Machine Learning Algorithm Selection

COSC 5010 – 03

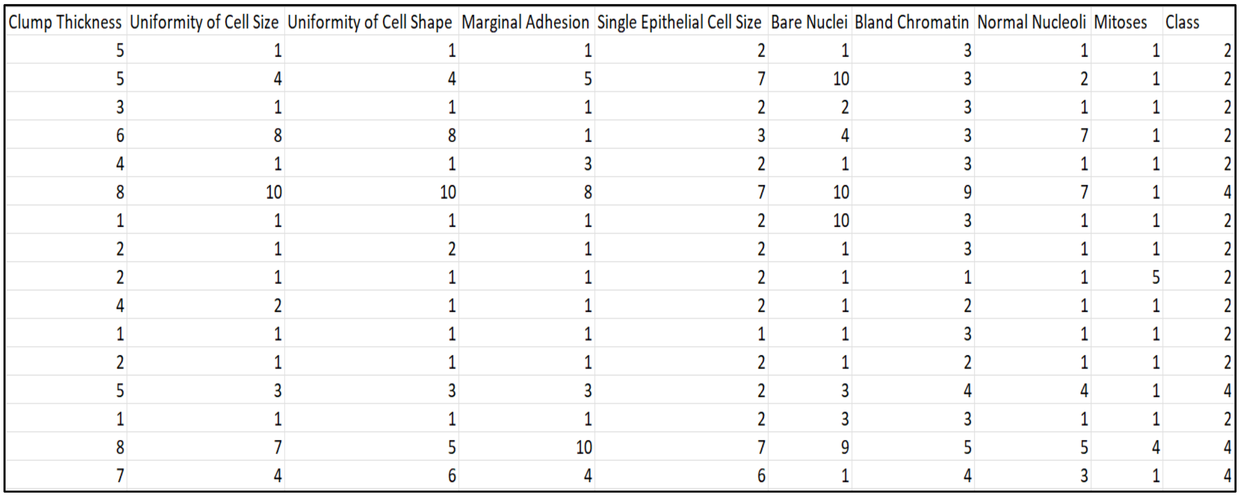
Isaac Baah

1.0 Introduction

The objective of the project is to build machine learning models and evaluate them using the necessary evaluation metrics, and then choose the best performing algorithm. In this study, I did both regression and classification problems using several machine learning algorithms like decision trees, k-nearest neighbors. However, in this report I will be presenting the results of only the classification models. Selecting a model is very important for a particular dataset. The data used has a binary response variable, hence when using parametric algorithms, we need to take into account their assumptions. For example, when we are dealing with this data, we can use logistic regression to make a prediction instead of linear regression or even multinomial logistic regression.

2.0 Data

The data used is from a machine learning tutorial I did on Udemy, the data consists of nine independent variables. All the variables are numeric. The dependent variable is the "Class". The data is trying to predict the whether the plant will grow "big or tall" or otherwise considering the characteristics of the plant. Class "2" means the plant will grow "big or tall" and Class "4" means "Otherwise".



Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
2	1	1	1	2	1	2	1	1	2
5	3	3	3	2	3	4	4	1	4
1	1	1	1	2	3	3	1	1	2
8	7	5	10	7	9	5	5	4	4
7	4	6	4	6	1	4	3	1	4

Figure 1. A snapshot of the data used in the project.

3.0 Experimental Setup

To predict based on the characteristics of the plant whether a plant will grow very well "big or tall" or grow otherwise, several models were developed. In this project, the following models were built; Decision Trees, Logistic Regression, Support Vector Machines, Random Forest Classifier, K-Nearest Neighbors, and Naïve Bayes. Though I have mostly been using python, I decided to challenge myself by using R programming language, specifically “mlr3verse” library. Using the decision tree as an example for fitting the data. The first thing was to define the task. In this case, the task is classification and then define the variable which is the class. The next step was to instantiate the machine learning algorithm, in the case of decision tree "classif.rpart" and the prediction type "prob".

The next step was to partition the data into training and testing datasets. The training data is used to train the machine learning algorithm and then the testing set is used to make a prediction on the train model. The final step was to compare the machine learning algorithms using classification accuracy, classification log loss, and then classification mbrier.

4.0 Results

After fitting the machine learning models on the training dataset and testing on the testing data, we now must select the best performing machine learning to be used for our model prediction. As stated in the experimental section, the models were compared against classification accuracy, log loss and mbrier. The results are show in the figures below.

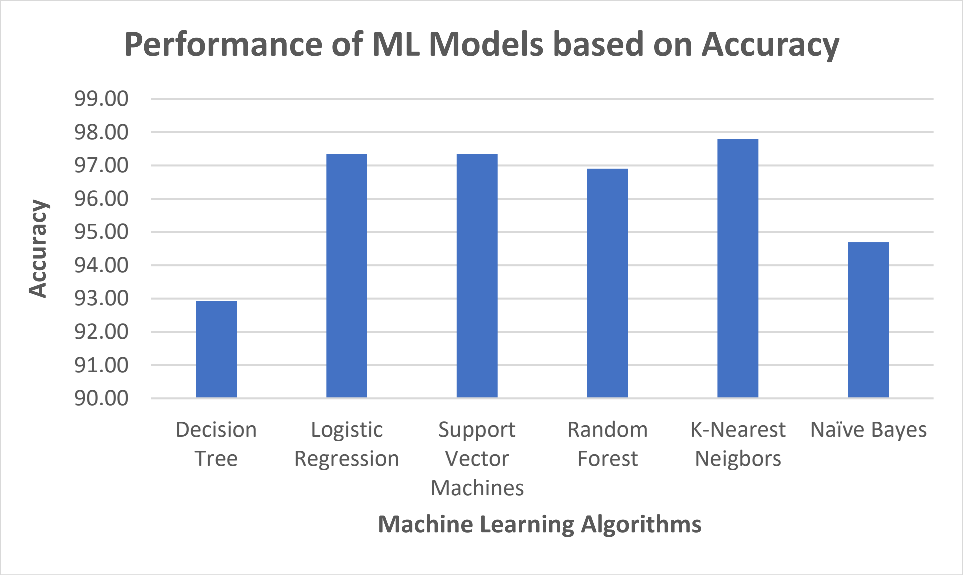


Figure 2. Comparison of the Performance of ML Models based on Accuracy.

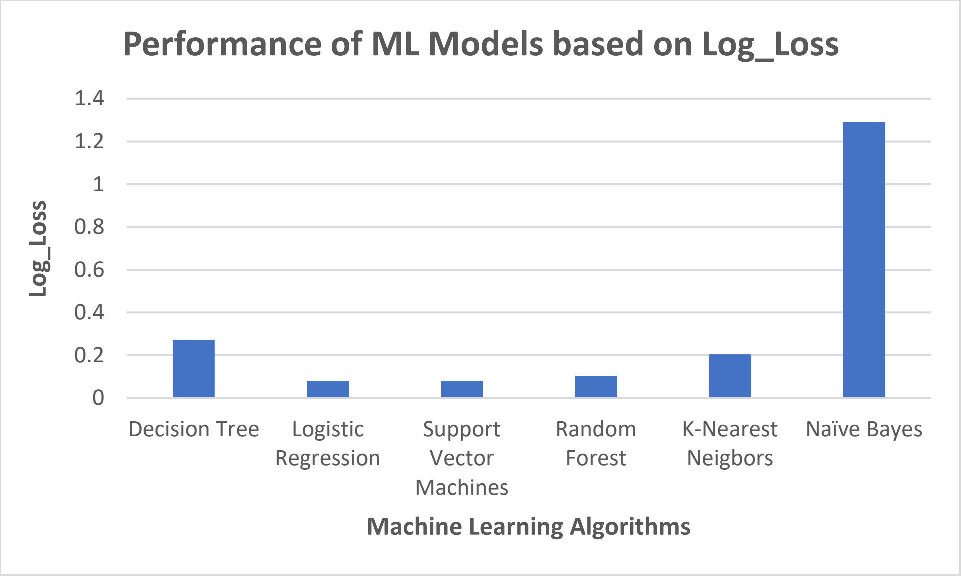


Figure 3. Comparison of the Performance of ML Models based on Log_Loss.

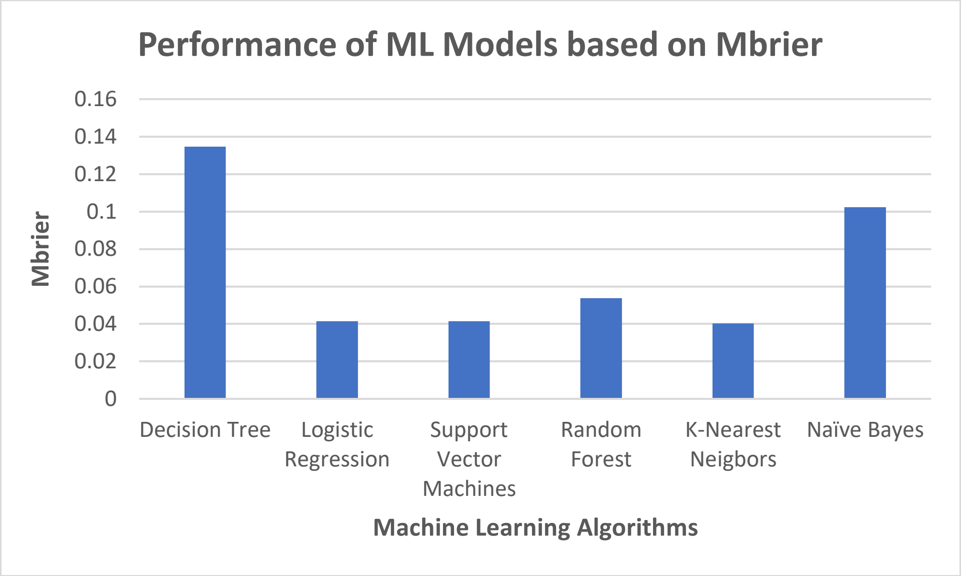


Figure 4. Comparison of the Performance of ML Models based on Mbrier.

The three Figures shown above compare the model performances based on the metric stated in the results and experimental sections of this project. The models were compared on separate charts due to their scales. From the above result, we can see that KNN performed well on accuracy and mbrier when compared to all the other models, however, logistic regression and SVM achieved the lowest log_loss. Based on the result, I will conclude that KNN Classification is the best model. Hence, K-Nearest Neighbors is chosen as the right algorithm to be used to make predictions as to whether based on the characteristics of the plant, it will grow well or not.

5.0 References

1. Lang et al., (2019). mlr3: A modern object-oriented machine learning framework in R. Journal of Open-Source Software, 4(44),1903, https://doi.org/10.21105/joss.01903.

2. Kotthoff L, Sonabend R, Foss N, Bischl B. (2024). Introduction and Overview. In Bischl B, Sonabend R, Kotthoff L, Lang M, (Eds.), Applied Machine Learning Using mlr3 in R. CRC Press. https://mlr3book.mlr-org.com/introduction_and_overview.html.

3. Foss N, Kotthoff L. (2024). Data and Basic Modeling. In Bischl B, Sonabend R, Kotthoff L, Lang M, (Eds.), Applied Machine Learning Using mlr3 in R. CRC Press. https://mlr3book.mlr-org.com/data_and_basic_modeling.html.

4. Hands-On Machine Learning with Python and R. https://www.udemy.com/course/machinelearning/. [Accessed, October 2023].