Report

Project Problem Statement:

Can I improve on the accuracy score of 0.996 by employing an ensemble (stacked) model?

Background:

Protein synthesis is an ever important part of the Biotechnology industry, especially for drug discovery. So it is important to synthesise proteins that fold correctly to ensure, for example, inhibitor affinity to its target. This is essential to discovering and making new therapies to manage and treat various diseases.

Many scientists will use species of bacteria to synthesise proteins for experimentation. It is useful to understand the codon bias of the species to ensure that the plasmids transferred into the bacteria contain the sequence that is faithful to that of the host's translation and protein synthesis machinery, this ensures better folding and higher efficiency.

So exploration of this area is essential to try and create tools to make some of the steps easier in research.

Dataset:

The data can be found on the UCI machine learning repository on the following link:

https://archive.ics.uci.edu/ml/datasets/Codon+usage

This dataset was curated by a team of researchers who produced a paper that can be found at the following link:

https://www.biorxiv.org/content/10.1101/2020.10.26.356295v1.full.pdf

The dataset is downloaded in csv format.

The data contains codon frequencies, including Kingdom of species, name of the species, and DNA type. DNA type is an incorrect naming as the codons are RNA codons, however I kept the DNA type naming for continuity.

Data Cleaning and EDA:

The data is downloaded in csv format. The SpeciesName column contains a few rows which have added commas. This created an issue during type changing, which was resolved by finding the rows which contained the errors, fixing the error in the raw csv file and renaming the fixed file.

This resolved the issue.

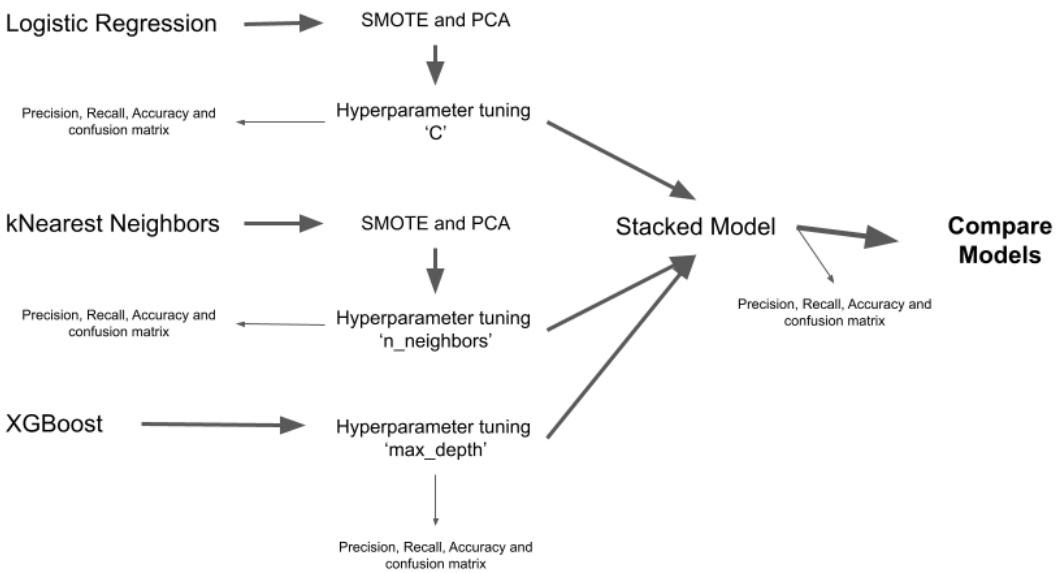
The dataset did contain a few null values which were easily replaced using the correct values which can be found using the CUTG database.

The dataset contained no duplicates.

In the Kingdom column, some of the kingdom classes were collapsed together as they made more sense as a group than as individual classes, one of the classes was discarded due to having too low value count.

In the DNAtype column, type 3 to 12 were discarded due to having very low value counts, only classes 0, 1, and 2 were kept for the classification.

Modelling outline:



Models were evaluated using the precision, recall and F1 score, including AUC score, and accuracy scores.

Models were selected for the Stacked model depending on the accuracy score.

Models were then compared by the evaluation metrics.

Conclusion:

The aim of using PCA to achieve an accuracy higher than 0.996 was not achieved. However, the Kingdom classification on 7 classes was achieved with an accuracy score of 0.95 (stacked model) and an AUC score of 1.0, so these classification models (stacked model and KNN model) can be used to classify in a more targeted way.