Machine Learning - Quiz 3

Name: ____

Directions: Write complete solutions with enough detail so that your reasoning is clear to Prof. Chakraborty.

Question 1 [5 points]

Consider the toy dataset below.

Obs.	Y	X
1	A	1.0
2	A	1.8
3	B	3.2
4	A	4
5	B	5
6	B	5.8

We want to evaluate a KNN classifier with $K = 3$ using Leave-One-Out Cross Validation (LOOCV). **Obtain the cross-validation accuracy.**

[Hint: For each round of the LOOCV process, one observation is left out as the validation fold and the model is built on all the other observations.]

Question 2

Prof. Chakraborty was tasked to classify whether a banknote is `authentic` or not (‘Yes’-‘No’ response) based on the following variables measured from banknote images:

- `variance`,
- `skewness`,
- `kurtosis`,
- `entropy`, and
- `old` - ‘Yes’ or ‘No’?

Step 1: The following outputs show the results of his data exploration phase.

```r
glimpse(banknote)
```

```
## Rows: 1,372
## Columns: 6
## $ variance  <dbl> 3.62160, 4.54590, 3.86600, 3.45660, NA, NA, 3.59120, 2.09220~
## $ skewness  <dbl> 8.6661, NA, -2.6383, 9.5228, -4.4552, 9.6718, 3.0129, -6.810~
## $ kurtosis  <dbl> -2.80730, -2.45860, 1.92420, -4.01120, 4.57180, -3.96060, 0.~
## $ entropy   <dbl> -0.44699, -1.46210, 0.10645, -3.59440, -0.98880, -3.16250, 0~
## $ old       <fct> No, Yes, Yes, No, Yes, No, No, No, Yes, No, Yes, Yes, No, Ye~
## $ authentic <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, ~
```

```r
summary(banknote)
```

```
##     variance          skewness          kurtosis          entropy       
##  Min.   :-7.0421   Min.   :-13.773   Min.   :-5.2861   Min.   :-8.5482  
##  1st Qu.:-1.7976   1st Qu.: -1.862   1st Qu.:-1.5572   1st Qu.:-2.3931  
##  Median : 0.4957   Median :  2.249   Median : 0.6286   Median :-0.5996  
##  Mean   : 0.4382   Mean   :  1.795   Mean   : 1.3621   Mean   :-1.1793  
##  3rd Qu.: 2.8297   3rd Qu.:  6.642   3rd Qu.: 3.0895   3rd Qu.: 0.4003  
##  Max.   : 6.8248   Max.   : 12.952   Max.   :17.9274   Max.   : 2.4495  
##  NA's   :203       NA's   :191       NA's   :179       NA's   :204      
##   old       authentic
##  No :696   Yes:610  
##  Yes:676   No :762  
##                     
##                     
##                     
##                     
##                     
```

```r
nearZeroVar(banknote, saveMetrics = TRUE)
```

```
##           freqRatio percentUnique zeroVar   nzv
## variance   1.333333    83.4548105   FALSE FALSE
## skewness   1.000000    79.2274052   FALSE FALSE
## kurtosis   1.250000    81.1953353   FALSE FALSE
## entropy    1.250000    73.6151603   FALSE FALSE
## old        1.029586     0.1457726   FALSE FALSE
## authentic  1.249180     0.1457726   FALSE FALSE
```

Step 2: He then did a 80-20 split of the data into training (1098 observations) and test sets (274 observations).

Step 3: The next step was to create the blueprint and obtain the baked train and test datasets.

What blueprint steps should he use for this dataset? Provide a brief explanation of each step. Also, mention the order in which the blueprint steps should be implemented. [**5 points**]

You don't need to write any code to answer this question, but provide sufficient explanation of your blueprint steps.

Step 4: With the appropriate blueprint, he then implemented 5-fold CV repeated 1 time for each of the models below using the **Accuracy** metric.

- Logistic regression;
- KNN classifier with a grid of $K = 1, 11, 21, 31, 41, 51$.

The following results show the output of the CV process.

```
logistic_cv$results   # CV results of logistic regression model
```

```
##   parameter  Accuracy     Kappa AccuracySD   KappaSD
## 1      none 0.8861727 0.7695544 0.01348234 0.0269773
```

```
knn_cv$results   # CV results of KNN
```

```
##    k  Accuracy     Kappa AccuracySD    KappaSD
## 1  1 0.9353425 0.8689438 0.00585488 0.01177374
## 2 11 0.9253010 0.8495100 0.01065018 0.02111586
## 3 21 0.9280365 0.8551785 0.02074984 0.04182280
## 4 31 0.9289456 0.8571491 0.01604748 0.03235076
## 5 41 0.9216604 0.8423971 0.01894791 0.03838448
## 6 51 0.9216563 0.8421290 0.02029523 0.04113513
```

What is the optimal value of K for the KNN classifier? Which model performs best in this context? Explain your choice. [3 points]

Approximately, how many observations are included in the validation fold and training fold at each round of this CV process? Show work to explain your answer. [2 points]

Step 5: Finally, with the optimal model, he obtained class label predictions on the test set (using a threshold of 0.5). The corresponding confusion matrix is shown below.

```
##          reference
## predicted Yes  No
##       Yes 112   4
##       No   10 148
```

Calculate the test set accuracy. [2 points]

Question 3 [3 points]

Indicate which of (i) through (iv) is correct. **Justify your answer in terms of the bias-variance trade-off and the ideas of overfitting and underfitting.**

The LASSO (regularization method), relative to least squares (ordinary regression), is:

(i) More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

(ii) More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

(iii) Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

(iv) Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.