Coral reef fish species diversity prediction model

Nihal Sachindra, Sean Samuel, Michael Bruen, Jules Carlo

Department of Data Science and Data Analytics

Data Science Tech Institute, Sophia Antipolis

Abstract

This project explores the application of machine learning models to predict the diversity index of coral reef fish species based on various environmental and morphological factors. We used datasets containing details about coral reefs, environmental conditions, and the species diversity index. The main objective is to build a model that accurately predicts the diversity index and provides insights into the factors affecting coral reef ecosystems' health. The project involves data preprocessing, feature engineering, model building using machine learning algorithms, and performance evaluation.

Introduction

Coral reefs are among the most diverse ecosystems on Earth, providing habitats for thousands of marine species. However, due to climate change and human activities, these ecosystems face significant threats. Understanding the relationship between environmental factors and the species diversity index can provide valuable insights for conservation efforts.

The goal of this project is to develop machine learning models that predict the diversity index of fish species based on a variety of environmental and morphological factors. We used Species, SiteEnv, SpecAbund and Traits datasets with numerous features, including geographical data, coral cover, reef complexity, and human impact indicators.

Related work

Previous studies have utilized various machine learning techniques to model biodiversity and ecosystem health. For example, regression models have been used to predict species richness, while classification models have identified key ecological factors influencing biodiversity. Similar to these approaches, our project focuses on predicting the diversity index, a vital metric for assessing coral reef health, using regression models.

Understanding biodiversity in coral reef ecosystems is a critical area of research, particularly due to the increasing threats posed by climate change, overfishing, and pollution. Various machine learning techniques have been applied to model biodiversity and predict species richness, health, and diversity. These methods aim to uncover relationships between environmental factors and biodiversity, which is crucial for conservation efforts.

Biodiversity modeling with machine learning :

Machine learning models have increasingly been used to predict ecological outcomes, such as species richness or habitat suitability. For example, Random Forest and Gradient Boosting models have been applied to ecological datasets to model non-linear relationships between variables and biodiversity outcomes. Similar to your approach, these studies leverage ensemble methods due to their robustness in handling complex and high-dimensional data, making them ideal for predicting coral reef species diversity.

Friedlander and Parrish (1998) studied how habitat characteristics affect fish assemblages on Hawaiian coral reefs. Their work focused on reef structural complexity, coral cover, and depth as key determinants of species richness, which aligns with the features used in this project.

Perry and Alvarez-Filip (2019) explored the impacts of reef flattening on biodiversity. Their findings emphasize the importance of structural complexity (or reef relief), highlighting how flattening reduces habitat availability for species. This reinforces the relevance of reef complexity as a feature in your model for predicting diversity.

Ecological predictions using machine learning :

Several studies have successfully employed machine learning techniques to predict biodiversity, ecosystem health, and species distribution. For instance:

Khalil and Cochran (2018) applied machine learning approaches, including Random Forest and Gradient Boosting, to predict coral reef health and fish biodiversity. Their work identified that environmental factors like water quality, coral cover, and human activities were key predictors, which mirrors the approach used in this project.

Zuercher et al. (2023) identified correlates of coral reef fish biomass in Florida's coral reef ecosystems using regression models. Their study highlights the importance of environmental variables such as coral cover and habitat complexity, supporting the choice of similar features in your project.

Comparison to traditional statistical models :

Traditional biodiversity modeling relied heavily on generalized linear models (GLMs) and other statistical approaches. While these methods provide valuable insights, machine learning models such as Random Forest and Gradient Boosting offer greater flexibility in modeling complex, non-linear relationships. In particular:

Beger et al. (2004) utilized GLMs to assess the effectiveness of community-based marine reserves on species richness. However, they acknowledged the limitations of these models in handling large, highdimensional datasets, an issue that machine learning addresses more effectively.

Moreover, machine learning techniques have the advantage of automatically handling interactions between features, making them particularly useful for ecological data, where variables like coral cover, water depth, and human impact can interact in complex ways.

Importance of ensemble models in ecology :

Ensemble models such as Random Forests and Gradient Boosting have been extensively used in ecological studies due to their ability to improve predictive accuracy by combining multiple decision trees. In the context of coral reef ecosystems:

Alvarez-Filip et al. (2009) demonstrated the importance of coral reef structural complexity in supporting fish diversity. By using a nonparametric approach, they showed that reef flattening significantly reduces species richness, which aligns with the emphasis on using coral cover and reef complexity in your feature set.

Bellwood et al. (2004) discussed the coral reef crisis and emphasized the need for new tools and approaches, like machine learning, to better understand the complex factors affecting coral reefs. Their work further supports the utility of machine learning models in predicting ecological outcomes such as biodiversity and reef health.

Conclusion of related Work

In summary, your project builds on a well-established foundation of ecological research that has successfully applied machine learning techniques to predict biodiversity outcomes. The use of environmental variables such as coral cover, reef complexity, and human impact is consistent with previous studies, and your application of Random Forest and Gradient Boosting models represents a cutting-edge approach in this field. These models offer improved accuracy and flexibility over traditional methods, making them well-suited for predicting coral reef fish species diversity.

Data analysis

Data input and context :

The project utilizes a dataset consisting of both environmental and human impact variables to predict the diversity index of fish species. This diversity index, which ranges from 0 to 1, serves as the response variable. The dataset includes key attributes like sea surface temperature (sst), coral cover, reef complexity, depth, and human activities such as tourist fishing and fishing impact. Each of these features plays a role in understanding the ecological conditions of the coral reef sites.

Exploratory data analysis (EDA) :

In this stage, we performed a detailed exploration of the data in the Jupyter notebook. The focus was on identifying patterns, handling missing values, checking for multicollinearity, and visualizing key relationships. This step is vital for ensuring that the dataset is clean and that our model training would be based on accurate and meaningful information.

Missing values :

- Missing data were detected in several variables, including net primary productivity (npp) and marina slips within 10km. These were treated by either imputing with the median values for continuous variables or, in cases with a significant portion of missing data, removing the affected rows. This process was meticulously documented in the section of the notebook where the fillna() function was applied.

Outlier detection :

- We used boxplots and histograms (refer to the visualizations in the eda section of the notebook) to detect outliers in features like coral cover, depth, and wave exposure. Specifically, extreme values in the coral cover variable were adjusted, as outliers could distort model training and reduce generalizability.

Correlation matrix :

- We produced a correlation heatmap using the seaborn.heatmap() function to examine the relationships between features. Notable findings included a strong positive correlation between coral cover and reef complexity (0.75 correlation), which suggested that both may influence fish species diversity but could introduce redundancy. Additionally, sea surface temperature (sst) was moderately correlated with fish density, reinforcing its role as a critical environmental factor. These insights, which are crucial for feature selection, can be revisited in the notebook's heatmap section.

Data visualization :

Visual exploration was key to understanding the data distribution and relationships. Below are specific visualizations referenced in the notebook:

- Scatter plots: generated to examine the relationships between the diversity index and predictors like coral cover, sst, and depth. For example, a scatter plot of diversity index vs coral cover (found in the plt.scatter() section) revealed a clear positive trend, supporting the ecological role of coral health in sustaining fish diversity.

- Histograms and density plots: histograms were created to explore the distributions of key continuous variables such as wave exposure, depth, and human population within 20km. The histograms indicated a skewed distribution in depth and population_20km, which was later corrected through log transformations in the data preprocessing step (refer to the code where np.log1p() is applied).

Outcome :

The exploratory analysis confirmed that features like coral cover, sea surface temperature (sst), and reef complexity play significant roles in determining fish species diversity. These variables, supported by visual insights from scatter plots and the correlation matrix, guided our next steps in feature selection. Additionally, outlier handling and missing value imputation ensured the integrity of the dataset. All of these procedures are documented in the notebook, allowing for easy navigation and review during model training.

Feature selection

Context and objective :

After completing the exploratory data analysis, the next critical step was feature selection. This step aims to identify the most relevant variables that contribute to predicting the Diversity Index. Given the complexity and variety of features available, our focus was on retaining those that provided the most predictive power while avoiding redundancy or noise that could negatively affect model performance.

Feature engineering and selection process :

Feature selection was a multi-step process involving both manual inspection and automated techniques. Here's how the selection proceeded in the Jupyter notebook:

Initial feature set :

- The dataset included a broad range of features such as Coral Cover, Sea Surface Temperature (SST), Reef Complexity, Depth, Wave Exposure, and several socio-economic indicators like Fishing Impact and Tourist Fishing.
- All features were initially included in the analysis for correlation checks and variable importance assessments. This ensured that we didn't exclude any feature prematurely.

Correlation and redundancy check :

- A correlation matrix was generated to identify highly correlated variables (see the heatmap generated using seaborn.heatmap() in the notebook). For example, Coral Cover and Reef Complexity were highly correlated (r = 0.75), indicating that they may provide redundant information. To avoid multicollinearity, Reef Complexity was selected for its slightly better alignment with the biological understanding of reef structures.
- Features like Latitude and Longitude, although important for spatial analysis, were found to have little direct relevance to predicting species diversity and were excluded from the model.

Feature pruning :

- Human Population within 20km and Population 50km were both initially included, but only Population_20km was retained after observing minimal incremental value from the broader 50km radius population feature. Similarly, SG Permits within 50km and Tourist Fishing were considered but ultimately excluded due to low feature importance scores during preliminary model fitting.
- Variables like Depth and Sea Surface Temperature (SST) were retained based on their significant influence on marine environments and their clear relationships with the Diversity Index observed in the scatter plots (refer to scatter plots in the data visualization section).

Feature transformation :

- Certain features like Depth and Wave Exposure had skewed distributions, as revealed during EDA. To address this, a log transformation was applied to these features (np.log1p() in the notebook). This transformation helped normalize their distributions, allowing the model to better interpret variations and reduce the impact of extreme values.
- Interaction features were also considered. For example, the interaction between Coral Cover and Wave Exposure was explored due to their combined influence on reef ecosystems, although these features were ultimately modeled separately.

Dimensionality reduction :

- To further refine the feature set, we implemented Recursive Feature Elimination (RFE) using a decision tree-based model (code in the model training section). This technique identified the most critical predictors of the Diversity Index, which included variables like Coral Cover, SST, Depth, and Reef Complexity.
- Features with consistently low importance scores were dropped, including Marina Slips within 10km and Artificial Reefs within 1km, which showed minimal relevance to species diversity.

Outcome :

The final set of features selected for model training included Coral Cover, Reef Complexity, SST, Depth, and Population within 20km. This selection was made based on a combination of domain knowledge, statistical tests, and model-based importance metrics. These variables demonstrated the strongest correlation with the Diversity Index while avoiding multicollinearity and retaining interpretability. All feature selection steps were executed and can be reviewed in the notebook’s feature engineering section.

Model selection

Context and Objective :

The objective of the model training phase was to build predictive models to estimate the Diversity Index based on the environmental, morphological, and anthropogenic features selected during the feature engineering phase. Several models were evaluated to identify the one that best generalized the relationship between these predictors and the diversity score. The model training process involved tuning, comparing multiple algorithms, and interpreting their performance to select the optimal approach.

Initial model selection:

We began by considering a range of models, including both linear and non-linear techniques, to account for potential complex relationships between the predictor variables and the Diversity Index. The models explored in the Jupyter notebook included:

1. Linear regression: A baseline model to assess how well a simple linear relationship could explain the variance in the diversity index.
2. Random forest regressor: An ensemble learning method that handles non-linear relationships and complex interactions between features. This model was chosen for its flexibility and robustness.
3. XGBoost Regressor: Another ensemble-based method known for its high performance in many regression tasks. It was used to test if its gradient-boosting approach would outperform the Random Forest model.

Hyperparameter tuning :

For each model, hyperparameters were tuned to optimize performance. The tuning process was performed using GridSearchCV, with cross-validation to ensure that the model was not overfitting to the training data.

- Linear regression: No hyperparameter tuning was needed for this model, as it serves as a straightforward benchmark.
- Random forest regressor: We performed grid search on parameters such as n_estimators (number of trees in the forest), max_depth (maximum depth of the tree), and min_samples_split (minimum number of samples required to split an internal node). The optimal values found in the notebook were n_estimators = 100 and max_depth = 15, which balanced performance with computational efficiency.
- XGBoost Regressor: Tuning focused on learning_rate, n_estimators, and max_depth. The best performance was achieved with a learning rate of 0.1, n_estimators = 100, and max_depth = 10.

Model comparison and selection :

The models were compared based on their performance using Mean Squared Error (MSE) and R-squared (R^2) as evaluation metrics. Below are prelimary results we observed:

Linear regression :

- MSE: 0.14
- R^2: 0.35

Random forest regressor :

- MSE: 0.11
- R^2: 0.65

XGBoost Regressor :

- MSE: 0.099
- R^2: 0.64

Outcome :

The XGBoost Regressor was selected as the final model due to its superior performance in predicting the Diversity Index. It effectively handled the complex relationships between variables and achieved the best balance between bias and variance. The code for training and comparing the models can be found in the “Model Training” section of the notebook, where the grid search and cross-validation processes are also detailed.

Model evaluation

Context and objective :

In this final phase of the project, the focus was on evaluating the trained models' performance using appropriate metrics. This evaluation allowed us to assess how well the models generalized to unseen data and determine the best approach for predicting the Diversity Index of coral reef fish species. The chosen metrics were selected based on their ability to reflect both the accuracy and robustness of the models.

Evaluation metrics :

We used two primary metrics to evaluate the performance of our models:

1. Mean squared error (mse): MSE provides an average of the squared differences between predicted and actual values, making it a straightforward measure of prediction accuracy. A lower MSE indicates better model performance.

2. R-squared (r^2): This metric represents the proportion of the variance in the dependent variable (Diversity Index) that is predictable from the independent variables. An R^2 value closer to 1 indicates a strong model that explains a large portion of the variance.

Both of these metrics were computed on the test data, ensuring that the evaluation reflects how the model will perform on new, unseen observations.

Evaluation result :

Linear regression :

- MSE: 0.12
- R^2: 0.45
- While the linear regression model provided a baseline for comparison, its relatively low R^2 score indicated that it could not capture the complex, non-linear relationships between the features and the Diversity Index.

Random forest regressor :

- MSE: 0.08
- R^2: 0.67

- The Random Forest model significantly improved the predictions over linear regression by capturing non-linear relationships and interactions between variables. It performed well in terms of reducing the prediction error (MSE) and explained a higher percentage of the variance in the Diversity Index.

XGBoost regressor :

- MSE: 0.07
- R^2: 0.72
- The XGBoost model performed the best overall. Its gradient-boosting mechanism allowed it to reduce the error further while capturing complex relationships in the data. The model's ability to minimize bias and variance made it the most suitable choice for predicting the Diversity Index.

Model insights and final selection :

From our model evaluation, it became evident that the XGBoost Regressor was the best model for predicting the Diversity Index, given its superior performance in both MAE and R^2 metrics. The Random Forest also provided a robust alternative, but XGBoost's gradient boosting offered a slight edge in capturing the complexity of the coral reef ecosystem data. The simplicity of linear regression, while useful for baseline comparison, was not sufficient to capture the intricate interactions in the dataset.

Methodology conclusion

The methodology from start to finish—from exploratory data analysis, to feature selection, model training, and finally, model evaluation—provided a comprehensive approach to tackling the problem of predicting coral reef fish species diversity. By carefully selecting relevant features such as Coral Cover, Reef Complexity, and Sea Surface Temperature, and by testing multiple models, we were able to choose the most effective approach. The XGBoost model, with an MAE of 0.06 and R^2 of 0.72, stands as the final model due to its ability to generalize well to new data, offering valuable insights for biodiversity conservation in coral reef ecosystems.

Discussion

The machine learning models demonstrated that depth, coral cover and reef complexity are the most significant predictors of fish species diversity. These results align with previous ecological studies, suggesting that maintaining coral cover and promoting structural complexity can help preserve biodiversity in coral reef ecosystems. However, human impact variables such as population density within 20 km of the reef had a lesser influence, possibly due to the limited resolution of the data.

Project conclusion

The project successfully developed a predictive model for coral reef fish species diversity using environmental data. The Random Forest and Gradient Boosting models outperformed linear regression, indicating the importance of using ensemble methods for ecological predictions. Future work could focus on incorporating additional environmental factors, such as water temperature and pollution levels, to further enhance model accuracy.