Project Proposal

Project Title : Oragnism Prediction based on Codon Usage

Dataset Information:

Dataset Characteristics	Multivariate
Subject Area	Biology
Associated Tasks	Classification, Clustering
Feature Type	Real
Number of Instances	13028
Number of Features	69
No of missing values	69

Dataset: **https://archive.ics.uci.edu/dataset/577/codon+usage**

Quick Summary of the Project:

There are a total of 65 numerical and 4 categorical attributes in the dataset. In total there are 63 dependent features and 1 independent feature. In total, there are 13,028 instances for the dataset whereby 80% (10,352) belongs to organisms classification kingdom bacteria, vertebrate, plants and virus and (2676) belongs to other organism classification kingdom types.

The project focuses on the DNA codon usage patterns of a large number of taxa of biological organisms and establishes the relationships of the taxa to one another through the DNA sequencing techniques. The dataset includes a number of multivariate attributes features associated with the kingdom, DNA, species id, codon frequencies and species names. Using this dataset since the targeting column in the case is the Prison Kingdom, this column has values of various species from 'arc' (archaea), 'bct' (bateria), 'phg' (bacteriophage), 'plm' (plasmid), 'pln' (plants), 'inv' (invertebrate), 'vrt' (vertebrate), to 'mam' (mammal), ‘rod’ (rodent), ‘pri’ (primate) to ‘vrl’ (virus). The objective of the study is to build a classification model that is capable of distinguishing between different kingdoms of an organism using its codon usage. This methodology prototypically enables the reconstruction of codon usage bias systematics and function across a taxon virtually in the genomic level correlating evolution with function.

Project Pipeline:

To build a machine learning project pipeline based on the dataset provided, the team will follow these steps:

Data Collection: Collected the data for our task from the UCI Machine Learning Re pository *(We would use **Codon Usage** from the **UCI Machine Learning** repository for our task/project/problem)*. Click here to download the data.

Data Pre-processing: Impute missing data; apply codon frequencies to a common scale, and incorporate nominal variables such as kingdom in DNA type etc.

Feature Selection: Use a correlation plot to determine factors that are most important in explaining dependent variables.

Model Selection: Compare different models of machine learning to select appropriate ones for the task of classifying objects.

Model Training: Apply the obtained models on the prepared dataset, dividing it into training and test parts.

Model Evaluation: Assess the classification performance of the model with such metrics as accuracy, precision, lost recall, and the F 1 score.

Results Interpretation: Analyze the results obtained in order to evaluate their biological meaning in relation to the patterns and relationships established between various organisms.

Following these stages, our goal is to build a machine-learning/deep learning pipeline appropriate for classifying the kingdom of an organism based on the analysis of alterations of the codon usage patterns.

Tools, Libraries, and Frameworks:

- Python libraries: NumPy, Pandas for Data Preprocessing
- Scikit-learn for model building and evaluation
- Correlation based feature selection
- Matplotlib, Seaborn for Visualization
- Machine learning frameworks: TensorFlow, PyTorch, Keras
- Jupyter Notebook for code development and analysis