BT3041 – Analysis and Interpretation of Biological Data

Assignment 2

You are given a dataset of Fossils with 8 features (uranium lead ratio, carbon 14 ratio, radioactive decay series, stratigraphic layer depth, inclusion of other fossils, isotopic composition, fossil size and fossil weight). Using these features, you are supposed to build a regression model to predict the age of the fossil.

1. Split the dataset into test and train data using at least 2 different splitting ratios. Use the train data to build regression models with and without regularization (choose one between L1 or L2) to predict the fossil age.

   (20 marks)
2. Using the developed regression models perform more than one cross-validation method to tune your lambda parameter in regularization and report your results.

   (15 marks)
3. Using different model metrics comment on the quality of models generated (with and without regularization) and suggest the best model.

   (10 marks)
4. Present the actual fossil age to the predicted age of your best model.

   (5 marks)

Instructions: This is a hybrid assignment with both coding and Interpretation components. Thus, you must submit the model implementation code files with a detailed report on your interpretation of results obtained. Discussion with fellow mates is allowed provided you acknowledge your peer. Any similarity in the report will be treated as plagiarism.