School of Engineering

Applied MSc in Data Analytics
Applied MSc in Data Science & Artificial Intelligence
Applied MSc in Data Engineering & Artificial Intelligence

Project: **Coral Reef Fish Species Assessment**

Instructor: **Pauline Salis**

Project Summary:

The dataset includes density data for 109 fish species included in the analysis, environmental data used for the density analysis, and a trait table which includes the response variables of association with relief and coral cover.

Below is the available information regarding the dataset attributes:

1) **site:** numeric site descriptor matching NOAA Reef Visual Census sites
2) **model:** factor used to subset data for two separate models
3) **Year:** year of RVC fish survey
4) **Month:** month of RVC fish survey
5) **Latitude:** latitude of RVC fish survey
6) **Longitude:** longitude of RVC fish survey
7) **Depth:** depth of RVC fish survey averaged for each surveyor
8) **Region:** jurisdiction of RVC fish survey
9) **Coral_cover:** percentage of benthos made up of living hard coral visually estimated by RVC surveyors
10) **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
11) **SST:** minimum monthly average sea surface temperature in Celsius derived from CoRTAD database from 2012-2016
12) **NPP:** net primary productivity derived from remotely sensed chlorophyll-a from the OSU VGPM model
13) **Wave_exposure:** exposure calculated using linear wave theory

14) **Habitat_type_classLV0:** habitat classification of each site according to the FWC Unified Reef Map level 0
15) **Habitat_type_classLV2:** habitat classification of each site according to the FWC Unified Reef Map level 2
16) **Coral_area_UFRTM_20km:** area classified as reef by Unified Reef Map level 0 within 20 km of each site
17) **Coral_area_UFRTM_200km:** area classified as reef by Unified Reef Map level 0 within 200 km of each site
18) **Depth_Sbrocco:** remotely sensed depth of survey sites
19) **Deepwater:** euclidean distance in meters over water to the 30-meter bathymetric line
20) **FSA:** euclidean distance in meters over water to the nearest fish spawning aggregation site
21) **Marina_slips_10km:** number of marina slips over 45 ft within 10 km of each site
22) **Marina_slips_25km:** number of marina slips over 45 ft within 25 km of each site
23) **Marine_reserve:** protected status of site; whether fishing was allowed or not
24) **Population_20km:** human population living within 20 km of reef sites derived from LandScan dataset
25) **Population_50km:** human population living within 50 km of reef sites derived from LandScan dataset
26) **Recreational_fishermen_50km:** number of recreational fishing licenses within 50 km of reef sites derived by zip code
27) **Tourist_fishing:** statistics from Johns et al. 2001 and publicly available dataset of hotel units in Florida
28) **Artificial_reefs_1km:** number of artificial reefs within 1 km
29) **SG_permits_50km:** number of commercial snapper-grouper fishing permits within 50 km
30) **SG_charter_permits_25km:** SG_permits_50km: number of recreational snapper-grouper fishing permits within 25 km
31) **Total_gravity_intercept:** number of people in population centers within 500 km divided by the square of travel time
32) **Total_gravity:** number of people in population centers within 500 km divided by the square of travel time
33) **Keys_Divisions:** sub-jurisdictions of Florida Keys including Upper, Middle, Lower Keys and Marquesas; NAs for non Florida Keys sites
34) **FKNMS:** Florida Keys National Marine Sanctuary sites; NAs for non Florida Keys sites
35) **DryTortugas:** Dry Tortugas sites; NAs for non-Dry Tortugas sites
36) **BNP:** Biscayne National Park sites; NAs for non-BNP sites
37) **CoralECA:** Coral Ecological Conservation Area sites; NAs for non-ECA sites (also known as SEFCRI)
38) **Nursery_seagrass:** connectivity of reef sites to continuous seagrass patches within 10 km
39) **Nursery_mangroves:** connectivity of reef sites to mangrove stands within 12 km
40) **connectivity:** number of larva from upstream modeled to a connectivity matrix
41) **Comm_engagement:** metrics of commercial engagement based on landings and permits provided by NOAA

42) **Comm_reliance:** metrics of commercial engagement based on landings and permits relative to size of fishing community provided by NOAA
43) **Rec_engagement:** metrics of recreational engagement based on landings and permits provided by NOAA
44) **Rec_reliance:** metrics of recreational engagement based on landings and permits relative to size of fishing community provided by NOAA
45) **Commercial_pounds_landed:** annual number of pounds of fish reported by commercial anglers
46) **Pop_per_area_reef_20km:** human population divided by area of reef within 20km
47) **Random:** random number assigned to each column
48) **impact:** impact: fishing impact variable
49) **YEAR:** year of RVC surveys
50) **HABITAT_CD:** habitat code used by NOAA RVC surveys to stratify sites
51) **REGION:** jurisdiction of RVC survey sites
52) **PCT_CORAL:** percent coral cover
53) **MAX_HARD_RELIEF:** maximum hard relief
54) **no.divers:** number of divers for RVC survey
55) **Diversity_index:** score that varies between 0 and 1. A high score indicates high diversity, and a low score indicates low diversity. When the diversity index is zero, the community contains only one species (i.e., no diversity). As the number of different species increases and the population distribution of species becomes more even, the diversity index increases and approaches one.

Project Objectives:

Using the provided dataset, you are asked to train a model that predicts the diversity index of fish species based on different factors (morphological, environmental, etc.). The project can be submitted as a Jupyter Notebook and should include exploratory analysis of the data, feature engineering and selection, model training and evaluation.
You may use additional resources as you see fit (provided you can justify how they can serve your solution). You can even consult similar solutions from the Internet. **However, this comes with a big responsibility: any submission that is over-plagiarized or does not reflect personal work will not be accepted**.

Project Evaluation:

The project will be evaluated using the following rubric. It contains the required items for a complete submission as well as bonus elements. The grading system is over 5 and the final grade will be transformed to a grade over 100.
● Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes) **[1 point]**
● Feature selection (feature engineering, feature pruning, choice justification) **[1 point]**
● Model training (motivation for selected model, comparison of different models) **[1 point]**

● Model evaluation (evaluation metric, results interpretation) **[1 point]**
● Project report (short report explaining the approach and results) **[1 point]**
● **BONUS:** Project reproducibility (requirements file with necessary packages, README file for running the project) **[1/2 point]**
● **BONUS:** Project hosting (Github, Docker, AWS, Heroku or any other method) **[1/2 point]**