Pune Institute of Computer Technology
Dhankawadi, Pune

DATA ANALYTICS MINI-PROJECT REPORT
ON
Predicting wine quality

SUBMITTED BY

Nikhil Jain	41425
Atharva Jagtap	41423
Vaibhav Marathe	41433

Class BE 4

Under the guidance of
Prof. D.D.Kadam

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2020-21

Contents

1 Problem Statement

Consider a labeled dataset belonging to an application domain. Apply suit-abledata preprocessing steps such as handling of null values, data reduction, discretiza-tion. For prediction of class labels of given data instances, build classifier modelsusing different techniques (minimum 3), analyze the confusion matrix and com-pare these models. Also apply cross validation while preparing the training andtesting datasets

2 Abstract

Nowadays, industries are using product quality certifications to promote their products. This is a time taking process and requires the assessment given by human experts which makes this process very expensive. This paper explores the usage of machine learning techniques such as linear regression, neural network and support vector machine for product quality in two ways. Firstly, determine the dependency of target variable on independent variables and secondly, predicting the value of target variable. In this paper, linear regression is used to determine the dependency of target variable on independent variables. On the basis of computed dependency, important variables are selected those make significant impact on dependent variable. Further, neural network and support vector machine are used to predict the values of dependent variable. All the experiments are performed on Red Wine and White Wine datasets. This paper proves that the better prediction can be made if selected features (variables) are being considered rather than considering all the features.

3 H/W & S/W Requirements:

OS: Windows 10/ Ubuntu
LTSRAM: 4GBHard
Drive: 500 GB
Eclipse
Required Java libraries

4 Introduction

Classification is a process of categorizing a given set of data into classes, Itcan be performed on both structured or unstructured data. The process startswith predicting the class of given data points. The classes are often referred to astarget, label or categories

The classification predictive modeling is the task of approximating the mappingfunction from input variables to discrete output variables. The main goal is toidentify which class/category the new data will fall into.

Since classification is a type of supervised learning, even the targets are alsoprovided with the input data. Let us get familiar with the classification in machinelearning terminologies.

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes. The most common classification problems are – speech recognition, face detection, handwriting recognition, documentclassification, etc. It can be either a binary classification problem or a multiclassproblem too. There are a bunch of machine learning algorithms for classification inmachine learning. Let us take a look at those classification algorithms in machinelearning.

It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with adichotomous variable meaning it will have only two possible outcomes.The goal of logistic regression is to find a best-fitting relationship between the de-pendent variable and a set of independent variables. It is better than other binaryclassification algorithms like nearest neighbor since it quantitatively explains thefactors leading to classification

5 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

6 Scope

Data mining nowadays is most important technique which is utilized for investigation of the archives. It looks at the information and produces the required yield. With the headway in the innovation it helps in playing the sound test in the market thus benefits the client.

7 Test Cases

```
Rows, columns: (1599, 12)
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Figure 1

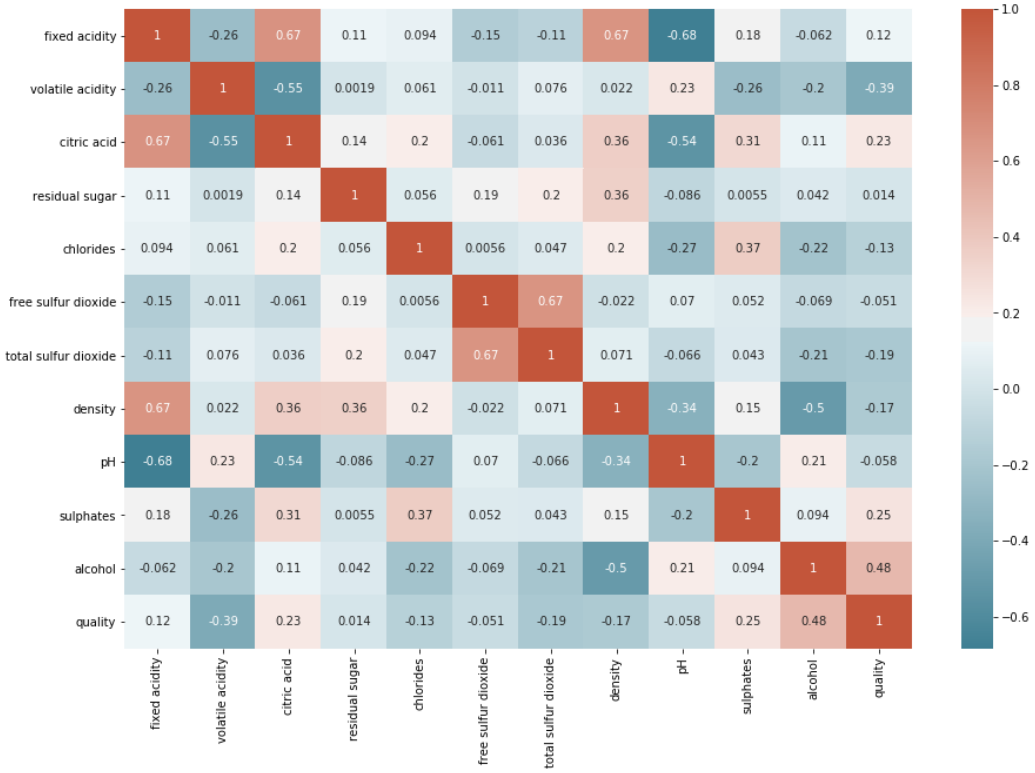

Figure 2

	precision	recall	f1-score	support
0	0.96	0.92	0.94	355
1	0.53	0.73	0.62	45
accuracy			0.90	400
macro avg	0.75	0.83	0.78	400
weighted avg	0.92	0.90	0.90	400

Figure 3

	precision	recall	f1-score	support
0	0.95	0.97	0.96	355
1	0.68	0.58	0.63	45
accuracy			0.92	400
macro avg	0.82	0.77	0.79	400
weighted avg	0.92	0.92	0.92	400

Figure 4

	precision	recall	f1-score	support
0	0.94	0.94	0.94	355
1	0.51	0.49	0.50	45
accuracy			0.89	400
macro avg	0.72	0.71	0.72	400
weighted avg	0.89	0.89	0.89	400

Figure 5

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	goodquality
count	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.000000	217.0
mean	8.847005	0.405530	0.376498	2.708756	0.075912	13.981567	34.889401	0.996030	3.288802	0.743456	11.518049	7.082949	1.0
std	1.999977	0.144963	0.194438	1.363026	0.028480	10.234615	32.572238	0.002201	0.154478	0.134038	0.998153	0.276443	0.0
min	4.900000	0.120000	0.000000	1.200000	0.012000	3.000000	7.000000	0.990640	2.880000	0.390000	9.200000	7.000000	1.0
25%	7.400000	0.300000	0.300000	2.000000	0.062000	6.000000	17.000000	0.994700	3.200000	0.650000	10.800000	7.000000	1.0
50%	8.700000	0.370000	0.400000	2.300000	0.073000	11.000000	27.000000	0.995720	3.270000	0.740000	11.600000	7.000000	1.0
75%	10.100000	0.490000	0.490000	2.700000	0.085000	18.000000	43.000000	0.997350	3.380000	0.820000	12.200000	7.000000	1.0
max	15.600000	0.915000	0.760000	8.900000	0.358000	54.000000	289.000000	1.003200	3.780000	1.360000	14.000000	8.000000	1.0

Figure 6

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	goodquality
count	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.000000	1382.0
mean	8.236831	0.547022	0.254407	2.512120	0.089281	16.172214	48.285818	0.996859	3.314616	0.644754	10.251037	5.408828	0.0
std	1.682726	0.176337	0.189665	1.415778	0.049113	10.467685	32.585604	0.001808	0.154135	0.170629	0.969664	0.601719	0.0
min	4.600000	0.160000	0.000000	0.900000	0.034000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000	0.0
25%	7.100000	0.420000	0.082500	1.900000	0.071000	8.000000	23.000000	0.995785	3.210000	0.540000	9.500000	5.000000	0.0
50%	7.800000	0.540000	0.240000	2.200000	0.080000	14.000000	39.500000	0.996800	3.310000	0.600000	10.000000	5.000000	0.0
75%	9.100000	0.650000	0.400000	2.600000	0.091000	22.000000	65.000000	0.997900	3.410000	0.700000	10.900000	6.000000	0.0
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	165.000000	1.003690	4.010000	2.000000	14.900000	6.000000	0.0

Figure 7

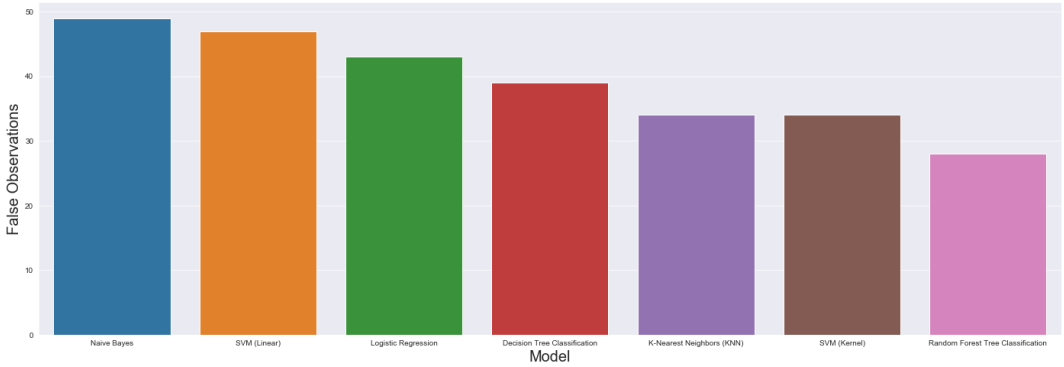

Figure 8

8 Result

Nowadays people used to consume red wine either as a necessity or for show off. All this results in health loss. Hence to preserve human health it became essential to predict the red wine quality before its consumption. Different machine learning algorithms are executed on the dataset in RStudio software. Accuracy is calculated and the best algorithm is predicted for a given dataset.

9 Conclusion

In this project we have implemented different classification models to see whichyields the highest accuracy and also compared different classification models. We have applied suitable data preprocessing steps such as handling of null values, datareduction, discretization.Also we have applied cross validation while preparing the training and testing datasets and analyzed the confusion matrix

References

[1] https://stackoverflow.com/
[2] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
[3] https://docs.python.org/3/library/