Mid (Fall 22) Machine Learning

Q1) True / False

1. Unsupervised Learning enables an agent to learn in an interactive environment through trial and error using the feedback of its own actions and experiences. False

2. Cross-validation splits the dataset into k-partitions (folds) and is preferable only for large datasets. False

3. Gradient descent is an optimization algorithm used to find the values of parameters of a function that minimizes a cost function. True

4. Logistic regression is an example of supervised learning, and it is used to calculate or predict the probability of a binary (yes/no) event (class) occurring. True

5. All learning algorithms require transforming labels (categorical feature) into numbers. False

6. Just removing data examples with missing features from a dataset could be an effective way to deal with missing values sometimes. True

7. Binning (also called bucketing) is the process of converting a continuous feature into multiple binary features called bins or buckets, typically based on value range. True

8. An underfitting problem happens when a model predicts very well the training data but poorly the data from a holdout set (e.g., testing set). False

9. Accuracy is not a useful metric when errors in predicting all classes are equally important. False

10. Model-based learning algorithms use the whole dataset as the model such as k-Nearest Neighbors (kNN). False

Q2)

1. How Logistic Regression is different from Linear Regression? Name two differences between them.

2. Name a significant disadvantage of Decision Tree models.

3. What is the purpose of the kernel function in SVM?

4. How does the k-Nearest Neighbors (kNN) algorithm calculate the distance between the data points?

5. Show how this dataset is transformed by the One-Hot encoding process.

Car_Brand	Car_Color
Toyota	White
Nissan	Black
Ford	White
Nissan	Red

Q3)

Suppose that we have built a model to classify student performance in a course into the following classes: Poor; Average, and Excellent. After evaluating the model on a test set, we got the following confusion matrix:

		Predicted		
		Poor	Average	Excellent
Actual	Poor	10	4	1
	Average	2	34	6
	Excellent	2	5	36

1. Calculate the accuracy of the classifier.
2. Calculate the precision of the "Poor" class.
3. Calculate the recall of the "Excellent" class.
4. Is the test set balanced or not? Justify your answer.