Assignment 4

Kubra Iqbal

Due Date: Saturday, November 3rd, by midnight

Total number of points: 50 points

Problem 1 (10 points):

A. (2 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 – The training phase of the algorithm consists of storing the feature vectors and class labels of the training samples. In the testing phase, a test point is classified by assigning the label which are most frequent among the k training samples nearest to that query point.

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

False

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.

Yes – that is correct. K-nearest-neighbors works well with small number of input variables and struggles when the number of input is very large.

B. (4 points) Would the following binary classifiers be able to correctly separate the training data (circles vs. triangles) given in Figure 1? Briefly explain your answer and show the decision boundary for each one of the two classifiers:

1. Decision tree classifier
2. 3-nearest neighbor classifier with the Euclidean distance

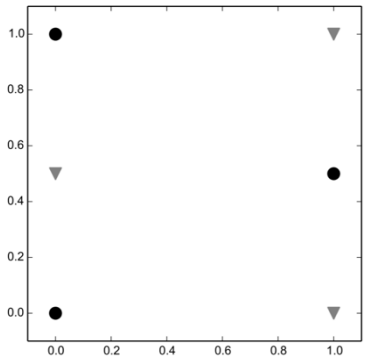

Figure 1: Training data

Decision tree classifier: yes – you can certainly partition the space with lines orthogonal to the axes in such a way that every sample ends up in a different region.

3 Nearest Neighbor classifier: The 3 nearest neighbors of any point in the training set are 1 of the same class and 2 of the opposite class. According to this, 3NN will be symmetrically wrong.

C. (4 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.

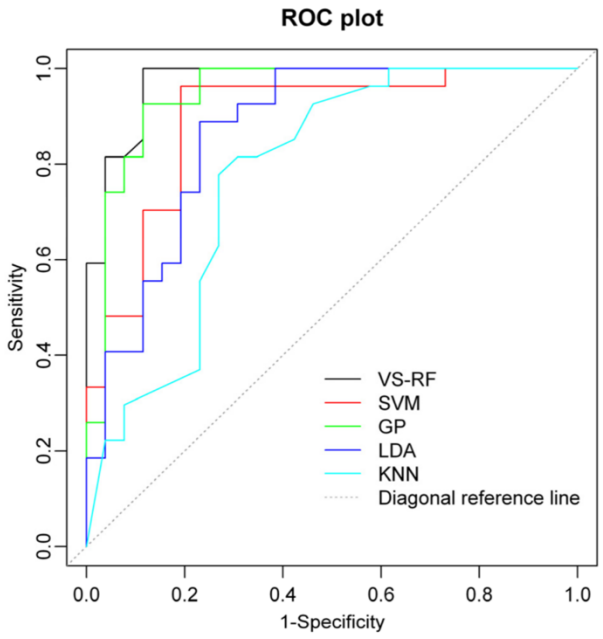

Figure 2: : ROC curves for several algorithms classifying molecules according to their action on HRSV, computed on a test set. Sensitivity = True Positive Rate. Specificity = 1 - False Positive Rate. VS-RF: Random Forest. SVM: Support Vector Machine. GP: Gaussian Process. LDA: Linear Discriminant Analysis. 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.

Random forest. (top line)

This is because 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. It also has the highest area under the ROCC curve.

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?

We want low false positive rate(so as to ensure there are most promising compounds among those that will be selected for further development, therapeutic development will end up being costly) which means it is high specificity. We are overall concerned on the left part of the curve.

Overall ROC curve is most useful in early stages of the evolution of a new test. Once the test is established – only a portion of the ROC curve is of interest. Similar to sensitivity and specificity, ROC curves are invariant to the prevalence of a disease but depend on the patient characteristics and the disease. The latter is not possible with sensitivity and specify measures because a change in the cut point to classify the test results as positive or negative could affect the two tests differently.

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. Steps to reduce overfitting will include: add more data, use data augmentation, reduce architecture complexity, use data that generalizes well.

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.

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 and minimizing the risk value for training and test. 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: 0.93for training and 0.78 for testing.

Classification

Sample	Observed	Predicted 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	397	0	0	0	0	0	1	0	2	0	0	0	2	0	0	1	0	0	1	1	0	1	0	0	0	0	97.8%
	B	0	358	1	1	0	1	3	1	6	0	1	0	1	1	1	0	0	3	2	0	0	2	1	0	0	0	93.5%
	C	2	0	322	0	1	1	4	2	3	0	0	0	1	1	2	2	3	4	2	2	1	2	0	2	0	1	89.9%
	D	0	4	0	364	0	1	1	6	3	2	1	0	0	3	4	0	0	3	0	2	0	0	0	1	0	0	92.2%
	E	1	1	4	0	349	3	4	2	3	1	1	3	2	1	0	0	0	1	3	0	0	0	0	3	0	0	91.4%
	F	0	2	0	3	0	359	1	2	4	0	0	0	0	0	1	8	0	2	4	1	0	0	0	0	1	0	92.5%
	G	3	1	1	2	1	0	336	0	5	2	0	0	1	0	2	0	5	1	2	1	1	2	1	1	1	0	91.1%
	H	1	1	1	5	1	2	2	339	6	1	3	0	1	0	7	1	2	8	0	0	1	1	0	0	1	0	88.3%
	I	0	0	1	0	1	1	0	0	383	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	98.2%
	J	2	0	0	0	0	2	0	1	9	325	1	1	0	1	2	4	2	0	3	0	0	0	0	2	1	1	91.0%
	K	1	0	0	0	0	0	4	6	0	1	334	1	2	1	3	0	1	6	3	1	0	0	3	5	1	0	89.5%
	L	0	2	1	1	3	0	1	1	1	1	0	364	0	1	1	2	1	0	5	1	0	0	1	0	0	1	93.8%
	M	1	1	0	2	0	3	1	2	2	0	2	0	359	0	2	0	0	4	1	0	0	0	2	0	0	0	94.0%
	N	2	1	1	1	1	0	1	1	1	3	1	0	2	347	2	0	0	5	0	1	1	0	4	0	0	0	92.5%
	O	1	0	2	2	0	0	2	0	2	0	0	0	0	2	345	1	4	1	0	0	4	1	2	0	0	0	93.5%
	P	0	1	0	0	0	10	1	0	4	0	0	0	0	0	3	397	0	3	1	0	1	0	2	0	0	0	93.9%
	Q	3	3	0	1	2	1	2	0	3	1	1	0	0	1	10	1	340	3	3	0	0	2	0	4	1	0	89.0%
	R	2	3	0	3	0	1	0	3	6	0	4	1	0	3	1	1	0	347	0	1	1	0	0	1	0	1	91.6%
	S	0	4	3	0	0	2	1	1	7	3	1	0	0	1	0	0	1	4	339	1	0	0	2	0	0	4	90.6%
	T	0	0	3	0	2	4	1	0	1	0	2	0	0	0	1	0	2	0	0	385	0	1	1	4	2	0	94.1%
	U	0	0	0	0	0	1	1	0	1	1	0	0	2	3	6	0	1	1	0	0	382	2	3	0	2	0	94.1%
	V	0	1	0	1	0	8	0	0	0	0	1	0	0	1	1	0	0	3	2	2	0	361	5	0	0	0	93.5%
	W	0	1	0	0	0	2	2	1	1	1	1	1	2	0	0	0	0	0	0	0	0	0	329	0	1	0	96.2%
	X	2	3	1	2	0	0	0	2	4	2	2	0	1	0	1	0	0	0	0	0	0	0	0	390	3	1	94.2%
	Y	0	0	0	1	0	4	0	3	1	2	0	0	0	0	0	0	1	2	0	2	0	1	1	0	380	1	95.2%
	Z	1	1	0	0	4	1	1	0	1	0	1	1	0	0	2	1	2	0	1	1	0	1	0	0	0	362	95.0%
	Overall Percentage	4.2%	3.9%	3.4%	3.9%	3.7%	4.1%	3.7%	3.7%	4.6%	3.5%	3.6%	3.7%	3.8%	3.7%	4.0%	4.2%	3.7%	4.0%	3.7%	4.0%	3.9%	3.8%	3.6%	4.1%	3.9%	3.8%	93.0%
Test	A	341	3	2	2	0	0	2	2	0	2	2	2	6	3	1	1	2	0	6	0	1	0	0	1	1	3	89.0%
	B	4	291	0	16	3	0	8	9	9	3	4	0	0	0	10	0	1	5	11	1	0	6	0	2	0	0	76.0%
	C	2	2	300	1	14	2	11	2	3	0	2	0	1	0	7	4	4	4	1	7	7	0	0	1	3	0	79.4%
	D	1	18	0	309	0	2	4	16	2	5	1	0	3	6	21	2	0	6	7	1	0	0	1	4	0	1	75.4%
	E	0	2	10	1	304	4	7	1	1	0	12	4	1	0	1	0	4	2	7	1	1	2	1	9	0	11	78.8%
	F	0	5	2	3	1	286	1	6	12	2	0	0	0	3	5	29	0	0	5	11	0	1	1	3	9	2	73.9%
	G	7	5	9	4	9	3	285	11	11	0	9	0	0	0	15	0	6	7	4	1	4	9	1	1	1	2	70.5%
	H	1	7	3	16	1	6	8	241	4	7	5	1	3	1	9	6	3	10	4	2	1	2	6	1	1	1	68.9%
	I	2	6	7	2	2	5	2	1	305	15	0	3	0	0	1	4	1	0	3	0	0	0	0	1	0	5	83.6%
	J	3	0	1	6	0	3	3	4	11	332	4	1	0	3	2	3	1	2	4	0	0	0	1	4	1	1	85.1%
	K	6	1	4	2	3	2	2	11	0	0	287	4	3	3	1	1	1	17	1	3	1	0	1	12	0	0	78.4%
	L	8	2	6	2	3	1	0	1	3	3	2	313	1	0	1	2	4	1	7	3	2	0	0	7	0	1	83.9%
	M	5	0	0	6	0	3	4	5	2	5	3	4	326	6	3	2	2	3	0	3	8	3	14	0	2	1	79.5%
	N	5	6	3	8	2	4	3	2	1	4	2	1	5	313	8	7	0	6	1	1	4	3	16	1	2	0	76.7%
	O	3	2	2	16	0	1	11	5	3	1	0	0	1	4	280	2	21	11	2	1	5	4	6	2	1	0	72.9%
	P	0	2	0	3	0	36	1	4	2	5	0	0	0	1	0	306	0	0	3	6	1	1	3	0	6	0	80.5%
	Q	1	0	2	5	11	1	11	5	12	2	0	2	1	3	27	1	296	4	6	1	2	1	1	3	3	0	73.8%
	R	3	16	1	9	5	2	2	19	7	3	4	4	1	11	4	0	5	269	3	0	1	2	2	4	0	2	71.0%
	S	4	8	4	6	4	5	6	3	8	4	3	3	3	2	4	5	4	4	272	7	1	1	1	3	1	8	72.7%
	T	0	0	1	2	3	7	1	0	1	6	2	0	2	1	2	1	2	1	5	325	1	4	1	4	12	3	84.0%
	U	0	1	7	2	2	1	5	5	2	1	3	1	7	6	11	2	1	4	2	1	332	5	2	1	2	1	81.6%
	V	0	3	0	1	0	13	0	4	1	1	0	0	0	2	2	4	1	0	2	3	4	320	11	0	5	1	84.7%
	W	1	1	0	2	2	1	2	0	1	0	0	0	12	3	2	3	1	1	0	0	7	7	358	0	6	0	87.3%
	X	3	1	5	0	22	8	0	7	4	1	5	1	3	1	0	2	2	3	9	3	0	0	0	291	0	2	78.0%
	Y	1	3	0	0	1	9	1	4	2	2	2	2	1	2	2	4	3	2	3	10	1	10	3	0	313	6	80.9%
	Z	0	1	1	3	10	2	3	4	3	0	2	1	0	1	4	1	2	0	5	8	0	0	0	8	0	294	83.3%
	Overall Percentage	4.0%	3.9%	3.7%	4.3%	4.0%	4.1%	3.8%	3.7%	4.1%	4.0%	3.5%	3.5%	3.8%	3.7%	4.2%	3.9%	3.7%	3.6%	3.7%	4.0%	3.8%	3.8%	4.3%	3.6%	3.7%	3.4%	78.8%

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, V10, V14, V13, V9, V15, V16, V17, V4, V3, V6, V2, V5
	Number of Nodes	1555
	Number of Terminal Nodes	778
	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.

Attached below is the misclassification matrix for the best tree configuration. The model was performed with and has high 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.

```
Class Distribution:
    789 A     766 B     736 C     805 D     768 E    775 F
773 G
    734 H     755 I     747 J     739 K     761 L    792 M
783 N
    753 O     803 P     783 Q     758 R     748 S    796 T
813 U
    764 V     752 W     787 X     786 Y     734 Z
```

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

The most important three attributes would be : xege, xybar and xegvy.

(The table for independent variable importance table being listed as the first three independent variables below)

Independent Variable	Importance	Normalized Importance
xege	0.319	100.00%
xybar	0.317	99.50%
xegvy	0.299	93.80%
xy2br	0.29	90.90%
yege	0.288	90.30%
x2ybr	0.288	90.20%
x2bar	0.28	87.70%
y2bar	0.276	86.50%
yegvx	0.246	77.10%
xbar	0.245	76.70%
ybar	0.235	73.80%
xbox	0.133	41.60%
ybox	0.13	40.80%
onpix	0.13	40.80%

Problem 3 (20points):

On the same data from Problem 1, 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 was no data transformation done because the data has already been transformed. (It explains that in the information section). It is better to determine before running a K-nearest neighbor classification if the independent variables need to be scaled. This process is done so it can prevent distance measures from being dominated by other attributes that are present in the data.

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

Applied crosstabs between predicted in the columns and observed on the rows for each k-value computed in spss.
The accuracy was calculated in excel.

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	1	0	0	1	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	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	0	2	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	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	0	2	1	0	0	0	0	1	723	0	5	0	0	0	0	4	0	2	0	0	753		
	P	0	0	0	4	3	30	0	0	1	0	0	0	3	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	0	1	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		
Sum of misscalulations	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			
Sum of all correctly classified	19058																													
Overall Accuracy	0.9529																													

K = 1

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	1	0	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	0	2	0	3	0	1	1	1	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	1	0	4	0	31	1	2	0	0	0	2	0	0	753	1	2	1	0	0	1	0	4	0	0	803		
	Q	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	18	5	751	2	0	0	0	0	0	2	1	783		
	R	0	23	0	12	0	0	0	1	6	0	0	5	2	1	2	0	1	1	696	1	1	0	6	0	0	0	758		
	S	0	5	0	3	7	2	0	0	0	1	1	0	0	0	0	0	0	0	3	717	3	1	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	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	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		
Sum of misscalulations	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			
Sum of all correctly classified	18975																													
Overall Accuracy	0.94875																													

K=3

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	0	2	0	0	0	0	0	10	1	6	0	0	0	0	0	1	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
Sum of misscalulations	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			
Sum of all correctly classified	18958																													
Overall Accuracy	0.9479																													

K=5

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
Sum of misscalulations	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			
Sum of all correctly classified	18881																													
Overall Accuracy	0.94405																													

K=7

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. K=1 is too small and k=3 would be the right balance to the most predictive model – as noticed that the right balance to the most predictive model as the accuracy value is decreasing.

If the big picture is seen and is compared to the original researches, the use of decision trees or k-nearest neighbor has shown that the acracy is high

K VALUE	ACCURACY
1	0.952
3	0.948
5	0.947
7	0.944

Submission Instructions

1. Answer the problems and write your answers in a Word document.
2. Submit your file online at the website at **http://d2l.depaul.edu** and check your submission
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.
5. Late submissions are allowed with a 5%, 10%, and 15% penalty for a one day, two days, and three days, respectively.
6. No late work will be accepted after three days since the assignment was due.