UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 5 Examination, January 2024

Module No: EE5253 **Module Name: Machine Learning**

Part-A
[**45 minutes**]

Instructions for candidates

- Write your index number on top of every page.
- Question paper contains 20 multiple choice questions.
- Answer all questions. Each question has only one answer.
- Each question carries 0.5 mark.
- Read the question and all answers before making the choice.
- **For each question, put an X mark on the letter: (a), (b), (c), or (d) which corresponds to the correct answer, by using a black or blue pen.**

---

1. Machine Learning is a subset of,

   (a) Deep learning.

   (b) Artificial intelligence.

   (c) Data science.

   (d) None of the above.

2. Correct way to perform the Standard Scaling on the train and test sets using Scikit-Learn library. Assume: Train set features: `X_train`, Test set features: **`X_test`**

   (a)

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.fit_transform(X_test)
```

   (b)

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

(c)

```
scaler = StandardScaler()
X_train_scaled = scaler.fit(X_train)
X_test_scaled = scaler.transform(X_test)
```

(d)

```
scaler = StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

3. What are the machine learning models that can be used for both regression and classification problems?

   (a) Decision trees.

   (b) Support vector machines.

   (c) K-Nearest Neighbors.

   [✓ (d) All of the above.]

4. Which of the following best describes overfitting in machine learning?

   [✓ (a) A model that performs well on training data but poorly on new, unseen data.]

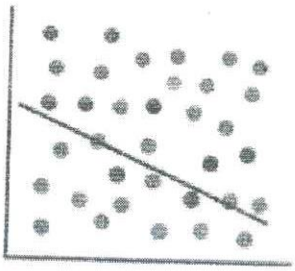
   (b) A model that consistently makes inaccurate predictions, even on training data.

   (c) A model that is too simple to capture the underlying patterns in the data.

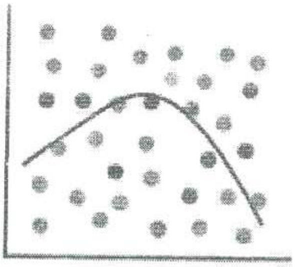
   (d) A model that is too computationally expensive to train efficiently.

5. Which figure below indicates an overfitting situation for a classification problem?

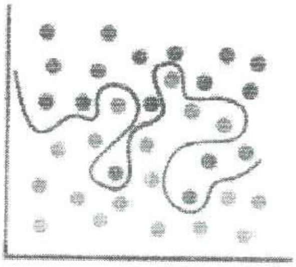
   (a)



(b)

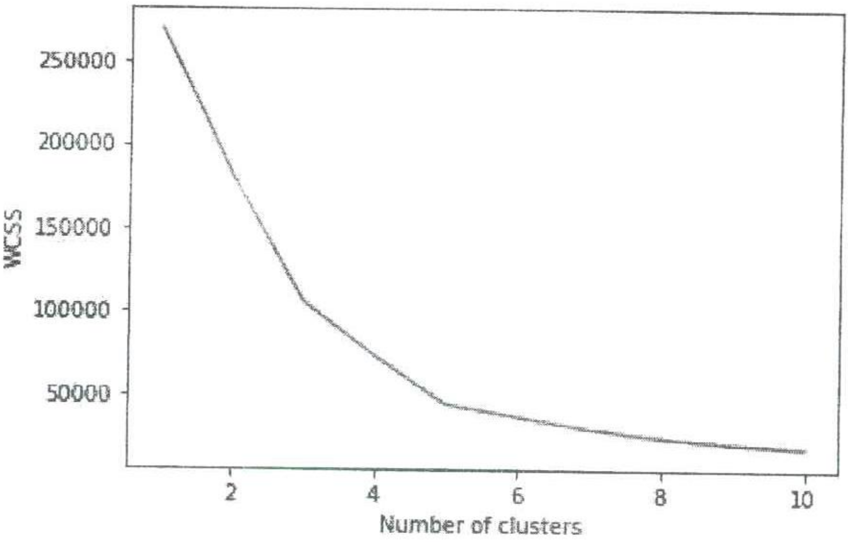


(c)



(d) None of the above.

6. Under the K-Means clustering problems, we can perform the Elbow method to identify the optimal number of clusters to perform the clustering. Based on the given plot, what should be the optimal number of clusters (Note: WCSS stands for Within-Cluster Sum of Square)?



(a) 3

(b) 4

[✓ (c) 5]

(d) 6

7. Which of the following best explains the term “Data Snooping” in machine learning?

[✓ (a) A form of statistical bias manipulating data or analysis to artificially get statistically significant results.]

(b) The use of specialized algorithms to uncover hidden patterns in large datasets with minimal human intervention.

(c) Removing irrelevant features from a dataset to improve model performance.

(d) Using a small sample size that doesn’t accurately represent the population.

8. The following table indicates two categorical features that have been extracted from a dataset. Select the most appropriate encoding scheme for each of the features.

Level	**Color**
Unique Values: Low, Medium, High, Other	Unique Values: Red, Green, Blue

(a) Label Encoding/Label Encoding

[✓ (b) Label Encoding/One Hot Encoding]

(c) One Hot Encoding/Label Encoding

(d) One Hot Encoding/One Hot Encoding

9. An instance of the support vector classifier has been indicated below. In the provided code snippet what does the ‘C’ hyper-parameter indicate?

```
from sklearn.svm import SVC
classifier = SVC(C = 1.0, verbose = False, probability = False)
```

(a) Tolerance for stopping criterion.

(b) Cache size of the kernel.

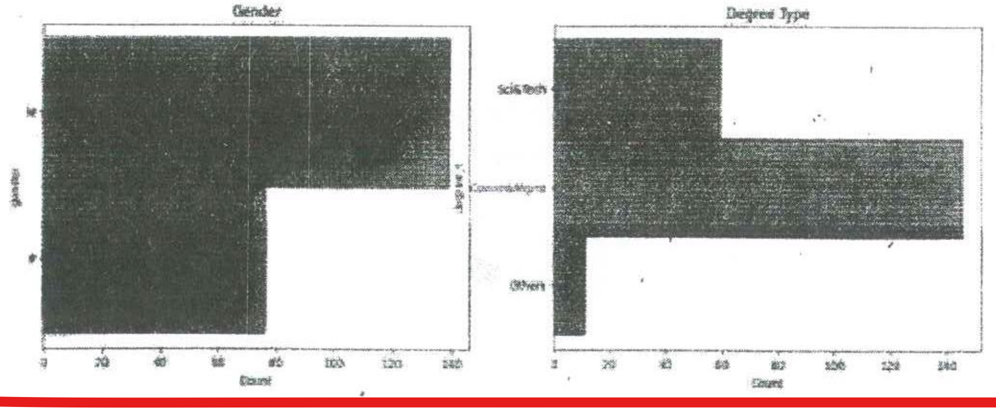
(c) Class weight.

[✓ (d) Regularization parameter.]

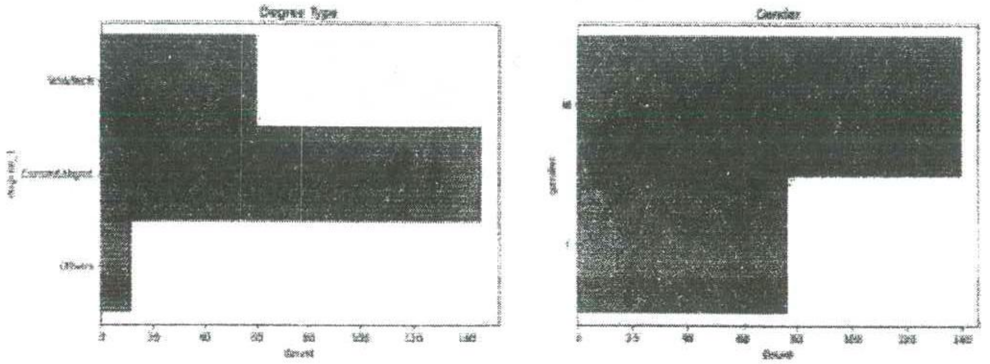
10. What is the expected plot from the following code snippet (consider df as a pandas data frame)?

```
import seaborn as sns
fig, axes = plt.subplots(1, 2, figsize=(15, 5))
sns.histplot(y=df.gender, ax = axes[0]).set(title = 'Gender')
sns.histplot(y=df.degree_t,ax = axes[1]).set(title ='Degree Type')
plt.show()
```

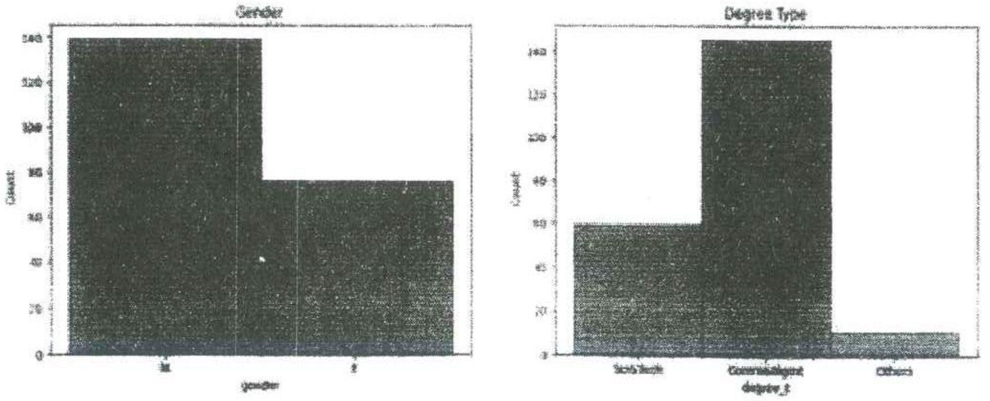
(a)


Gender
M
F
gender
0
20
40
60
80
100
120
140
Count
Degree Type
Sci&Tech
Comm&Mgmt
Others
degree_t
0
20
40
60
80
100
120
140
Count

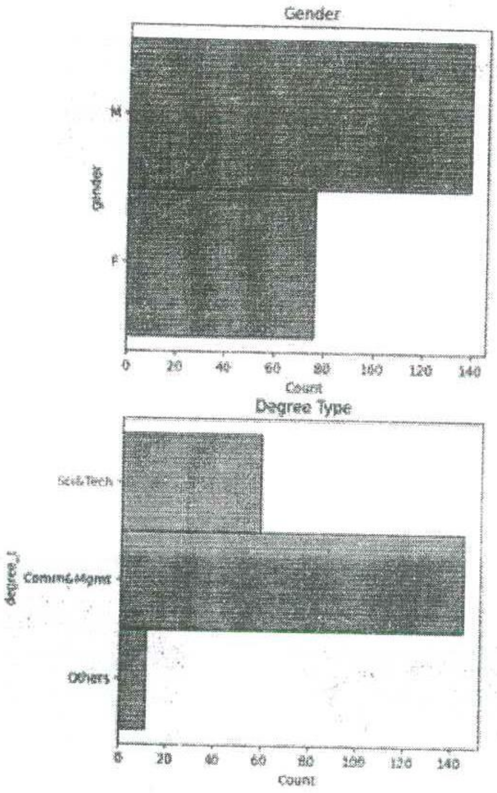
(b)


Degree Type
Sci&Tech
Comm&Mgmt
Others
degree_t
Count
Gender
M
F
gender
Count

(c)


Gender
0
20
40
60
80
100
120
140
Count
M
F
gender
Degree Type
0
20
40
60
80
100
120
140
Count
Sci&Tech
Comm&Mgmt
Others
degree_t

(d)



11. Which of the following is not a performance metric for a classification problem?

    (a) Accuracy

    (b) Precision

    (c) Recall

    [✓ (d) Mean Squared Error (MSE)]

12. For an imbalanced dataset, which of the following is the most suitable performance metric?

    (a) Accuracy

    [✓ (b) F1-score]

    (c) Precision

    (d) Recall

13. Which of the following best explains the "cost function" in the context of machine learning?

    (a) It defines the data structure used to store model parameters.

    (b) It calculates the accuracy of the model on the training data.

    [✓ (c) It quantifies the difference between predicted and actual outputs, guiding model improvement.]

(d) It determines the computational resources required to train the model.

14. What is the main usage of Principal Component Analysis (PCA) in the machine learning domain?

    (a) To create complex decision rules for classification tasks.

    (b) To automatically label the unlabeled data points.

    [✓ (c) To reduce the dimensionality of high-dimensional datasets while preserving variance.]

    (d) To tune hyperparameters of machine learning models.

15. What is the correct mathematical equation to perform "sklearn.preprocessing.StandardScaler" for a numerical data sample (x), where μ is the mean of the training samples and σ is the standard deviation?

    [✓ (a) $z = (x - \mu)/\sigma$.]

    (b) $z = x - \mu/\sigma$.

    (c) $z = (x^2 - \mu^2)/\sigma$.

    (d) None of the above.

16. Which of the following is not a clustering algorithm?

    (a) Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

    [✓ (b) Support Vector Machines (SVM)]

    (c) K-Means

    (d) Hierarchical Clustering

17. What is the primary purpose of cross-validation in machine learning model development?

    (a) To measure the accuracy of the model on the training data.

    (b) To compare the performance of different machine learning algorithms.

    (c) To identify the optimal hyperparameters for the model.

    [✓ (d) To estimate the generalizability of the model on unseen data.]

18. Which one of the following best explains the purpose of GridSearchCV?

    (a) It automates the process of training a machine learning model on multiple datasets.

    (b) It visualizes the relationship between different hyperparameters and model performance.

    [✓ (c) It systematically searches through a predefined grid of hyperparameter values and identifies the combination that optimizes a specified performance metric.]

    (d) It reduces the dimensionality of a dataset by selecting the most informative features.

19. What is the main advantage of using SMOTE (Synthetic Minority Over-sampling Technique) for imbalanced datasets compared to other data balancing approaches?

    (a) It significantly improves the accuracy of the model on both majority and minority classes.

    (b) It avoids overfitting by preventing the model from simply memorizing the minority class examples.

    (c) It is computationally efficient and requires minimal parameter tuning.

    [✓ (d) It generates new and plausible minority class examples, enriching the training data and potentially improving model performance for the minority class.]

20. Which of the following best explains the reason behind data balancing in machine learning models?

    [✓ (a) To address the issue of bias towards the majority class in imbalanced datasets, improving the performance for the minority class.]

    (b) To simplify the analysis by reducing the number of data points in the dataset.

    (c) To ensure all features within the dataset have equal variance.

    (d) To increase the overall size of the dataset and improve model accuracy.

UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 5 Examination in Engineering: January 2024

Module Number: EE5253 **Module Name: Machine Learning**

[1 Hour and 15 minutes]

[Answer **all questions**, each question carries **10** marks]

Attach Question Paper to the Answer Script

Q1 a) A machine learning problem is set up to predict the price of laptops. It utilizes 1-5 variables given below to predict the price (given by variable 6). A student proposes to use logistic regression to solve the above problem. Do you think it is a good choice? Justify your answer.

1) Inches – Numeric - Screen Size
2) ScreenResolution – String - Screen Resolution
3) Cpu - String -Central Processing Unit (CPU)
4) Ram – String - Laptop RAM
5) Memory – String - Hard Disk / SSD Memory
6) Price_euros – Numeric - Price (Euro)

[2.0 Marks]

b) (i) Pruning is often used in Decision Trees to avoid overfitting the training set. Briefly describe what pruning is.

(ii) Which regression model given below (Model I or Model II) is more appropriate to fit the training data better? Justify your answer.

Model I: y = ax + e
Model II: y = ax + bx^2 + e

[2.0 Marks]

c) Table Q1c shows whether students will pass or fail EE5253 based on whether or not they attended class, studied, and slept well before the exam. You are given the following data for five students. The column "Result" shows the label we want to predict.

Table Q1c

	Attended Class?	Studied?	Slept?	Result
Student 1	Yes	No	No	Passed
Student 2	Yes	No	Yes	Failed
Student 3	No	Yes	No	Failed
Student 4	Yes	Yes	Yes	Failed
Student 5	Yes	Yes	No	Passed

(i) What is the entropy H(Result) at the root node? Show your workings.

(ii) Draw the decision tree where every split maximizes the information gain. Show your workings.

[2.0 Marks]

d) Consider the data points in 2-D Euclidean space shown in Table Q1d.

Table Q1d

x	y	Class
-1	1	1
0	1	2
0	2	1
1	-1	1
1	0	2
1	2	2
2	2	1
2	3	2

(i) What is the prediction of the 3-nearest neighbour classifier at point (2,4)?
(ii) What is the prediction of the 5-nearest neighbour classifier at point (1,1)?
(iii) What is the prediction of the 7-nearest neighbour classifier at point (1,1)?
(iv) What is the prediction of the 1-nearest neighbour classifier at point (2,-1)?

[2.0 Marks]

e) You are required to train a Support Vector Machine (SVM) on a tiny dataset with 4 points shown in Figure Q1e. This dataset consists of two examples with class label –1 (–), and two examples with class label +1 (+).

(i) Find the weight vector w and bias b. What is the equation corresponding to the decision boundary?

(ii) Circle the support vectors and draw the decision boundary on Figure Q1e provided.

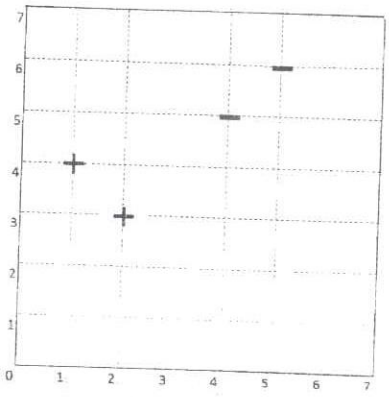


Figure Q1e

[2.0 Marks]

Q2 a) (i) List the three (3) main types of gradient descent based on the amount of data that is used.

(ii) Show the derivation for the update equations for $J(\theta_1, \theta_2) = {\theta_1}^2 + {\theta_2}^2$ using multivariate gradient descent.

[2.0 Marks]

b) Figure Q2b shows four plots with data showing low and high variance and low and high bias. Answer questions (i) to (iv) based on Figure Q2b by choosing from A, B, C or D as necessary.

(i) Which plot or plots have high variance?
(ii) Which plot or plots have high bias?

(iii) Which plot or plots have low variance?

(iv) Which plot or plots have low bias?

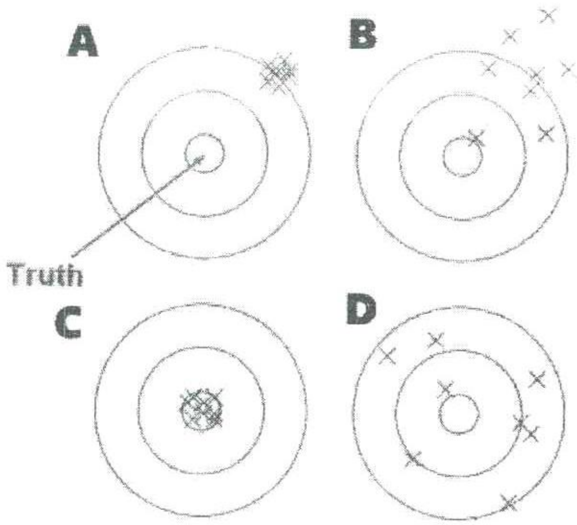


Figure Q2b

[2.0 Marks]

c) Explain stratified sampling. What is the purpose of stratified sampling in machine learning?

[2.0 Marks]

d) Briefly explain the importance of four (4) methods used for data pre-processing.

[2.0 Marks]

e) Answer the following questions regarding principal component analysis (PCA).

(i) Briefly explain giving a graphical example how outliers are removed using PCA.

(ii) Give two (2) instances when NOT to use PCA in a dataset?

[2.0 Marks]