This dataset comes from the University of California, Irvine's massive archive of Machine Learning Datasets.

The Wine Quality Dataset that focuses on the Portuguese "Vinho Verde" Wine. The datasets for white wine has 4898 entries over 12 features. The features are:

* 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
* 12 - quality (score between 0 and 10)

The quality of the wine, the dependent variable, is a classification marking on the label (range between 1 -10) and is identified by patrons.

`From our data set, we are interested in determining the following:

* **What can you comment on the variables?**
* Data are collected on 12 different properties of the wines one of which is Quality, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc.
* All chemical properties of wines are continuous variables.
* Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best).
* 11 predictors and 1 output attribute
* Quality has most values concentrated in the categories 5, 6 and 7. Only a small proportion is in the categories [3,4] and [8,9] and none in the categories [1,2,10].
* Fixed acidity, volatile acidity, citric acid, chlorides, free\_sulfur\_dioxide, total\_sulfur\_dioxide, pH and sulphates have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.
* Residual sugar has a positively skewed distribution; even after eliminating the outliers distribution will remain skewed.
* Some of the variables, e.g density, have a few outliers but these are very different from the rest.
* Mostly outliers are on the larger side.
* Alcohol has an irregular shaped distribution but it does not have pronounced outliers.
* The classes are ordered and not balanced e.g. there are much more normal wines than excellent or poor ones.
* Outlier detection algorithms could be used to detect the few excellent or poor wines.
* Several of the attributes may be correlated, thus it makes sense to apply some sort of feature selection.
* **Do we need to do outlier treatment? What is the most apt method for this?**
* Possibly the most important step in data preparation is to identify outliers. Since this is a multivariate data, we consider only those points which do not have any predictor variable value to be outside of limits constructed by boxplots.
* The following rule is applied: **z-score and boxplots.**
* The intuition behind Z-score is to describe any data point by finding their relationship with the Standard Deviation and Mean of the group of data points. Z-score is finding the distribution of data where mean is 0 and standard deviation is 1 i.e. normal distribution.
* A predictor value is considered to be an outlier only if the **z-score** value is greater than 3. Application of this rule reduces the data size from 4898 to 4487.
* **Does it make a business case to predict the quality of the wine?**
* Yes, it does make business case. Since its affecting the quality of wine. On these features we can predict the quality of wine and make business decisions depending on it.
* **Are we able to classify the type of wine given the predictors?**
* We can classify the type of wine into two classifiers based on the predictors – whether it is bad[0] or good[1].
* **Which among the below methods will be your recommendation? Does this make business sense? how will this help in the real world?**

|  |  |
| --- | --- |
| Decision Tree | 81.40% |
| KNN | 85.55% |
| Naives Bayes | 73.4% |

* **Using hyperparameter tuning, we could find that the KNN algorithm was giving the best accuracy with 85.55%.**
* **And this algorithm approach will be the best use in the real world as it predicts with the best possible accuracy based on the given predictors.**
* **Lower Dimensionality: KNN is suited for lower dimensional data. Since the dataset is having lower dimensionality, using this algorithm will be the best in the real world.**