Data 3444 - Lab 1 - Classification using k-NN - Jessie Reynolds

Our wine dataset included 6,497 examples with 12 features. A wine's fixed acidity is measured by the amount of acids that do not evaporate when boiled, while the volatile acidity is measured by the amount of gaseous acids in the wine. Fixed acids contribute more to a wine's texture, while volatile acids affect a wine's taste. Citric acids can be added to wines to enhance acidity and taste, which is more common in white wines than red wines. Residual sugar is measured by the amount of sugar left in the wine after fermentation. Chlorides play a large role in the saltiness of wine, these levels have to do with the type of grapes used to make the wine. Free sulfur dioxide measures the amount of sulfur dioxide that is able to protect wine from oxidizing and spoiling, whereas total sulfur dioxide is the sum of free sulfur dioxide and the sulfur dioxide bonded to other chemicals. Total sulfur dioxide does not give us as much information about wine types as free sulfur dioxide levels do. Density helps determine the alcohol levels in wine, they are similar for both red and white wines. pH measures the level of acidity in the wine and often falls between 2.8 and 4. White wines are typically more acidic and fall between 3.0 and 3.4, while red wines are typically less acidic and fall between 3.3 and 3.6. Similar to sulfur dioxide, sulphates act as preservatives and fight against oxidation and spoilage in wine. Alcohol values tell us the percentage of alcohol in the wines. Finally, quality is determined by chemical composition and characteristics, this value does not tell us a lot about determining the type of wine present.

In our dataset, there were 1599 red wines (24.6% of the total) and 4898 white wines (75.45% of the total) in our dataset. Based on the histograms and boxplots, most of the features in our dataset are skewed to the right. Some of our features are on small scales, like citric acid which varies from 0 to 1.6, whereas some features have huge scales, like free sulfur dioxide which ranges from 1 to 289. This tells us we will need to normalize our dataset in the next part before running the algorithm and cross tables.

	wines_n_pred		
wines_test_labels	Red	White	Row Total
Red	312	16	328
	0.951	0.049	0.252
	0.981	0.016	
	0.240	0.012	
White	6	966	972
	0.006	0.994	0.748
	0.019	0.984	
	0.005	0.743	
Column Total	318	982	1300
	0.245	0.755	

***Cross Table 1*: Normalized dataset with k value of 72**

Based on ***Cross Table 1*** we can see that there were 16 wines that were predicted white, but were actually classified as red. There were also 6 wines that were predicted red, but were actually classified as white. In total there were 22/1300 wines predicted incorrectly for this cross table.

	wines_z_pred		
wines_test_labels	Red	White	Row Total
Red	320	8	328
	0.976	0.024	0.252
	0.985	0.008	
	0.246	0.006	
White	5	967	972
	0.005	0.995	0.748
	0.015	0.992	
	0.004	0.744	
Column Total	325	975	1300
	0.250	0.750	

***Cross Table 2*: Z-scores dataset with k value of 72**

Based on ***Cross Table 2*** we can see there were 8 wines that were predicted white, but were actually classified as red. There were 5 wines that were predicted red, but were actually classified as white. In total there were 13/1300 wines predicted incorrectly for this cross table.

Based on my results, the z-score dataset predicted less wines incorrectly than the normalized dataset. This was also true for Cooper's results and others in the class.

To improve my model, I reran the algorithm for the normalized dataset, with k values 50, 25 and 10 represented in ***Cross Tables 3, 4 and 5***, respectively. When the k value was 50, there were 19 total wines predicted incorrectly. When the k value was 25, there were 15 wines predicted incorrectly. When the k value was 10, there were 12 wines predicted incorrectly. As the k value got smaller, less wines were predicted incorrectly.

	wines_n1_pred		
wines_test_labels	Red	White	Row Total
Red	314	14	328
	0.957	0.043	0.252
	0.984	0.014	
	0.242	0.011	
White	5	967	972
	0.005	0.995	0.748
	0.016	0.986	
	0.004	0.744	
Column Total	319	981	1300
	0.245	0.755	

Cross Table 3: **Normalized dataset with k value of 50**

	wines_n2_pred		
wines_test_labels	Red	White	Row Total
Red	318	10	328
	0.970	0.030	0.252
	0.985	0.010	
	0.245	0.008	
White	5	967	972
	0.005	0.995	0.748
	0.015	0.990	
	0.004	0.744	
Column Total	323	977	1300
	0.248	0.752	

Cross Table 4: **Normalized dataset with k value of 25**

	wines_n3_pred		
wines_test_labels	Red	White	Row Total
Red	319	9	328
	0.973	0.027	0.252
	0.991	0.009	
	0.245	0.007	
White	3	969	972
	0.003	0.997	0.748
	0.009	0.991	
	0.002	0.745	
Column Total	322	978	1300
	0.248	0.752	

Cross Table 5: **Normalized dataset with k value of 10**

Similarly, for the z-score dataset I reran the algorithm again with the same k values, 50, 25 and 10 represented in ***Cross Tables 6, 7 and 8***, respectively. When the k value is 50, there were 13 total wines predicted incorrectly. When the k value was 25, there were 12 wines predicted incorrectly. When the k value was 10, there were 11 wines predicted incorrectly. As the k value got smaller, less wines were predicted incorrectly similar to the normalized dataset.

	wines_z1_pred		
wines_test_labels	Red	White	Row Total
Red	320	8	328
	0.976	0.024	0.252
	0.985	0.008	
	0.246	0.006	
White	5	967	972
	0.005	0.995	0.748
	0.015	0.992	
	0.004	0.744	
Column Total	325	975	1300
	0.250	0.750	

***Cross Table 6*: Z-scores dataset with k value of 50**

	wines_z2_pred		
wines_test_labels	Red	White	Row Total
Red	321	7	328
	0.979	0.021	0.252
	0.985	0.007	
	0.247	0.005	
White	5	967	972
	0.005	0.995	0.748
	0.015	0.993	
	0.004	0.744	
Column Total	326	974	1300
	0.251	0.749	

***Cross Table 7*: Z-scores dataset with k value of 25**

	wines_z3_pred		
wines_test_labels	Red	White	Row Total
Red	321	7	328
	0.979	0.021	0.252
	0.988	0.007	
	0.247	0.005	
White	4	968	972
	0.004	0.996	0.748
	0.012	0.993	
	0.003	0.745	
Column Total	325	975	1300
	0.250	0.750	

***Cross Table 8*: Z-scores dataset with k value of 10**

My script is under Lab 1 - Reynolds in PositCloud