Enzymes classification using kNN, SVC and Decision Tree algorithms

Amaury

Abstract

In this research is used a dataset with information about proteins that I called "rcsb_pdb_custom_report_enzymes”. We used Machine Learning classification method (algorithms KNN,SVC and Decision Tree) to predict the type of enzyme into six major knowns groups (classes). First, we investigated how well is the performance of this algorithms using only the feature Amino Acid Sequence and second how well can the performance be improved adding features like temperature and pH (Hydrogen Ion Concentration) for the prediction. We found an overall acceptable performance (Accuracy, Precision, Recall,F1-Score) for kNN and moderate performance for SVC and Decision tree. The performance did not improve when added pH and Temperature. The algorithms used performed better in the enzyme classes Hydrolase, Oxidoreductase and Transferase than the classes Lyase, Isomerase and Ligase.

Motivation

Researcher are continually discovering new proteins and obtaining the Amino Acid sequences they are composed. Enzymes are a type of protein; classified them into known groups help to understand better its characteristics, functions and possible applications. This is particularly important in industries like drug pharmaceuticals and food/drinks processing.

Machine Learning classification algorithms could help them to find a method for classified these enzymes using the Amino Acid sequence obtained in the laboratory to save cost and time in comparison with used traditional laboratory methods. This is my main motivation por this project, that could be used as a first approach and as a base for further researches.

My second motivation is to apply the learned Machine Learning knowledges in a real-life project.

Dataset(s)

I customized a dataset by myself creating a custom report using the Advanced Search Query Builder of the RCSB PDB (Protein Data Bank). This search generated .csv files that were concatenated in a single data set using Python language. I called this data set " rcsb_pdb_custom_report_enzymes".

The data set has 332556 rows and 22 columns. Contain information about different features for Proteins, ARN and DNA. The most important features are:

Entry ID= "Entry Identifier for the container".

pH= "The pH at which the crystal was grown".

Temp(K)="The temperature in Kelvins at which the crystal was grown".

Entity Polymer Type= " A coarse-grained polymer entity type".

Structure Keywords= "Terms characterizing the macromolecular structure." (protein/enzyme type)

Sequence= Amino Acid sequence , Chain Length= "The monomer length of the sample sequence".

Molecular Weight= " Formula mass(Kda) of the entity.

Data Preparation and Cleaning

- Many .csv files of data acquired so we concatenated them into 1 dataframe with 22 columns.
- Much of unneeded data so we created a new dataframe with the features ['Entry ID', 'pH', 'Temp (K)','Entity Polymer Type', 'Stucture Keywords', 'Sequence', 'Chain Length']
- Eliminate 222.053 rows with NaN values. Result a data frame of 110.503 rows
- Applied method str.upper() to ensure columns with string data were all in uppercase.
- Deleted rows with duplicate values in columns 'Sequence' and 'Structure Keywords'. Result a final data frame with 51.624 rows.
- Created 20 new columns, one for each type of Amino acid ('A','R','N','D','C','Q','E','G','H','I','L','K','M','F','P','S','T','W','Y','V'). For each row is calculated the frequency of appearance of each Amino Acid (# of appearance of letter A,R,N,etc/length of the sequence) present in the Amino Acid Sequence (column 'Sequence') and put the result in each of these new columns.

Data Preparation and Cleaning (Continue)

- Chose column ‘Stucture Keyword’ as a target for our predictions and create new column with values resulted of apply LabelEncoder()(from Scikit-Learn library) to the feature ‘Stucture Keyword’. The purpose of LabelEncoder() is convert categorical data (type of enzyme) into numbers (six in this case) that the models can understand better. This will be our Target feature. Stored target in variable “y”.

- Created data frame (X), for option 1, with features selected as inputs for answer research question 1. Resulted 20 features corresponding to the Amino Acid composition ['A','R','N','D','C','Q','E','G','H','I','L','K','M','F','P','S','T','W','Y','V’].

- For answer research question 2, it was created data frame(X), for Option 2, we add the features [‘pH’, ‘Temp(K)’] to the features of option 1. Total features = 22.

- Split data into Training set (67%) and Test Set(33%) with the purpose of have different data for training and testing our models.

- Scaling data with StandardScaler() technique to set the features in the same scale.

- Applied SMOTE (Synthetic Minority Oversampling Technique) method to balance our training data set due that the classes Lyase, Isomerase and Ligase had small quantities of samples.

Research Question(s)

Enzymes are a type of protein. Based on the dataset, how well is the performance on KNN, SVM and Decision Tree machine learning classification algorithms for predicting enzymes type into six major groups (Hydrolase, Oxidoreductase, Transferase, Lyase, Isomerase and Ligase) using only the amino acid sequence feature?

How well can the performance be improved adding features like temperature and pH(Hydrogen Ion Concentration) for the prediction?

Methods

We used Machine Learning classification methods imported from Scikit-Learn library that are appropriated to predict categorical variables (class label) like the type of enzymes (Hidrolase, Oxidoreductase, Transferase, Lyase, Isomerase and Ligase). We used Standard Scaling technique to set the features in the same scale. It was used SMOTE method to balance our training data set due that the classes Lyase, Isomerase and Ligase had small quantities of samples.

The three-classification methods are:

1.Decision Tree Classifier: Classify samples according to its closed neighbors. **Param**: Max Depth= None (Nodes expanded until leaves have minimum samples or leaves are pure), criterion: ‘gini’ for Gini impurity, max_leaf_nodes=3000.

2.kNN (k Nearest Neighbors): Choose according to multiple classification paths**. Param:** n_neighborsk=3, metric= ‘minkoswski’ and p=2 that is equivalent to Eucledian distance, algorithm=auto (decide better among methods: ‘Ball_Tree’, ‘kd_tree’ and ‘Brute’.

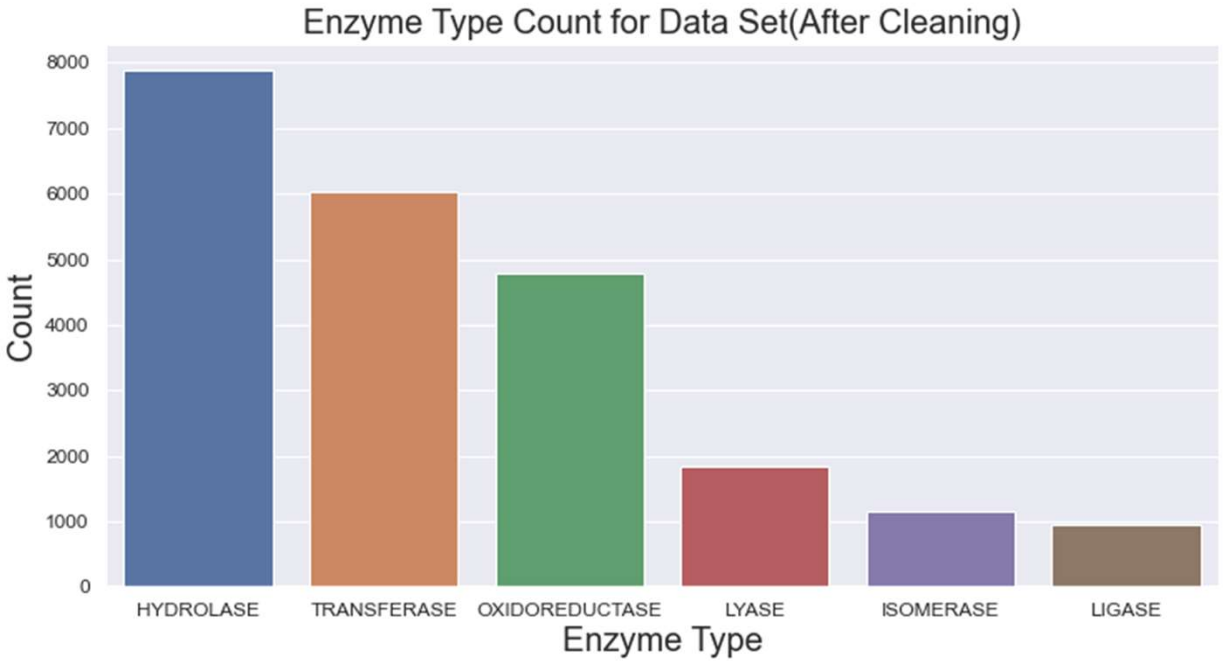
3.SVC(Support Vector Classification): Choose class creating hyperplanes to separate data**. Param:** kernel:’rbf’(radial basis function) used for non-linear problems, gamma='scale’ (it use 1/(n_features*X.var()) as value of gamma)

Methods (Continue)

For all three algorithms, first we follow this process for data frame Option 1 and then for data frame option 2 as the input variables for our models:

- Split data into Training set (67%) and Test Set(33%).
- Building the model: We used the training data set (67% of the data) to fit the model.
- Testing the Model: Apply learned Model with the Test data set (33% of the data).This data is different to the training data. Here we make predictions and then compare that predictions with the known's outputs of the test set. We use Confusion matrix and Classification reports to evaluate and compare our models.

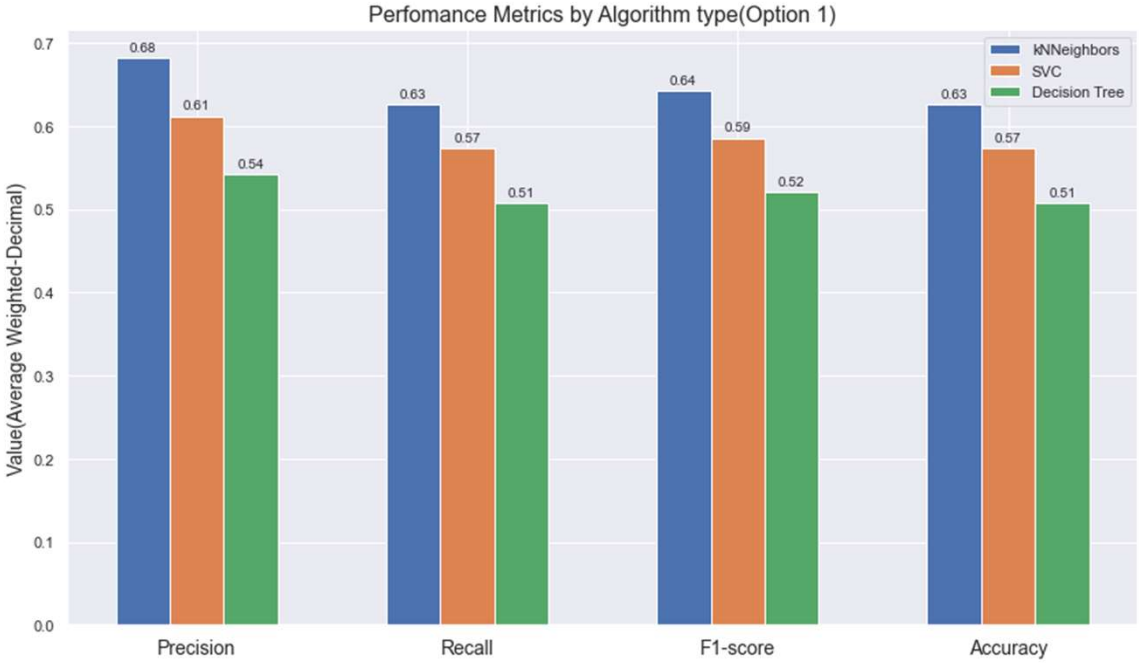
Findings (Continue)



From the graph you can see that enzyme classes Hydrolase, Transferase and Oxidoreductase have more samples than classes Lyase, Isomerase and Ligase.

This is important ahead when we will analyze performance by enzyme type for each algorithm.

Findings (Continue)

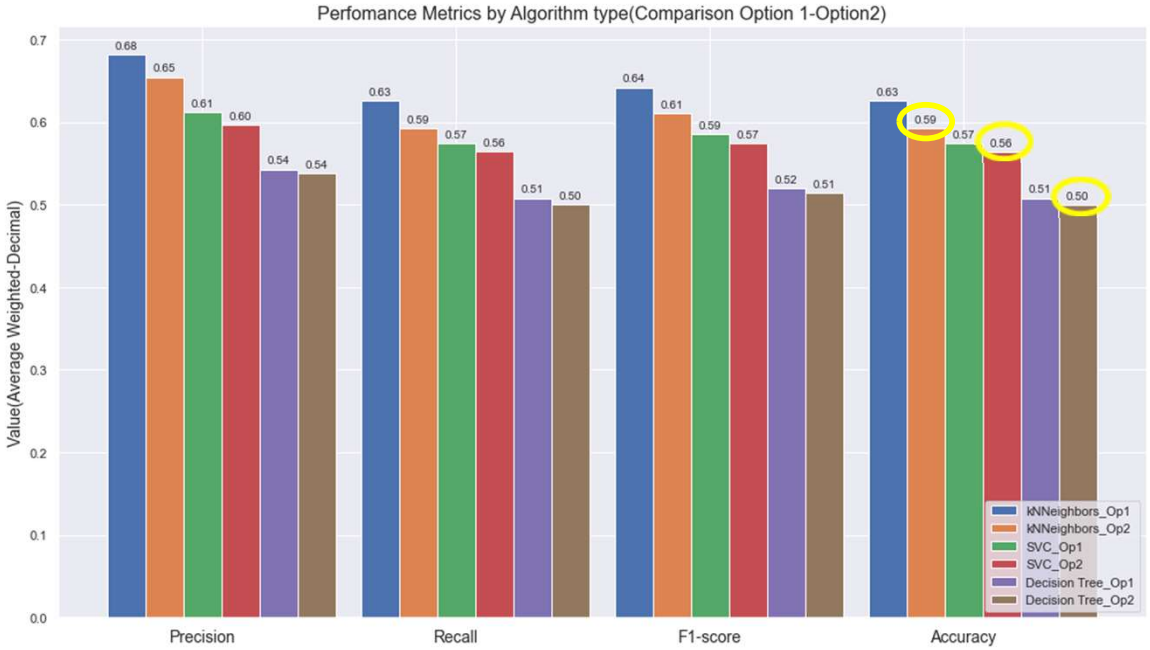


Related to the research question 1 , from the graphs you can see kNN algorithm obtained an acceptable performance greater than 60% in all metrics: Precision 68%,Recall 63%,F1-score 64% and an overall Accuracy of 63%.

The second-best performance was for SVC with a moderate performance , overall accuracy of 57% and Precision 61%,Recall 57%, F1-score 52%.

The worst performance was for Decision Tree with a moderate performance with a value for all metrics between 51% and 54 %.

Findings (Continue)

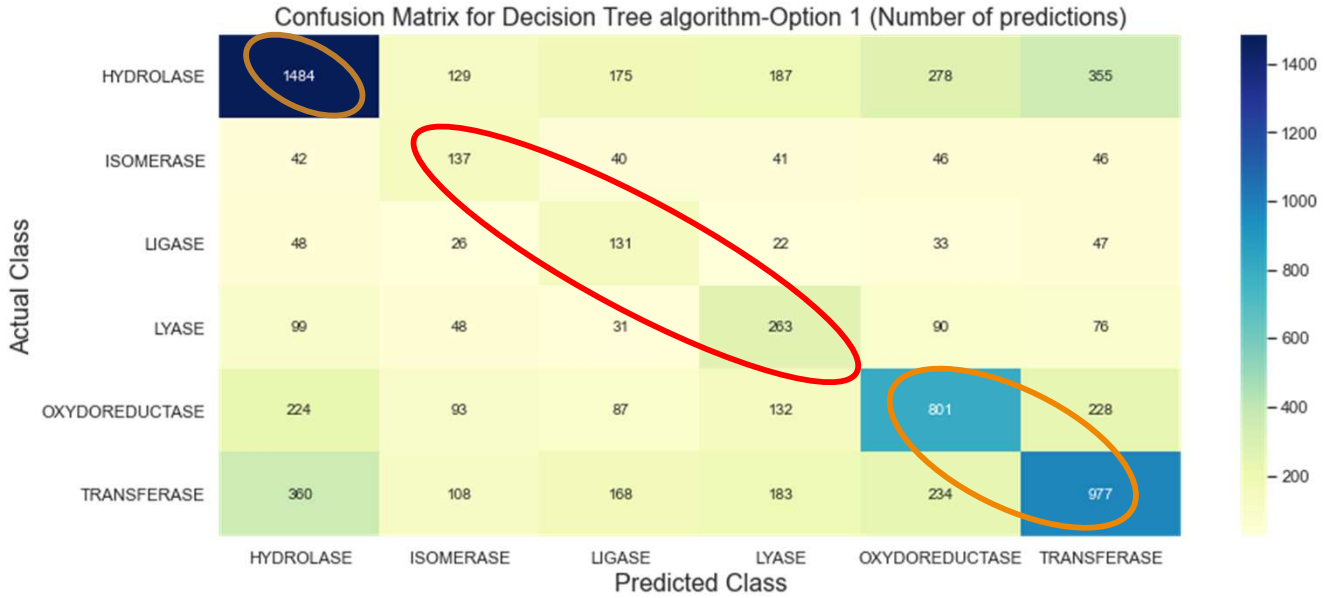


Abording our second question, the graph suggests that, for option 2, all the algorithms performance metrics are slightly worse when the features pH and Temperature (K) were added to the Amino Acid Composition features that were used in option 1.This behavior is observed for all the metrics: Precision, Recall,F1-Score and Accuracy.

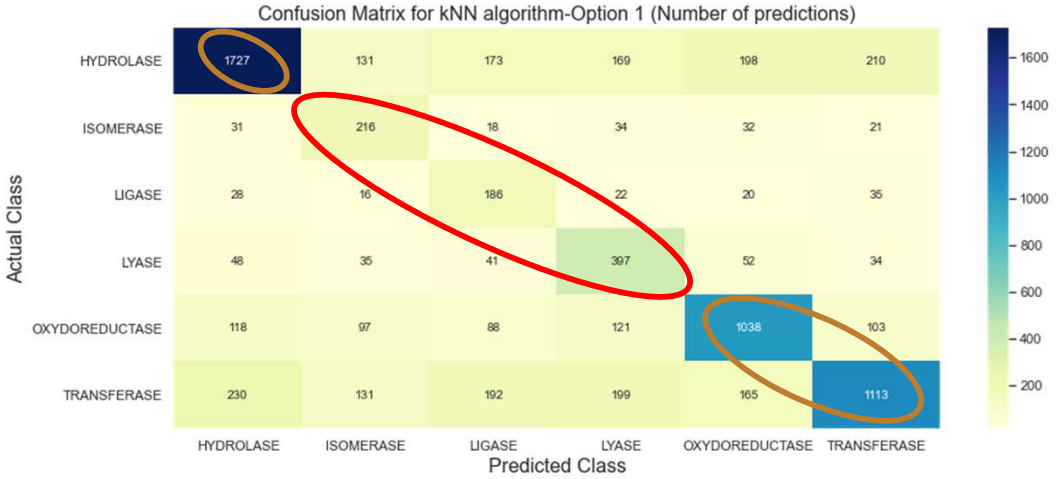
In term of Accuracy: kNN declined from 63% to 59%, SVC from 57% to 56% and Decision Tree from 51% to 50%

Findings (Continue)

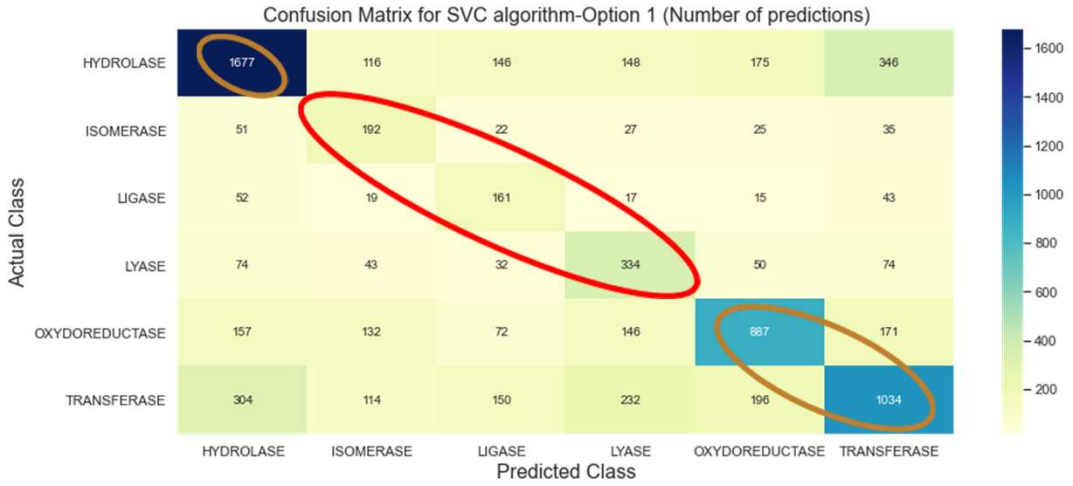
Now we'll see how is the behavior of the algorithms for option 1 that had better result that option 2, this time to the class level. From the the following graphs corresponding to the confusion matrices, it is evident that the number of predictions correctly classified for classes Isomerase, Ligase and Lyase (red oval) are smaller than the predictions for Hydrolase, Transferase and Oxidoreductase (orange ovals).As the number of samples in the test sets is also less for types Isomerase , Ligase and Lyase we need to go more deeper, checking the performance metric for each algorithm by each of the six enzyme types.



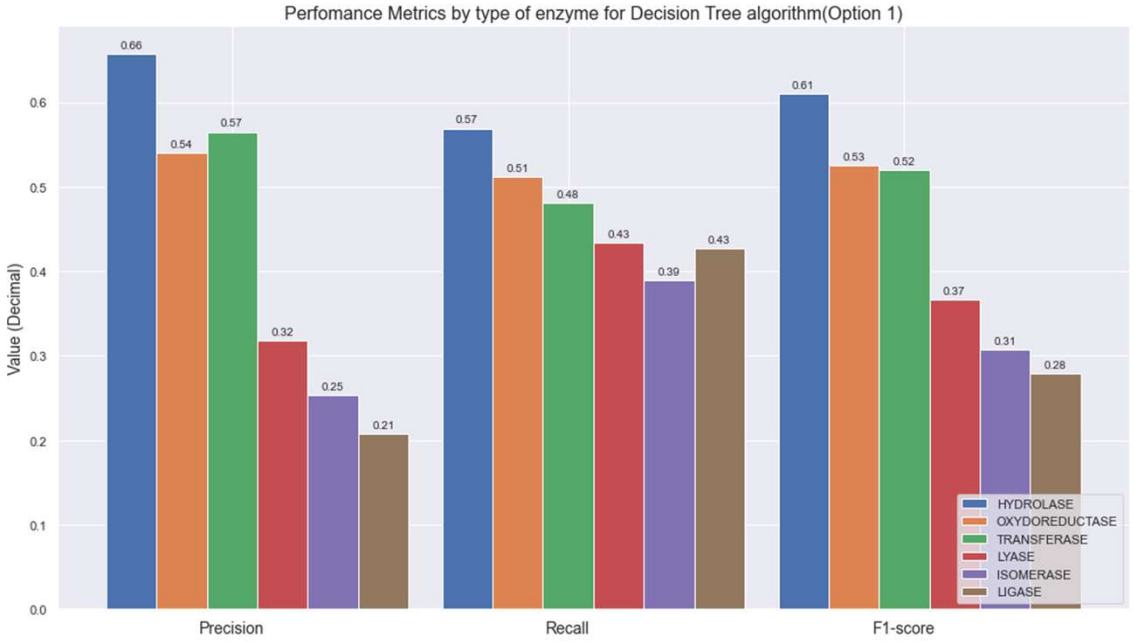
Findings (Continue)



Comparing the diagonals of the three confusion matrices (From upper left corner to lower right corner) it is evident that the number of predictions correctly classified for kNN algorithm by each class is the higher among the three algorithms. For example, it correctly classified 1727 samples while SVC had 1677 and Decision Tree 1484.



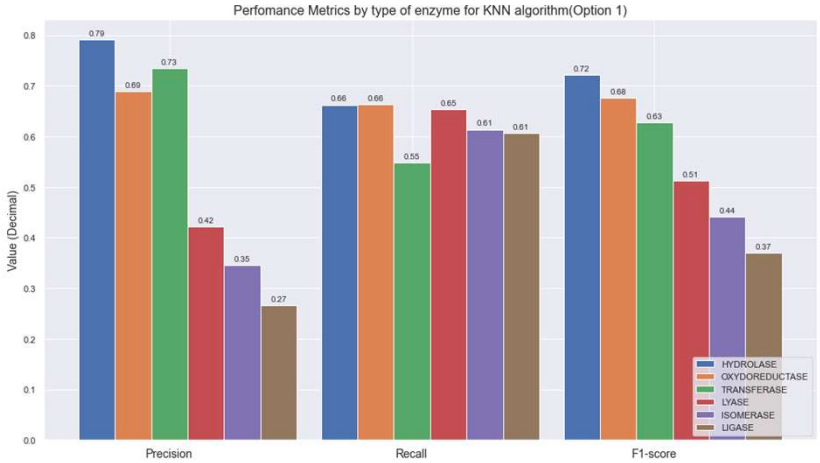
Findings (Continue)



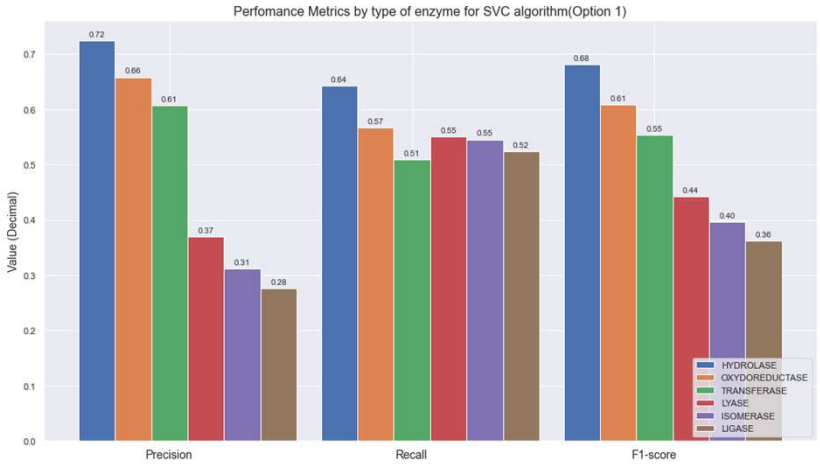
From the graph you can see Decision Tree obtained better results for Precision, Recall and F1-Score for classes Hydrolase, Oxidoreductase, Transferase than for classes Lyase, Isomerase and Ligase.

Using the F1-Score metric that is a harmonic mean of Precision and Recall ; we can see that for Hydrolase has a value of 61% followed by Oxidoreductase 53% and transferase 52%.

Findings (Continue)



From the graph you can see kNN Tree obtained better results for Precision and F1-Score for classes Hydrolase, Oxidoreductase and Transferase than for classes Lyase, Isomerase and Ligase. For Hydrolase and Oxidoreductase, the Recall was better between all classes and Transferase was the worst.
Using the F1-Score we can see that Hydrolase has a value of 72% followed by Oxidoreductase 68% and transferase 63%. This is the best result of all three algorithms.



From the graph you can see SVC obtained better results for Precision and F1-Score for classes Hydrolase, Oxidoreductase and Transferase than for classes Lyase, Isomerase and Ligase. For Hydrolase and Oxidoreductase, the Recall was better between all classes and Transferase was the worst.
Using the F1-Score we can see that Hydrolase has a value of 68% followed by Oxidoreductase 61% and transferase 55%.

Limitations

We choose the approach of using the Amino Acid composition frequency as input for our models. This method don’t have into account the position of each Amino Acid in the sequence neither the presence of subsets of a combination of two, three or more Amino acid (for example % of presence of ‘ACQ’, ‘RHI’,Etc.) that could lead to a future research with better performance results.

Also, we had data with less samples for classes Ligase, Lyase and Isomerase, so we artificially incremented that samples using SMOTE. The result show worse performance for this classes, until new research determined the cause , this limit the use of our research in comparison with other research with better balanced data set.

Conclusions

Taking into account that the data set had smaller sample for classes Isomerase, Ligase and Lyase, we can conclude that kNN had an overall acceptable performance greater than 60% in all the metrics (Accuracy, Precision, Recall and F1-Score) and moderate performance between 50% and 60%for SVC and Decision tree.

With respect to the second research question, the performance did not improve when added pH and Temperature. All metrics slightly worsen for second option.

The algorithms used performed better with the enzyme classes Hydrolase, Oxidoreductase and Transferase than the classes Lyase, Isomerase and Ligase. One possible cause could be that we had less data for these last types, future investigations are necessary to confirm cause.

Acknowledgements

The data was obtained from rcsb.org. (RCSB Protein Data Bank).

References

(RCSB PDB). H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Mohammed, A., & Guda, C. (2011). Computational approaches for automated classification of enzyme sequences. Journal of proteomics & bioinformatics, 4, 147.

Saidi R, Maddouri M, Mephu Nguifo E, Mephu Nguifo E. Protein sequences classification by means of feature extraction with substitution matrices. BMC Bioinformatics. 2010 Apr;11:175. DOI: 10.1186/1471-2105-11-175.