Module Code & Module Title

CU6051NT Artificial Intelligence

Assessment Weightage & Type

80% Individual Coursework

Year and Semester

2020-21 Autumn Year Long

Student Name: Girija Tamang

London Met ID: 18030995

Title: Red Wine Quality Prediction

Assignment Due Date: 07 Feb 2021

Assignment Submission Date: 07 Feb 2021

Module Tutor: Weenit Maharjan

I confirm that I understand my coursework needs to be submitted online via Google Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

Abstract

This report is about “Red Wine Quality Prediction” using random forest algorithm. The report includes literature review of similar projects and briefly explains the AI concepts and terminologies used in the proposed solution. The flowchart, pseudocode, development tool and the explanation of used algorithms with work analysis, and the future work is presented in the report. The main aim of this project is to predict the quality of red wine and help wine industries in checking the quality of wines for making good wine products in the future.

Table of Contents

Tables of Figures

1. Introduction

1.1 Introduction to the AI concepts used.

Machine Learning

Machine Learning at its most basic level is the practice of using algorithms to parse data, learn from it and then make a determination or prediction about something in the world. Machine learning is the branch of Artificial intelligence where a system is trained with a set of datasets or patterns using various mathematical models and algorithms to make a machine capable of making decisions or predictions independently without being explicitly programmed for performing the given task (Rouse, 2019).

Random Forest Algorithm

Random forest is a popular supervised learning algorithm used for both classification and regression problems in machine learning. Random Forest is a classifier that includes a number of decision trees on different subsets of the dataset specified and takes the average to increase the predictive accuracy of the dataset. It is an ensemble technique that is better than a single decision tree because by averaging the result, it decreases the over-fitting.

Here are some points illustrating why the algorithm of the Random Forest is used:

- As compared to other algorithms, it takes less training time.
- Even for the large dataset it manages effectively, it predicts highly accurate results.
- It can also preserve precision when a significant proportion of the data is missing.
- Random forest has less variance than a single decision tree (Tutorials Point, 2019).

1.2 Introduction of the chosen problem domain

Red Wine Quality Prediction

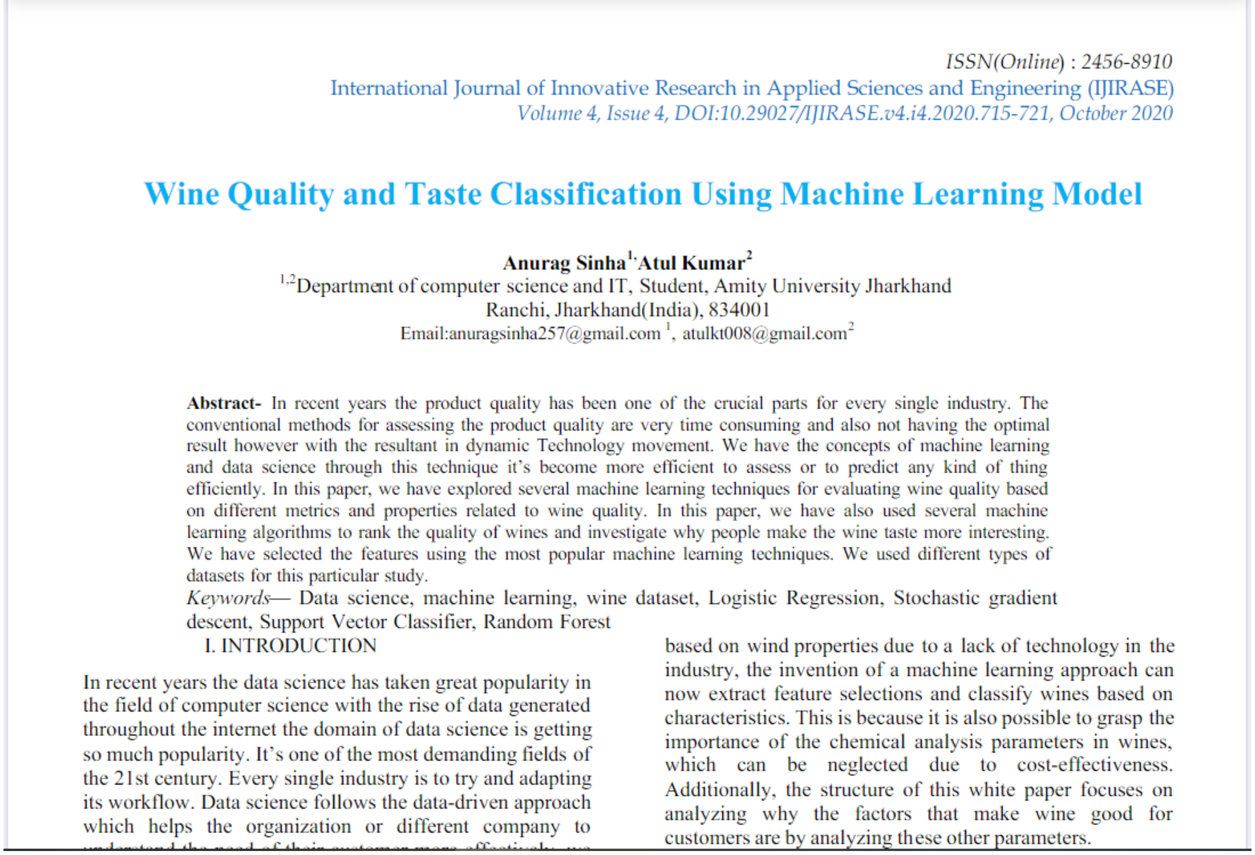
Product quality has been one of the essential components of any single industry in the recent years. Since the last decade, the wine industry has been rising well in the market. As wine demand has risen in recent years, the consumption of wine has also increased. With increasing demand, the quality checking of wine has been the major problem faced by wine industries. Wine quality is generally measured by professional tasters in the wine industry who render their decision based on several sensory criteria, such as color, taste, and odor, which is very complex and time-consuming since wine demands have been increasing worldwide.

Understanding the criteria of wine quality testing in industries can be a challenging activity for a laboratory with a wide variety of analyses and residues to track. Different kinds of machine learning algorithms should be implemented by wine industries for analyzing taste and other properties in wine. Machine learning makes it more effective to test or predict any kind of thing effectively, to find the quality of wine in a short time without the need for any human expertise.

2. Background

Some significant work has been done in the field of wine quality prediction. In the past decade, academic papers on the very subject have proliferated. To offer a clear understanding of wine quality prediction techniques and their implementation, this proposal analyses the various papers written.

1. Wine Quality and Taste Classification Using Machine Learning Model



ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

Wine Quality and Taste Classification Using Machine Learning Model

Anurag Sinha[1],**Atul Kumar**[2]
[1,2]Department of computer science and IT, Student, Amity University Jharkhand
Ranchi, Jharkhand(India), 834001
Email:anuragsinha257@gmail.com [1], atulkt008@gmail.com[2]

Abstract- In recent years the product quality has been one of the crucial parts for every single industry. The conventional methods for assessing the product quality are very time consuming and also not having the optimal result however with the resultant in dynamic Technology movement. We have the concepts of machine learning and data science through this technique it's become more efficient to assess or to predict any kind of thing efficiently. In this paper, we have explored several machine learning techniques for evaluating wine quality based on different metrics and properties related to wine quality. In this paper, we have also used several machine learning algorithms to rank the quality of wines and investigate why people make the wine taste more interesting. We have selected the features using the most popular machine learning techniques. We used different types of datasets for this particular study.

Keywords— Data science, machine learning, wine dataset, Logistic Regression, Stochastic gradient descent, Support Vector Classifier, Random Forest

I. INTRODUCTION

In recent years the data science has taken great popularity in the field of computer science with the rise of data generated throughout the internet the domain of data science is getting so much popularity. It's one of the most demanding fields of the 21st century. Every single industry is to try and adapting its workflow. Data science follows the data-driven approach which helps the organization or different company to understand the need of their customer more effectively, we

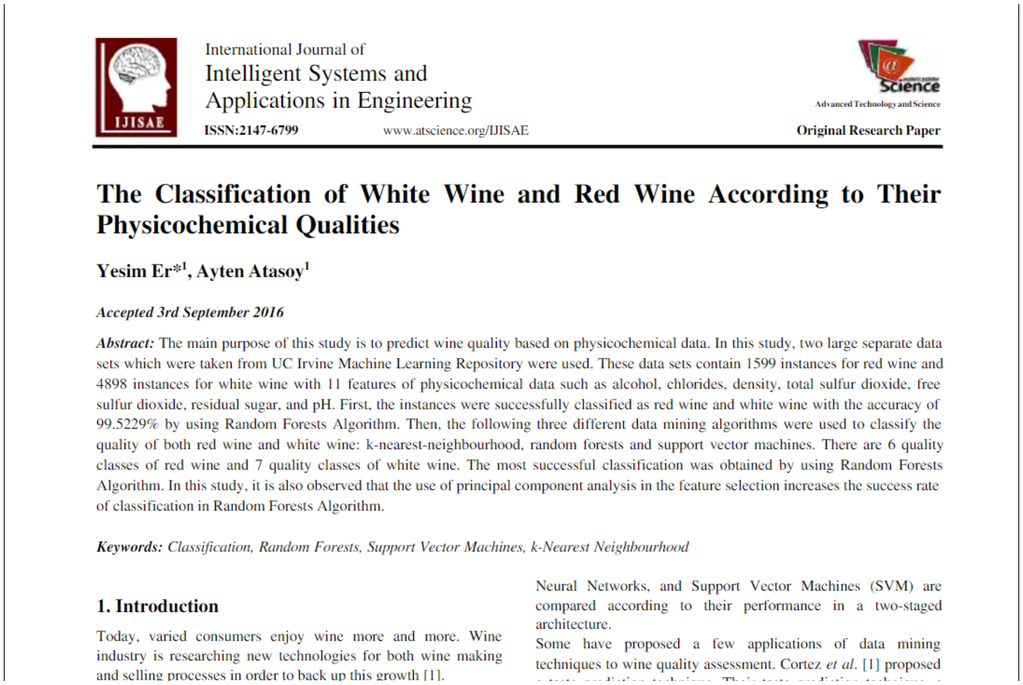
based on wind properties due to a lack of technology in the industry, the invention of a machine learning approach can now extract feature selections and classify wines based on characteristics. This is because it is also possible to grasp the importance of the chemical analysis parameters in wines, which can be neglected due to cost-effectiveness. Additionally, the structure of this white paper focuses on analyzing why the factors that make wine good for customers are by analyzing these other parameters.

Figure 1: Article on Wine Quality and Taste Classification Using Machine Learning Model

Research on wine quality and taste classification using machine learning has been done by Anurag Sinha and Atul Kumar. In this paper, several machine learning techniques were explored to determine wine quality based on different parameters and properties related to wine quality. In this paper, they have used logistic regression, Stochastic descent of the gradient, support Vector classifier and Random forest machine learning algorithms for classification of wines.

Logistic regression and Random Forest provided 86% and 87.33% accuracy in predicting quality of wines. High quality of wine is usually associated with low levels of volatile acidity (Sinha & Kumar, 2010). They have used various datasets of wines for doing this research which was actually a good idea for assuring wine quality and taste classification. In this paper, they have described the used dataset clearly but there was no proper description of the algorithm used. This paper helps me to understand and select a dataset to predict the quality of red wine.

2. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities



International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

The Classification of White Wine and Red Wine According to Their Physicochemical Qualities

Yesim Er*[1], Ayten Atasoy[1]

Accepted 3rd September 2016

Abstract: The main purpose of this study is to predict wine quality based on physicochemical data. In this study, two large separate data sets which were taken from UC Irvine Machine Learning Repository were used. These data sets contain 1599 instances for red wine and 4898 instances for white wine with 11 features of physicochemical data such as alcohol, chlorides, density, total sulfur dioxide, free sulfur dioxide, residual sugar, and pH. First, the instances were successfully classified as red wine and white wine with the accuracy of 99.5229% by using Random Forests Algorithm. Then, the following three different data mining algorithms were used to classify the quality of both red wine and white wine: k-nearest-neighbourhood, random forests and support vector machines. There are 6 quality classes of red wine and 7 quality classes of white wine. The most successful classification was obtained by using Random Forests Algorithm. In this study, it is also observed that the use of principal component analysis in the feature selection increases the success rate of classification in Random Forests Algorithm.

Keywords: *Classification, Random Forests, Support Vector Machines, k-Nearest Neighbourhood*

1. Introduction

Today, varied consumers enjoy wine more and more. Wine industry is researching new technologies for both wine making and selling processes in order to back up this growth [1].

Neural Networks, and Support Vector Machines (SVM) are compared according to their performance in a two-staged architecture.

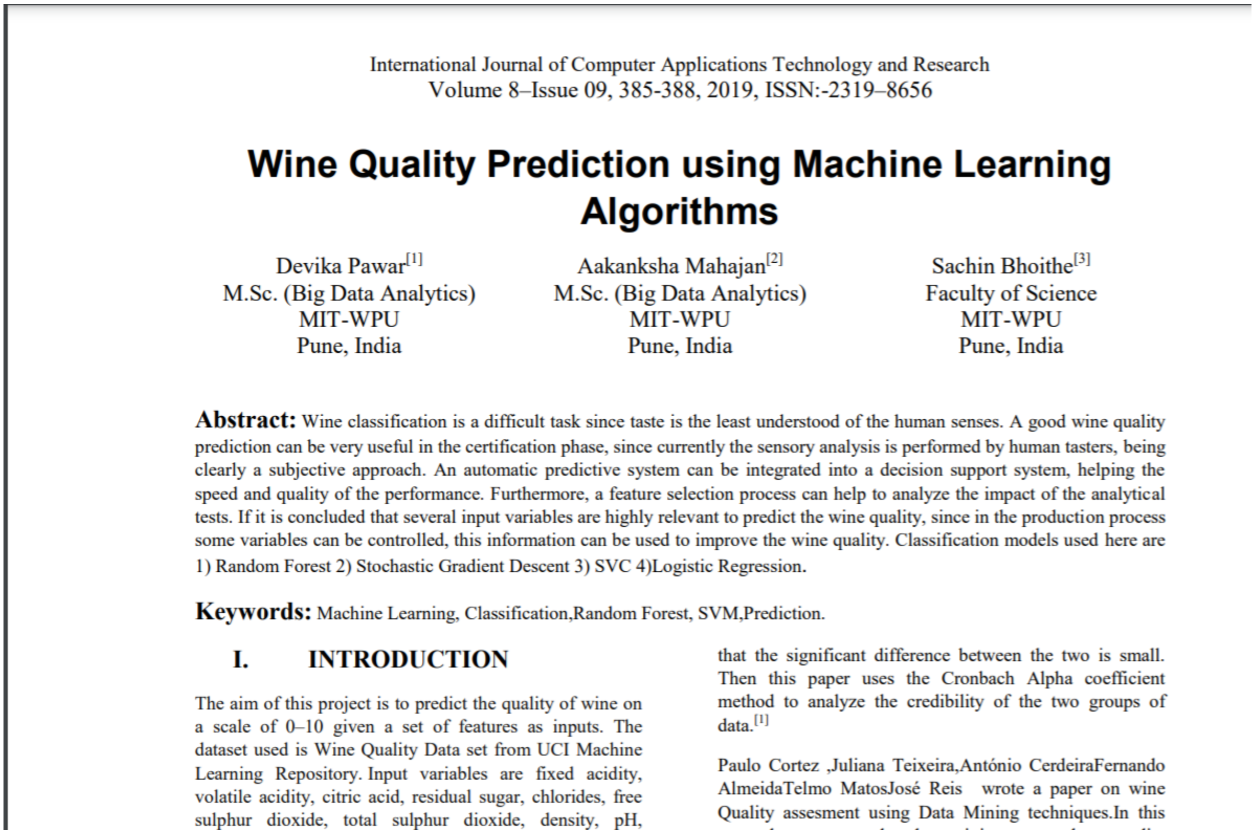
Some have proposed a few applications of data mining techniques to wine quality assessment. Cortez *et al.* [1] proposed

Figure 2: Research Paper on The Classification of White Wine and Red Wine According to Their Physicochemical Qualities.

Research of this article was performed by Yesim Er and Ayten Atasoy in the International Journal of Intelligent Systems and Applications in Engineering. Predicting the quality of wine based on physicochemical data was the main objective of this research paper. In this study, two different large data sets taken from the UC Irvine Machine Learning Repository were used. They have used both red and white dataset for classification of wine. In this report they have used k-nearest-neighborhood, random forests, and support vector machines classifier for evaluating the datasets of both red and white wine.

Using the Random Forests Algorithm, the cases were successfully categorized as red wine and white wine with an accuracy of 99.5229 percent (Er & Atasoy, 2016). In this article they have properly described the dataset and all the used machine learning algorithms. This article helps me to choose a random forest algorithm for quality prediction of red wine.

3. Wine Quality Prediction using Machine Learning Algorithms



International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 385-388, 2019, ISSN:-2319–8656

Wine Quality Prediction using Machine Learning Algorithms

Devika Pawar[1]
M.Sc. (Big Data Analytics)
MIT-WPU
Pune, India

Aakanksha Mahajan[2]
M.Sc. (Big Data Analytics)
MIT-WPU
Pune, India

Sachin Bhoithe[3]
Faculty of Science
MIT-WPU
Pune, India

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. Classification models used here are 1) Random Forest 2) Stochastic Gradient Descent 3) SVC 4)Logistic Regression.

Keywords: Machine Learning, Classification,Random Forest, SVM,Prediction.

I. INTRODUCTION

The aim of this project is to predict the quality of wine on a scale of 0–10 given a set of features as inputs. The dataset used is Wine Quality Data set from UCI Machine Learning Repository. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH,

that the significant difference between the two is small. Then this paper uses the Cronbach Alpha coefficient method to analyze the credibility of the two groups of data.[1]

Paulo Cortez ,Juliana Teixeira,António CerdeiraFernando AlmeidaTelmo MatosJosé Reis wrote a paper on wine Quality assesment using Data Mining techniques.In this

Figure 3: Article on Wine Quality Prediction using Machine Learning Algorithms.

Devika Pawar, Aakanksha Mahajan and Sachi Bhoithe wrote a paper on wine quality prediction using machine learning algorithms. Random Forest, Stochastic Gradient Descent, Logistic Regression are the classification models they have used for predicting the wine quality. They were able to achieve optimum precision using random forests of 88 percent where Stochastic gradient, SVC, and logistic regression were able to provide 81 %, 85%, and 86% accuracy, respectively.

This article is well prepared for learning basic machine learning concepts with various algorithms (Pawar, et al., 2019). This paper presented a proper explanation of the data set and past work done to predict the quality of wine. There was a lack of a proper explanation of how machine learning algorithms operate when predicting the quality of wine.

3. Solution

3.1 Explanation of the proposed solution

The checking and predicting of wine quality have been the major problem faced by wine industries. After doing a lot of research, I found that there are many machine learning algorithms to determine wine quality based on various wine quality parameters and properties. I have chosen a random forest algorithm for red wine quality prediction. Here are the steps that are followed while testing quality of wines:

1. Data Collection of Red Wine from public datasets.

2. Data preparation for building models.

3. Feature selection

4. Implementing machine learning techniques

5. Comparison of performance.

6. Interpretation of results

Data set Information

The dataset is related to red variants of the Portuguese "Vinho Verde" wine. Only physicochemical (inputs) and sensory (output) variables are accessible due to privacy and logistical problems (example: there is no data about grape types, wine brand, wine selling price, etc.). The classes are ordered and not balanced. This dataset can be viewed as classification or regression tasks (P. Cortez, 2019).

Attribute Information: Input variables (based on physicochemical tests):

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

10. sulphates

11. alcohol

Output variable (based on sensory data):

12. quality (score between 0 and 10).

After finding the dataset, I have performed initial visual analysis of data present in the dataset. So, I got a concise summary of the data frame like information from the data set, whether there are null values available or not in the data frame, shows the data types and detailed information of the columns. I will visualize the relationship between quality and the other columns using subplot. By using label encoding, I will categories the label data of quality of good or bad. I will assign 1 to good and 0 to bad quality of wine. After that I will separate the dataset as response variables and feature variables. I will split the data to both testing and training data sets and then I will standardize the data. Our training and testing data will be ready after standardization. I will perform classification using a random forest classifier machine learning algorithm and see how the developed model will perform.

3.2 Explanation of the AI algorithm used.

The random forest classifier is used for predicting the quality of red wine after preparing training and testing data from the red wine dataset. Random Forest is a popular algorithm for machine learning based on the principle of ensemble learning, which is a method of combining multiple classifiers to solve a complex problem and improve the model's output. Random Forest is a classifier that comprises a number of decision trees and takes the average to boost the predictive accuracy of that dataset on different subsets of the given dataset. The random forest takes the prediction from each tree and is based on the majority votes of predictions rather than depending on a decision tree and predicts the final output (JavaTpoint, 2019).

In two phases, Random Forest operates, first by combing the N decision tree to create the random forest, and secondly by making predictions for each tree generated in the first phase. The Working process can be explained in the below steps:

Step-1: Choose a random K data point from the training set.

Step-2: Create decision trees associated with the data points selected (Subsets)

Step-3: Choose the number N for the decision trees you want to create.

Step-4: Repeat step 1 and step 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes (Sharma, 2019).

The following diagram will illustrate the working of a random forest algorithm.

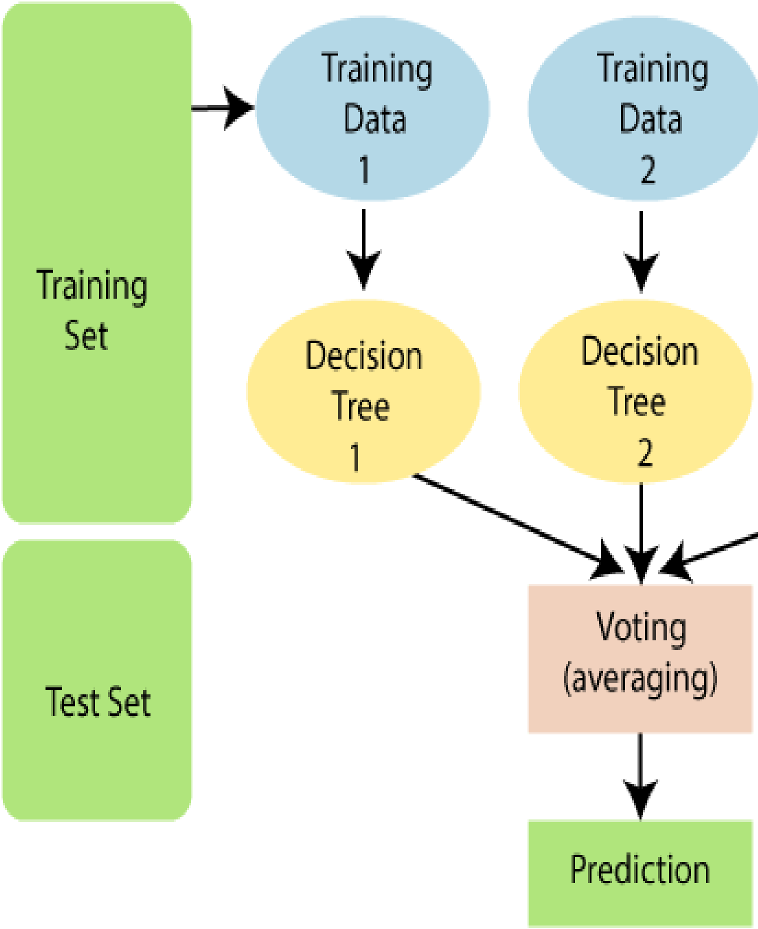


Figure 4:Working of Random Forest Algorithm.

3.3 Pseudocode of the solution

Step 1: Import Necessary Libraries (pandas, numpy, matplotlib, sklearn)

Step 2: Load the data set for getting the data of wine.

df = pd.read_csv('filename.csv')

Step 3: Check how the data set looks or distributed by using dot info attribute and to know how the columns of data are distributed in the dataset, do some plotting.

Step 4: Divide wine as good and bad by giving the limit for the quality and assign a label to the quality variable using label encoding.

Step 5: Separate the data as response variables and feature variables.

Example: X = wine.drop('quality', axis = 1)

y = wine['quality']

Step 6: Split the data into training and testing sets and apply standard scaling to get optimized results. The training and testing data will be ready to perform with machine learning algorithms.

Step 7: Now use a random forest classifier for predicting the quality of red wine.

Example: RF = RandomForestClassifier(n_estimators=100)

RF.fit(X_train, Y_train)

pred_RF = RF.predict(X_test)

Step 8: Calculate the accuracy score with the target variable and with the predicted target variable using random forest classifier.

3.4 Diagrammatical representations of the solution

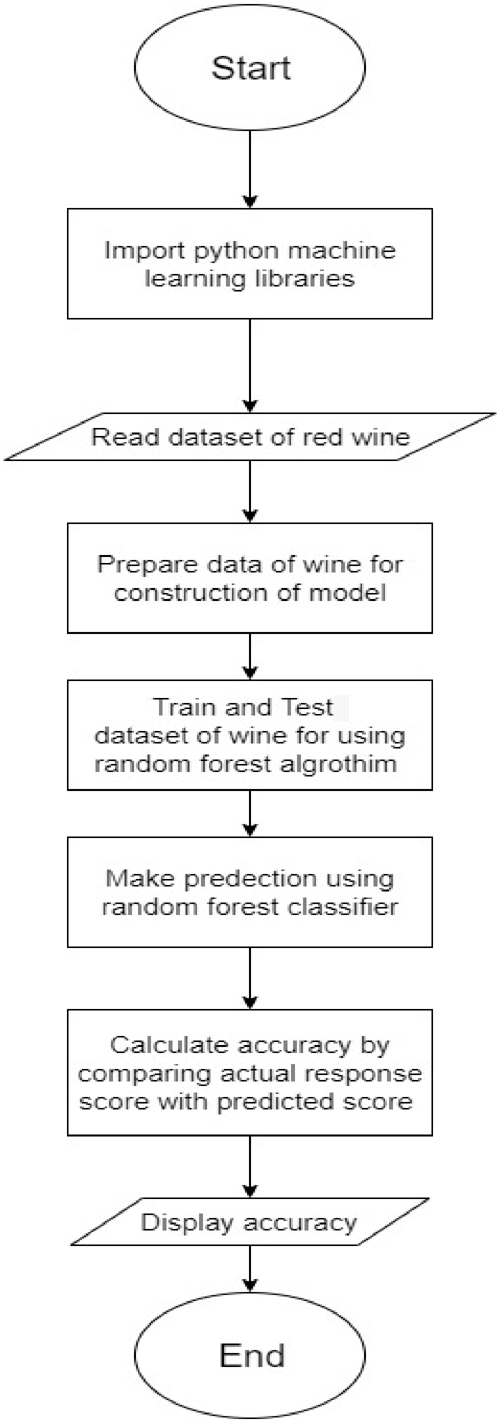

Start
Import python machine learning libraries
Read dataset of red wine
Prepare data of wine for construction of model
Train and Test dataset of wine for using random forest algrothim
Make predection using random forest classifier
Calculate accuracy by comparing actual response score with predicted score
Display accuracy
End

Figure 5: Flowchart of red wine quality prediction system.

3.5 Explanation of the development, tools and technologies used.

As supervision learning is the selected approach, the development process starts by collecting labeled dataset to prediction of sentiment. Below is all tools and libraries used for development this project:

Libraries/Tools	**Description**
Pandas	It is a python package for data manipulation and analysis.
Numpy	Numpy is a Python library used for working with arrays. For calculations such as means, medians, square roots, etc., Numpy is required.
Scikit-learn.	It is a library in python with many supervised and unsupervised algorithms.

For better understanding the development process of training the model and predicting the quality of red wines was divided into following steps:

1. Data Collection of Red Wine from public datasets.

From Kaggle.com, which is a massive online group of data scientists and computer learners, a dataset containing red wine knowledge was extracted. There are several published datasets, and one of the datasets published on this website was the dataset used in this project.

2. Data reading and data cleaning.

Pandas module is used for reading files. The data in '.csv' format.'read_csv()' function is used for loading the data. By using head() method data is check how it looks. Dataset data were checked for duplicates, null values, missing values, etc. And then check how the data set looks or distributed by using dot info attribute and to know how the columns of data are distributed in the dataset, do some plotting.

3. Data preparation for building models.

By using bins the quality of wine is divided as good and bad by giving the limit for the quality and assign a label to the quality variable using label encoding. Bad wine quality becomes 0 and good wine quality becomes 1.

4. Split the prepared data into training and testing set.

After the data-pre-processing is complete. Two data frames are created one(x) will be the original data frame with the target variable quality being removed and other (y) will contain the series df quality. Then the data are split to both testing and training data. The training and testing set are separated with test size of 0.2. using Trane Tesla (Train_test_split from Scikitlearn) which means we are getting 80% training data and then the remaining 20% will be the testing data. Then standardize the data is done.

5. Fit the training data set into model.

After successfully splitting data, classification will be done using random forest algorithms. Random forest classifier was instantiated as rf with the number of trees being hundred. Rf is used to fit training set and the target variable of the training data set, and I will predict this for the testing data set.

6. Test data to calculate prediction.

After the model has been trained the test data is used to calculate the accuracy score. An accuracy score is calculated using random forest classifier with the target variable and with the predicted target variable.

Tools used for the development:

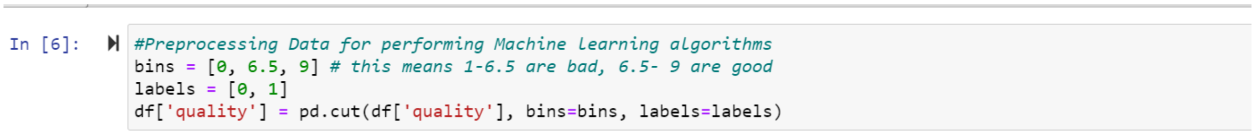
• Anaconda • Jupyter Lap Notebook

• Python 3

Above is the complete explanation of development steps and the tools/libraries used for development. It explains the process of creating a wine predicting model using machine learning algorithm.

3.6 Achieved Result.

Gathering the training and testing data.

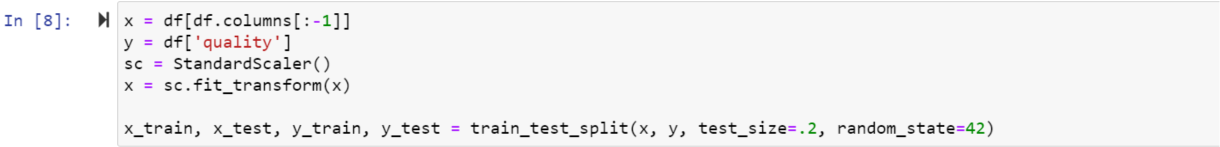


```
#Preprocessing Data for performing Machine Learning algorithms
bins = [0, 6.5, 9] # this means 1-6.5 are bad, 6.5- 9 are good
labels = [0, 1]
df['quality'] = pd.cut(df['quality'], bins=bins, labels=labels)
```

Figure 6: Gathering the training and testing data.

Here the data are preprocessing for performing the machine learning algorithms. All the wine are divided in good or bad.

Splitting data into training and testing data.



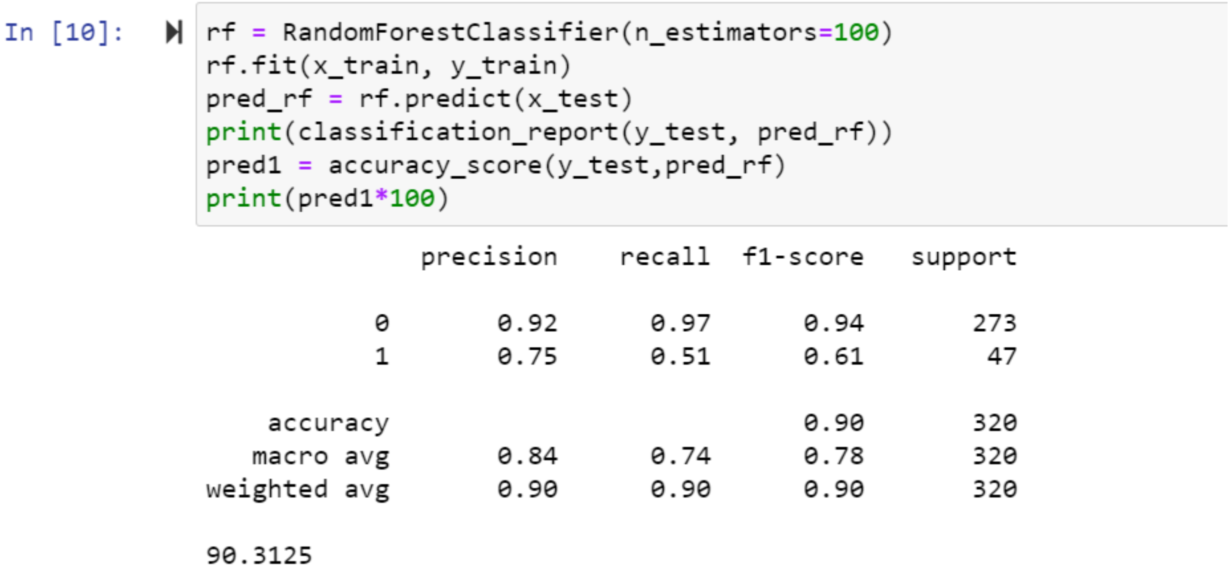
```
x = df[df.columns[:-1]]
y = df['quality']
sc = StandardScaler()
x = sc.fit_transform(x)

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=.2, random_state=42)
```

Figure 7: Splitting data.

Here the data are split to both testing and training data. The training and testing set are separated with test size of 0.2. using Trane Tesla (Train_test_split from Scikitlearn) which means we are getting 80% training data and then the remaining 20% will be the testing data. Then standardize the data is done.

Predicting accuracy using classifier

A random forest classifier is used for predicting the quality of red wine.



```
In [10]:
rf = RandomForestClassifier(n_estimators=100)
rf.fit(x_train, y_train)
pred_rf = rf.predict(x_test)
print(classification_report(y_test, pred_rf))
pred1 = accuracy_score(y_test,pred_rf)
print(pred1*100)
```

```
              precision    recall  f1-score   support

           0       0.92      0.97      0.94       273
           1       0.75      0.51      0.61        47

    accuracy                           0.90       320
   macro avg       0.84      0.74      0.78       320
weighted avg       0.90      0.90      0.90       320

90.3125
```

Figure 8:Using random forest classifier for predicting the accuracy score.

4. Conclusion

4.1 Analysis of the work done.

With the growth of data generated around the internet, the field of data science has gained so much attention in the field of computer science. Data science has become very popular. A brief description of AI is given in this article, highlighting its influence on other distinct fields. How machine learning techniques is used to makes machine or a software achieve the highest accuracy for predicting the quality of wine is explained in this report.

Wine quality prediction has been implemented in the introductory section of this project; with an overview of the methods, it takes to address various problem areas. Some research works conducted in wine quality prediction has been included in the background section with taken procedures and the result of their research.

In this report, the red wine quality prediction system is proposed which may help wine industries for checking and predicting wine quality. The random forest algorithm is used for classification of red wine. This report offered a good understanding of the value of quality prediction attributes using functions selected on the algorithm, which was time consuming and costly when performed in the conventional way. This report has the solution and dataset information for predicting the red wine quality with the proper description of the algorithm that is used. In the solution section a clear concept of machine learning algorithms is described to develop an application for predicting red wine quality.Pseudocode and flowchart of the algorithm have been included in the report, which can be used during actual implementation of the algorithm.

This project may face some limitation such as:

1. Since the prediction is entirely data-based, accuracy is highly dependent on it.

2. Model was not checked with real-time data on red wines and this project has not focused on white wine.

3. The system may be ideal for factories and not for local citizens.

4.2 How the solution addresses real world problems.

This paper discussed the use of machine learning techniques to predict the quality of Red Wine. The feature selection algorithm provided a clear idea about the importance of the attributes for prediction of quality, which was time consuming and expensive when done in the traditional way. Random forest machine learning algorithms should be implemented by wine industries for analyzing taste and other properties in wine. It makes it more effective to test, predict or find the quality of wine in a short time without the need for any human expertise.

By leveraging Machine Learning and data analysis of wine quality datasets through training, prediction and evaluation using Random Forests, this paper helps to solve key problems and predicts the quality of each wine sample that may be poor, medium, or high. Leading wine brands companies will be able to work faster and predict wine quality more accurately by integrating machine learning algorithms into their current systems and analytics. With the help of this application the wine industry will be to evaluate taste and other properties of wine in order to stay competitive in the future. This project is very useful for any other companies to track product performance, identify necessary changes and all kinds of insights These are only some real-world areas that wine analysis can benefit or has been benefiting. It can be applied to many other aspects of business, from brand monitoring to product analytics, from customer service to market research.

4.3 Further work

This report has only touched the surface of red wine quality prediction. In the future, this project could be improved in several respects. To some degree, the prediction made by this system is reliable, but the prediction of systems can be made more precise by gathering more real-time data and other training attributes that influence it.

In the future, I will try other machine learning methods for a better comparison of outcomes and try to solve the limitation of this project. I will try to add other machine learning algorithms for predicting the quality of the different types of wines based on certain attributes which will be helpful for industries to make good products of wine in the future. This is very useful for any other companies to track product performance, identify necessary changes and all kinds of insights.

5. References

Er, Y. & Atasoy, A., 2016. The Classification of White Wine and Red Wine According to TheirPhysicochemical Qualities. *Intelligent Systems andApplications in Engineering,* I(2147), pp. 23-26.

JavaTpoint, 2019. *Random Forest Algorithm.* [Online] Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm [Accessed 06 January 2021].

P. Cortez, A. C. F. A. T. M. a. J. R., 2019. *Wine Quality Data Set.* [Online] Available at: https://archive.ics.uci.edu/ml/datasets/wine+quality [Accessed 13 January 2021].

Pawar, D., Mahajan, A. & Bhoithe, S., 2019. Wine Quality Prediction using Machine Learning. *International Journal of Computer Applications Technology and Research,* 8(09), pp. 385-388.

Rouse, M., 2019. *Machine Learning.* [Online] Available at: https://searchenterpriseai.techtarget.com/definition/machine-learning-ML [Accessed 5 January 2021].

Sharma, N., 2019. Quality Prediction of Red Wine based on Different. *Department of Computer Science & Engineering,* 4(5), pp. 20-34.

Sinha, A. & Kumar, A., 2010. Wine Quality and Taste Classification Using Machine Learning Model. *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE),* 4(4), pp. .715-721.

Tutorials Point, 2019. *Classification Algorithms - Random Forest.* [Online] Available at: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm
[Accessed 5 January 2021].