STAT 6340 Statistical Machine Learning

Mini Project 2

Author:
Suarez Rodes, Randy

Supervisor:
Choudhary, Pankaj Ph.D.

February 27, 2021

1 Answers

Question 1.a We explored the correlations of our variables, both predictors and response on Figure 1. Visualizing the correlation matrix we can see that predictors Oakiness and Clarity have a very low correlation with Quality. We infer that these variables are not much relevant to predict Quality and are the most likely to be removed from future models. On the other hand predictors Aroma, Flavor and Region has a high correlation with Quality, making them good candidates for our models.
The variable Body presents some correlation with Quality, although, it is not as high as other predictors, we require further analysis to determine its relevance for the response variable. Finally, predictors Aroma, Body, Flavor and Region are highly correlated. This can cause overfitting issues, since the data would be over explained, so with further analysis one or some of these predictors might be dropped since they can be explained by other predictors.

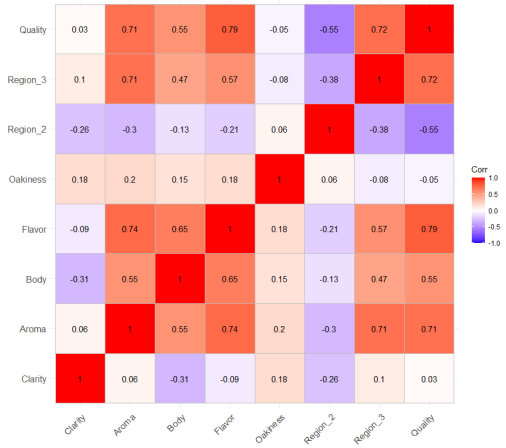

Figure 1: Correlation Matrix

We present the scatterplot of our variables on Figure 2. As expected from the correlation matrix we can notice that the predictors Aroma, Body and Flavor exhibit a positive linear relation with Quality. On Figure 3 we present the Quality boxplots by Region. We see a clear distribution of the Quality determine by the Region (high on 3, medium on 1, low on 2). This fact evidences the importance of this predictor for determining the Quality of a wine.

Figure 2: Scatter Plots

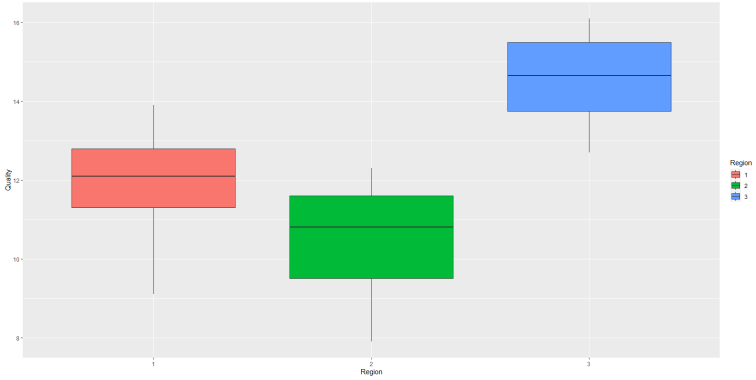

Figure 3: Quality by Region

Question 1.b We explore the variable Quality using boxplots to identify possible outliers (Figure 4). We did not encounter any evident outlier. We also explore the distribution of Quality using histograms (Figure 5). We can appreciate that the data distribution looks approximately normal. We also try several transformations (sqrt, log, normalization) without obtaining any substantial improvement in our correlations or scatterplots. At this moment we consider that a transformation is not required and we will check model assumptions once we build an appropriate model.

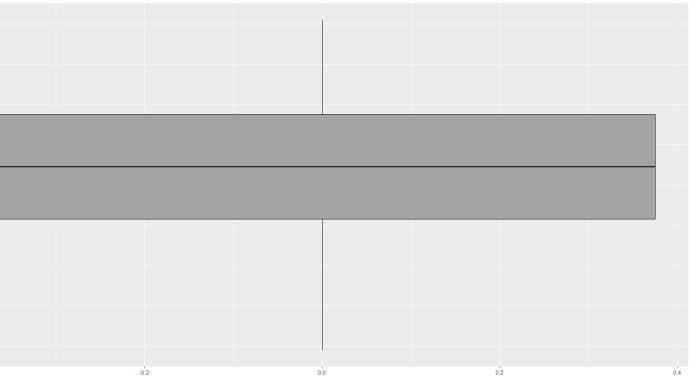

Figure 4: Quality Boxplot

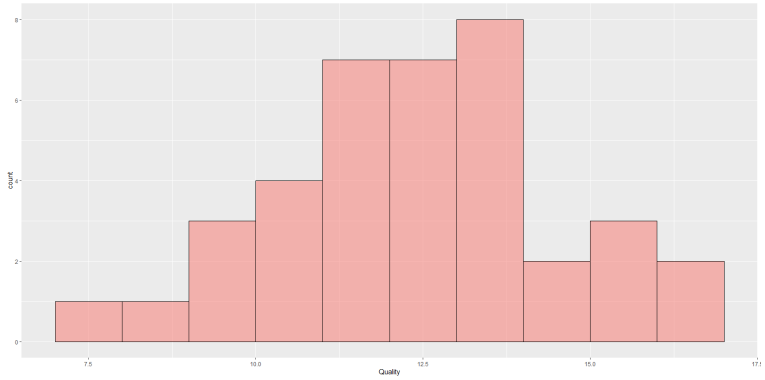

Figure 5: Quality Histogram

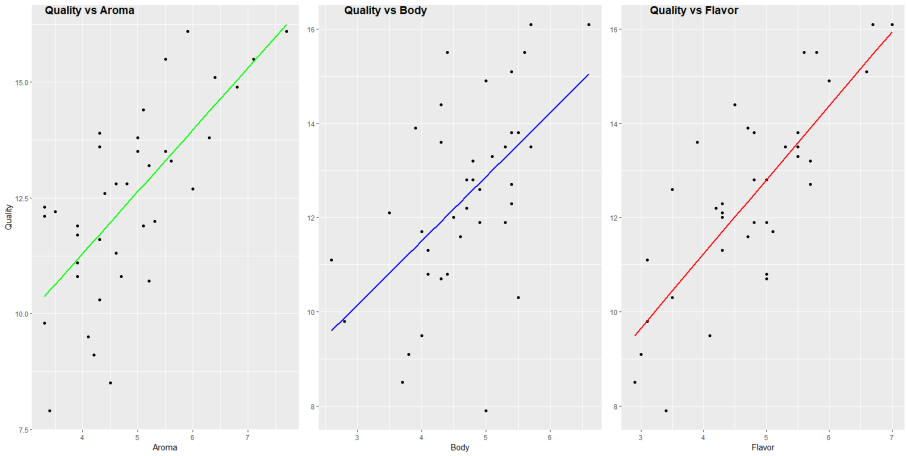

Figure 6: Quality vs Continuous Predictors

Question 1.c We fitted a simple linear regression model for every predictor (Tables 1 to 6). Taking a look to the p-values of every model we can affirm that there is a statistically significant association between the predictor and the response variable on all models except for the models with predictors Clarity (p-value = 0.865) and Oakiness (p-value = 0.779). This is consistent with our previous exploration. We can see on Figure 6 how there is a clear linear relation for predictors Flavor, Aroma and Body. Also we discussed on question 1.a how Region is relevant to predict Quality (Figure 3).

Table 1:

	Estimate	Pr(>|t|)
(Intercept)	12.0034	3.89e-5
Clarity	0.4692	0.865

Table 2:

	Estimate	Pr(>|t|)
(Intercept)	5.9583	4.51e-6
Aroma	1.3365	6.87e-7

Table 3:

	Estimate	Pr(>|t|)
(Intercept)	6.0580	0.0007
Body	1.3618	0.0004

Table 4:

	Estimate	Pr(>|t|)
(Intercept)	4.9414	1.57e-5
Flavor	1.5719	3.68e-9

Table 5:

	Estimate	Pr(>|t|)
(Intercept)	12.9916	1.4e-7
Oakiness	-0.1304	0.779

Table 6:

	Estimate	Pr(>|t|)
(Intercept)	11.9765	2e-16
Region2	-1.5320	0.00757
Region3	2.6069	7.01e-6

Question 1.d From the summary of the full model we can reject the null hypothesis for the Flavor and Region predictors, meaning that this predictors are relevant in order to predict Quality.

Question 1.e To start building our model, we will proceed to drop one variable at a time from the full model starting from the highest p-value predictor and rechecking the test hypothesis results for every new model until all predictors remaining are relevant for the response. After this process we ended up with the model Quality ~Flavor + Region (Table 8). The anova test(Table 9) between the full model and the current model confirms our findings. With a p-value of 0.6528, we fail to

Table 7: Full Model

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept)	7.8144	1.9694	3.97	0.0004
Clarity	0.0171	1.4563	0.01	0.9907
Aroma	0.0890	0.2525	0.35	0.7269
Body	0.0797	0.2677	0.30	0.7681
Flavor	1.1172	0.2403	4.65	6.25e-5
Oakiness	-0.3464	0.2330	-1.49	0.1475
Region2	-1.5129	0.3923	-3.86	0.0006
Region3	0.9726	0.5102	1.91	0.0662

Table 8: Resulting Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0943	0.7912	8.97	1.76e-10
Flavor	1.1155	0.1738	6.42	2.49e-07
Region2	-1.5335	0.3688	-4.16	0.0002
Region3	1.2234	0.4003	3.06	0.0043

reject the null hypothesis, this is all extra predictors are 0.

able 9: Anova with Full Model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
34	27.21				
30	25.14	4	2.07	0.62	0.6528

Table 10: Anova vs Interactions

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	34	27.21				
2	32	25.43	2	1.78	1.12	0.3378

Now we will proceed to explore interactions between these two predictors. We will compare the resulting model with the interactions model Quality ~Flavor + Region + Flavor:Region. By carrying out an anova test we obtained a p-value of 0.3378, meaning that we fail to reject the null hypothesis, therefore the interactions between Flavor and Region are not meaningful for our model.

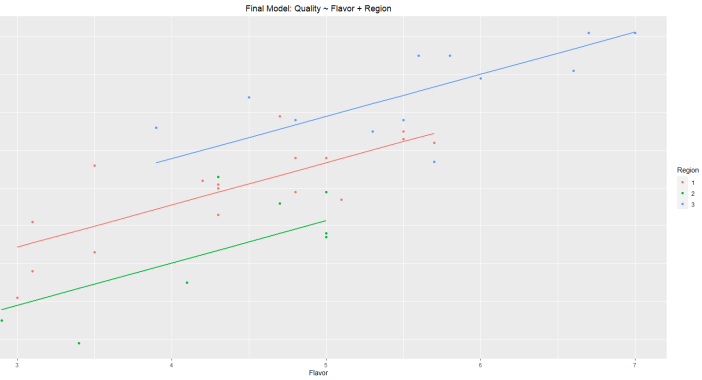

Figure 7: Final Model

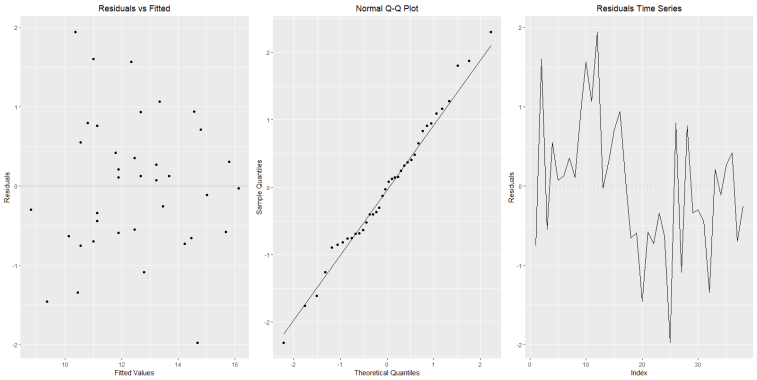

Figure 8: Model Assumptions

We can appreciate the three resulting regression lines of our model for each region on Figure 7. Clearly our model captures the linear trend on each region.
Also we present several diagnostics plots on Figure 8 in order to check for model assumptions. We can see in the Residuals vs Fitted plot how the points are scattered around zero and there is no a clear pattern, verifying that our errors have an approximate zero mean and constant variance. For the Q-Q Plot, observe how for an exception of a few points the errors are close to be normally distributed. Finally checking on the Residuals Time Series, there is not a clear dependence of the residuals, verifying the independence assumption.

Question 1.f Final model:
Quality = 7.0943 + 1.1155(Flavor) -1.53351$\mathbb{1}_{(Region2)}$ + 1.22341$\mathbb{1}_{(Region3)}$

Where $\mathbb{1}$ is an indicator function. We can interpret our model as follows: Approximately every unit change in Flavor will directly change the Quality by one unit. Region 1 wines has approximately a base quality of 7, Region 3 offers an additional unit (quality base around 8) and Region 2 a decrease of 1.5 (quality base approximately of 5.5).

Question 1.g Having an average Flavor wine from our sample and assuming Region 1, our prediction of Quality is: 12.41371.
We obtained a 95% confidence interval of [11.95152, 12.8759], this is that given an average Flavor

wine and assuming it is from Region 1, 95% of intervals of this form will contain the true value of the wine Quality.
Also we obtained a 95% prediction interval of [10.53775, 14.28967], meaning that we can be 95% confident that the Quality of an average Flavor wine from Region 1 will lie in this range.

Question 2.a On Figure 9 we present several plots to explore how the GMAT and GPA can help to predict the group of an applicant. We can observe in the graph of GMAT vs GPA, how the GPA predictor can clearly separate the three groups. Moreover in the GPA boxplot we have a clear distribution of the GPA by group, evidence of how much this predictor contributes to identify the correct group. For the GMAT boxplot there is some confusion for applicants of group 2 and 3, but a clear distinction for applicants of group 1, meaning that GMAT also helps us characterizes the possible group of an applicant.

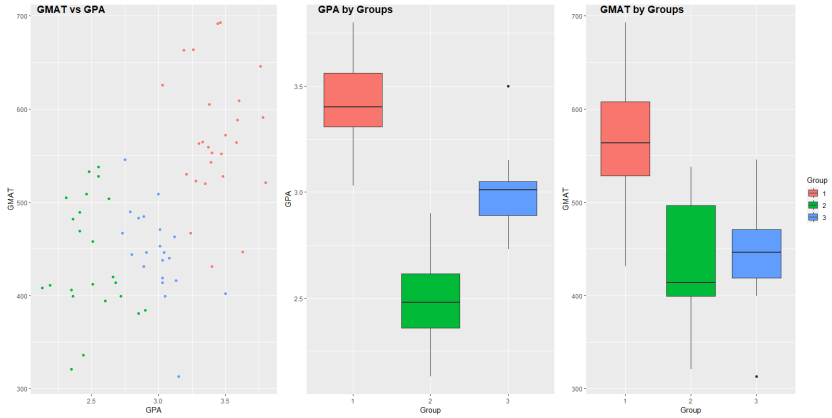

Figure 9:

Question 2.b We fitted an LDA model for the training set and superimposed the decision boundary on these data (Figure 10). There is misclassification between the boundary of group 1 and 3 and the boundary of 2 and 3. With the amount of data used, we consider the decision boundary is sensible for new data points. We present the confusion matrix for training and testing sets on Tables 11 and 12. The misclassification error rate for training set is around 8.5% while it jumps to 20% for the testing set confirming the sensibility of our decision boundary.

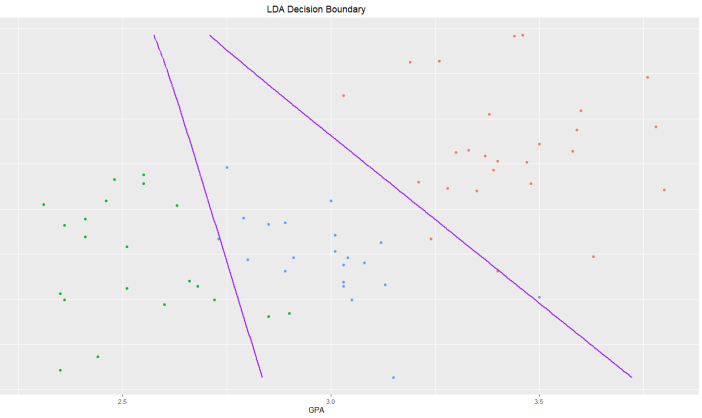

Figure 10: LDA Decision Boundary

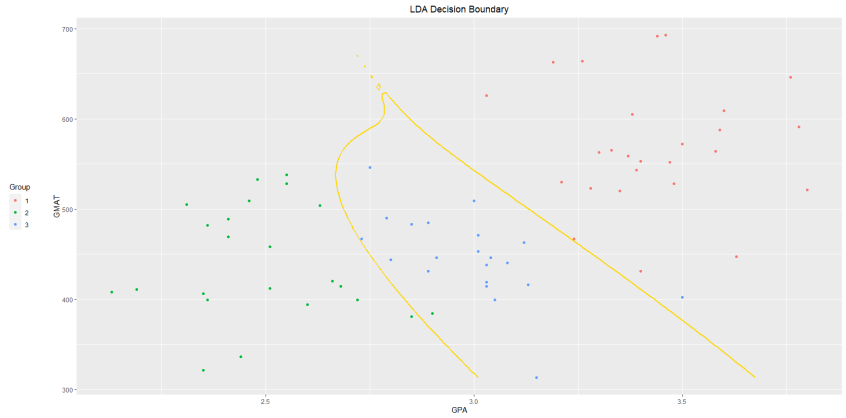

Figure 11: QDA Decision Boundary

Question 2.c We fitted a QDA model for the training set and superimposed the decision boundary on these data (Figure 11) We can see how QDA manage to correctly classify the challenging observations that LDA missed. The QDA decision boundary is less sensible than LDA's. We present the confusion matrix for training and testing sets on Tables 13 and 14. The misclassification error rate for training set is around 2.8% and 6.7% for testing set. We concluded that QDA decision boundary is not as much sensible to new data points.

ble 11: LDA Training
Actual

		1	2	3
	1	24	0	1
	2	0	21	1
dicted	3	2	2	19

Error Rate: 0.08571

Table 12: LDA Testing
Actual

		1	2	3
	1	2	0	0
	2	0	5	0
Predicted	3	3	0	5

Class Error Rate: 0.2

Table 13: QDA Training
Actual

		1	2	3
	1	26	0	1
	2	0	22	0
Predicted	3	0	1	20

Class Error Rate: 0.02857

Table 14: QDA Testing
Actual

		1	2	3
	1	4	0	0
	2	0	5	0
Predicted	3	1	0	5

Class Error Rate: 0.06667

Question 2.d We recommend QDA over LDA as the classifier for our data. QDA performed very well on both training and testing data and offered a less sensible decision boundary than LDA.

Question 3.a The data is unbalanced having 1316 observations labelled as 0(non diabetic) and 684 as 1(diabetic). This fact can cause that classification algorithms may have a high accuracy but a low sensitivity (setting 1 as the positive response). We explored several boxplots to investigate how our predictors affect the classes (Figure 12, Figure 13). Predictors Glucose and Age separates the two classes in some degree (Figure 14).

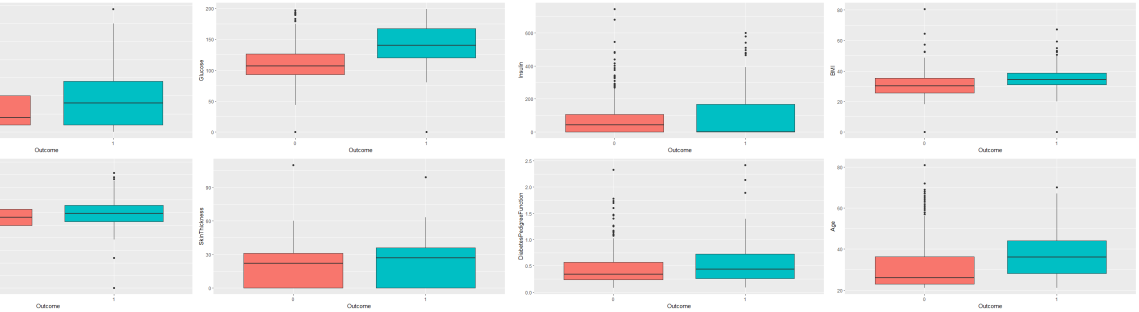

Figure 12:

Figure 13:

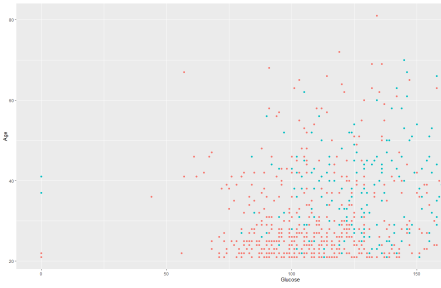

Figure 14: Age vs Glucose

Question 3.b and Question 3.c The confusion matrix for LDA and QDA are presented on Tables 15 and 16 respectively. On Table 17 we present the sensitivity, specificity, and overall misclassification rate of both methods. We can observed that even if both methods had an error close to 20%, both exhibit a low sensitivity, this fact is caused by the unbalanced of both classes like we pointed out during our exploration. The ROC graphs for both methods are very similar. Both methods perform significantly better than a random guess for all possible cutoffs.

Table 15: LDA Classifier
Actual

		0	1
	0	1174	298
Predicted	1	142	386

Table 16: QDA Classifier
Actual

		0	1
	0	1135	290
Predicted	1	181	394

Table 17:

	Error.Rate	Sensitivity	Specificity
LDA	0.22	0.5643275	0.8920973
QDA	0.2355	0.5760234	0.8624620

Question 3.d We recommend the use of the LDA classifier, since it has a slightly greater AUC of .8369 vs .8354 of QDA and also an error rate of 22% inferior to the 23.5% of QDA. We want to improve the sensitivity without sacrifying much specificity and overall accuracy. So we decided to find a cutoff of the best trade-off between sensitivity and specificity using Youden's J statistic. This is the cutoff that maximizes the value of J = sensitivity + specificity - 1. It's geometrical interpretation is the maximum distance from the random guess which is a desirable result. The pROC package offers a way to automatically find this cutoff using the function coords(x,"best"). The cutoff obtained for this method is 0.2906709 and the associated error, sensitivity and specificity is presented on Table 1. We can see how our error slightly grew but we improve our sensitivity without giving away much specificity.

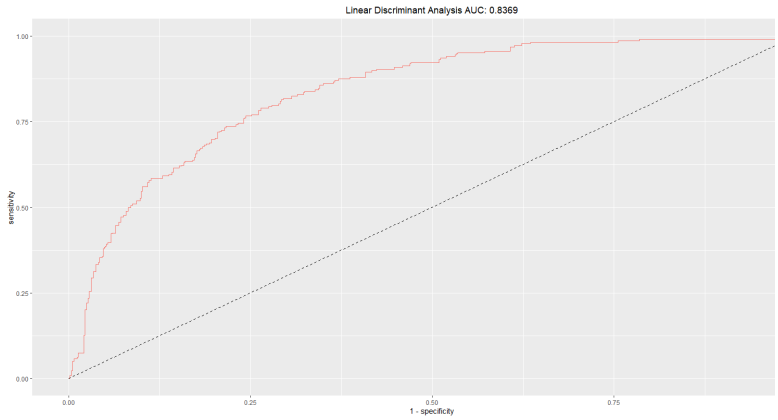

Figure 15: ROC LDA

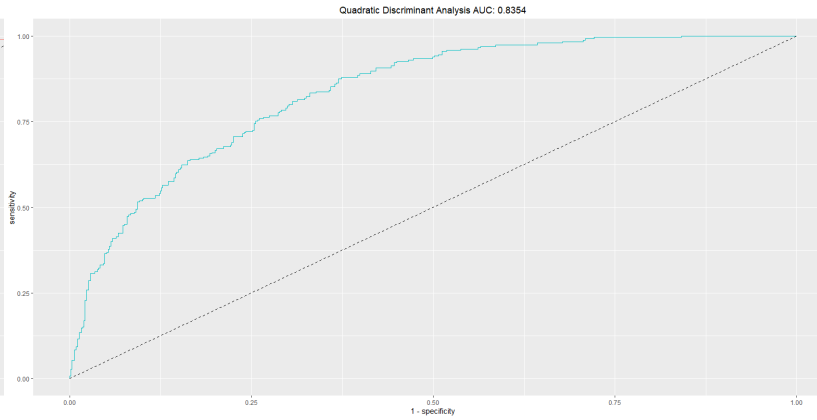

Figure 16: ROC QDA

Table 18:

	Error.Rate	Sensitivity	Specificity
New Cutoff	0.246	0.7894737	0.7355623

2 Code