DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF BARISHAL

FINAL EXAMINATION-2022
Course Title: Machine Learning
Course Code: CSE-4213
4th Year 2nd Semester
Session: 2018-19

Time: 3 hours **Marks:** 60

Answer any five Questions from the followings.

1. a) Define machine learning. How is it different from traditional programming? [4]

   b) What are the factors affecting the performance of machine learning algorithm? [4]

   c) Explain semi supervised and reinforcement learning? [4]

2. a) Explain the difference between classification and regression in machine learning with example. [3]

   b) Why data preprocessing is important in machine learning? Describe three popular data preprocessing techniques in machine learning. [3]

   c) You are evaluating the performance of a machine learning model on a classification problem. You are given the following data: [6]

Instance	True Label	Predicted Label
1	A	A
2	B	A
3	A	B
4	B	B
5	A	A
6	B	B
7	A	B
8	B	A

Calculate the following performance metrics for the classification problem:

I. Accuracy
II. Precision
III. Recall
IV. F1 Score

3. a) Given the dataset for a classification task, use the k Nearest Neighbor (kNN) algorithm with k=3 to classify the data object with features (t1=1,t2=2,t3=3,t4=4,t5=4). Apply Euclidean distance as the distance metric. [6]

t_1	t_2	t_3	t_4	t_5	Class
2	3	4	5	5	Buy
0	1	2	3	5	Buy
2	2	2	2	4	Sell
3	3	3	3	3	Sell
4	2	4	4	4	Sell

   b) Distinguish between classification and Clustering. List the applications of clustering and identify advantages and disadvantages of clustering algorithm. [3]

   c) Estimate the problems associated with clustering large data. [3]

4. a. You are given a dataset containing information about customer purchases at a retail store. Each customer is represented by two features: the amount spent on electronics and the amount spent on clothing. The dataset is as follows: [8]

Customer	Electronics	Clothing
1	200	50
2	150	60
3	300	200
4	250	180
5	120	30
6	350	300
7	400	250
8	180	70

Now, group these customers into k=2 clusters. You can use K-means clustering algorithm.

b) Explain the primary difference between K-means clustering and Principal Component Analysis (PCA) in the context of unsupervised learning. Provide one application where each method is typically used. [4]

5. a) Define overfitting and underfitting in machine learning. Provide a visual explanation for each case using a polynomial regression model. [3]

b) A model achieves 95% accuracy on training data but only 60% on test data. Is this an example of overfitting or underfitting? Justify your answer. [3]

c) Imagine we have a dataset where we classify emails as "Spam" or "Not Spam" based on whether they contain specific words. [6]

Email	Word1	Word2	Class
1	Yes	No	Spam
2	Yes	Yes	Not Spam
3	No	Yes	Spam
4	No	No	Not Spam

Now, classify a new email that contains "Word1" but not "Word2" by using Naïve Bayesian classification.

6. a) Compare and contrast decision trees, random forests, and support vector machines as classification algorithms. [4]

b) Explain the concept of the kernel trick in SVM. How does the kernel trick allow SVMs to perform classification in non-linearly separable data spaces? Provide examples of common kernel functions used in SVMs. [4]

c) Write down the mathematical formulation of the SVM optimization problem for a binary classification task. Describe the objective function and the constraints involved. [4]

7. a) What are ensemble methods in machine learning? Explain the difference between bagging and boosting with examples. [4]

b) Describe how the Random Forest algorithm works as an ensemble method. How does it improve upon individual decision trees? [4]

c) You are given the predictions from three models on a binary classification task. [4]

Model 1 predicts: [0, 1, 0, 1],
Model 2 predicts: [0, 1, 1, 1],
Model 3 predicts: [1, 1, 0, 0].

Using majority voting, compute the final ensemble prediction for each case.

8. You are provided with a dataset containing information about loan applications. Each application has the following features: [12]

Application	Credit Score	Employment Status	Income	Loan Approved
1	High	Employed	High	Yes
2	Medium	Employed	Medium	Yes
3	Low	Unemployed	Low	No
4	High	Employed	Medium	Yes
5	Medium	Unemployed	High	No
6	Low	Employed	Low	No
7	High	Unemployed	High	Yes
8	Medium	Employed	High	Yes
9	Low	Unemployed	Medium	No
10	High	Employed	Low	No

Now solve the following problem:

I. Calculate the Entropy for the entire dataset.
II. Calculate the Information Gain for each feature (Credit Score, Employment Status, and Income).
III. Calculate the Gini Index for each feature (Credit Score, Employment Status, and Income)

Good Luck!!!