American International University-Bangladesh (AIUB)

Department of Computer Science
Faculty of Science & Technology (FST)

Introduction to Data Science

Paper Review

Submitted to

Tohedul Islam

Assistant Professor, CSE

Submitted By

Semester: Spring_22_23	**Section: D**
Student Name	Student ID
AHNAF, MD. ALI	20-42378-1

Prediction of Wine Quality: Comparing Machine Learning Models in R Programming

In the past, wine quality was determined purely by taste or physical inspections, but in the current era, data science and machine learning are required for a comprehensive knowledge of the complicated data analysis of wine quality evaluation characteristics. Programming in R is not only informative in the academic realm, but it also yields enhanced outcomes for optimal forecasts in the commercial realm; thus, R is also employed for machine learning forecast.

With the development of diverse algorithms for ML models, machine learning methods have grown simpler, making it possible for researchers to produce accurate predictions using the appropriate and readily available tools. R, among other programming languages, is particularly good for predictive analytics.

Despite the numerous studies cited in the literature review, R as a statistical and machine learning tool has been frequently overlooked in earlier research. It was mentioned that the power of R is strongly tied to the availability of functions, algorithms, and flexible packages. In addition, during the years, many in industries have been hooked to other tools and have not explored the beauty of R; hence, bringing R's machine learning capabilities to the forefront of this paper is the key objective of this study.

The dataset collected from Kaggle comprises thirteen columns and is organized into three subsections describing the input variables as physio-chemical tests, the output variables as sensory data, and the label on the wine as wine ID. Figures in this study depict the dataset, including the data kinds, central tendency measure, and other characteristics. Id is just a label which was removed in the data pre-processing segment where the dataset is clean with no NA or voided spaces. After cleansing the data, a zero-linear connection between quality and other covariables was identified. When a linear link between an output and the inputs is not established, linear models are not thoroughly evaluated.

For making machine learning easier to apply, several R packages were utilized, including naivebayes, ggplot2, lattice, caret, dplyr, and psych. Due to the unbalanced nature of the data, it was divided into two parts: 20% for data validations and the remaining 80% for training and testing the models. In this work, the algorithms were run using k fold cross validation approach where the usual value of k = 10 is chosen hence this approach splits the data set in ten different parts with nine for training and one for testing. Basically, each of the k parts in turn is used as a test set and the other k − 1 parts are used as a training set.

The outputs data and input data were visualized via Histogram only for the wine quality where the data exploration method is Univariate. Further, for density plots for quality of wine and feature plots of the dataset, Multivariate plots have been used to explore the dataset. Naïve bayes classification results were not up to the mark because this classifier had higher margin of error range. Lastly, in Ten Fold (k) cross validation, LDA, CART, KNN, SVM with linear kernel Random Forest (RF) were considered in this papers work. After validating the approaches through Ten-Fold classifier it is evident that Random Forest (RF) has the highest confidence level which is 0.95 out of 1 and KNN had the least of accuracy and kappa values. In the following paper the training set was selected, and RF classifier had predicted the better in quality of wine. Here confusion matrix was generated which helped to predict overall accuracy on unseen instances which is often used to breakdown a classifiers performance. Here Random

Forest over the ten-fold classification had the highest accuracy compared to other models having the highest true positives. Therefore, R has easiest and simpler ways to perform good accuracy in predictive analytics which always gets overlooked.