A MINI PROJECT REPORT

ON

“Wine Quality Testing”

Submitted by

Ankita Chaudhari (Roll No. : 18)
Tanishka Borade (Roll No. : 17)

Under the guidance of

Prof. Vishal Patil

For the Subject

Laboratory Practice II (410247) -

Data Mining and Warehousing (410244 (D))

Submitted in partial fulfilment of the requirements
for the award of the degree of

Bachelor in Computer Engineering

THE MET LEAGUE OF COLLEGES
MET
AS SHARP AS YOU CAN GET
Bhujbal Knowledge City

Institute of Engineering
Department of Computer Engineering

Academic Year 2021-22

CONTENTS

1 PROBLEM STATEMENT

Consider a labeled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

2 ABSTRACT

Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality

Keywords - ***Machine Learning, Classification,Random Forest, SVM,Prediction.***

3 INTRODUCTION

As the quarantine continues, I've picked up a number of hobbies and interests… including WINE. Recently, I've acquired a taste for wines, although I don't really know what makes a good wine. Therefore, I decided to apply some machine learning models to figure out what makes a good quality wine!

For this project, I used Kaggle's Red Wine Quality dataset to build various classification models to predict whether a particular red wine is "good quality" or not. Each wine in this dataset is given a "quality" score between 0 and 10. For the purpose of this project, I converted the output to a binary output where each wine is either "good quality" (a score of 7 or higher) or not (a score below 7). The quality of a wine is determined by 11 input variables:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH

4 OBJECTIVE

The objectives of this project are as follows

1. To experiment with different classification methods to see which yields the highest accuracy
2. To determine which features are the most indicative of a good quality wine.

5 MODELLING

For this project, I wanted to compare five different machine learning models: decision trees, random forests, AdaBoost, Gradient Boost, and XGBoost. For the purpose of this project, I wanted to compare these models by their accuracy.

Model 1: Decision Tree

Decision trees are a popular model, used in operations research, strategic planning, and machine learning. Each square above is called a node, and the more nodes you have, the more accurate your decision tree will be (generally). The last nodes of the decision tree, where a decision is made, are called the leaves of the tree. Decision trees are intuitive and easy to build but fall short when it comes to accuracy.

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

Model 2: Random Forest

Random forests are an ensemble learning technique that builds off of decision trees. Random forests involve creating multiple decision trees using bootstrapped datasets of the original data and randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode of all of the predictions of each decision tree. What's the point of this? By relying on a "majority wins" model, it reduces the risk of error from an individual tree.

Model 3: AdaBoost

The next three models are boosting algorithms that take weak learners and turn them into strong ones. I don't want to get sidetracked and explain the differences between the three because it's quite complicated and intricate. That being said, I'll leave some resources where you can learn about AdaBoost, Gradient Boosting, and XGBoosting.

6 DATA PROCESSING METHODS

For making automated decisions on model selection we need to quantify the performance of our model and give it a score. For that reason, for the classifiers, Precision which expresses how accurate the model was on predicting a certain class and Recall which expresses the inverse of the regret of missing out instances which are misclassified. Since we have multiple classes we have multiple F1 scores. We will be using the unweighted mean of the F1 scores for our final scoring.

This is a business decision because we want our models to be optimized to classify instances that belong to the minority side, such as wine quality of 3 or 8 equally well with the rest of the qualities that are represented in a larger number. For the regression task we are scoring based on the coefficient of determination, which is basically a measurement of whether the predictions and the actual values are highly correlated. The larger this coefficient the better. For regressors we can also get an F1 score if we first round our prediction. Splitting for Testing : We are keeping 20% of our dataset to treat it as unseen data and be able and test the performance of our models. We are splitting our dataset in a way such that all of the wine qualities are represented proportionally equally in both training and testing dataset.

7 CONCLUSION

Based on the bar plots plotted we come to an conclusion that not all input features are essential and affect the data, for example from the bar plot against quality and residual sugar we see that as the quality increases residual sugar is moderate and does not change drastically. So this feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection.

We were able to achieve maximum accuracy using a random forest of 88%. Stochastic gradient descent giving an accuracy of 81% .SVC has an accuracy of 85% and logistic regression of 86%

8 REFERENCES

1. Yunhui Zeng1 , Yingxia Liu1 , Lubin Wu1 , Hanjiang Dong1. “Evaluation and Analysis Model of Wine Quality Based on Mathematical Model ISSN 2330-2038 E-ISSN 2330-2046,Jinan University, Zhuhai,China.
2. Paulo Cortez1, Juliana Teixeira1, Ant´onio Cerdeira2.“Using Data Mining for Wine Quality Assessment”.
3. Yesim Er*1 , Ayten Atasoy1. “The Classification of White Wine and Red Wine According to Their Physicochemical Qualities”,ISSN 2147-67992147-6799,3rd September 2016