Classification of Family Types of Proteins

Name:	**Anshuman Dangwal**
Registration No./Roll No.:	21044
Institute/University Name:	IISER Bhopal
Program/Stream:	Data Science and Engineering
Problem Release date:	August 17, 2023
Date of Submission:	September 17, 2023

1 Introduction

The aim of this project is to identify potential features of protein structures and develop a framework to classify proteins to their family types based on those features. In the given dataset, we have 125452 training instances and 13 features(protein structure information) which are classified to 2160 classes(protein families) out of which, 2 class labels 'TRANSFERASE' and 'OXIDOREDUCTASE' are much more frequent than other with 14405 and 11326 instances respectively. The dataset also has lots of missing values (explained my dealings with them in the next section).

2 Methods

Our problem is a classification problem as our target variable is discrete, hence we would be using classification Machine Learning algorithms. Before starting with the classification techniques, I plan on selecting the salient features of the dataset using various feature selection methods like **Information gain**, **Chi-Square test and low-variance thresholding [1]**. Low-variance thresholding will eliminate features below a particular variance. I would be determining the optimum threshold value by trial and error.I wish to impute the missing values with simple statistics like **mean, median or mode**. For classification, I would be using machine learning model from the Scikitlearn python library. The models are **k-Nearest Neighbours(kNN), Decision Trees, Random Forest Classifier and Naive Bayes Classifier** and would select the best performing framework. If possible I would also like to try some other classification methods like **tweak on kNN [2], Support Vector Machine(SVM)** and some others as I learn. For measuring the performance of the model, the evaluation parameters I would be using are **accuracy, precision, recall, F-score, confusion matrix and logarithmic-loss** .

References

[1] sklearn.feature$_s$*election.variancethreshold. https* : *//scikit* − *learn.org/stable/modules/generated/sklearn.feature*$_s$*election.VarianceThreshold.html.*

[2] Tanmay Basu, CA Murthy, and Himadri Chakrabarty. A tweak on k-nearest neighbor decision rule. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2012.