DSC441- Spring 2019 - 2020, Assignment #4,

Problem 1 (10 points):

A. (4 points) Which of the following statements are true? Briefly explain your answer.

1. Training a k-nearest-neighbors classifier takes less computational time than testing it.

True, because training a KNN classifier only requires storing the attribute and labels of the training set. Testing the KNN classifier requires assigning a label to each data point based on the label that is most frequently shown in the K training samples closes to that point, therefore it requires a higher computation.

2. The more training examples, the more accurate the prediction of a k-nearest-neighbors.

False, prediction of a k-nearest-neighbors does not only depend on training examples

3. k-nearest-neighbors cannot be used for regression.

False, it can be used for regression

4. A k-nearest-neighbors is sensitive to the number of features.

True, because KNN works well with a smaller number of attributes and struggles when the number of attributes is large.

B. (6 points) Figure 2 presents the performance of several algorithms applied to the problem of classifying molecules in two classes: those that inhibit Human Respiratory Syncytial Virus (HRSV), and those that do not. HRSV is the most frequent cause of respiratory tract infections in small children, with a worldwide estimated prevalence of about 34 million cases per year among children under 5 years of age.

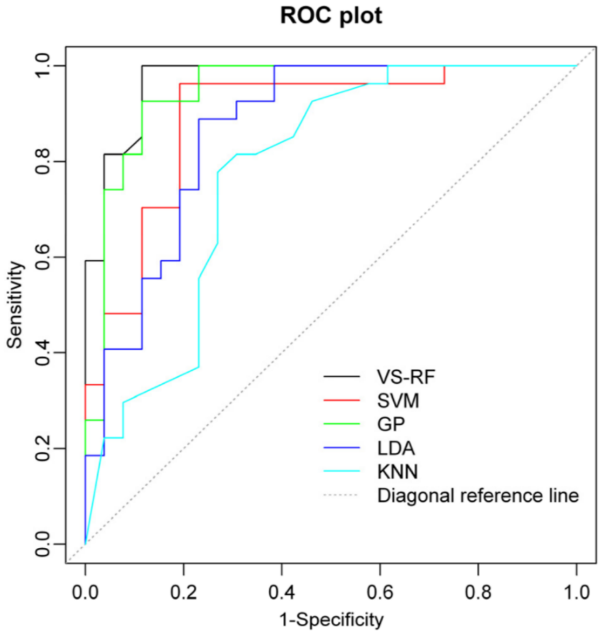


Figure 1: ROC curves for several algorithms classifying molecules according to their action on HRSV, computed on a test set. Sensitivity = True Positive Rate. Specificity = 1 - False Positive Rate. VS-RF: Random Forest. SVM: Support Vector Machine. GP: Gaussian Process. LDA: Linear Discriminant Analysis. kNN: k-Nearest Neighbors. Source: M. Hao, Y. Li, Y. Wang, and S. Zhang, Int. J. Mol. Sci. 2011, 12(2), 1259-1280.

1. Which method gives the best performance? Explain your answer.

The Random Forest provides the best performing classification model as it has the highest area under the ROC curve. Random forests require almost no input preparation. They can handle binary features, categorical features and numerical features without any need for scaling. They also perform implicit feature selection and provide a good indicator of feature importance.

2. The goal of this study is to develop an algorithm that can be used to suggest, among a large collection of several millions of molecules, those that should be experimentally tested for activity against HRSV. Compounds that are active against HSRV are good leads from which to develop new medical treatments against infections caused by this virus. In this context, is it preferable to have a high sensitivity or a high specificity? Which part of the ROC curve is the most interesting?

It is preferable to have low false positive rate to have a high sensitivity rate as this represents the number of accurately classified positive results. The specificity would represent the incorrectly classified positive results, which we would want to minimize.

3. In this study, the authors have represented the molecules based on 777 descriptors. Those descriptors include the number of oxygen atoms, the molecular weights, the number of rotatable bonds, or the estimated solubility of the molecule. They have fewer samples (216) than descriptors. What is the danger here? How would you solve this issue?

Overfitting is an issue here.

To solve this issue, the researches would either have to collect many more data points or reduce the number of descriptors used in the model.

Problem 2 (20 points):

Download the letter recognition data from: **http://archive.ics.uci.edu/ml/datasets/Letter+Recognition**

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. Below is the attribute information, but more information on the data and how it was used for data mining research can be found in the paper:

P. W. Frey and D. J. Slate. "Letter Recognition Using Holland-style Adaptive Classifiers". (Machine Learning Vol 6 #2 March 91)

Attribute Information:

1. lettr capital letter (26 values from A to Z)
2. x-box horizontal position of box (integer)
3. y-box vertical position of box (integer)
4. width width of box (integer)
5. high height of box (integer)
6. onpix total # on pixels (integer)
7. x-bar mean x of on pixels in box (integer)
8. y-bar mean y of on pixels in box (integer)
9. x2bar mean x variance (integer)
10. y2bar mean y variance (integer)
11. xybar mean x y correlation (integer)
12. x2ybr mean of x * x * y (integer)
13. xy2br mean of x * y * y (integer)
14. x-ege mean edge count left to right (integer)
15. xegvy correlation of x-ege with y (integer)
16. y-ege mean edge count bottom to top (integer)
17. yegvx correlation of y-ege with x (integer)

Create a classification model for letter recognition using decision trees as a classification method with a holdout partitioning technique for splitting the data into training versus testing.

a. (15 points) Changing the values for the depth, number of cases per parent and number of cases per leaf produces different tree configurations with different accuracies for training and testing. Choose at least five different configurations and report the accuracy for training and testing for each one of them. Which configuration will you choose as the best model? Explain your answer.

Model	Configuration			Result				
	Tree Depth	Parent Nodes	Child Nodes	Depth	Nodes	Terminal Nodes	Training Accuracy	Testing Accuracy
1	10	100	50	11	143	72	63.5%	62.1%
2	20	6	2	20	1857	929	93.4%	81.6%
3	10	200	100	11	91	45	58.9%	58.4%
4	10	120	60	11	125	63	63.7%	61.5%

After trying different configurations for value changes for depth, parent and child/leaf cases. The best model that I was able to create was a depth of 20, parent minimum cases of 6 and child cases of 2. This appears to be the best configuration at the overall training and test percentage correct values were the highest compared to the other groups. On the other hand, I tried to go higher and lower with values to ensure I was not over or under fitting the values. The overall accuracy for the best model was: 0.934 for training and 0.816 for testing.

Test	A	2...	2	3	0	0	0	3	2	1	0	0	2	1	1	1	0	1	0	6	0	1	1	0	0	0	1	88.5%
	B	0	181	0	6	1	2	8	7	2	1	0	0	1	2	9	0	1	7	3	1	0	3	0	2	1	1	75.7%
	C	1	0	183	0	3	3	5	0	0	3	1	2	1	1	5	0	1	1	1	2	1	0	0	0	0	1	85.1%
	D	0	8	0	194	0	0	3	8	2	1	1	0	1	8	11	5	3	9	1	1	1	0	0	5	1	1	73.5%
	E	0	0	4	0	184	1	10	2	1	2	0	4	0	0	0	1	2	0	8	0	1	0	0	3	0	4	81.1%
	F	0	5	1	2	1	185	1	1	1	3	0	0	0	1	1	8	0	0	5	9	2	6	0	2	2	0	78.4%
	G	4	4	5	3	5	0	1...	3	0	0	4	6	0	1	11	0	2	4	4	3	1	1	2	0	1	1	73.8%
	H	0	8	1	7	3	3	0	142	1	2	0	1	6	0	5	4	0	5	2	1	4	1	0	1	0	1	71.7%
	I	0	2	0	2	1	3	0	1	194	5	2	1	0	0	1	3	0	1	1	2	0	0	0	5	0	1	86.2%
	J	1	0	1	3	3	2	1	3	11	187	1	0	0	1	3	0	1	3	0	1	2	0	0	2	1	0	82.4%
	K	5	0	1	0	6	3	0	9	0	0	189	0	0	4	0	0	0	7	2	4	2	0	0	3	0	0	80.4%
	L	2	2	2	1	2	0	0	3	0	3	1	186	0	1	1	0	0	0	5	2	1	0	0	3	1	0	86.1%
	M	5	3	1	1	0	1	2	2	0	0	3	1	190	7	2	0	1	6	0	0	2	3	3	1	1	0	80.9%
	N	2	2	0	2	0	1	1	5	0	0	2	0	9	201	5	0	1	8	0	0	5	2	0	0	0	0	81.7%
	O	0	0	6	9	0	2	1	2	0	3	0	2	3	3	170	4	6	5	1	1	4	0	2	0	2	0	75.2%
	P	0	4	1	2	1	11	2	3	0	3	0	0	1	0	2	183	1	0	1	1	0	3	0	0	1	0	83.2%
	Q	2	1	0	1	2	0	3	1	4	2	0	6	2	2	13	0	205	1	2	2	3	0	0	0	1	1	80.7%
	R	1	6	1	8	3	1	0	7	3	0	10	4	2	0	4	1	3	156	2	0	0	0	0	1	0	0	73.2%
	S	3	7	1	3	3	2	1	4	1	1	3	0	2	1	0	0	1	0	181	1	0	0	1	3	1	6	80.1%
	T	0	0	1	0	3	1	0	1	1	1	1	1	1	0	1	2	1	0	0	205	0	1	0	3	11	0	87.2%
	U	0	1	0	0	3	0	2	1	1	1	2	1	6	7	7	0	0	2	0	0	207	3	2	0	0	0	84.1%
	V	0	4	0	0	3	2	1	1	0	0	0	0	3	2	2	3	0	0	2	3	1	187	2	0	3	0	85.4%
	W	0	1	0	0	0	0	0	0	1	0	0	0	5	8	2	1	1	0	0	0	1	5	224	0	0	0	90.0%
	X	0	0	0	0	9	5	1	2	1	1	5	1	0	0	2	1	0	1	2	1	0	0	1	187	0	1	84.6%
	Y	0	3	0	1	1	3	2	0	1	1	0	0	0	1	2	2	2	0	1	2	0	6	1	0	204	0	87.6%
	Z	1	0	0	2	2	3	1	2	1	0	0	0	2	1	0	1	2	1	7	4	0	0	0	1	1	173	84.4%
	Overall Percentage	4%	4.1%	3.5%	4%	4.0%	3.9%	4%	4%	4%	3.7%	4%	3.6%	3.9%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	3%	81.6%

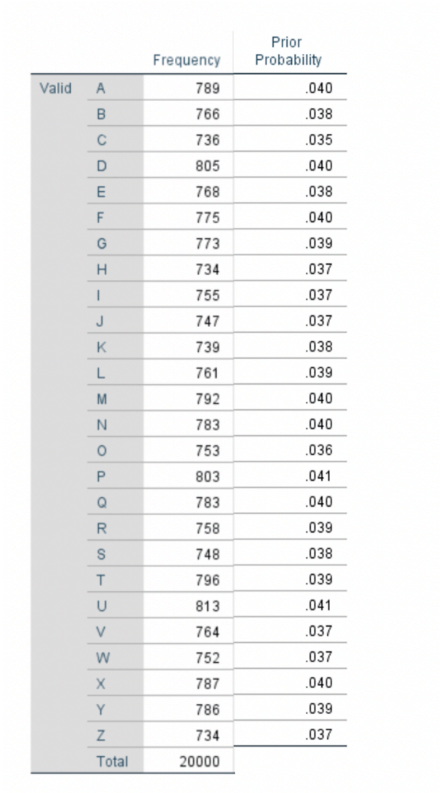
Classification

		Predicted																									
Sample	Observed	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Percent Correct
Training	A	5...	3	1	0	1	0	2	2	1	0	1	2	1	1	5	0	0	1	2	2	0	0	0	1	0	0	95.4%
	B	0	487	0	2	0	2	6	2	3	0	0	3	0	1	11	1	0	6	2	0	0	0	1	0	0	0	92.4%
	C	2	2	488	0	3	1	3	1	3	1	0	3	1	2	2	1	1	0	2	3	1	0	0	0	1	0	93.7%
	D	0	2	0	499	1	1	0	3	2	1	0	1	1	8	9	0	1	5	3	4	0	0	0	0	0	0	92.2%
	E	0	1	3	1	497	2	3	1	3	3	2	4	1	0	0	0	2	3	8	1	0	0	0	1	2	3	91.9%
	F	0	4	0	0	0	502	2	1	1	2	1	1	1	1	2	9	0	1	1	5	0	1	1	0	2	1	93.1%
	G	1	3	8	1	3	1	4...	1	1	2	1	1	0	2	11	0	6	1	4	1	0	0	2	0	0	0	90.5%
	H	1	10	1	3	0	2	4	476	3	2	8	1	4	3	7	0	1	6	1	1	1	0	0	1	0	0	88.8%
	I	0	0	0	0	1	3	0	2	516	2	0	1	0	0	0	0	0	0	0	2	1	0	1	0	0	1	97.4%
	J	3	1	0	4	1	6	0	0	10	477	1	1	2	2	2	0	1	1	0	2	2	0	0	0	2	2	91.7%
	K	1	1	0	0	3	2	0	6	0	0	473	2	3	2	0	1	0	2	1	5	0	0	0	2	0	0	93.8%
	L	0	2	0	0	2	1	6	3	1	3	1	515	0	1	0	0	3	1	2	2	1	0	0	0	0	1	94.5%
	M	5	2	1	1	0	0	0	2	0	0	1	0	536	3	0	0	0	0	0	0	1	1	3	0	1	0	96.2%
	N	2	2	0	1	0	0	0	2	0	0	0	0	3	516	3	0	0	1	1	0	1	2	2	0	1	0	96.1%
	O	0	2	3	7	2	0	2	1	0	2	0	0	4	1	484	0	7	2	1	2	4	1	2	0	0	0	91.8%
	P	0	2	1	1	2	9	1	1	2	2	0	0	2	2	4	539	0	2	3	0	1	3	3	0	2	1	92.5%
	Q	1	3	2	3	2	1	5	1	1	2	0	3	1	0	4	0	478	4	6	0	1	1	2	5	2	1	90.4%
	R	0	12	0	1	0	0	1	3	1	0	3	1	3	4	8	0	1	505	0	1	0	0	0	0	0	1	92.7%
	S	5	10	1	1	3	5	0	2	2	5	2	3	2	3	2	0	2	3	457	2	0	1	1	3	0	7	87.5%
	T	1	1	0	0	4	2	2	0	0	0	1	3	0	0	0	0	0	0	0	536	1	1	0	3	5	1	95.5%
	U	1	1	1	0	1	1	0	1	2	0	0	0	3	8	3	1	0	0	0	0	539	0	4	0	1	0	95.1%
	V	0	1	0	0	0	3	1	1	0	0	0	0	2	4	0	4	0	1	1	4	0	517	2	0	4	0	94.9%
	W	0	2	0	0	0	2	0	1	1	0	0	0	4	5	2	0	0	1	0	0	1	0	482	1	1	0	95.8%
	X	2	3	1	0	3	1	1	1	1	0	2	0	0	1	8	0	0	3	1	0	0	0	0	537	0	1	94.9%
	Y	1	1	0	0	0	2	0	0	0	1	0	1	1	2	4	1	0	0	0	4	1	5	1	0	528	0	95.5%
	Z	1	2	1	4	9	3	2	0	2	2	0	0	0	0	0	1	0	1	3	2	0	0	0	1	0	495	93.6%
	Overall Percentage	4%	4.0%	3.7%	4%	3.8%	3.9%	4%	4%	4%	3.6%	4%	3.9%	4.1%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	93.4%

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V1
	Independent Variables	V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	6
	Minimum Cases in Child Node	2
Results	Independent Variables Included	V12, V8, V11, V7, V2, V4, V10, V14, V13, V9, V15, V16, V17, V6, V3, V5
	Number of Nodes	1857
	Number of Terminal Nodes	929
	Depth	20

b. (4 points) For the best tree configuration, report the misclassification matrix and interpret it. In your opinion, is accuracy a good way to interpret the performance of the model? If not, suggest other measures.



		Frequency	Prior Probability
Valid	A	789	.040
	B	766	.038
	C	736	.035
	D	805	.040
	E	768	.038
	F	775	.040
	G	773	.039
	H	734	.037
	I	755	.037
	J	747	.037
	K	739	.038
	L	761	.039
	M	792	.040
	N	783	.040
	O	753	.036
	P	803	.041
	Q	783	.040
	R	758	.039
	S	748	.038
	T	796	.039
	U	813	.041
	V	764	.037
	W	752	.037
	X	787	.040
	Y	786	.039
	Z	734	.037
	Total	20000	

Interpretation:

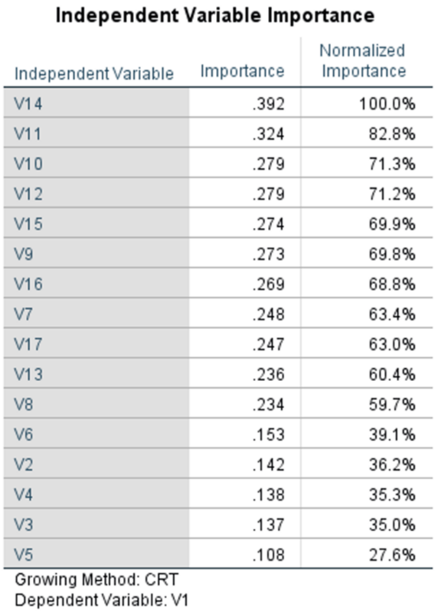
Training accuracy is 93.4% and testing accuracy is 81.6%

The model was performed with high performance percentages for both: training and testing data sets. The matrix also shows that there was a good fair distributions mixture of the difference classes within each of the datasets for training and testing.

Accuracy is one of the best interpretation methods of the model performance, but other evaluations need to be considered as well, such as the balance in the data. In the letter recognition data, the letters are very balanced so there is no one letter that dominates over the others to impact the classification.

b. (1 point) What are the most important three attributes for recognizing the letters?

x-ege, xybar, y2bar are the attributes with with high importance



Independent Variable Importance

Independent Variable	Importance	Normalized Importance
V14	.392	100.0%
V11	.324	82.8%
V10	.279	71.3%
V12	.279	71.2%
V15	.274	69.9%
V9	.273	69.8%
V16	.269	68.8%
V7	.248	63.4%
V17	.247	63.0%
V13	.236	60.4%
V8	.234	59.7%
V6	.153	39.1%
V2	.142	36.2%
V4	.138	35.3%
V3	.137	35.0%
V5	.108	27.6%

Growing Method: CRT
Dependent Variable: V1

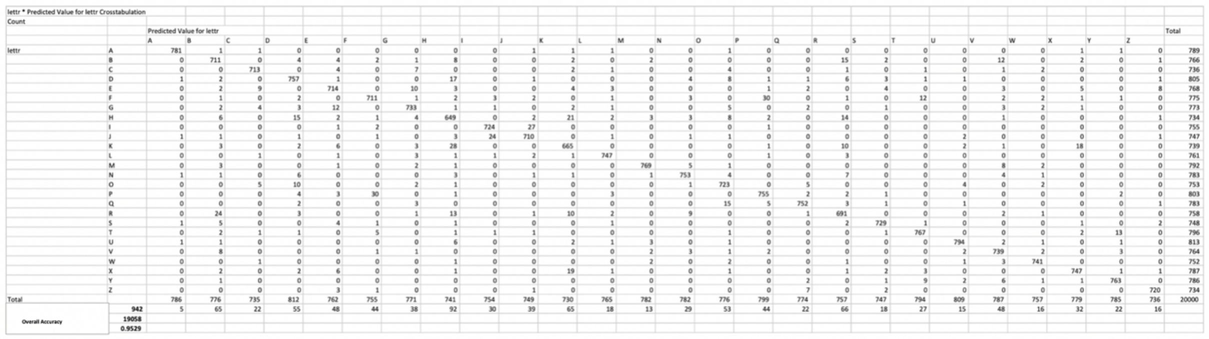
Problem 3 (20points):
On the same data from Problem 2, apply a K-nearest neighbor classifier to classify the data.
Report the following:

1. (2 points) If you are doing any data transformation, explain the transformation and why it is needed.

There is no need of data transformation because the data is already scaled to fit into a range between 0 and 15 integer values so there are no additional transformations required.

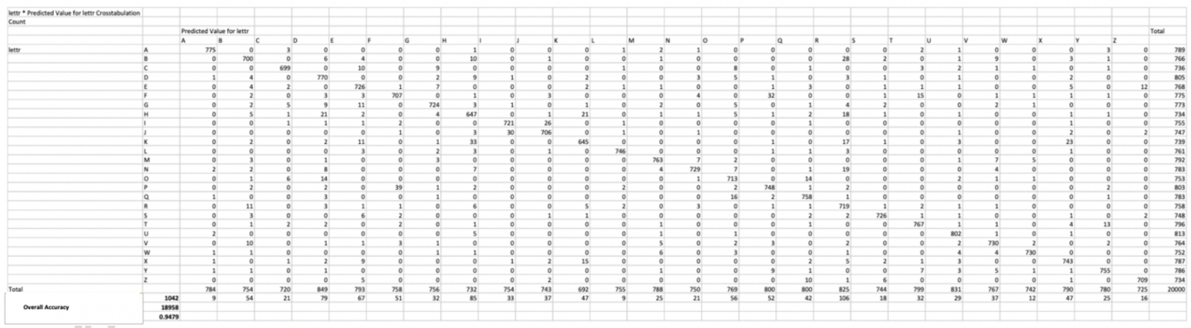
2. (16 points) Report the misclassification matrix and the appropriate performance metrics for different values of K (K=1, 3, 5, and 7).

K =1 Overall Accuracy = 95.29%



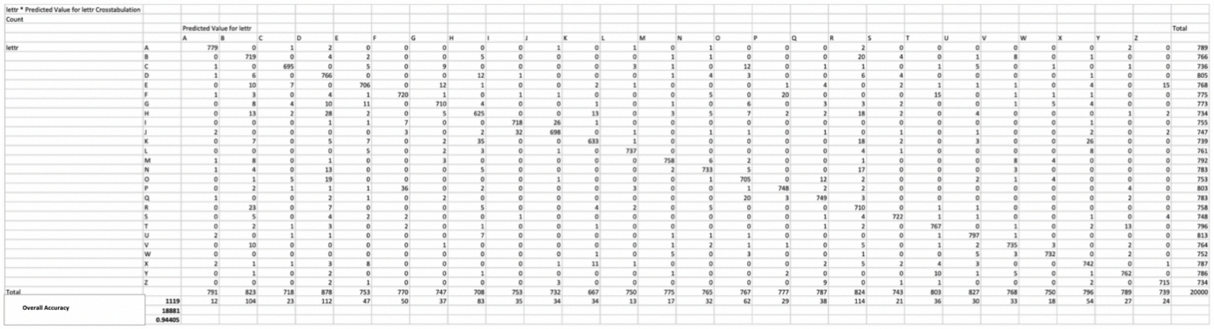
lettr * Predicted Value for lettr Crosstabulation																												
Count																												
		Predicted Value for lettr																										Total
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
lettr	A	781	1	1	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	789
	B	0	711	0	4	4	2	1	8	0	0	2	0	2	0	0	0	0	15	2	0	0	12	0	2	0	1	766
	C	0	0	713	0	4	0	7	0	0	0	2	1	0	0	4	0	0	1	0	1	0	1	2	0	0	0	736
	D	1	2	0	757	1	0	0	17	0	1	0	0	0	4	8	1	1	6	3	1	1	0	0	0	0	1	805
	E	0	2	9	0	714	0	10	3	0	0	4	3	0	0	0	1	2	0	4	0	0	3	0	5	0	8	768
	F	0	1	0	2	0	711	1	2	3	2	0	1	0	3	0	30	0	1	0	12	0	2	2	1	1	0	775
	G	0	2	4	3	12	0	733	1	1	0	2	1	0	0	5	0	2	0	1	0	0	3	2	1	0	0	773
	H	0	6	0	15	2	1	4	649	0	2	21	2	3	3	8	2	0	14	0	0	0	1	0	0	0	1	734
	I	0	0	0	0	1	2	0	0	724	27	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	755
	J	1	1	0	1	0	1	0	3	24	710	0	1	0	1	1	0	0	0	0	0	2	0	0	0	0	1	747
	K	0	3	0	2	6	0	3	28	0	0	665	0	0	0	0	1	0	10	0	0	2	1	0	18	0	0	739
	L	0	0	1	0	1	0	3	1	1	2	1	747	0	0	0	1	0	3	0	0	0	0	0	0	0	0	761
	M	0	3	0	0	1	0	2	1	0	0	0	0	769	5	1	0	0	0	0	0	0	8	2	0	0	0	792
	N	1	1	0	6	0	0	0	3	0	1	1	0	1	753	4	0	0	7	0	0	0	4	1	0	0	0	783
	O	0	0	5	10	0	0	2	1	0	0	0	0	0	1	723	0	5	0	0	0	4	0	2	0	0	0	753
	P	0	0	0	4	3	30	0	1	0	0	0	3	0	0	0	755	2	2	1	0	0	0	0	0	2	0	803
	Q	0	0	0	2	0	0	3	0	0	0	0	0	0	0	15	5	752	3	1	0	1	0	0	0	0	1	783
	R	0	24	0	3	0	0	1	13	0	1	10	2	0	9	0	0	1	691	0	0	0	2	1	0	0	0	758
	S	1	5	0	0	4	1	0	1	0	0	0	1	0	0	0	0	0	2	729	1	0	0	0	1	0	2	748
	T	0	2	1	1	0	5	0	1	1	1	0	0	0	0	1	0	0	0	1	767	0	0	0	2	13	0	796
	U	1	1	0	0	0	0	0	6	0	0	2	1	3	0	1	0	0	0	0	0	794	2	1	0	1	0	813
	V	0	8	0	0	0	1	1	0	0	0	0	0	2	3	1	2	0	0	0	0	2	739	2	0	3	0	764
	W	0	0	1	0	0	0	0	1	0	0	0	0	2	0	2	0	0	1	0	0	1	3	741	0	0	0	752
	X	0	2	0	2	6	0	0	1	0	0	19	1	0	0	1	0	0	1	2	3	0	0	0	747	1	1	787
	Y	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	9	2	6	1	1	763	0	786
	Z	0	0	0	0	3	1	0	0	0	1	0	0	0	0	0	0	7	0	2	0	0	0	0	0	0	720	734
Total		786	776	735	812	762	755	771	741	754	749	730	765	782	782	776	799	774	757	747	794	809	787	757	779	785	736	20000
Overall Accuracy	942	5	65	22	55	48	44	38	92	30	39	65	18	13	29	53	44	22	66	18	27	15	48	16	32	22	16	
	19058																											
	0.9529																											

K =3 overall accuracy = 74.79



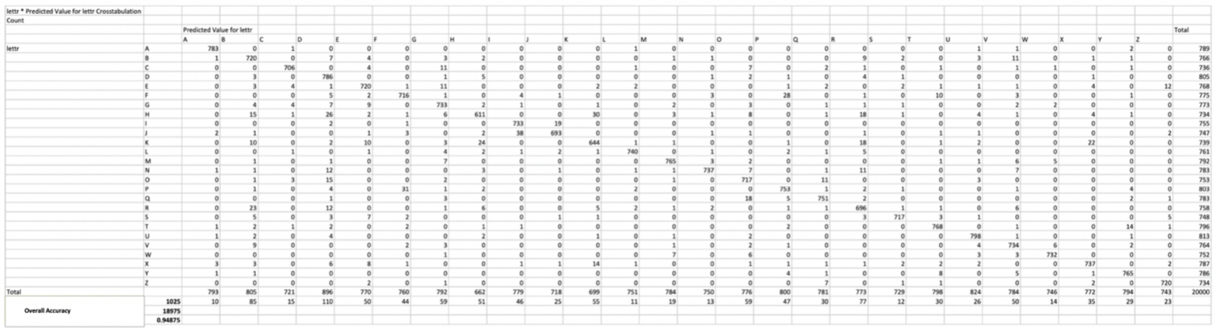
lettr * Predicted Value for lettr Crosstabulation																												
Count																												
		Predicted Value for lettr																										Total
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
lettr	A	775	0	3	0	0	0	0	1	0	0	0	1	2	1	0	0	0	0	0	2	1	0	0	0	3	0	789
	B	0	700	0	6	4	0	0	10	0	1	0	0	1	0	0	0	0	28	2	0	1	9	0	3	1	0	766
	C	0	0	699	0	10	0	9	0	0	0	0	1	0	0	8	0	1	0	0	3	2	1	1	0	1	0	736
	D	1	4	0	770	0	0	2	9	1	0	2	0	0	3	5	1	0	3	1	0	1	0	0	2	0	0	805
	E	0	4	2	0	726	1	7	0	0	0	2	1	1	0	0	1	3	0	1	1	1	0	0	5	0	12	768
	F	0	2	0	3	3	707	0	1	0	3	0	0	0	4	0	32	0	0	1	15	0	1	1	1	1	0	775
	G	0	2	5	9	11	0	724	3	1	0	1	0	2	0	5	0	1	4	2	0	0	2	1	0	0	0	773
	H	0	5	1	21	2	0	4	647	0	1	21	0	1	1	5	1	2	18	1	0	1	0	0	1	1	0	734
	I	0	0	1	1	1	2	0	0	721	26	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	755
	J	0	0	0	0	0	1	0	3	30	706	0	1	0	1	0	0	0	0	0	0	1	0	0	2	0	2	747
	K	0	2	0	2	11	0	1	33	0	0	645	0	0	0	0	1	0	17	1	0	3	0	0	23	0	0	739
	L	0	0	0	0	3	0	2	3	0	1	0	746	0	0	0	1	1	3	0	0	0	0	0	1	0	0	761
	M	0	3	0	1	0	0	3	0	0	0	0	0	763	7	2	0	0	0	0	0	1	7	5	0	0	0	792
	N	2	2	0	8	0	0	0	7	0	0	0	0	4	729	7	0	1	19	0	0	0	4	0	0	0	0	783
	O	0	1	6	14	0	0	0	0	0	0	0	0	0	1	713	0	14	0	0	0	2	1	1	0	0	0	753
	P	0	2	0	2	0	39	1	2	0	0	0	2	0	0	2	748	1	2	0	0	0	0	0	0	2	0	803
	Q	1	0	0	3	0	0	1	0	0	0	0	0	0	0	16	2	758	1	0	0	0	0	0	0	1	0	783
	R	0	11	0	3	1	1	0	6	0	0	5	2	0	3	0	1	1	719	1	2	1	1	0	0	0	0	758
	S	0	3	0	0	6	2	0	0	0	1	1	0	0	0	0	0	2	2	726	1	1	0	0	1	0	2	748
	T	0	1	2	2	0	2	0	1	0	0	0	0	1	0	0	0	1	0	0	767	1	1	0	4	13	0	796
	U	2	0	0	0	0	0	0	5	0	0	0	0	1	0	1	0	0	0	0	0	802	1	0	1	0	0	813
	V	0	10	0	1	1	3	1	0	0	0	0	0	5	0	2	3	0	2	0	0	2	730	2	0	2	0	764
	W	1	1	0	0	0	0	1	1	0	0	0	0	6	0	3	0	0	1	0	0	4	4	730	0	0	0	752
	X	1	0	1	2	9	0	0	0	1	2	15	0	0	0	0	0	2	5	2	1	3	0	0	743	0	0	787
	Y	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	9	1	0	0	7	3	5	1	1	755	0	786
	Z	0	0	0	0	5	0	0	0	0	2	0	0	0	0	0	0	10	1	6	0	0	0	0	1	0	709	734
Total		784	754	720	849	793	758	756	732	754	743	692	755	788	750	769	800	800	825	744	799	831	767	742	790	780	725	20000
Overall Accuracy	1042	9	54	21	79	67	51	32	85	33	37	47	9	25	21	56	52	42	106	18	32	29	37	12	47	25	16	
	18958																											
	0.9479																											

K=5 overall accuracy = 94.4%



lettr * Predicted Value for lettr Crosstabulation																												
Count																												
		Predicted Value for lettr																										Total
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
lettr	A	779	0	1	2	0	0	0	0	0	1	0	1	0	1	0	0	0	2	0	0	0	0	0	0	2	0	789
	B	0	719	0	4	2	0	0	5	0	0	0	0	1	1	0	0	0	20	4	0	1	8	0	1	0	0	766
	C	1	0	695	0	5	0	9	0	0	0	0	3	1	0	12	0	1	1	0	1	5	0	1	0	1	0	736
	D	1	6	0	766	0	0	0	12	1	0	0	0	1	4	3	0	0	6	4	0	0	0	0	1	0	0	805
	E	0	10	7	0	706	0	12	1	0	0	2	1	0	0	0	1	4	0	2	1	1	1	0	4	0	15	768
	F	1	3	0	4	1	720	1	0	1	1	0	0	0	5	0	20	0	0	0	15	0	1	1	1	0	0	775
	G	0	8	4	10	11	0	710	4	0	0	1	0	1	0	6	0	3	3	2	0	0	1	5	4	0	0	773
	H	0	13	2	28	2	0	5	625	0	0	13	0	3	5	7	2	2	18	2	0	4	0	0	0	1	2	734
	I	0	0	0	1	1	7	0	0	718	26	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	755
	J	2	0	0	0	0	3	0	2	32	698	0	1	0	1	1	0	1	0	1	0	1	0	0	2	0	2	747
	K	0	7	0	5	7	0	2	35	0	0	633	1	0	0	0	0	0	18	2	0	3	0	0	26	0	0	739
	L	0	0	0	0	5	0	2	3	0	1	0	737	0	0	0	0	0	4	1	0	0	0	0	8	0	0	761
	M	1	8	0	1	0	0	3	0	0	0	0	0	758	6	2	0	0	1	0	0	0	8	4	0	0	0	792
	N	1	4	0	13	0	0	0	5	0	0	0	0	2	733	5	0	0	17	0	0	0	3	0	0	0	0	783
	O	0	1	5	19	0	0	0	0	0	1	0	0	0	1	705	0	12	2	0	0	2	1	4	0	0	0	753
	P	0	2	1	1	1	36	0	2	0	0	0	3	0	0	1	748	2	2	0	0	0	0	0	0	4	0	803
	Q	1	0	0	2	1	0	2	0	0	0	0	0	0	0	20	3	749	3	0	0	0	0	0	0	2	0	783
	R	0	23	0	7	0	0	0	5	0	0	4	2	0	5	0	0	0	710	0	1	1	0	0	0	0	0	758
	S	0	5	0	4	2	2	0	0	1	0	0	0	0	0	0	0	1	4	722	1	1	0	0	1	0	4	748
	T	0	2	1	3	0	2	0	1	0	0	1	0	0	0	0	0	1	2	0	767	0	1	0	2	13	0	796
	U	2	0	1	1	0	0	0	7	0	0	0	0	1	1	1	0	0	0	0	1	797	1	0	0	0	0	813
	V	0	10	0	0	0	0	1	0	0	0	0	0	1	2	1	1	0	5	0	1	2	735	3	0	2	0	764
	W	0	0	0	0	0	0	0	0	0	0	1	0	5	0	3	0	0	1	0	0	5	3	732	0	2	0	752
	X	2	1	1	3	8	0	0	0	0	1	11	1	0	0	0	0	2	5	2	4	3	0	0	742	0	1	787
	Y	0	1	0	2	0	0	0	1	0	0	0	0	1	0	0	2	0	0	0	10	1	5	0	1	762	0	786
	Z	0	0	0	2	1	0	0	0	0	3	0	0	0	0	0	0	9	0	1	1	0	0	0	2	0	715	734
Total		791	823	718	878	753	770	747	708	753	732	667	750	775	765	767	777	787	824	743	803	827	768	750	796	789	739	20000
Overall Accuracy	1119	12	104	23	112	47	50	37	83	35	34	34	13	17	32	62	29	38	114	21	36	30	33	18	54	27	24	
	18881																											
	0.94405																											

K =7 overall accuracy is 94.875%



lettr * Predicted Value for lettr Crosstabulation																												
Count																												
		Predicted Value for lettr																										Total
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
lettr	A	783	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	2	0	789
	B	1	720	0	7	4	0	3	2	0	0	0	0	1	1	0	0	0	9	2	0	3	11	0	1	1	0	766
	C	0	0	706	0	4	0	11	0	0	0	0	1	0	0	7	0	2	1	0	1	0	1	1	0	1	0	736
	D	0	3	0	786	0	0	1	5	0	0	0	0	0	1	2	1	0	4	1	0	0	0	0	1	0	0	805
	E	0	3	4	1	720	1	11	0	0	0	2	2	0	0	0	1	2	0	2	1	1	1	0	4	0	12	768
	F	0	0	0	5	2	716	1	0	4	1	0	0	0	3	0	28	0	1	0	10	0	3	0	0	1	0	775
	G	0	4	4	7	9	0	733	2	1	0	1	0	2	0	3	0	1	1	1	0	0	2	2	0	0	0	773
	H	0	15	1	26	2	1	6	611	0	0	30	0	3	1	8	0	1	18	1	0	4	1	0	4	1	0	734
	I	0	0	0	2	0	1	0	0	733	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	755
	J	2	1	0	0	1	3	0	2	38	693	0	0	0	1	1	0	0	1	0	1	1	0	0	0	0	2	747
	K	0	10	0	2	10	0	3	24	0	0	644	1	1	0	0	1	0	18	0	1	2	0	0	22	0	0	739
	L	0	0	1	0	1	0	4	2	1	2	1	740	0	1	0	2	1	5	0	0	0	0	0	0	0	0	761
	M	0	1	0	1	0	0	7	0	0	0	0	0	765	3	2	0	0	0	0	1	1	6	5	0	0	0	792
	N	1	1	0	12	0	0	0	3	0	1	0	1	1	737	7	0	1	11	0	0	0	7	0	0	0	0	783
	O	0	1	3	15	0	0	2	0	0	0	0	0	1	0	717	0	11	0	0	0	3	0	0	0	0	0	753
	P	0	1	0	4	0	31	1	2	0	0	0	2	0	0	0	753	1	2	1	0	0	1	0	0	4	0	803
	Q	0	0	0	1	0	0	3	0	0	0	0	0	0	0	18	5	751	2	0	0	0	0	0	0	2	1	783
	R	0	23	0	12	0	0	1	6	0	0	5	2	1	2	0	1	1	696	1	1	0	6	0	0	0	0	758
	S	0	5	0	3	7	2	0	0	0	1	1	0	0	0	0	0	0	3	717	3	1	0	0	0	0	5	748
	T	1	2	1	2	0	2	0	1	1	0	0	0	0	0	0	2	0	0	0	768	0	1	0	0	14	1	796
	U	1	2	0	4	0	0	0	2	0	0	1	0	1	0	2	0	0	0	0	0	798	1	0	0	1	0	813
	V	0	9	0	0	0	2	3	0	0	0	0	0	1	0	2	1	0	0	0	0	4	734	6	0	2	0	764
	W	0	0	0	0	0	0	1	0	0	0	0	0	7	0	6	0	0	0	0	0	3	3	732	0	0	0	752
	X	3	3	0	6	8	1	0	0	1	1	14	1	0	0	1	1	1	1	2	2	2	0	0	737	0	2	787
	Y	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	8	0	5	0	1	765	0	786
	Z	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	7	0	1	1	0	0	0	2	0	720	734
Total		793	805	721	896	770	760	792	662	779	718	699	751	784	750	776	800	781	773	729	798	824	784	746	772	794	743	20000
Overall Accuracy	1025	10	85	15	110	50	44	59	51	46	25	55	11	19	13	59	47	30	77	12	30	26	50	14	35	29	23	
	18975																											
	0.94875																											

3. (2 points) Interpret the results and also compare them with the ones obtained by using the decision trees

If compared with the decision trees results, for the values for the k-nearest neighbor have a higher accuracy result for every k-value that was being tested But they are almost equal to decision tree.