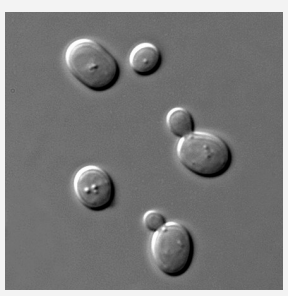
Data Science Poster

What are some means of performing classification on a dataset and how do they compare in terms of performance?

Data Science - BPINFOR-33 - Poster Workshop

Dataset

This dataset is about predicting the cellular localization sites of proteins. It is a multivariate dataset in the subject area of biology, primarily used for classification tasks. The dataset contains 1484 instances with 8 real-valued features.

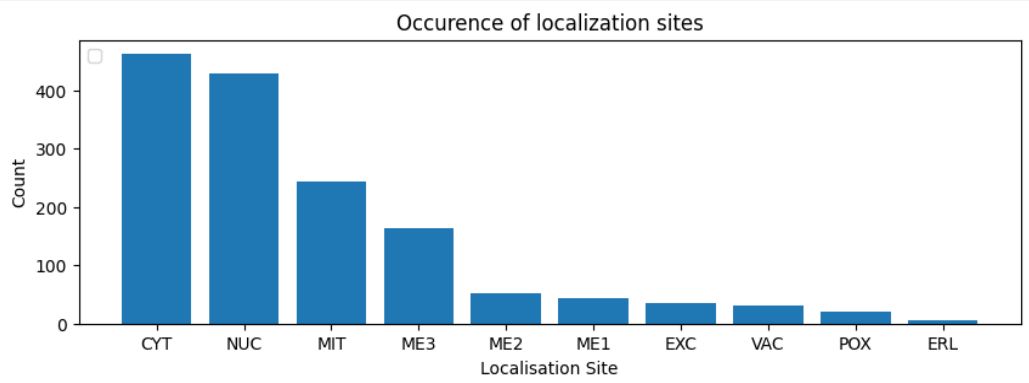


The goal of this poster is to predict our target "localization site" based on the 8 features given. We will perform classification using multiple different models and evaluate the performance of each model based on its accuracy and precision.

Exploration

Class Imbalance

As we can see the dataset suffers from class imbalance. The distribution of localization sites is skewed. This means that some instances (like CYT and NUC) have significantly more datapoints.



Model Selection

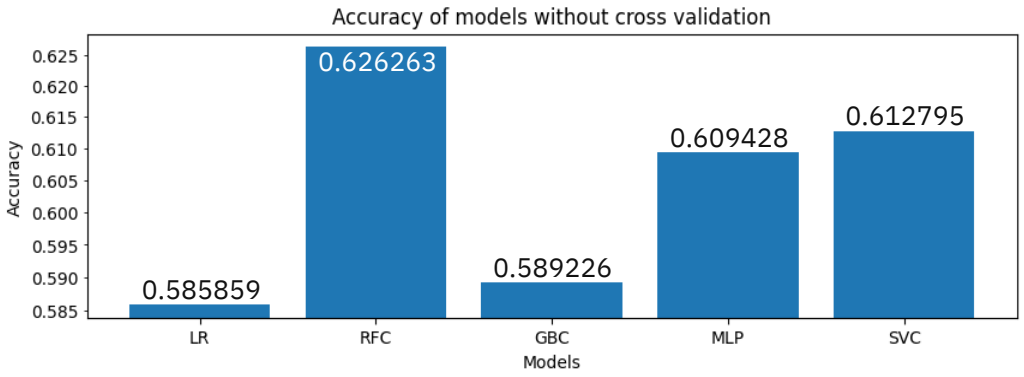
Different machine learning models vary in their performance for multivariate classification data. Here we will be using:

Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, Neural Network Classifier, Support Vector Classification

Performance

To run these models, we perform a 80/20 split between testing and training sets. We are using the accuracy metric of each prediction to compare the models.

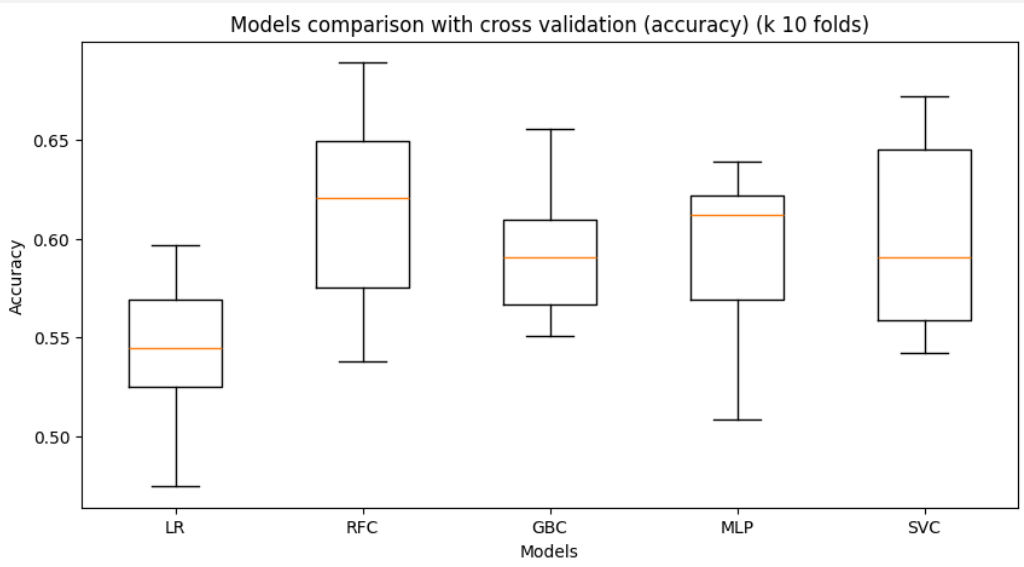
Accuracy of prediction by model



We see with the accuracy scores above we may have reached a certain threshold with our models. We can further explore the limitation by using cross validation.

Accuracy using Cross Validation

We now try again using cross validation method and using a k value of 10. We see that the previous trends stay quite similar.



Conclusion

In conclusion, our best performing model is the Random Forest Classifier with a mean accuracy of 0.614927 using cross validation. We can see that there is a notable outlier: Logistic Regression. It seams that it cannot perform as well as the other models.

Other models such as Neural Network Classifier and Support Vector Classifier also perform well but cannot match the accuracy of of Random Forest. This may be different with different datasets.

We can see that a certain maximum threshhold around 0.6 accuracy. This is probably a limitation of this dataset. We can speculate that more instances could increase our accuracy.

Limitations

It was not possible to test all the different kinds of models, however this selection shows differences between some major models and their varied usefulness for classification tasks. In a real world scenario, this analisys would only be one part of a larger data science pipeline, and many different models could be tested. Hopefully you have gained some insight into some means of performing classification and how different classification models perform.

Yeast Dataset - https://archive.ics.uci.edu/dataset/110/yeast