Course: CSE 847 - Machine Learning
Instructor: Jiayu Zhou
Students: Hao Yuan, Height Yan, and Raghav Gaur

Genome Based Characteristic Classification and Feature Importance Ranking

I. Introduction and Problem Description

The past decade has seen Machine Learning (ML) systems find a wider range of applications than any time preceding. The power of ML systems lay in their ability to recognize patterns (referred to as ‘classes’) and manipulate the inter–class variation to appropriately classify these patterns. Such pattern recognition and classification can be used for trivial tasks such as simple image classification to more ambitious tasks such as species classification.. It was this latter task which drew our group’s interest.

In the time between the initial proposal and this intermediate report, our focus had shifted away from the classification of different types of hominids (and the report of their most distinct set of traits) toward an intra–species (salmo salar, or more simply, salmon) classifier. That is, we have moved toward building a classifier which studies salmon gene–sequences and determines their origin – lakes (freshwater) or seas/oceans (saltwater).

This was done for practical purposes, human data for individuals with genetic disorders is, by its nature, not made public. Whatever data was made public was insufficient for the scope of this project. Hence, we were made to change the species being studied in order to build a more robust and nuanced ML system.

For salmon particularly, this is a deceptively complicated task, as even the fishermen who catch such fish are unable to determine the difference by sight alone. Rather, we must probe into the genetics of the fish themselves and determine a decision boundary that will classify the salmon. With this we can restate our problem as:

“Given a (salmon) gene sequence, can we determine if the individual fish originated in fresh-water or saltwater?”

Ultimately, this is *still* a comparative genomics problem. The input data is a pattern of SNP markers. We will also try to find potential mutations related to local environment adaptation.

II. Description of Used Data Sets

To test the power of machine learning in species/population delimitation and detecting adaptive mutation, we collected two set of SNP data from Bourret et al. and Milano et al. Bourret et al. sampled 1295 wild Atlantic salmon (Salmo Salar) from Europe and North America, which include 1063 anadormous (fish that, while born and spawned in freshwater, spend most of their lives in saltwater) and 232 fresh–water individuals. 6176 SNP markers were selected for analysis.

We will test the power of models in delineating individuals occupying different niches, and find potential mutations resulting in the adaptation. For data from Milano et al., 850 European hakes from Atlantic and Mediterranean populations were sampled. 381 SNPs will be fed into models. We will try to train models to discern individuals from two populations, and also find loci related to divergence. For both datasets, biallelic genotypic data was one-hot encoded. To avoid detecting structural variants specific to subpopulations, multiple subpopulations were included in either class.

III. What Has Been Done So Far

Working on the project we have determined some pros and cons to our approach:

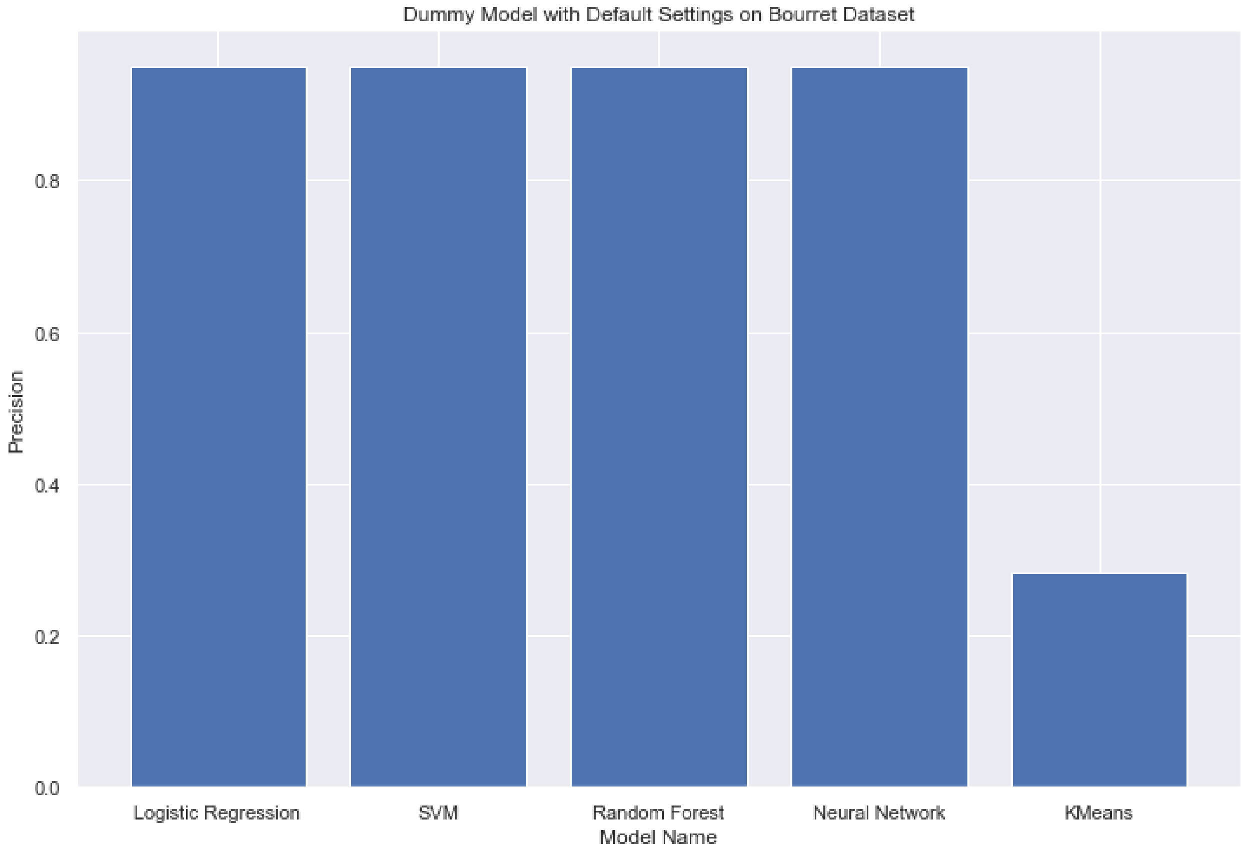
Pros:

- Easy to discern the origin of individuals after building a concrete model, no need to rerun the entire analysis.
- Our method is applicable to the whole genome, including non-coding regions.
- An efficient paradigm to similar questions. Other than SNPs, various types of input data are suitable to the model, e.g. k-mer, microsatellite, expression data, ATAC peak, etc.

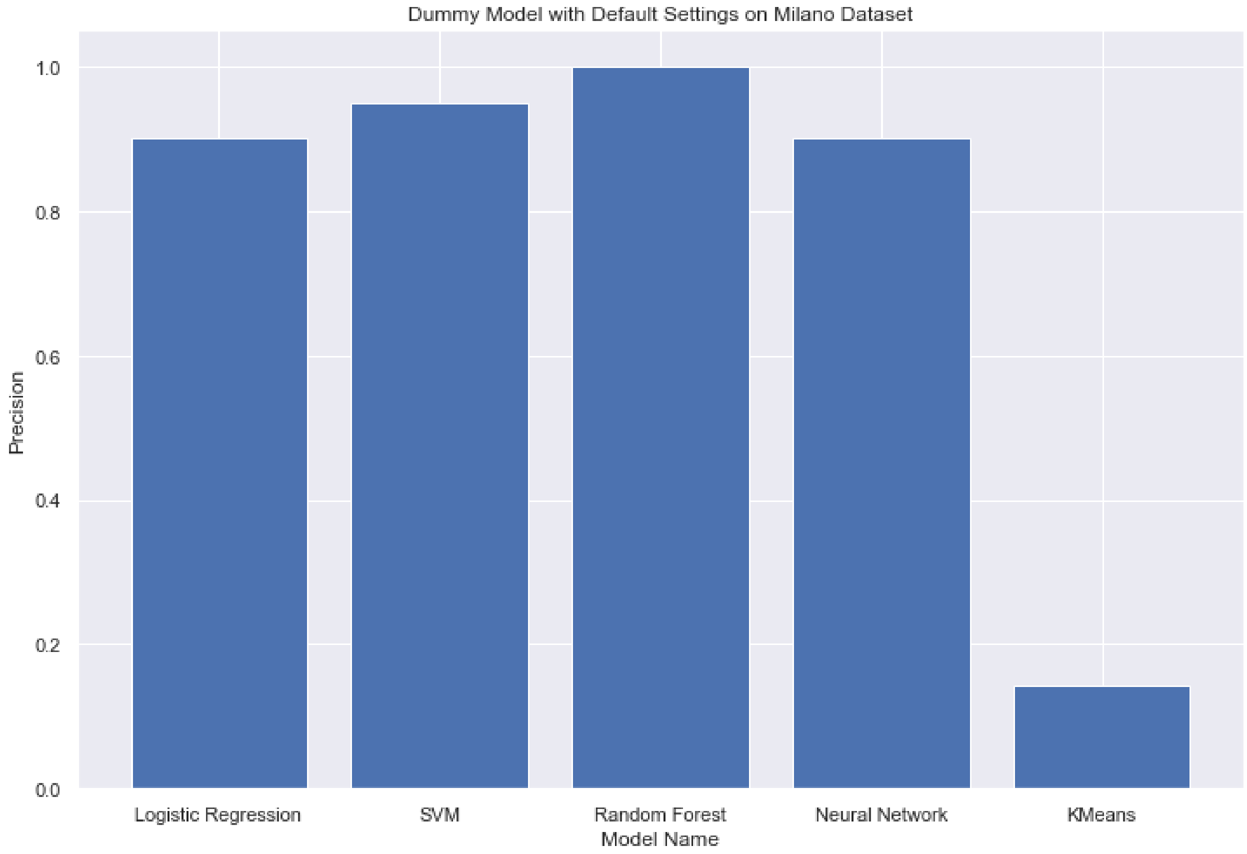
Cons:

- Ideally, modeling requires a substantial amount of data.
- There are always more features than instances.
- Sample numbers from different classes are not always balanced.

We conducted preliminary experiments on the datasets, including data preprocessing, model training and evaluating, and hyperparameter tuning. The average performances (10-fold validation) of the models with their default settings are as follows:

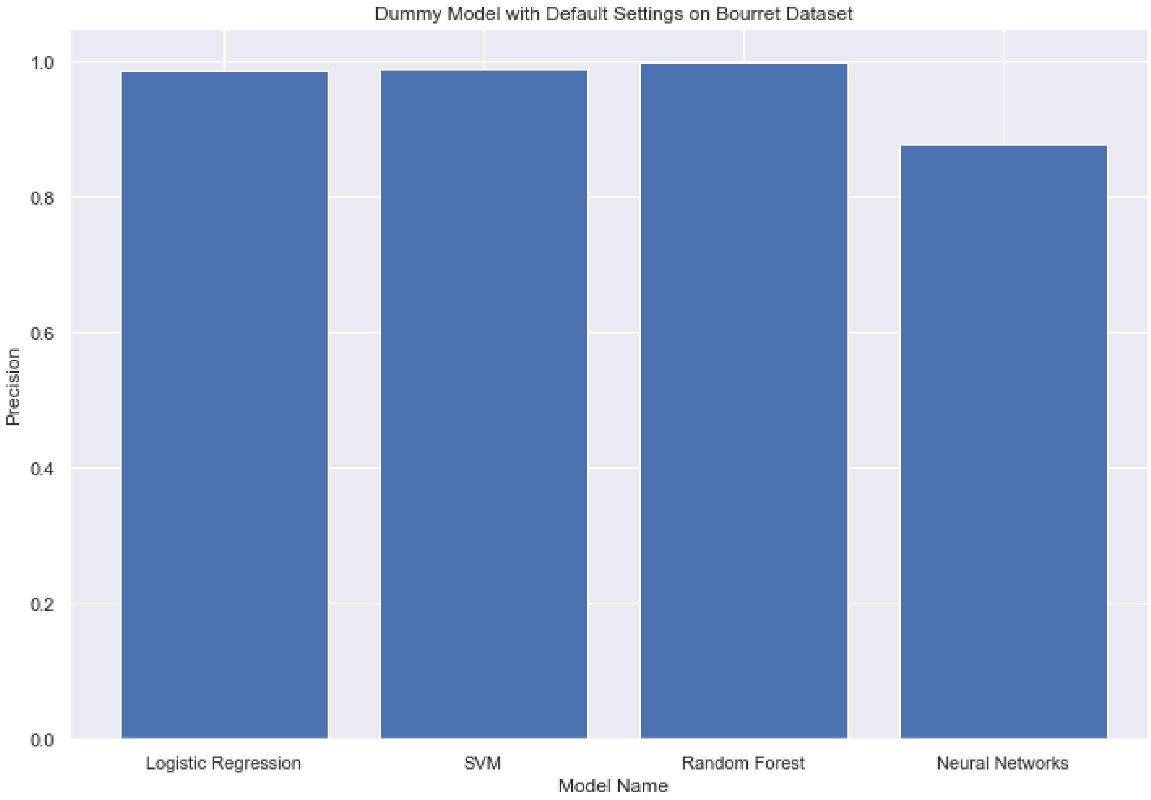


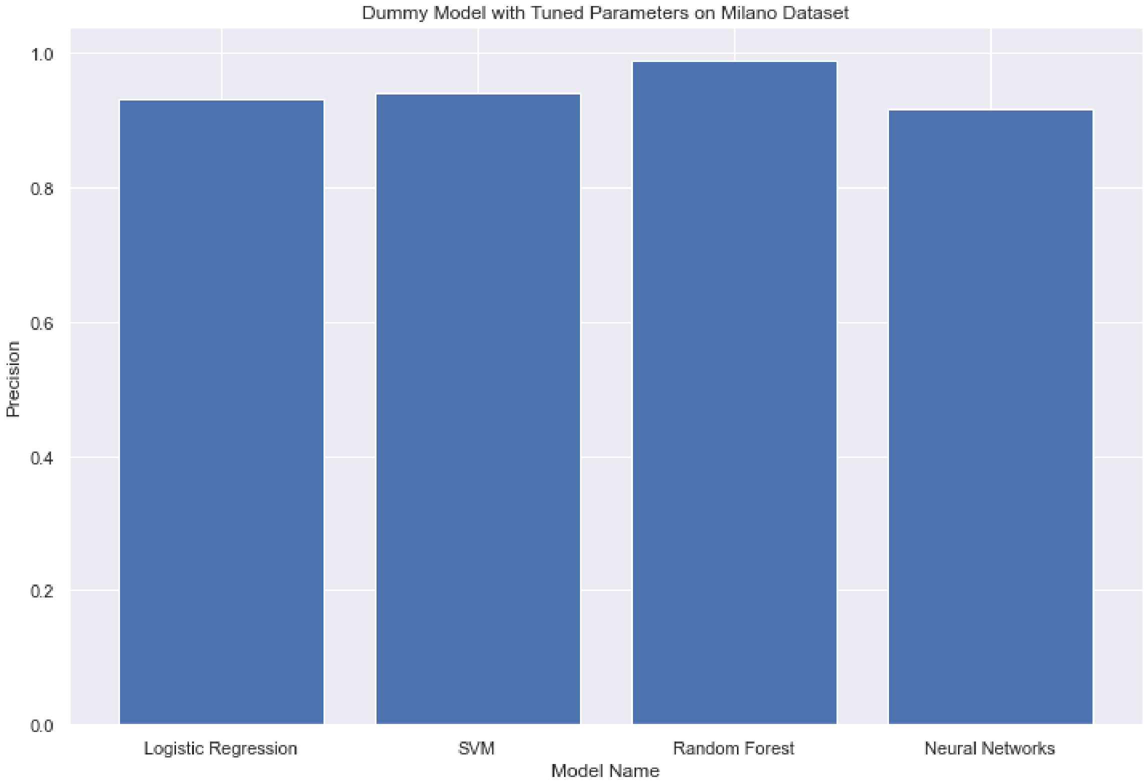
Dummy Model with Default Settings on Bourret Dataset



Dummy Model with Default Settings on Milano Dataset

After tuned the TOP models on each dataset, we got the following results:





Our result conveys that traditional machine learning classifiers can already achieve high performance on classifying the fish species. Tuned Random Forest in both datasets have best performances. (For details please refer to “Data Exploration.html” attached with this report).

IV. What Remains to be Done

- Our data are relatively small, in which they might not provide sufficient information to train a well-generalized model. Therefore, we may have to find more datasets.
- In case of insufficient data, we may consider implementing the techniques of Active Learning to rank the features by predict_proba.
- Our preliminary experiments show that Random Forest has the best performance. We would like to find out why.