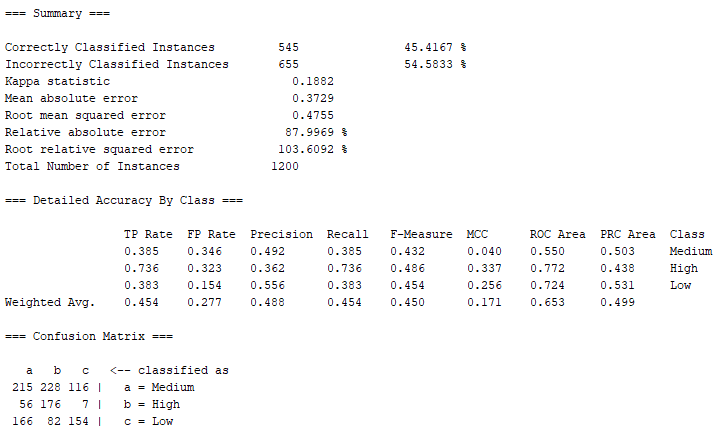
**Classification Models – predicting quality**

**Naïve Bayes** model was run in Weka on using the dataset with 4,000 instances (split 70.0% train, remainder test) and 12 attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, Quality and Alcohol content).

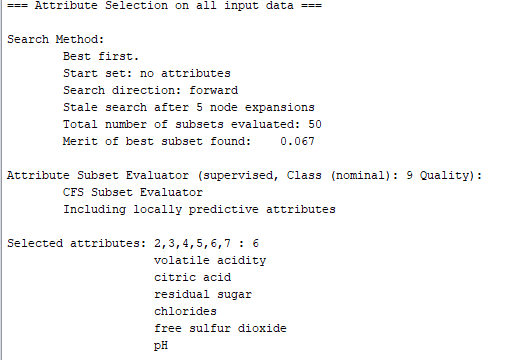
The accuracy of that model was only 45.4% (Fig 1.1).



*Figure 1.1 Naïve Bayes classifier output (Weka)*

The same model was run after removing total sulfur dioxide attribute (correlated to free sulfur dioxide) and slight improvement in accuracy level was observed – 46.4%.

In the next step we used the correlation-based feature selection (CFS) algorithm in Weka to further eliminate the correlated redundant features from the dataset (Fig. 1.2).

**

*Figure 1.2. Attributes selected by applying the CFS algorithm in Weka*

However, after executing the Naïve Bayes model in Weka using only attributes selected by CFS algorithm (volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, pH and Quality) we did not observe any major improvement in accuracy. The accuracy for this model was 45.5%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Class** | |  |  |
| **Attribute** | | **High** | **Medium** | **Low** |
|  | | 0.21 | 0.45 | 0.34 |
| **fixed acidity** | |  |  |  |
| mean | | 6.6968 | 6.8267 | 6.9602 |
| std. dev. | | 0.7846 | 0.8485 | 0.9121 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.1552 | 0.1552 | 0.1552 |
| **volatile acidity** | |  |  |  |
| mean | | 0.2689 | 0.2615 | 0.3121 |
| std. dev. | | 0.0946 | 0.0895 | 0.1167 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0082 | 0.0082 | 0.0082 |
| **citric acid** | |  |  |  |
| mean | | 0.3306 | 0.3394 | 0.3329 |
| std. dev. | | 0.0828 | 0.1206 | 0.1436 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0193 | 0.0193 | 0.0193 |
| **residual sugar** | |  |  |  |
| mean | | 4.6206 | 6.0033 | 6.6776 |
| std. dev. | | 3.7365 | 4.9864 | 5.1673 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.211 | 0.211 | 0.211 |
| **chlorides** | |  |  |  |
| mean | | 0.0374 | 0.0452 | 0.0522 |
| std. dev. | | 0.0108 | 0.0209 | 0.0288 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0021 | 0.0021 | 0.0021 |
| **free sulfur dioxide** | | | | |
| mean | | 34.1524 | 35.313 | 34.7731 |
| std. dev. | | 14.2492 | 15.7009 | 20.5611 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 2.1908 | 2.1908 | 2.1908 |
| **pH** | |  |  |  |
| mean | | 3.2288 | 3.1951 | 3.1728 |
| std. dev. | | 0.1536 | 0.1512 | 0.147 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0108 | 0.0108 | 0.0108 |
| **sulphates** | |  |  |  |
| mean | | 0.5 | 0.4921 | 0.4816 |
| std. dev. | | 0.1345 | 0.1118 | 0.1008 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.011 | 0.011 | 0.011 |
| **Alcohol content** | |  |  |  |
| High | | 300 | 564 | 473 |
| Medium | | 467 | 1008 | 777 |
| Low | | 70 | 237 | 113 |
| [total] | | 837 | 1809 | 1363 |

*Table 1.1 Naïve Bayes classifier model*

It would appear that the quality is decreasing with higher level of fixed acidity, residual sugar and chlorides and it is improved with higher pH level. It appears that higher quality wines have lower alcohol level.