CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Worksheet 11

Name (Print):________________________________ Section:______________

1. Fill in the corresponding formulas:

Metric	Formula	Description
Recall		Fraction of actual positives that are correctly identified.
Precision		Fraction of positive predictions that are correct.
Accuracy		Fraction of all predictions (positive and negative) that are correct.

2. A spam detection system is evaluated on a dataset of 100 emails. The actual and predicted labels are summarized in the confusion matrix below:

	Predicted: Not Spam	**Predicted: Spam**
Actual: Not Spam	40	10
Actual: Spam	5	45

Fill in the following values based on the table:

(a) True Positives (TP):

(b) True Negatives (TN):

(c) False Positives (FP):

(d) False Negatives (FN):

3. True/False
Accuracy is always the best metric in imbalanced datasets

4. The goal of SVM is to find the ________________ that separates data points from different classes with the largest ________________.

5. True/False
SVMs can only be used for linearly separable data

6. What does the margin in SVM represent?
   (a) The distance between two data points
   (b) The distance between the hyperplane and the nearest data points
   (c) The width of the dataset

7. True/False
   SVM is effective for high-dimensional datasets

8. The k-Nearest Neighbors algorithm predicts the target value for a given query point based on the ________________ of the k ________________ neighbors

9. Why do you choose an odd number for k?

10. Why is it important to use feature scaling (e.g., normalization or standardization) when using k-NN?

11. Which method is often used to determine the optimal k?
    (a) Elbow
    (b) Euclidean distance
    (c) Cross-validation

12. In K-fold cross-validation, the dataset is split into ________________ (hint: how many) equal-sized subsets called folds

13. True/False
    LOOCV is computationally efficient for large datasets

14. True/False
    In K-fold cross-validation, the number of training samples is reduced by the size of one fold in each iteration