Comparative Analysis of Supervised Learning Algorithms for Predicting the Cellular Localization Sites of Proteins with Yeast Dataset

D. Induja

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract*—** *The examination of protein restriction locales is a significant errand in bioinformatics. Foreseeing the yeast protein restriction locales is a promising space among various exploration techniques in view of the yeast protein estimation information which have numerous records/highlights. Proteins are a significant piece of the organic entity and are engaged with pretty much every cycle in the cell. This research paper presents a comparative analysis of two supervised learning algorithms, Multilayer Perceptron (MLP) and Logistic Regression, for predicting the cellular localization sites of proteins. The algorithms were evaluated based on their accuracy, precision, and recall. The results and their implications are discussed, leading to a conclusion about the effectiveness of each algorithm in this predictive modeling task.*

I. INTRODUCTION

A biochemical component is protein. It has one or more polypeptides that have been folded into a fiber or spherical. As a biological process, it operates. Peptide connections between the carboxyl groups and the amino acids of nearby residues are what hold together polypeptides, which are linear strands of amino acids [5]. The Yeast Protein Localization is the data of protein localization patterns in the yeast (Saccharomyces cerevisiae). Learning the functions and roles of yeast proteins involved in all cellular processes is essential to predict the yeast protein localization sites. The localization sites can also be used to evaluate protein information indicated from gene data. Additionally, we can infer which pathway an enzyme belongs to by its proper localization sites [6]. In light of the importance of predicting the yeast protein localization sites, many researchers involved in biology and computer science have been making effort to explore the prediction methods in the domain, and a great deal of research methods and results have emerged. Despite recent technical advances in the prediction of the yeast protein localization sites, experimental results have low accuracy and experimental determination remains time-consuming and labour-intensive.

The order of genes encoded in the genetic code determines the amino acid sequence of the protein. The genetic code typically consists of the conventional twenty amino acids [6] and the pattern of protein folding varies depending on the organism or the use of cells. Importantly, these sorting proteins can also be applied in medicinal settings. Additionally, the information can be used to help with genetic modification, boost protein quality, or meet requirements.

II. METHODOLOGY

In this way, the paper proposed Logistic Regression and Multilayer Perceptron for predicting the cellular localization sites of proteins.

2.1 Logistic Regression

Logistic Regression at times called the strategic model or logit model, dissects the connection between numerous free factors and an all-out subordinate variable, and evaluations the likelihood of event of an occasion by fitting information to a strategic bend [1][2]. There are two models of calculated relapse, paired strategic relapse and multinomial strategic relapse. Parallel strategic relapse is regularly utilized when the reliant variable is dichotomous and the free factors are either ceaseless or downright. At the point when the reliant variable isn't dichotomous and is contained beyond what two classifications, a multinomial calculated relapse can be utilized [3][4][8]. Since strategic relapse computes the likelihood of an occasion happening over the likelihood of an occasion not happening, the effect of free factors is generally made sense of regarding chances. With calculated relapse the mean of the reaction variable p as far as an informative variable x is displayed relating p

and x through the situation p=α+βx. The basis of logistic regression is the logistic function, also called the sigmoid function, which takes in any real valued number and maps it to a value between 0 and 1.

Sigmoid Function $y = 1/1 + e^{-x}$

2.2 Multilayer Perceptron (MLP)

MLP is a subset of AI and at the center of profound learning calculations, are likewise alluded to as artificial neural networks (ANNs). Their design and classification are designed according to the human cerebrum, reflecting the correspondence between natural neurons. PCs can utilize this to fabricate a versatile framework that assists them with ceaselessly improving by gaining from their disappointments. Thus, counterfeit brain networks try to handle testing issues like summing up archives or distinguishing faces [7].

We can group and bunch information utilizing brain organizations, which can be seen as a layer of grouping and characterization on top of the information you oversee and store. At the point when given a named dataset to prepare on, they assist with ordering information by placing unlabeled information into bunches in light of likenesses between model information sources.

A multi-facet perceptron is a completely convolutional network that makes an assortment of results from a bunch of sources of info[1]. A coordinated diagram interfacing the info and result layers of a MLP is comprised of different layers of information hubs.

III. EXPERIMENTAL RESULTS

The experiments were conducted using the Python programming language, utilizing the powerful Scikit-learn library for data representation, manipulation, and analysis. For this study, the yeast protein dataset from the UCI data repository of AI datasets was employed [9]. This dataset consists of 1484 instances, each containing eight features (attributes) are used: mcg, gvh, alm, mit, erl, pox, vac, nuc. And proteins are classified into 10 classes are shown in the figure-1 as mentioned class wise label count.

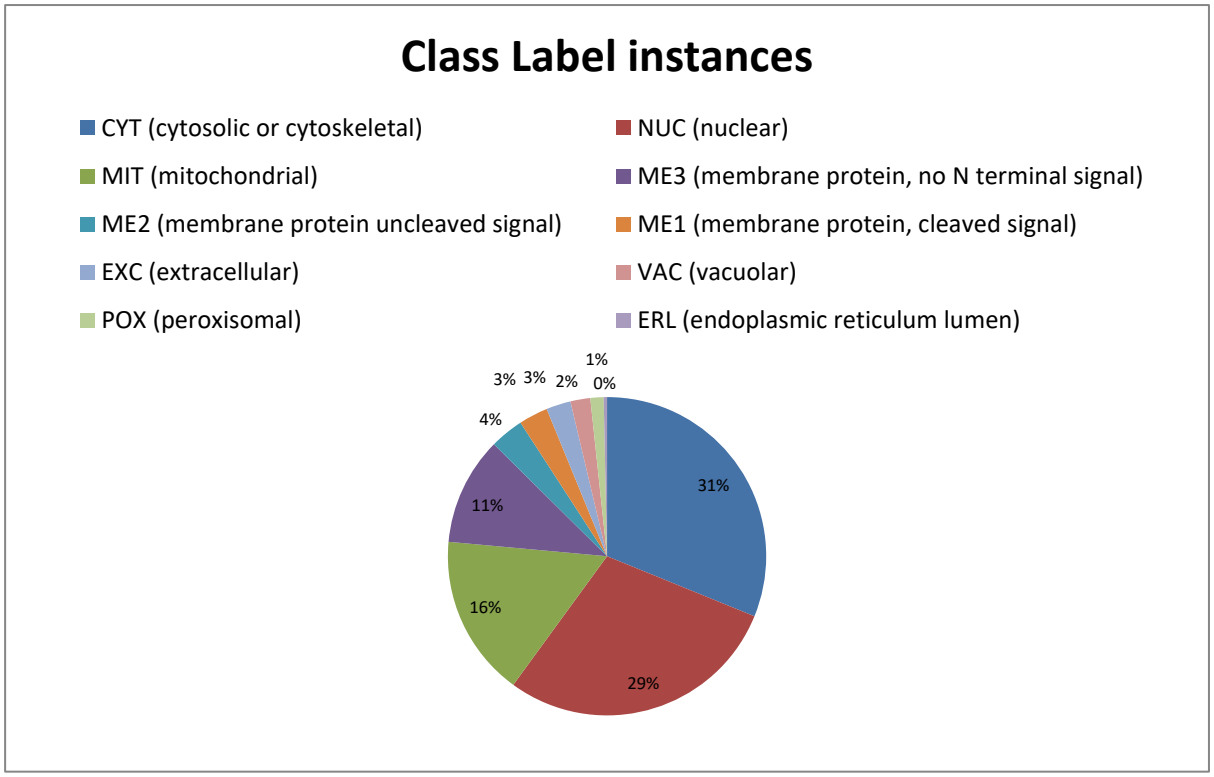

Figure-1: Class label instances

In this study, two machine learning algorithms, namely Logistic Regression and Multilayer Perceptron, were applied to the yeast protein dataset. The performance of each algorithm was evaluated using accuracy, precision, and recall as evaluation metrics. The experimental results are summarized in the table-1 and same shown in the figure-2:

Table-1
Performance of Classifier

Algorithm	Accuracy	Precision	Recall
Multilayer Perceptron	84.5	84.7	84.5
Logistic Regression	81.23	81.4	81.2

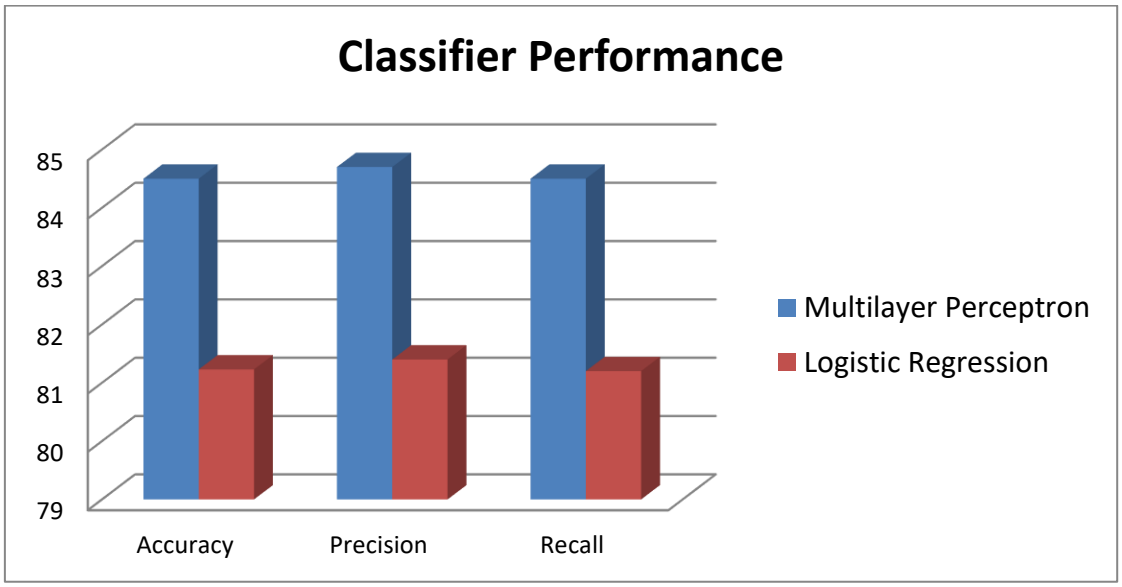

Figure-1: Classifier Results

We observe in the figure-2, the results demonstrate that the Multilayer Perceptron achieved higher accuracy, precision, and recall compared to Logistic Regression. With an accuracy of 84.5%, precision of 84.7%, and recall of 84.5%, the Multilayer Perceptron exhibits superior performance in predicting the cellular localization sites of proteins.

The higher accuracy of the Multilayer Perceptron suggests that it has a better ability to correctly classify proteins into their respective cellular localization sites. The higher precision indicates that the Multilayer Perceptron has a lower rate of false positives, making it more reliable in identifying proteins correctly. Similarly, the higher recall suggests that the Multilayer Perceptron has a lower rate of false negatives, implying that it can effectively capture proteins belonging to different cellular localization sites.

The results highlight the effectiveness of the Multilayer Perceptron as a powerful tool for predicting protein cellular localization sites. Its ability to capture complex relationships within the data, thanks to its multi-layer architecture, likely contributes to its superior performance compared to Logistic Regression. The findings of this study have practical implications for bioinformatics and protein research. Accurate prediction of cellular localization sites is crucial for understanding protein functions, interactions, and their roles in biological processes. The Multilayer Perceptron can aid researchers in identifying the correct cellular localization site of proteins, enabling further investigations and insights into their functional properties.

IV. CONCLUSION

In conclusion, the comparative analysis of the Multilayer Perceptron and Logistic Regression for predicting the cellular localization sites of proteins indicates that the Multilayer Perceptron outperforms Logistic Regression in terms of accuracy, precision, and recall. The results affirm the effectiveness of the Multilayer Perceptron as a robust algorithm for this specific predictive modeling task. Researchers and practitioners in bioinformatics can benefit from leveraging the Multilayer Perceptron for accurate protein localization predictions, contributing to advancements in protein research and related fields.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G. Ravi Kumar, S. Rahamat Basha, Surya Bhupal Rao, "A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, ISSN:0973-8975.

[3] G. Ravi Kumar, K. Tirupathaiah and B. Krishna Reddy, "Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques", International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019

[4] J.Han and M.Kamber,"Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nded.San Mateo, CA; Morgan Kaufmann, 2006.

[5] M. Al Ahmad, Z. Al Natour, S. Attoub and A. H. Hassan, "Monitoring of the Budding Yeast Cell Cycle Using Electrical Parameters", *IEEE Access*, vol. 6, pp. 19231-19237, 2018.

[6] M. Kopecká, "Yeast and fungal cell-wall polysaccharides can self-assemble in vitro into an ultrastructure resembling in vivo yeast cell walls", *Microscopy*, vol. 62, no. 3, pp. 327-339, June 2013.

[7] N.Michael, "Artificial Intelligence – A Guide to Intelligent Systems", 2nd Edition, Addison Wesley 2005

[8] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[9] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/ datasets/Yeast.