



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?

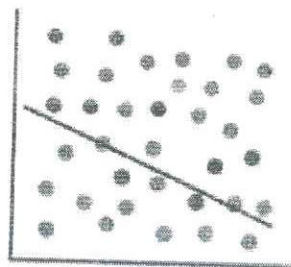
- (a) Decision trees.
- (b) Support vector machines.
- (c) K-Nearest Neighbors.
- (d) All of the above.

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

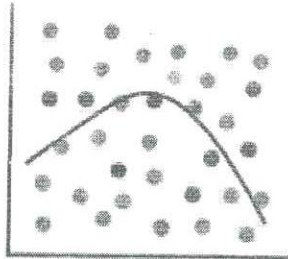
- (a) A model that performs well on training data but poorly on new, unseen data.
- (b) A model that consistently makes inaccurate predictions, even on training data.
- (c) A model that is too simple to capture the underlying patterns in the data.
- (d) A model that is too computationally expensive to train efficiently.

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

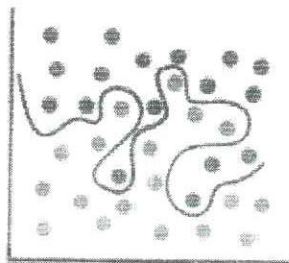
(a)



(b)

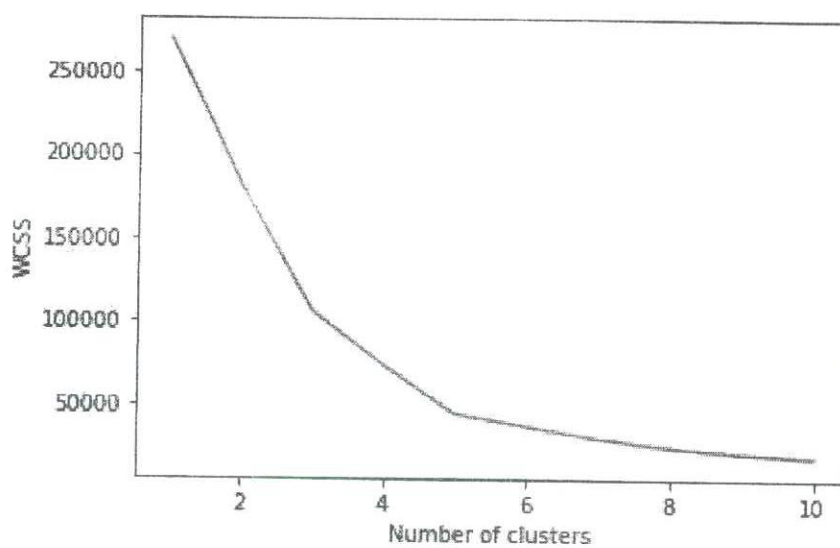


(c)



(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

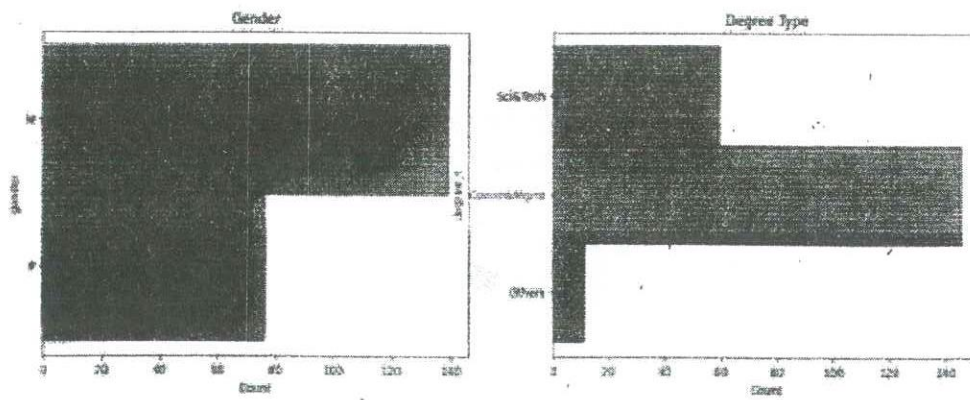
(d) 6

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.

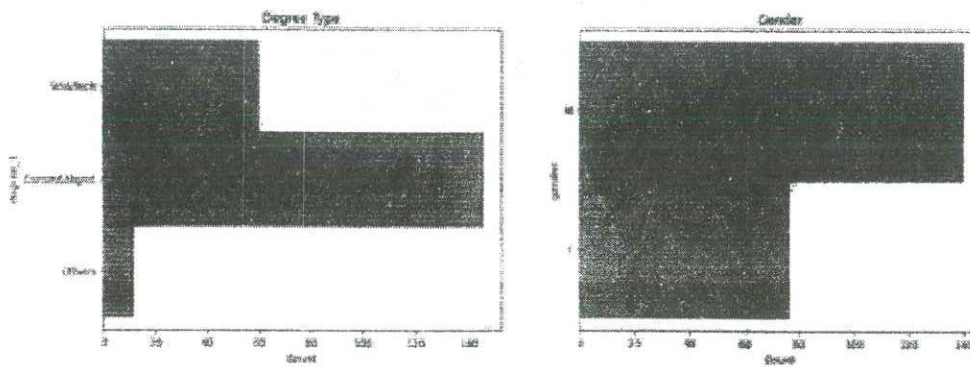
Level	Color
Unique Values: Low, Medium, High, Other	Unique Values: Red, Green, Blue

- (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()
```

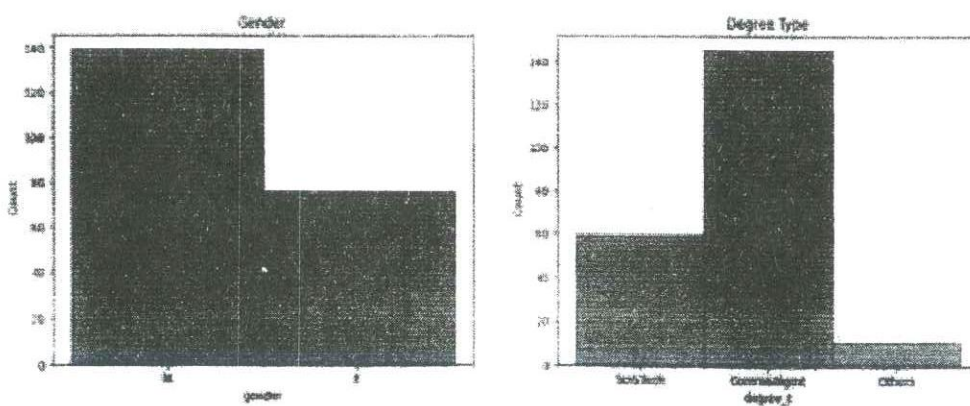

(a)



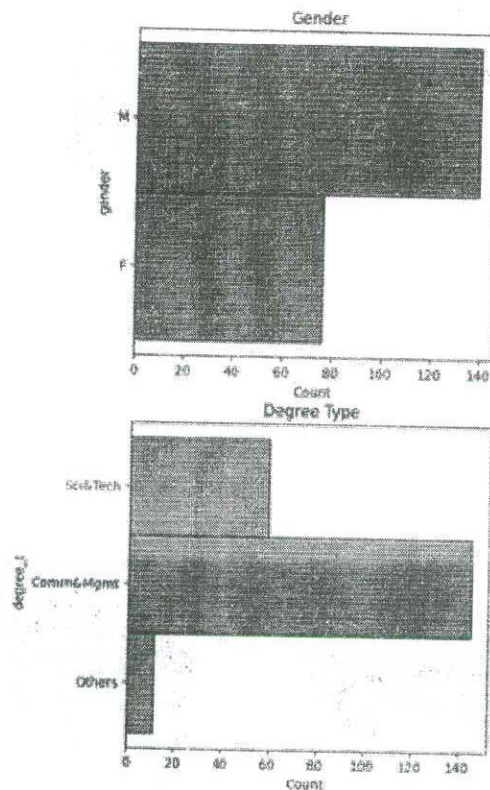
(b)



(c)



(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.