELIEF:

Reason: Hard relief is another indicator of reef complexity. Since complex structures provide more habitats, it's crucial for predicting biodiversity.

15. Coral_area_UFRTM_200km:

Reason: This variable extends the area of coral reef consideration, which could influence regional biodiversity beyond the immediate surroundings. Include this to account for larger-scale ecosystem impacts.

16. Pop_per_area_reef_20km:

Reason: Population pressure on reefs, when normalized by reef area, can affect biodiversity by influencing fishing intensity and pollution. This is a meaningful human impact variable.

Variables to Exclude:

1. Year, Month, Site, Model, Random,Unnamed: 0:

Reason: These are identifiers or time-based variables that do not directly influence the diversity index. They are used for data organization and model splitting, not prediction.

2. Keys_Divisions, FKNMS, DryTortugas, BNP, CoralECA:

Reason: These are location-specific jurisdictions. Latitude and longitude already cover geographical variation, so these are redundant.

3. Tourist_fishing, SG_permits, SG_charter_permits, Marina_slips:

Reason: While human activity can affect species diversity, these specific metrics are indirect and likely less impactful compared to population density or reserve status.

4. Comm_engagement, Rec_engagement, Comm_reliance, Rec_reliance:

Reason: These variables relate to socio-economic factors, which are useful for fisheries management but are less directly related to predicting ecological diversity.

5. Artificial_reefs_1km:

Reason: Artificial reefs may have local effects, but coral reef area is a more comprehensive measure of natural habitat availability.

6. Deepwater, FSA, Distance to Fish Spawning Aggregations:

Reason: While potentially relevant, the inclusion of reef complexity, coral cover, and depth already provides sufficient environmental context for diversity.

7. Impact, Total_gravity:

Reason: Impact or human pressure is already captured by other more relevant variables like population, marine reserve status, and fishing impact.

8. PCT_CORAL (Percent Coral Cover):

Reason: Similar to Coral_cover

9.Depth_Sbrocco:

Reason: Since Depth is already included, and depth is measured directly in other ways, this remotely sensed depth measure is redundant.

10. Marina_slips_10km and Marina_slips_25km:

Reason: Marina slips might indicate human activity, but Population_20km or Pop_per_area_reef_20km already provides a stronger proxy for human impact, making these variables less necessary.

11. Recreational_fishermen_50km:

Reason: While recreational fishing can influence fish populations, it's less directly related to fish diversity than other human impact variables like Population_20km or Marine_reserve.

12. Total_gravity_intercept:

Reason: This is a more complex socio-economic measure and is indirectly related to the diversity index. Other, more direct measures of human pressure, such as population density, are already included.

13. connectivity:

Reason: This variable is useful in a larval dispersion context but is less directly linked to fish species diversity compared to habitat and environmental variables.

14. Commercial_pounds_landed:

Reason: The amount of fish commercially landed focuses on economic activity rather than ecological diversity. It can be excluded as its impact is indirect.

15. HABITAT_CD:

Reason: This is a code used for site stratification. Since habitat complexity and coral cover are already included in more detailed forms, this variable can be excluded.

16. no.divers:

Reason: The number of divers does not impact fish diversity; it is simply a survey parameter. Therefore, it is unnecessary for predicting diversity.

2. **Handling Missing Values**: Missing data were handled using the median for numerical columns.

3. **Handling Duplicate Values :** Remove Duplicate values

4. **Outlier Detection**: Box plots were used to detect outliers, and the IQR method was applied for treatment.

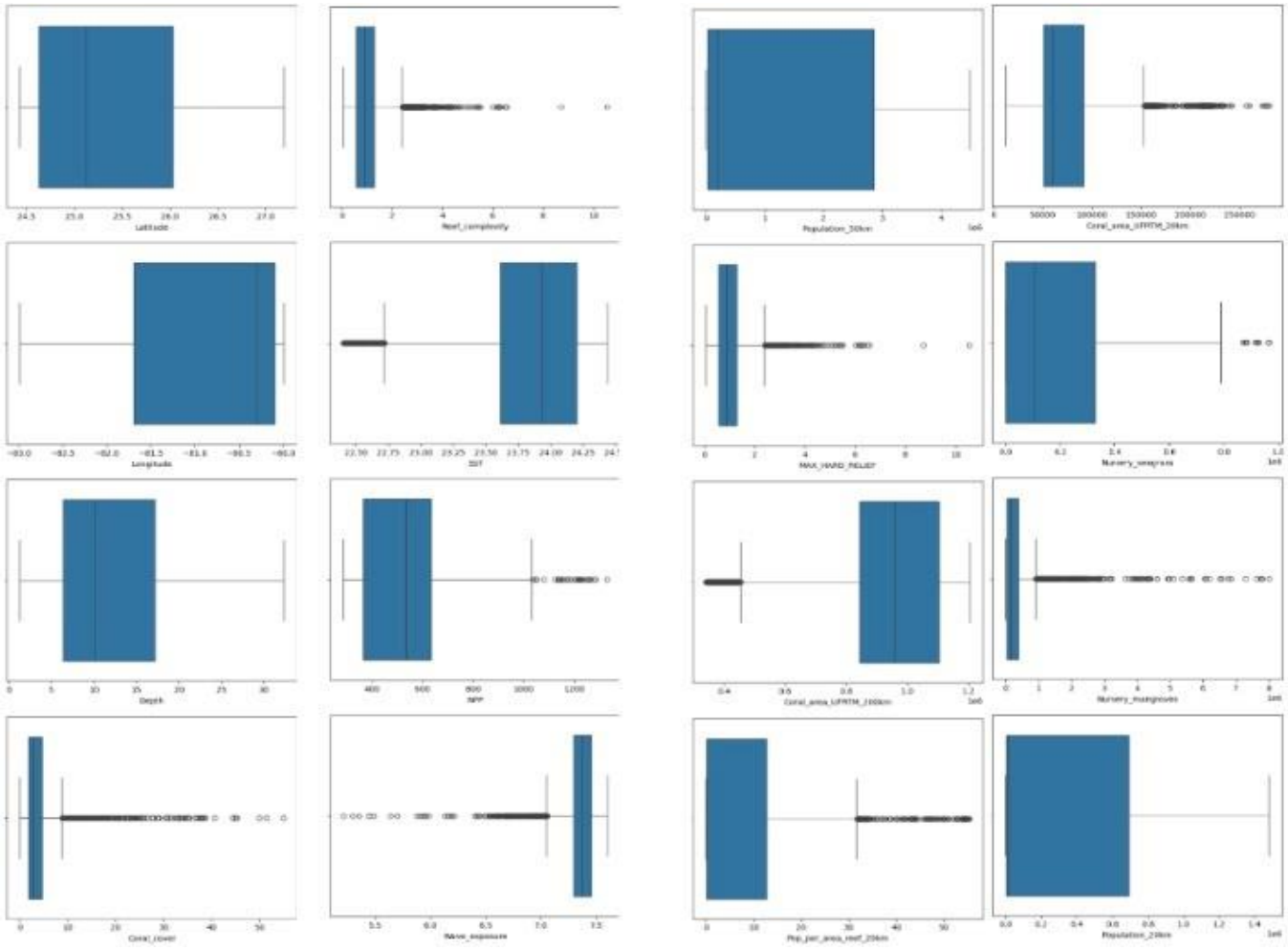

Figure 1.Outliers

5. Exploratory Data Analysis (EDA)

In this section, we analyze the dataset to understand its structure, identify patterns, and detect any anomalies. EDA was performed to understand the distribution and relationships within the data. The following visualizations and statistical analyses were conducted to explore the data:

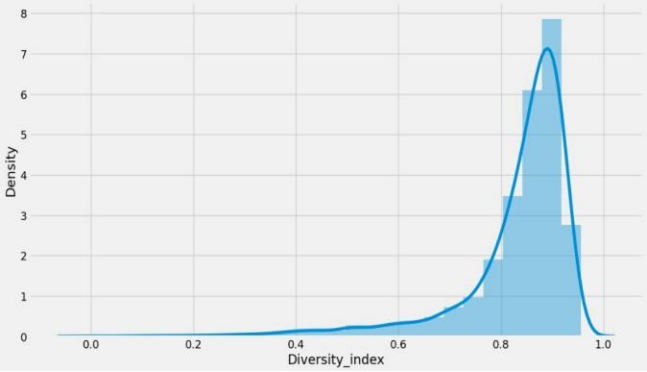

Figure 2. Displot for Diversity_index

The distribution of the Diversity_index variable in the provided plot appears to be right-skewed, with most values concentrated between 0.7 and 1.0. The density peaks near 0.85, indicating that higher diversity index values are more common. The sharp drop after the peak suggests fewer instances with diversity indices closer to 1.0, while lower diversity index values (below 0.5) are rare.

- **Univariate Analysis**

  Histograms revealed that some variables were skewed, and multimodal distributions suggested the presence of distinct groups.

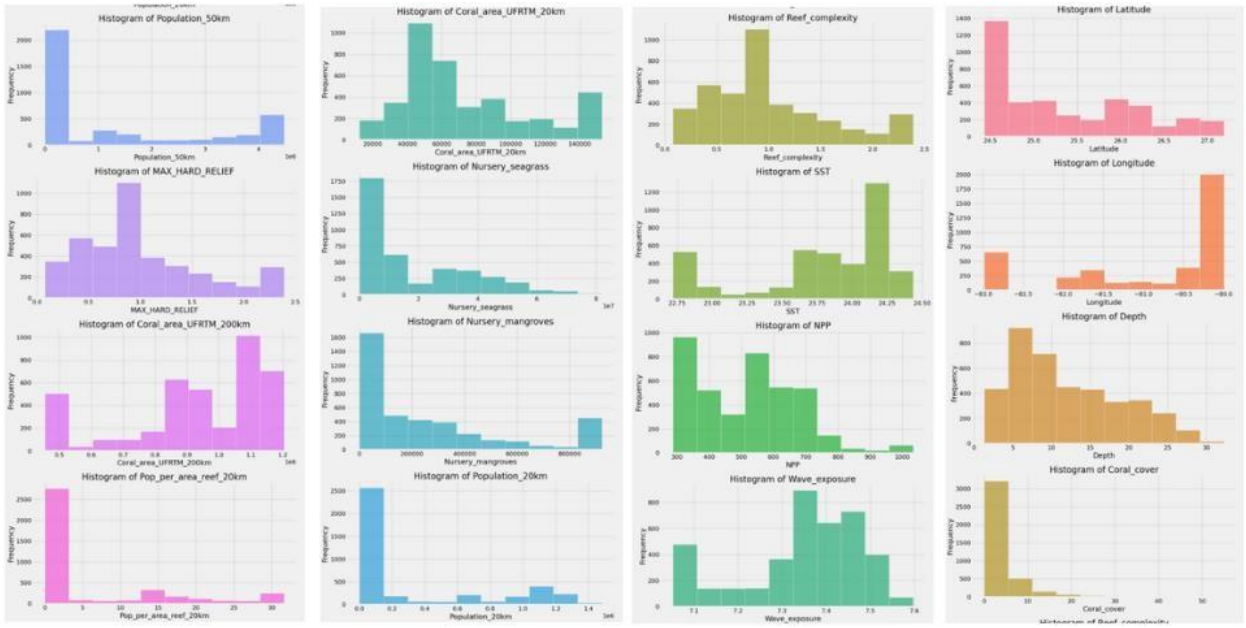

Figure 3. Histograms

The histograms in the image provide an overview of the distributions for various features in Our dataset. Here are some key conclusions:

- Skewed Distributions: Many variables, such as Population_50km, Coral_area_UFRTM_20km, Wave_exposure, and Coral_cover, show skewed distributions. These may require transformation (e.g., log or square root) to normalize their distributions for better modeling performance.

- Multimodal Distributions: Some features, such as Reef_complexity, SST (Sea Surface Temperature), and NPP (Net Primary Productivity), exhibit multimodal distributions, indicating the presence of multiple distinct groups or regions within the data.

- Sparse Data: Certain variables, such as Pop_per_area_reef_20km and Coral_area_UFRTM_200km, show sparse regions, where a large number of instances have lower values, while only a few instances have higher values. This might influence the modeling process.

- Potential Outliers: Features like Population_50km and Pop_per_area_reef_20km have extreme values that might be considered outliers. These should be examined to decide if they are legitimate values or if they need special treatment.

- **Bivariate Analysis**

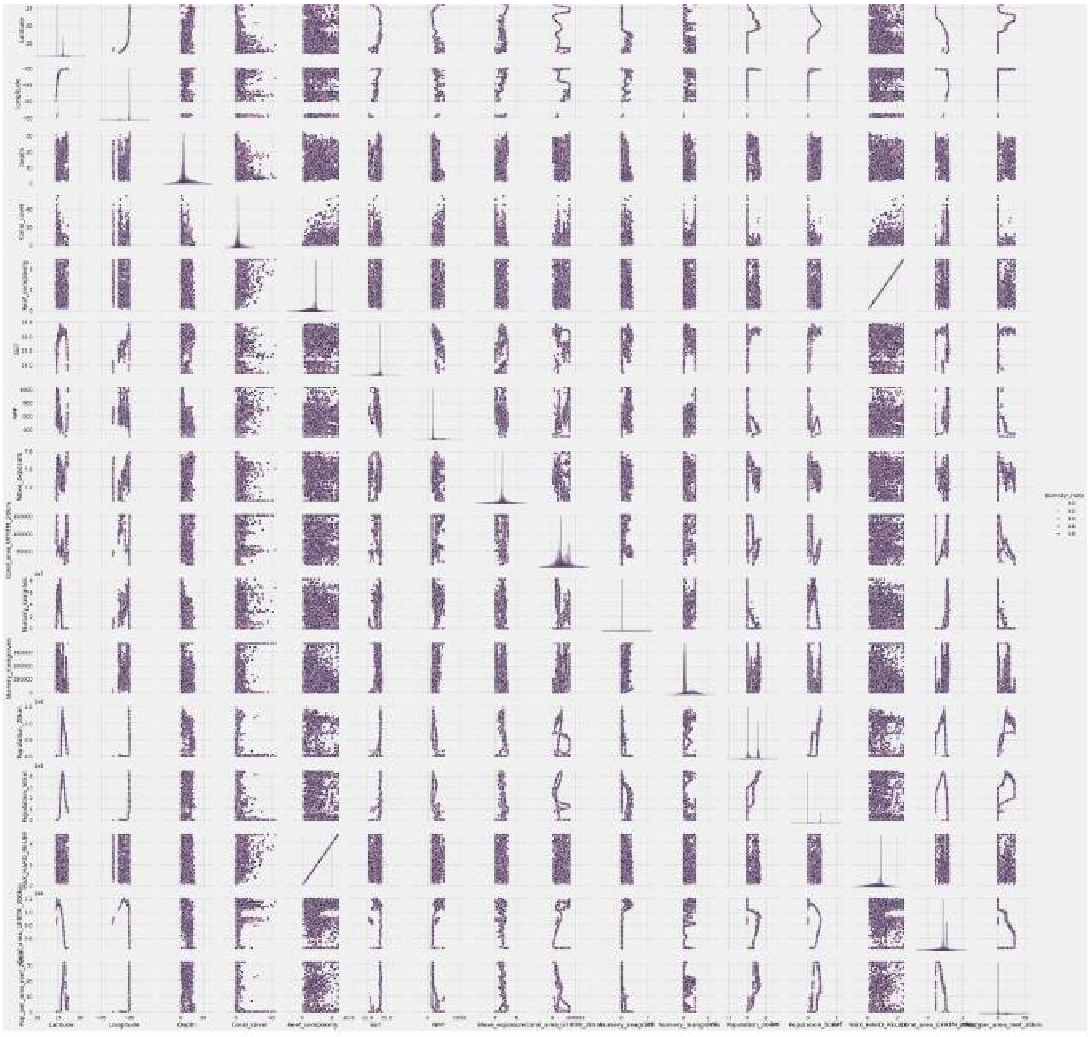

Figure 4.Pirplot

The pairplot provides insights into the relationships and distribution between various features in our data set.

6. FEATURE ENGINEERING

Here we use **One -Hot -Encoding** method to convert our Categorical columns to Numerical columns. After that we do Feature scaling process using StandScaler().

7. FEATURE SELECTION

- Correlation plot of variables

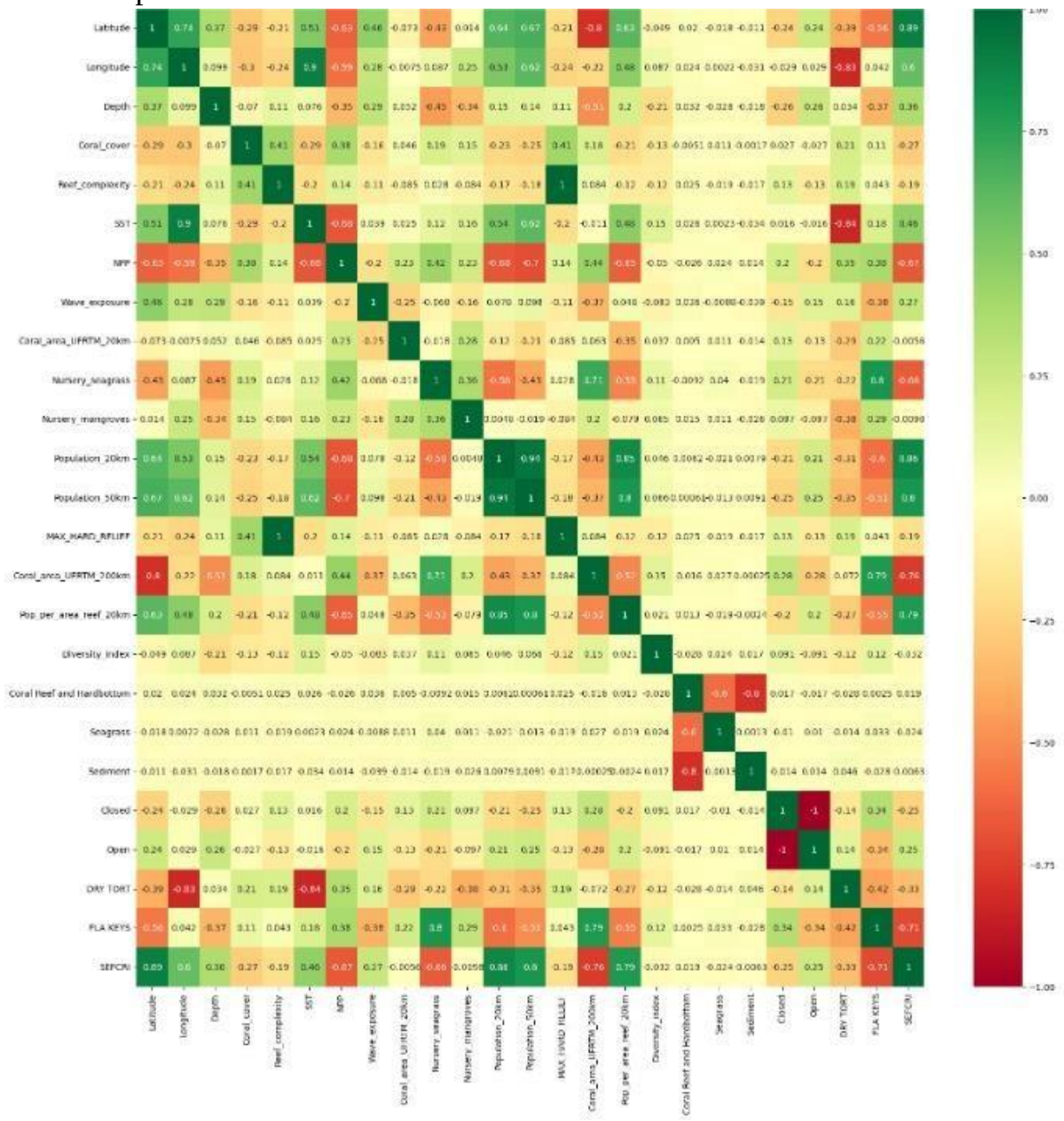

Figure 5.Correlation Matrix

This plot shows the correlation between the explanatory variables. Feature selection was based on correlation with the target variable. Variables with little or no correlation to the diversity index were removed. A correlation matrix was used to identify significant variables.

8. MODEL BUILDING

In this problem is the Regression Problem. So we used Regression models to develop best model for that problem. Here we use,

- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- Support Vector Regression (SVR)
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression
- K-Nearest Neighbors (KNN) Regression
- Bayesian Ridge Regression

I. Model Performance Comparisons

To select the best model, we evaluated each based on key performance metrics such as **R-squared (R^2)**, **Mean Squared Error (MSE)** using cross-validation. The model with the highest accuracy and lowest error, while maintaining generalization across unseen data, was ultimately chosen as the best-performing model.

IMPORTANT RESULTS IN ADVANCED ANALYSIS

MODEL	R^2	MSE
Gradient Boosting	0.153290	0.011105
Support Vector Regression	0.11565	0.011565
Random Forest Regression	0.096904	0.011768
Linear Regression	0.072688	0.012192
Ridge Regression	0.072681	0.012192
Bayesian Ridge Regression	0.071865	0.012203

K-Nearest Neighbors (KNN) Regression	0.022419	0.012794
Lasso Regression	-0.002815	0.013184
ElasticNet Regression	-0.002815	0.013184
Decision Tree Regression	-0.684659	0.022350

Table 2 - Model Performance

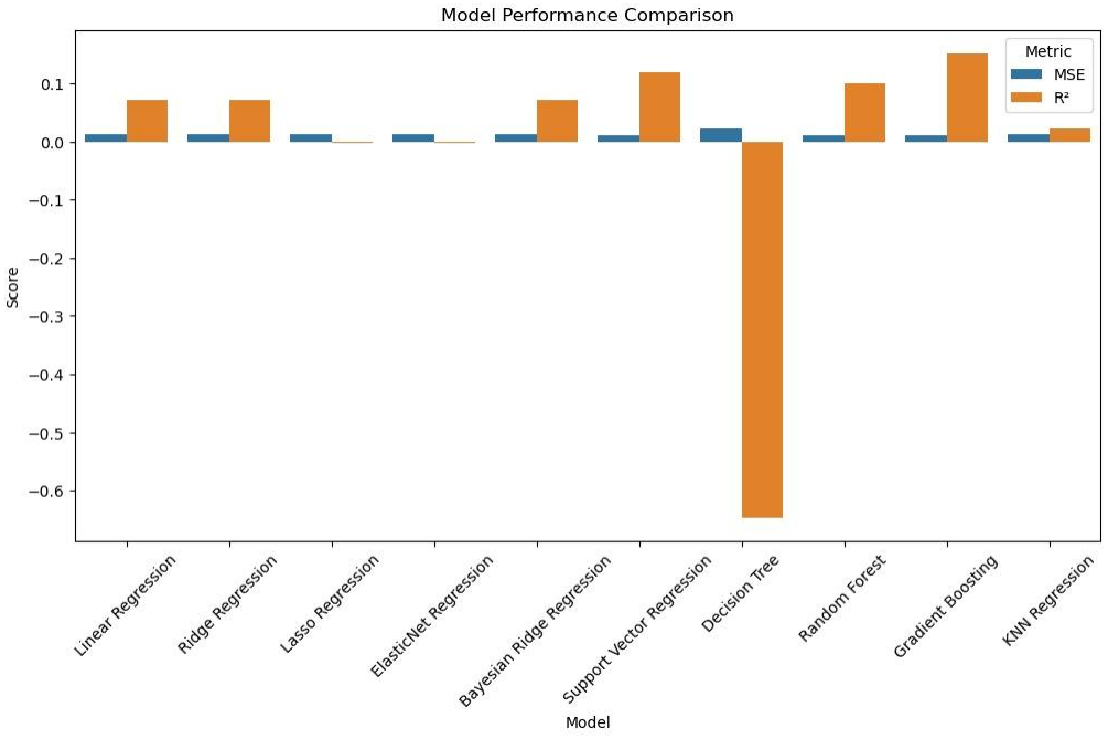

Figure 6.Model Performance

- After using the R^2 and the MSE values, We want to identify what is the best method for for predict our Diversity_index.According to the R^2 and the MSE values Gradient Boosting Regression is the best model for our data set.

II. Hyperparameter Tuning for Gradient Boosting Regression

After identifying Gradient Boosting Regression as the best-performing model, we focused on optimizing its performance by tuning the hyperparameters. Hyperparameter tuning is a critical

step in improving the model's accuracy and generalization by finding the most suitable combination of settings for the algorithm.

Key Hyperparameters in Gradient Boosting Regression

The performance of Gradient Boosting Regression can be influenced by several hyperparameters:

- **n_estimators**: The number of boosting stages or trees to be used. More trees can increase accuracy but also raise the risk of overfitting.
- **learning_rate**: Controls the contribution of each tree. Lower values make the model more robust but require a higher number of trees.
- **max_depth**: The maximum depth of each tree. Shallower trees may underfit, while deeper trees can overfit the data.
- **min_samples_split**: The minimum number of samples required to split an internal node.
- **min_samples_leaf**: The minimum number of samples required to be at a leaf node.
- **subsample**: The fraction of samples to be used for fitting individual trees, which helps prevent overfitting.
- **max_features**: The number of features to consider when looking for the best split.

Hyperparameter Tuning Strategy

To optimize the Gradient Boosting Regression model, we used **Grid Search Cross-Validation (GridSearchCV)**. This method exhaustively tests all combinations of the hyperparameters to find the best set for the model, ensuring the highest accuracy and generalization capability.

GridSearchCV was applied with 5-fold cross-validation to prevent overfitting while evaluating the performance across different combinations.

Best Hyperparameter Combination

After running GridSearchCV, the optimal hyperparameters were:

- **n_estimators**: 300
- **learning_rate**: 0.01
- **max_depth**: 3
- **min_samples_split**: 5
- **subsample**: 0.8

These settings provided the best performance based on cross-validation, with further improvement in both the **R^2** and **MSE** values.

Results After Hyperparameter Tuning

The performance of the Gradient Boosting model improved after tuning:

- **R^2**: 0.19208749921895585
- **MSE**: 0.010384370892983158

Through hyperparameter tuning, the Gradient Boosting model's performance increased. The fine-tuned model showed better accuracy in predicting the diversity index, with improved generalization to unseen data.

9. CONCLUSIONS

- The **Gradient Boosting Regression** model, which had the highest R^2 score and lowest MSE after hyperparameter tuning, proved to be the best model for predicting the fish species diversity index.
- The final Gradient Boosting model identified the following key variables as most influential in predicting the diversity index: **Depth, Coral Cover, Reef Complexity, Sea Surface Temperature (SST), Nursery_segrass,MAX_HARD_RELIEF,Coral_area_UFRTM_200km,DRY TORT,FLAKEYS**. These factors play a crucial role in determining the diversity of fish species across various reef sites.

- Through careful feature selection, we reduced the number of input variables from an initial set of 56 to a more refined set of 10 key features. This reduction not only improved the efficiency of the model but also enhanced its interpretability, providing clearer insights into the factors most critical for predicting biodiversity.

10. ISSUES ENCOUNTED AND SOLUTIONS

- **Imbalanced Dataset**: The dataset had varying distributions of fish species diversity, with a majority of observations concentrated between 0.7 and 1.0 on the diversity index scale. This imbalance may pose challenges for model performance, as fewer instances represent lower diversity indices. By combining techniques such as **targeted resampling** with domain knowledge, we ensured that our predictions remained meaningful and relevant for understanding fish species diversity across reef ecosystems.

- **Large Number of Variables**: The dataset initially included 56 variables, many of which were indirect or socio-economic in nature. To streamline the analysis, we relied on **ecological knowledge** to focus on variables most likely to affect biodiversity. Features like **Latitude**, **Wave Exposure**, and **Marine Reserve**

Status were prioritized, as they are known to influence coral reef ecosystems, leading to better model interpretability and biological relevance.