Python Machine Learning Labs -S24

Project 2: Coral Reef Species Assessment

Group Members: Susmitha Ann Alex - Wafa Bouakez - Afra Muhammad - Marie Winter

1. Summary

Evaluating biodiversity on the basis of a diversity index can be useful for addressing ecological issues, such as the relationship between biodiversity and ecosystem function (Loiseau *et al,*). It appears as a complex concept because it is related to multicomponent (*e.g.,* species traits, environment characteristics, population density). **In this project we aimed to predict the diversity index based on the data from Kochan *et al*, 2023.** To do it we used the data sets generated by the authors of the study. They contain information about the fish's traits (trait_combined_2023),study's site characteristics (SiteEnv, species), and the number of fish observed at each study's site (SpecAbund). The figure 1 illustrates the diversity_index repartition across the study's site.

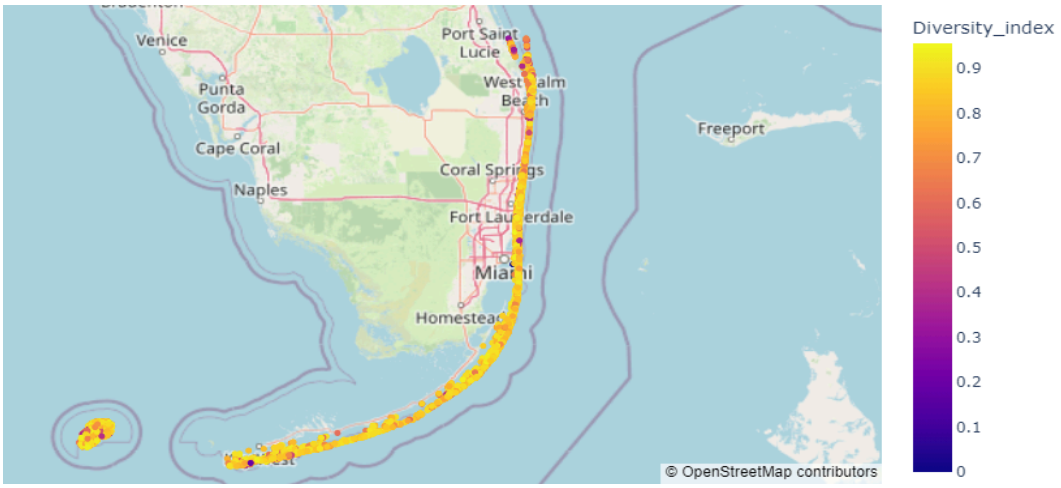

Figure 1. Repartition of the diversity_index values across the study's site. The diversity_index value is represented by the scale at the right of the pictures. Each dot represents the coordinates of the studies' site.

2. Exploratory Data Analysis

The exploratory data analysis (EDA) starts with the species dataset which contains our target the diversity_index. It is a continuous variable indicating that we have to answer a regression problem. Then we explore the dataset trait_combined_2023 because it is recognized in the literature that the fish trait can help to predict the diversity index of a sea environment (Loiseau *et al,*).

To visualize the repartition of the observation inside the categorical feature we choose to use a pie chart allowing to have the proportion of each category of the feature. To look at the relation between the categorical feature and the target we choose to use a violin plot. This type of plot highlights the way the observation can be spread (concentrate around a value or spread along a high range). The dense and larger part shows where the observations are concentrated. The queue shows which side of the median/mean value the

outlier could be. At the end to identify redundant features we choose to use a scatter plot to plot the 2 features we suspect to be similar.

To look at the relationship between the categorical data and the target we used the python library Dython. This library calculates the correlation/strength of association by a ratio of the correlation score (Pearson method between the continuous, the Cramer's method between the categorical, and the ratio of the both between categorical/quantitative). To measure the linear correlation between the numerical features and the target we used Pearson's method. To estimate if the relation between two features or a feature and the target is real we used the random features. In fact the random features have no sense with the other data, in that way the correlation score between it and a feature or the target give us a threshold for accepting or not a linear relation as real.

To classify the features and estimate their importance for predicting the diversity index we choose to use the package Borruta in the R software. We choose to do it with R because this package in python is low developed and causes many trouble with our package version. This algorithm classifies the features importance by comparing them iteratively to random features created by randomizing the observation of the datasets. The results of the analysis are developed in the EDA_report.ipynb file and the figures are in the folder "Result". Briefly, the numerical features as well as the categorical features Habitat_type_classLV2 and Region of the species are all informative for predicting the diversity_index (Fig2A), for the trait features it's more difficult to judge (Fig2B,C).

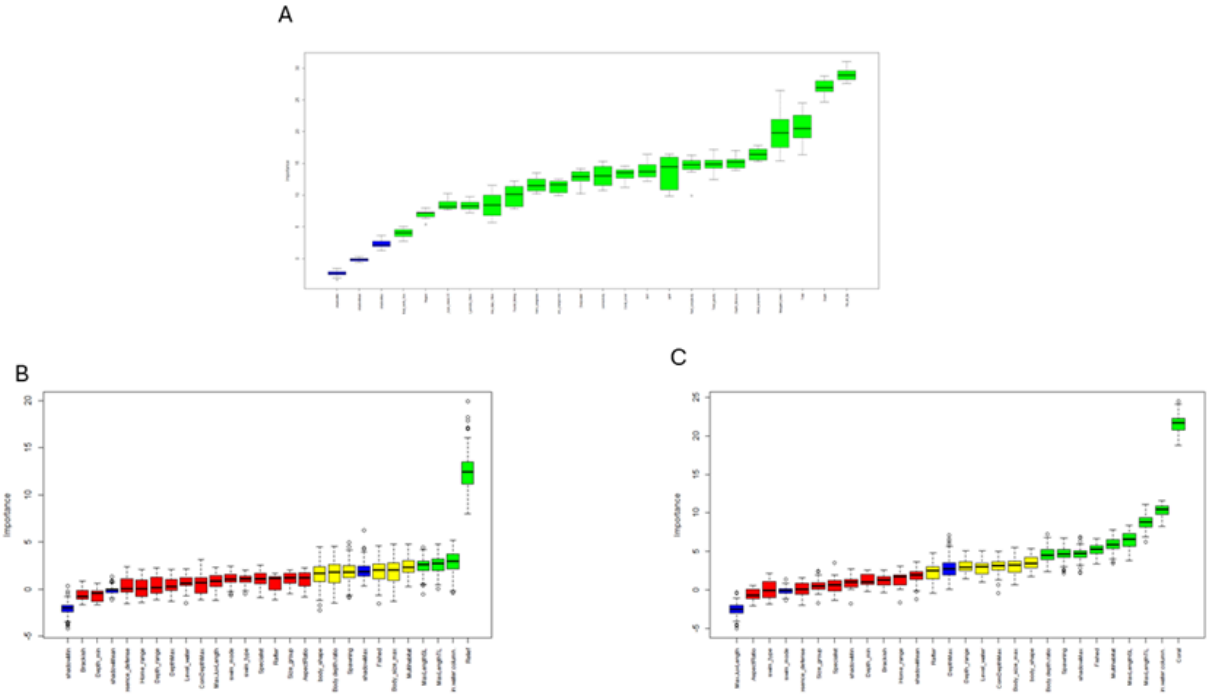

Figure 2.Borruta analysis on the species and traits features. A. Classification of the features from the species table to predict the diversity_index, B. Classification of the features from the traits table to predict the coral target, and C. Classification of the features from the traits table to predict the reef targets. In green the accepted features, in yellow the intermediate meaning the features we have to decide if they have an interest or not, and in red the rejected features. The Shadow min, mean, and max are the Z score of the randomized values derived from the

values of all the features in the table. The features are accepted if their score is higher than the shadows features and intermediate if the features' score are not significatively higher than the shadows. The reject has a score between the mean and the highest shadows scores.

3. Features engineering and modeling

01-Model_1

In this approach, initially the species dataset was used. Basic regression models such as Linear and Polynomial regression (2nd degree polynomial: y = ß0 + ß1x + ß2x2) were built. But the results were disappointing since the dataset did not have species related features.

Moving on, the SpecAbund dataset was used along with the 'Diversity Index' feature merged from the species dataset. The missing values were filled with mean values of the features' based on 3 categorical features: Region, Habitat_type_classLV0 and number of divers. Outlier treatment was performed using the 99th percentile method to handle extreme values. Feature selection was performed using VIF (variance Inflation factor). Then basic regression models were built and they still yielded sub-par results. The model performance was better for ensemble based regressors such as Random forest regressor and XGBooster regressor. Advanced ensemble techniques such as voting and stacking also produced very promising results. Fine tuning also marginally improved the results for the Random forest model.

After this 120 columns were added to the dataset based on on the following logic:
For any species if the value in fished column =1 (in the traits dataset), then the count was divided by 2 in the related species column of SpecAbund (Considering fished=1 as the species being fished, which will negatively impact the fish count) and if the value in fished column =0, then the count was multiplied by 2 in the related species column of SpecAbund (Considering fished=0 as the species being not fished, which will positively impact the fish count). After adding these columns the best performing model from the previous dataset was rebuilt and model performance did not improve.

Summary table of the best models used in this approach (Jupyter Notebook = ML_1):

Model	**R^2**	**MSE**	**RMSE**	**Description**
XGBoost 1 (xgb_model in the Jupyter Notebook)	0.70	0.004	0.063	Standard XGBoost model with Scaled Data
Stacking Regressor 1 (stacking_model5 in the Jupyter Notebook)	**0.76**	0.003	0.058	Built with Linear, Ridge, KNN, Random forest, XGBoost as estimators and Ridge regressor as the final estimator. | Mean cross-validated R2: 0.6851
Stacking Regressor 2 (stacking_model6 in the Jupyter Notebook)	0.74	0.003	0.059	Built with Linear, Ridge, KNN, Random forest (Fine tuned), XGBoost as estimators and Ridge regressor as the final estimator. | Mean cross-validated R2: 0.6787

Conclusion: A regressor built based on the Stacking ensemble technique built with Linear, Ridge, KNN, Random forest, XGBoost as estimators and Ridge regression as the final estimator performed the best inferring that the **Stacking ensemble technique is a viable approach**. Adding feature engineered columns did not work in this approach but might work with different logics.

02-Model_2

Trait-based analysis has emerged as a successful approach for determining species diversity in numerous ecological and taxonomic studies including marine ecosystems, among others. This led us to utilize the trait-combined dataframe in this analysis exploring techniques beyond those typically described in the literature.

-Feature selection: was primarily based on exploratory analysis.

-Handling Missing Values: missing values were filled using the MICE technique (multiple imputations by chained equations :iterative imputer) for numerical values and the mode for categorical values.

-Data Standardization : Datasets were standardized before models training and the scaling was performed whether using the Standard Scaler.

-Prediction models: the specabund dataframe was merged with the diversity index from species dataframe based on the site column.It was then merged with the dataframe containing the Margalef Index , total fish count , total species per site used in previous models , also based on site. Afterwards , a column combining all species per site was created and then exploded into one species per row before finally proceeding to merge with the trait combined dataframe upon the species column.

The result was a large dataset with more than 13.000 rows, after all the pre-processing step (feature selection , dealing with missing values, encoding and scaling the data) , predictions were made using a Random Forest Regressor achieving very good results that were a sign of overfitting.

However, random sampling produced good results with Random Forest Regressor and XGBoost models (R2 of 0.9 and 0.8 respectively),and then checked if the sample was representative of the whole dataset.

-Challenges: Further investigation is needed to determine the representativeness of the sample, as well as to address potential data leakage and overfitting concerns.

-Summary table for Models Predictions :

Model	**R^2**	**MSE**	**RMSE**	**Description**
Random Forest Regressor (sample)	**0.92**	**0.0011**	**0.0325**	-Standard random forest model with scaled data (via Standard Scaler). -No tuning was used (the previously used SMOGN and GridSearch were omitted).

XGBoost Regressor (sample)	**0.82**	**0.0025**	**0.0037**	-Scaled data with Standard Scaler. -No hyperparameter tuning.

03-Model_3

In this approach, the initial dataset used was the “species.csv” file. After exploring the data, we began by performing essential preprocessing steps, including handling missing values by imputing the mean, applying one-hot encoding for categorical variables, and removing irrelevant features. Outlier treatment was conducted using appropriate statistical techniques (99th percentile method) to improve data quality, and Min-Max Scaling was used to normalize the feature values.

Next, basic regression models such as Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor were built. However, the initial performance of these models was suboptimal, as the dataset lacked certain important features. To enhance the predictive capability of the models, “data augmentation” was performed by merging additional datasets, specifically the “species.csv” and “”specabund” datasets. This augmentation introduced more relevant features, which were then followed by feature scaling using Standard Scaler.

Post-augmentation, the performance of the models improved significantly. Further improvements were achieved by tuning the Gradient Boosting Regressor using RandomizedSearchCV. The tuning process optimized the model’s parameters and led to a moderate increase in accuracy. Gradient Boosting model and Random forest model showing notable improvement over the initial models.

The combination of data augmentation, feature engineering, and hyperparameter tuning demonstrated a substantial improvement in model performance, particularly for ensemble methods like Random Forest and Gradient Boosting. This process highlighted the importance of a comprehensive approach to data preparation and model refinement in achieving better predictive outcomes.

Model	**R^2**	**MSE**	**RMSE**	**Description**
Linear Regression (Baseline)	0.15	0.0132	0.115	/
Linear Regression Optimized	0.46	0.0073	0.0859	Data augmentation and feature scaling (Standard Scalar) and Outlier treatment
Random Forest	0.16	0.0131	0.1143	Feature scaling and transformation (Minmax Scalar), Outliers treatment

Random Forest Optimized	0.69	0.0043	0.0656	Data augmentation and feature scaling (Standard Scalar) and Outlier treatment
Gradient Boosting Regressor	0.16	0.0131	0.1146	Feature scaling and transformation (Minmax Scalar), Outliers treatment
Gradient Boosting Regressor Optimized	0.67	0.0044	0.067	Data augmentation and feature scaling (Standard Scalar), Outlier treatment and model tuning (RandomizedSearchCV)

Conclusion:Based on these results, the Random Forest Optimized model appears to be the most promising and Gradient Boosting regressor optimized also performed well. The model's performance was further analyzed by plotting various graphs, such as Actual vs. Predicted Values and the Histogram of Residuals. Commentary on these plots can be found in the notebook, providing deeper insights into the model's accuracy and error distribution.

04-Model_4:

To fill the missing values identified during the EDA we choose to create a function allowing us to get the mean of the features' values in the function of 2 categorical features: Region and Habitat_type_classLV2. The same strategy was chosen in the EDA for allowing to do a Borruta analysis. To integrate the information in the SpecAbund table, we calculate the total number of fish observed at each site (Total), the number of different species observed at each site (Nb_dif_Sp) by excel. After multiple tries with python it was most easy for us to do it with excel. Those features allow us to create a third one: the Margalef's index knowing to describe the fish diversity at a site (Bonjoru R. *et al,*). With excel it also creates a list with all the fish observed at each site with their number baser on the SpecAbund table. The 3 features described above are added to the species table, then with the list of fish per site, the species and trait table are linked together. The EDA highlights that the numerical data shows a heavy tail at the right or left or the median, we used a power transform strategy to scale them. For the categorical data we choose to use a one_hot encoding strategy because the categories haven't order and/or importance in the table. After these feature engineering we test here 4 types of regression models: linear regression, ridge regression, random forest and XGBoost. Summary table of the models used in this approach to predict the diversity_index:

Model	**R^2**	**MSE**	**RMSE**	**Description**
Linear Regression	0.32	0.008	0.09	/
Ridge (1st model)	0.33	0.008	0.09	Tuning using Grid search and cross validation
Random forest (1st model)	0.40	0.007	0.08	Tuning using Grid search and cross validation

XGBoost (1st model)	**0.48**	0.006	0.08	Tuning using Grid search and cross validation

To conclude, the XGBoost gave the highest performance with R^2 equal to 0.48 and the lowest MSE (0.006). This performance is in accordance with the previous study in the same field (Boll et *al.*, Cao et *al.*). In accordance with the literature, the presence of trait features in that context help to predict the diversity index. In biology and life science often the phenomenon isn't fixed and depends on multiple factors which can be limited to represent in a table and the information needs to be split in several tables. It was one of the difficulties of this project to group the information from different tables and find links between the different types of features. Maybe, it could be helped to use techniques such as data mining and graph databases to find hidden links between data, improve the features selection, and have more flexibility for linking the information together.

4. Bibliography

Boll *et al,* Fish size structure as an indicator of fish diversity: a study of 40 lakes in Türkiye, water, 2023.

Cao *et al,* Weighting effective number of species measures by abundance weakens detection of diversity responses, Journal of Applied Ecology, 2018.

Loiseau *et al*, Indices for assessing coral reef fish biodiversity: the need for a change in habits, Ecology and Evolution 2015; 5(18):4018-4027.

Kochan D.P. *et al,* Winners and losers of reef flattening: an assessment of coral reef fish species and traits, Oikos, 2023:e10011.

Bonjoru R. *et al,* Diversity and abundance of fish species in some selected riverine wetlands of upper benue river basin, Nigeria, IOSR Journal of environmental Science, Toxicology and Food Technology, 2019, IOSR Journal of Environmental Science, Toxicology and Food Technology (IOSR-JESTFT) , DOI: 10.9790/2402-1308021418.

5. Packages

Borutta package: https://cran.r-project.org/web/packages/Boruta/Boruta.pdf

Dython python package: https://shakedzy.xyz/dython/