# Prediction of Wine Quality Using Machine Learning

# Problem Definition

# Wine is an alcoholic drink which is made of fermented grapes. We have to determine the important features that are used to predict a quality wine. For this, we have to do data analysis on all the features with respect to target feature i.e. quality of wine. The quality of wine depends on several chemical factors like alcoholic content, sugar content, citric acid content etc. Here, we have to build a machine learning model that will take all the chemical parameters as its inputs and determine the quality of wine.

# Data Analysis

# Fixed Acidity – Primary fixed acids found in wine are tartaric, succinic, citric and malic.

# Volatile Acidity – It is a measure of wine’s volatile acids. The primary volatile acid in wine is acetic acid.

# Citric Acid – It is a colourless weak organic acid which is naturally found in citrus fruits.

# Residual Sugar – This is from natural grape sugars leftover in a wine after the alcoholic fermentation finishes.

# Chlorides – Wine contains 2-4grams of L-1 of salts of mineral acids and these chlorides are responsible for salty taste of wine.

# Free Sulphur Dioxide – SO2 is used in preventing microbial growth and the oxidation of wine.

# Total Sulphur Dioxide – It is the portion of SO2 that is free in the wine and the portion that is bound to other chemicals in the wine such as aldehydes.

# Density – The density of wine is decided by the amount of yeast added to it.

# ph – It is the measure of acidity or alkalinity of a solution.

# Sulphates – It is one of the by-product that is created during the fermentation process of wine.

# Alcohol – It is produced by fermentation of grains, fruits or other sweet sources of sugar that acts as a drug.

# Exploratory Data Analysis

# Fixed Acidity

# 

# Values of fixed acidity are in the range of 4.6 to 15.9.

# As we can see from the histogram plot, most of the fixed acidity values are around 7.

# 

# From the above lineplot between quality and fixed acidity, we can see that if the fixed acidity is between 8.5 to 9 then the quality of wine will be good.

# Volatile Acidity

# 

# Values of volatile acidity are in the range of 0.12 to 1.58.

# As we can see from the histogram plot, most of the volatile acidity values are around 0.6.

# 

# From the above lineplot between quality and volatile acidity, we can see that if the volatile acidity is between 0.4 to 0.5 then the quality of wine will be good.

# Citric Acid

# 

# Values of citric acid are in the range of 0.0 to 1.0.

# As we can see from the histogram plot, most of the citric acid values are around 0.2.

# 

# From the above lineplot between quality and citric acid, we can see that if the citric acid is between 0.3 to 0.4 then the quality of wine will be good.

# Residual Sugar

# 

# Values of residual sugar are in the range of 0.9 to 15.5.

# As we can see from the histogram plot, most of the residual sugar values are around 2.

# 

# From the above lineplot between quality and residual sugar, we can see that if the residual sugar is between 2.5 to 2.7 then the quality of wine will be good.

# Chlorides

# 

# Values of chlorides are in the range of 0.012 to 0.611.

# As we can see from the histogram plot, most of the chlorides values are around 0.08.

# 

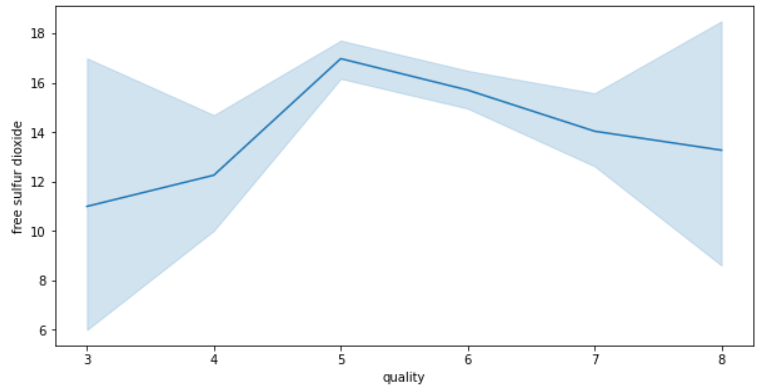
# From the above lineplot between quality and chlorides, we can see that if the chlorides is between 0.07 to 0.08 then the quality of wine will be good.

# Free sulphur dioxide

# 

# Values of free sulphur dioxide are in the range of 1.0 to 72.

# As we can see from the histogram plot, most of the free sulphur dioxide values are around 12.



# From the above lineplot between quality and free sulphur dioxide, we can see that if the free sulphur dioxide is between 13 to 14 then the quality of wine will be good.

# Total sulphur dioxide

# 

# Values of total sulphur dioxide are in the range of 6 to 289.

# As we can see from the histogram plot, most of the total sulphur dioxide values are around 30.

# 

# From the above lineplot between quality and total sulphur dioxide, we can see that if the total sulphur dioxide is between 30 to 40 then the quality of wine will be good.

# Density

# 

# Values of density are in the range of 0.9907 to 1.00369.

# As we can see from the histogram plot, most of the density values are around 0.997.

# 

# From the above lineplot between quality and density, we can see that if the density is between 0.995 to 00996 then the quality of wine will be good.

# pH

# 

# Values of pH are in the range of 2.74 to 4.01.

# As we can see from the histogram plot, most of the pH values are around 3.3.

# 

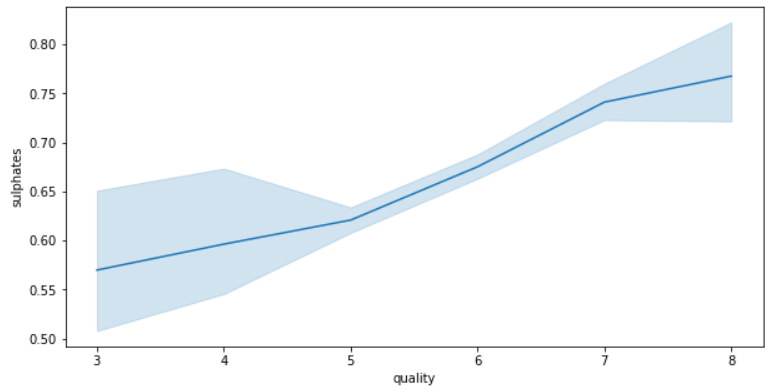
# From the above lineplot between quality and pH, we can see that if the pH is between 3.25 to 3.30 then the quality of wine will be good.

# Sulphates

# 

# Values of sulphates are in the range of 0.33 to 2.0.

# As we can see from the histogram plot, most of the sulphate values are around 0.6.



# From the above lineplot between quality and sulphates, we can see that if the sulphates is between 0.7 to 0.75 then the quality of wine will be good.

# Alcohol

# 

# Values of alcohol are in the range of 8.4 to 14.9.

# As we can see from the histogram plot, most of the alcohol values are around 9.5.

# 

# From the above lineplot between quality and alcohol, we can see that if the alcohol is between 11.5 to 12 then the quality of wine will be good.

# Quality

# There are 6 values namely 3,4,5,6,7,8 that represents the quality of wine.

# For simplification, we are dividing 3,4,5 as bad quality wine and 6,7,8 as good quality wine.

# Correlation

# Correlation gives us the exact value on how much the independent features are related to the target variable.

# 

# From the above correlation plot, we can see alcohol, citric acid and volatile acidity are highly correlated to quality of wine.

# We can see a multi-collinearity among density, fixed acidity and citric acid. This multi-collinearity causes in over fitting.

# Among the three features, fixed acidity has less correlation with the target feature i.e. quality. So, we can drop the fixed acidity feature from dataframe.

# Pre-processing

# Finding Missing Values

# 

# There are no missing values present in the given data.

# If any, we need to replace those missing values with mean or median.

# Outliers Removal

# 

# The above code remove all the values whose zscore is greater than 3 i.e. outliers and saves the data into a new dataframe n\_df.

# 

# If the data lost from the actual dataframe is more than 7.5%, then we should continue with the old dataframe.

# Skewness Removal

# 

# We have to remove skewness from the columns whose skew value is not between -0.5 to +0.5.

# 

# In the above code, by using the power transformer, we have removed the skewness present in the data.

# Splitting data to i/p and o/p dataframes.

# 

# Now in the x dataframe, we have all the independent featues and y dataframe, we have the target feature.

# Scaling

# In Scaling, we put our variables in the same range and in the same scale so that no any variable dominates the other variable.

# 

# By using the standard scalar, we have kept all the variables in the same scale.

# Sampling

# Sampling is done, when we have a class imbalance in the target column.

# 

# By doing sampling on the target column, we have balanced the 2 classes present in the column.

# vii.) Splitting dataset to training and test set

# We divide our dataset into train and test sets, so that we will make the algorithms to learn from the train dataset and apply its learning on the test dataset.

# 

# We have taken test size as 0.2 i.e. 20% of the total data goes to test dataset.

# Machine Learning Models

# This is the step where we train different machine learning models with the train data and select a better model based on the accuracy.

# 

# 

# 

# 

# 

# We can see that the accuracy score and Cross validation score of KNN classifier is almost same. So, we will choose KNN classifier as our model.

# Now, by using GridSearchCV we should do hyper parameter tuning to KNN model to increase the accuracy.

# 

# We have increased our accuracy from 77.4% to 85.67% by using GridSearchCV method i.e. Hyper parameter tuning.

# 

# From confusion matrix, we can see out of 342, our machine learning model is predicting 293 correctly and the area under the curve is 0.858.

# 

# In the final step, we have saved our machine learning model.

# Conclusion

# Machine Learning is one of the top growing technologies in the world. It has applications in almost all other fields of studies because machine learning solves problems that are difficult and time consuming to humans to solve.

# The above article is an example of one of the problem to predict the quality of wine by using machine learning model. The above mentioned steps are most commonly used steps while solving a machine learning problem.

# 