I COBBIND
Creating connections between biotechnology and industrial sustainability
August 25 to 28, 2024
Costão do Santinho Resort, Florianópolis, SC, Brazil
BIOPROCESS ENGINEERING

EXPLORING THE POTENTIAL AND LIMITATIONS OF MACHINE LEARNING MODELS IN DESCRIBING β-GALACTOSIDASE IMMOBILIZATION PARAMETERS

Raimundo V. A. Maia[1], Andréa S. Pereira[1] & Luciana R. B. Gonçalves[1*]

[1]*Center of Technology, Chemical Engineering Department, Federal University of Ceará, Fortaleza, Brazil.*
** email: LRG@ufc.br*

ABSTRACT

In this paper, a dataset was assembled containing 100 observations taken from the literature, correlating the immobilization parameters of the β-galactosidase enzyme with its optimum temperature, optimum pH and the Michaelis-Menten constant (Km). Three different models were used, including Random Forest, Ridge Regression, and Multiple Linear Regression, which were compared using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). The Random Forest model seemed to perform the best, as it presented R^2 values of 0.818, 0.781 and 0.963 for the optimal temperature, optimal pH and Km, respectively. The MAE, MSE and RMSE values also indicate that it is possible to use this model to predict experimental data. However, an extension of the dataset and a possible tuning of hyperparameters are necessary to improve the predictive capacity of the algorithm.

Keywords: Immobilization. Model. β-galactosidase. Machine Learning.

1 INTRODUCTION

β-galactosidase is an enzyme belonging to the hydrolase class, and therefore, can convert lactose into glucose and galactose. This means that this biocatalyst has several applications, mainly in the food industry, for supplementation for people who are lactose intolerant, and dessert fabrication. The literature studies[1] have also shown that it is possible to use the enzyme to produce sugars such as lactulose (which contains prebiotic properties), aggregating even more industrial value to it.

Unfortunately, this soluble biocatalyst also has some disadvantages, such as low thermal, operational and storage stability. Therefore, enzyme immobilization appears as an interesting alternative to overcome these difficulties, but it is also necessary to pay attention to some immobilization parameters[2], such as: pH, enzyme concentration, the temperature at which the procedure is carried out, immobilization time, etc. Consequently, dependent variables such as the optimum temperature, optimum pH and the Michaelis-Menten constant (Km) can be used to measure whether the process has been successful and whether it is applicable. However, conducting experimental studies on enzyme immobilization is a complex, costly and time-consuming process. On the other hand, modeling techniques can be employed to accelerate the process, allowing the identification of patterns and the selection of the most critical variables. These methods offer an efficient and effective alternative to traditional experimental approaches

Machine learning algorithms and models, for instance, can be very useful for discovering relationships and the impact of variables on prediction accuracy in various scientific areas, as has already been demonstrated by other authors.[3] On that account, the objective of this work is to compare the predictive performance of three different machine learning models (Random Forest, Ridge Regression and Multiple Linear Regression, respectively), using a dataset consisting of 50 observations extracted from the literature to predict the optimal temperature, optimal pH and Michaelis-Menten constant.

2 MATERIAL & METHODS

The dataset used in this study has 100 observations, and was assembled from data extracted from scientific literature from sources such as Scopus, PubMed and Google Scholar. The input variables to be analyzed by the three models were: the immobilization method, buffer ionic strength, buffer's ion, enzyme source, concentration of the enzymatic solution, pH of the medium, immobilization temperature and immobilization time. The ambient temperature was considered to be 25 °C and the overnight time was 12 hours, as this is the most reported time for enzymatic immobilization. The output variables were temperature optimum, pH optimum and Km. K-fold Cross Validation[4] and Grid Search[4] were used in order to better train the models and search for the better parameters of each one. In addition, all the data was normalized, since there were different ranges between the variables.

It is worth noting that algorithms were coded in Phyton for the Random Forest[3], Ridge[5] and Multiple Linear Regression[5] models, with the best one being chosen according to the values of the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). All the independent variables were used during the test.

3 RESULTS & DISCUSSION

The tables indicating the metrics for all the models were shown, whereas the scatter plot or the best model was displayed. As can be observed, the Random Forest model presented, on average, the best results, because, it had a best coefficient of determination (R^2) when compared to the Ridge Regression and Multiple Linear Regression. The values of MAE, MSE, and, consequently, RMSE indicate that there is greater proximity between the predicted values and those obtained experimentally. It is also possible to state that the Multiple Linear Regression model failed to capture the complexity of the relationships between the variables, which is expected, since this model assumes a linear relationship between inputs and outputs, which does not seem to be the case.

Table 1 Table for temperature optimum (all models)

Indicator	Random Forest	Ridge Regression	Multiple Linear Regression
R^2	0.852	0.306	0.250
MAE (°C)	2.828	7.002	7.247
MSE (°C^2)	14.990	70.403	76.087
RMSE (°C)	3.872	8.391	8.723

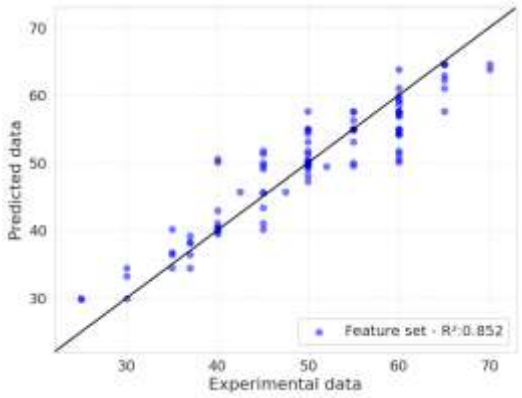

Figure 1 Scatter plot for temperature optimum (°C) (Random Forest model)

The same seems to be valid for the optimum pH, as the Random Forest appears to be the model that captures the best relationship of the dataset. This seems to be due to the fact that many independent decision trees are created, before the average result of prediction is taken, which helps to understand how this model could be useful in solving biotechnology practical obstacles.

Table 2 Table for pH optimum (all models)

Indicator	Random Forest	Ridge Regression	Multiple Linear Regression
R^2	0.845	0.530	0.471
MAE	0.392	0.720	0.738
MSE	0.255	0.767	0.868
RMSE	0.505	0.876	0.932

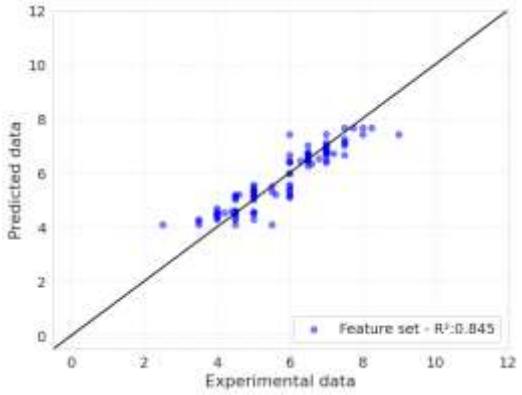

Figure 2 Scatter plot for pH optimum (Random Forest model)

The Michaelis-Menten constant, at last, follows the same trend, but the Random Forest had the best performance, followed by the Ridge Model and the Multiple Linear Regression model. This is probably due to the fact that the average of the trees fit the data better. However, it is still necessary to increase the number of rows of the dataset, and consequently to check if the model is not tending to overfit. The predictions values could be more trustworthy regarding the experimental ones.

Table 3 Table for Michaelis-Menten constant (all models)

Indicator	Random Forest	Ridge Regression	Multiple Linear Regression
R^2	0.981	0.120	0.074
MAE (mM)	4.055	31.538	34.684
MSE (mM^2)	45.412	2157.077	2266.526
RMSE (mM)	6.739	46.444	47.608

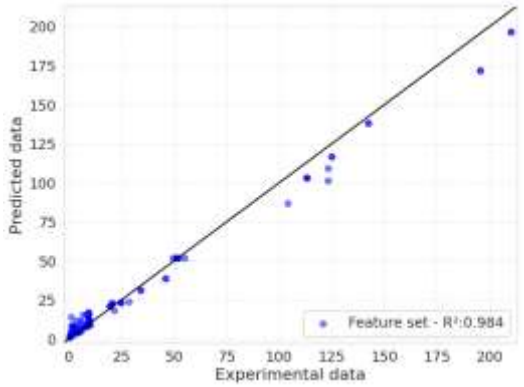

Figure 3 Scatter plot for Michaelis-Menten constant (Random Forest model)

4 CONCLUSION

In conclusion, a reliable predictive model can be obtained using supervised machine learning techniques to correlate β-galactosidase immobilization parameters in a given dataset. The results are within the expected margin of experimental error, which indicates that such models can be helpful in practical applications. However, to improve the accuracy and dependability, it is essential to collect more data and identify the other significant variables that can be incorporated into the analysis. Overall, this study provides valuable insights into the potential of machine learning techniques in predicting the behavior of β-galactosidase immobilization, which could have implications for several industrial and scientific applications.

REFERENCES

1 ALBUQUERQUE, T. L., GOMES, S. D. L., D'ALMEIDA, A. P., LAFUENTE, R. F., GONÇALVES, L. R. B. 2018. P. Biochem. 73. 65-73.
2 SASS, A. C., JÖRDENING, H. J. 2020. A. Biochem. Biotech. 191. 1155-1170.
3 CHAI, M., MORADI, S., ERFANI, E., ASADNIA, M., CHEN, V., RAZMJOU, A. 2021. Chem. Mat. 33 (22). 8666-8676.
4 BEHESHTI, N. 2022. Cross Validation and Grid Search. *In:* Towards Data Science.
5 POLAT, E., GUNAY, S., 2015. J. of Data Science. 13 (4). 663-692

ACKNOWLEDGEMENTS

The authors are grateful to UFC (Universidade Federal do Ceará), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil), and FUNCAPES (Fundação Cearense de Apoio ao Desenvolvimento Cientifico e Tecnológico) for the financial support provided.