Data Science Machine Learning Project

Wine Quality Prediction

Introduction

> *” Wine is the most healthful and most hygienic of beverages “*
> - Louis Pasteur

This is my first time writing a blog on any project. According to experts, wine is differentiated according to its **smell**, **flavor**, and **color,** but we are not wine experts to say that wine is good or bad. What will we do then? Here’s the use of **Machine Learning** comes, yes now we are using machine learning to check wine quality. ML has some techniques that will predict wine quality.

Now, I’m very excited to show my R&D over one of my favorite projects. Here, I have Wine Quality Dataset. In this dataset, we have to find the best quality wine from the collection of different wines. Let’s begin.

Problem Statement

A company is producing many wines and he wants to know how the quality of the wine is good? The company approaches a Data Scientist like me and I see there are different chemical parameters like fixed acidity, volatile acidity, citric acid, etc. I used these chemical parameters to build a Machine Learning algorithm to predict the best wine. First, describe the dataset:

- **volatile acidity:** Volatile acidity *is the* gaseous acids present in wine.
- **fixed acidity:** Primary **fixed acids** found in wine are **tartaric**, **succinic**, **citric**, and **malic**
- **residual sugar:** Amount of sugar left after fermentation.
- **citric acid:** It is a weak organic acid, found in citrus fruits naturally.
- **chlorides:** Amount of salt present in wine.
- **free sulfur dioxide:** So2 is used for the prevention of wine by oxidation and microbial spoilage.
- **total sulfur dioxide**
- **pH:** In wine, pH is used for checking the acidity
- **density**
- **sulphates**: Added sulfites preserve freshness and protect **wine** from oxidation, and bacteria.
- **alcohol:** Percent of alcohol present in wine.

Data Analysis And EDA

We first import necessary libraries like NumPy, pandas, seaborn, matplot, etc. And sklearn.model_selection, Random Forest Classifier, metrics, accuracy score, etc. for further machine learning model also. After this, we load our wine dataset from the

web. Here we see there are different features present to check the wine quality. There are 12 columns where quality is the target and the rest 11 are different chemical parameters. By using these 11 features we have to calculate the target variable. 11 features are as follows: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, ph, sulfates, alcohol. These are the chemical parameters that affect wine quality. In the quality column, we see different numeric values given for their quality like 1 for poor and 10 for the finest quality, and so on. Now we use the 'describe' method which finds the total count value, mean value, standard deviation, the minimum and maximum value of columns, and also finds quartile values like 25%,50%,75%, etc. Now by using the "value count" method on the quality column we find that medium-quality wine is more than low or high-quality wine. Now we do some R&D that which column is affect wine quality more. Here we find if volatile acidity is increased the wine quality decreases or we can say that both are inversely proportional. Now we check the relation between citric acid and wine quality, here we find direct proportion relation. If the wine contains more and more citric acid then the quality of wine is good. Now we check the correlation between all columns. Correlation shows two types of relation, one is a positive correlation and the second is a negative correlation. A negative correlation shows inversely proportion between columns and a positive correlation shows direct proportion. We show this correlation by seaborn heatmap, here darker color shows a positive correlation and lighter color shows a negative correlation. How cool is it? Here we also see there is a diagonal which is the darkest region but we do not consider it because it shows the same column relation. While we see the quality column and others that show the relation with the help of dark color and light color. This shows the real relation between the target and the rest of the columns.

Data Pre-Processing

Data preprocessing is the process when we split our data into target and data(non-variable) columns. Here we take our quality column on the x-axis and rest columns on the y-axis. Now we change our y value means target one into a binary value. There is six value in the quality column so, we change them into only good or bad. For this, we use the Lambda function. We take 0 if the values are less than 7 and if greater than or equal to 7 we take 1. 0 for low quality and 1 for high quality. This is called Label Encoding or binarization.

Building Machine Learning Model

Now first we split our data into train and test data. For this we use sklearn.model_selection and import train_test_split function. We split x data into train and test and also y data into train and test. By this, we first train our model by the train_test_split function that trains our data into specific proportions like 80–20% or 75–25%, etc. Now if we check the shape of our x data, x-train data, x-test data and so as y data we find that all train data and test data are in the proportion to our desire.

Model Training: Random Forest Classifier

Random Forest Algorithm: Let's understand our model with an example- When we want to purchase a new car, will we walk up to the first car shop and purchase one based on the advice of the dealer? It's highly unlikely.

We would likely browser a few web portals where people have posted their reviews and compare different car models, checking for their features and prices. We will also probably ask your friends and colleagues for their opinion. In short, we wouldn't directly conclude, but will instead make a decision considering the opinions of other people as well.

Ensemble models in machine learning operate on a similar idea. They combine the decisions from multiple models to improve the overall performance. Here we use Random Forest Algorithm which works as an ensemble model.

Random forest is a *Supervised Machine Learning Algorithm* that is *used widely in Classification and Regression problems*. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing *continuous variables* as in the case of regression and *categorical variables* as in the case of classification. It performs better results for classification problems.

Now, after understanding this we fit our training data x and y data to Random Forest Classifier. Now, we evaluate our model by *accuracy score*. Our model predicts the test data. Here we compare the prediction and actual score. If our accuracy score shows 75–95% then our model is working very well.

Building a Predictive System

We build a predictive system that predicts the wine quality in terms of 0 or 1 while we use any data randomly. For this, we take any random row except its quality value and then change it into a NumPy array with reshaping (1,-1). We reshape this value so that model can not confuse into many rows. Now we take the prediction model to predict. If the result is 1 then our model works great and prints good quality wine and if it gives 0 prints bad quality wine.

Saving Model

We use the pickle function to save our model. So, at this step, our Machine Learning Prediction is over.

Remarks

This is one of the interesting projects that I have working on. We see in this project that our data has to be handled with an ML algorithm.

I write here in basic language to explain every one. Hope you all enjoy it.