BUILDING A PREDICTIVE MODEL USING DIFFERENT CLASSIFIERS ON THE SAME DATASET

By

RHOGIEL V. ANDOY

Advance Database Management System Final Project

To be submitted to Doc. Jocelyn B. Barbosa PhD

University of Science and Technology of Southern Philippines - CDO

January 27, 2023

TEAM MEMBERS

NAME	ROLE
RHOGIEL V. ANDOY	I did all the work alone since I did not have a team or a group. I performed the following tasks: • Finding a dataset; • Cleaning of the dataset to be used in the training of the classifiers; • Performing six (6) machine learning algorithms (classifiers) using K-fold cross validation (10 folds) on the same dataset; • Comparing performance issues (accuracy, precision, recall, specificity, sensitivity and F1-score); • Documentation; • Recording a video demo;

RATIONALE

In this project, a comparison of six (6) different classifiers was performed to determine the best classifier to use in order to develop a machine vision system to distinguish between two different variety of raisins (Kecimen and Besni) grown in Turkey. The dataset used is the Raisin Dataset containing a total of 900 instances or pieces of raisin grains that were obtained from an equal number of both varieties (450 each class). Six (6) different classifiers were performed to determine the suitable classifier to be used in the dataset, namely, Decision Tree/Classification

Tree, Regularized Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbor. The classification achieved 80.89% with Decision Tree, 83.44% with K-Nearest Neighbor, 83.89% with Naive Bayes, 85.67% with Support Vector Machine, 86.56% with Regularized Logistic Regression, and 86.67% with the highest classification accuracy obtained in the experiment with Random Forest.

DATASET REPORT

As mentioned above, the dataset used was a collection of data consisting of different image properties and measurements of raisins, the Raisin Dataset. The dataset shape is approximately 900 rows x 8 columns. With a balanced size from both varieties, 450 instances each. There are 7 columns, the morphological features and 1 column that identifies their variety , namely; 'Area', 'MajorAxisLength', 'MinorAxisLength', 'Eccentricity', 'ConvexArea', 'Extent', 'Perimeter', 'Class'. The dependent or target variable is the ‘Class’ column, the rest are the independent or the feature variables. The dataset was downloaded from Kaggle Dataset: **https://www.kaggle.com/datasets/shrutisaxena/raisin-dataset?select=Raisin_Dataset.xlsx**

RESULTS AND DISCUSSION

The downloaded dataset undergone data checking and data cleansing. Data Checking was performed to check for any data outliers to prevent the trained model from overfitting and underfitting. Data Cleansing was also perform to eliminate any outlier data found in the data checking process and to finalize the dataset for model training. K-fold cross validation with the *n* value of 10 (n=10) was performed to evaluate the six (6) classifiers; Decision Tree/Classification

Tree, Random Forest, K-Nearest Neighbor, Regularized Logistic Regression, Support Vector Machine, and Naive Bayes.

After the execution of six (6) classifiers using K-fold cross validation, the figure below shows the performance measures.

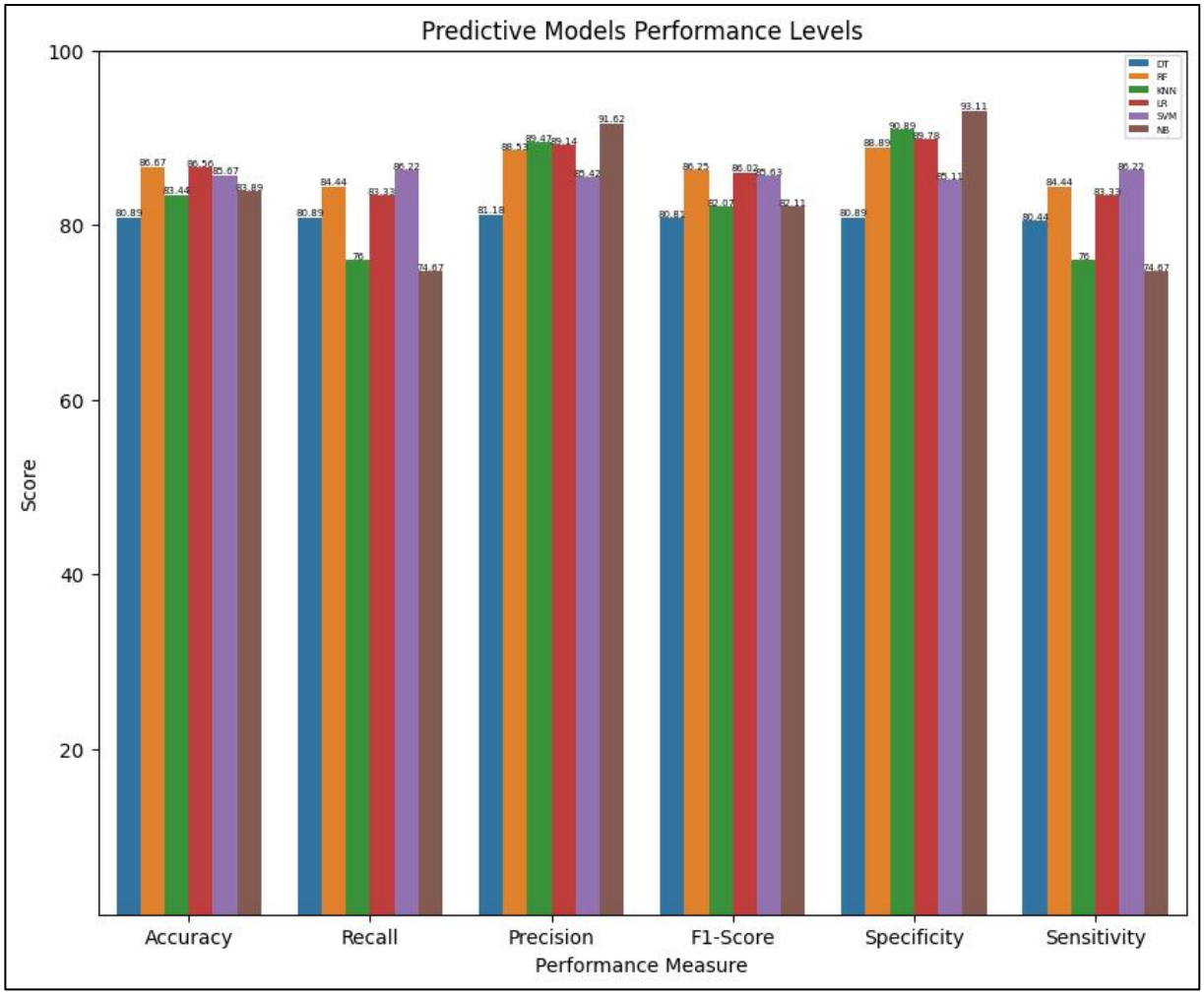

Figure 1.0: Performance Measure of Six Classifiers

In this bar plot, we can see the summary of the performed experiment on the mentioned classifiers above. Accuracy (ACC) measures the fraction of correct predictions. It is defined as "the ratio of correct predictions to total predictions made" (Pant. R, 2020). In terms of accuracy, the Random Forest classifier came out on top with 86.67% accuracy followed by the Regularized

Logistic Regression with 86.56% accuracy. However, classification accuracy does not bring out the detail you need to diagnose the performance of your model. This can be brought out by using a confusion matrix (Pant, R. 2020). The confusion matrix has 3 metric scoring; precision, recall, and F1-score. Recall calculates the ability of a classifier to find positive observations in the dataset. In Recall, Support Vector Machine out performed the group with 86.22% recall followed by Random Forest with 84.44%. Precision calculates the ability of a classifier to not label a true negative observation as positive. In Precision, Naive Bayes topped the group with 91.62% followed by K-Nearest Neighbor with 89.47%. In order to compare any two models, we use F1-Score. It is difficult to compare two models with low precision and high recall or vice versa. F1-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more (Pant, R. 2020). In F1-score, Random Forest topped the group again with 86.25% followed by Regularized Logistic Regression with 86.02%. In Specificity, Naive Bayes once again topped the group with 93.11% followed by 90.89%. Lastly, in Sensitivity, Support Vector Machine topped the group with 86.22% followed by Random Forest with 84.44%.

CONCLUSION AND RECOMMENDATION

After conducting the experiment and created models from each six different classifier, I came into conclusion. One of the highlights of this experiment is that, I found out that training Support Vector Machines can be very time consuming especially when your computer unit or laptop is low end based on the quality of the specifications. It is more time consuming and slow especially if the dataset that you used is not scaled first before training/cross validation. The bar plot in Figure 1.0 shows that the classifier with the highest accuracy rate is the Random Forest.

In the submitted notebook, the visualization of the 10-fold cross validation accuracy performance of the classifiers are shown and it only shows that Random Forest is the closest to perfection in terms of accuracy. Therefore, the better classifier for this dataset to be used for the system to be built or developed is, Random Forest.

REFERENCES

CINAR I., KOKLU M. and TASDEMIR S., (2020). Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods, Gazi Journal of Engineering Sciences, vol. 6, no. 3, pp. 200-209, December, 2020, DOI: https://doi.org/10.30855/gmbd.2020.03.03

Pant, R. (2020, December 1). Accuracy and its shortcomings: Precision, recall to the rescue. Analytics Vidhya. Retrieved January 27, 2023, from https://www.analyticsvidhya.com/blog/2020/12/accuracy-and-its-shortcomings-precision-recall-to-the-rescue/