OR iii. Consider the following dataset containing the measurements of petals and sepals of five different flowers: **7**

Flower	Petal Length (cm)	Petal Width (cm)	Sepal Length (cm)	Sepal Width (cm)
1	1.4	0.2	5.1	3.5
2	1.3	0.2	4.9	3.0
3	1.5	0.2	4.7	3.2
4	4.7	1.4	6.2	2.9
5	4.9	1.5	6.7	3.1

Perform K-means clustering on this dataset with K=2. Show the steps of the algorithm and determine the final centroids of the clusters. Provide the cluster assignments for each data point.

Q.5 i. Explain the fundamental concept of a neural network and its role in modern artificial intelligence. Discuss its basic architecture, including its layers, nodes, and activation functions. **4**

ii. Explicate the fundamental concepts of feedforward neural networks and recurrent neural networks (RNNs). Compare and contrast their architectures, highlighting their respective strengths and applications. **6**

OR iii. Describe the following Python libraries used for machine learning, highlighting their key features and relationships: Keras, and TensorFlow. **6**

Q.6 Attempt any two:

i. What are the advantages and disadvantages of ensemble methods in machine learning, and how do techniques like bagging, contribute to improving model performance and generalization? **5**

ii. Provide a real-world example where semi-supervised learning could be effectively applied, and describe the potential challenges associated with implementing such a method. **5**

iii. Describe the key components of a reinforcement learning system and how they interact to achieve optimal decision-making. **5**

******

Total No. of Questions: 6 *Total No. of Printed Pages:4*

Enrollment No.....

Faculty of Engineering
End Sem Examination May-2024
CS3EL15 Machine Learning

Programme: B.Tech. Branch/Specialisation: CSE All

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d. Assume suitable data if necessary. Notations and symbols have their usual meaning.

Q.1 i. What is the primary objective of the cost function in linear regression? **1**
(a) Maximize the accuracy of predictions
(b) Minimize the number of misclassifications
(c) Minimize the difference between predicted and actual values
(d) Maximize the margin between support vectors

ii. In logistic regression, what is the function used to model the probability of a binary outcome? **1**
(a) Sigmoid function (b) Gaussian function
(c) Exponential function (d) Hyperbolic tangent function

iii. What does the bias of a machine learning model refer to? **1**
(a) The error due to fluctuations in the training data
(b) The error introduced by approximating a real-world problem
(c) The error due to overly simplistic assumptions in the learning algorithm
(d) The error due to noise in the training data

iv. Which machine learning algorithm is based on the principle of maximum a posteriori (MAP) estimation? **1**
(a) Naïve bayes (b) Support vector machines
(c) Kernel methods (d) Decision trees

v. Which algorithm is commonly used for clustering in unsupervised learning? **1**
(a) K-means (b) Decision trees
(c) Support vector machines (d) Linear regression

vi. Which of the following is a technique for reducing the dimensionality of data? **1**
(a) K-means clustering
(b) PCA (Principal Component Analysis)
(c) Hierarchical clustering
(d) DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

vii. In the context of neural networks, perceptron is- **1**
(a) A type of activation function
(b) A single layer neural network
(c) A machine learning algorithm for regression
(d) A deep learning

viii. Which type of neural network architecture is designed to handle sequential data effectively? **1**
(a) Convolutional Neural Network (CNN)
(b) Multilayer Perceptron (MLP)
(c) Recurrent Neural Network (RNN)
(d) Generative Adversarial Network (GAN)

ix. Which ensemble method involves training multiple models independently on different subsets of the training data and then combining their predictions? **1**
(a) Bagging (b) Boosting
(c) Random forests (d) Semi-supervised learning

x. Which technique combines labeled and unlabeled data to improve the learning accuracy of a model? **1**
(a) Supervised learning (b) Unsupervised learning
(c) Semi-supervised learning (d) Deep learning

Q.2 i. What are the fundamental differences between supervised and unsupervised learning algorithms? Explain with examples. **2**

ii. Explicate the concept of the k-nearest neighbors (KNN) algorithm and how it is used for classification tasks. Discuss the influence of the value of k on the performance of the KNN algorithm. **3**

iii. How is machine learning utilized in the field of healthcare? Provide specific examples of applications and discuss the potential impact on patient care and medical research. **5**

OR iv. Discuss the working principle of logistic regression as a classification algorithm. Explain how the logistic regression model is trained using the cost function and optimized using gradient descent. Provide a step-by-step explanation with mathematical formulas. **5**

Q.3 i. Discuss the techniques used to address overfitting and underfitting in machine learning models. Provide examples of when each technique would be appropriate. **2**

ii. Given customers' purchasing behaviour at a supermarket. The dataset consists of the following attributes: Age (continuous), Income (continuous), Gender (categorical: Male/Female), Marital Status (categorical: Single/Married), and Purchase Decision (categorical: Yes/No). **8**

Age	Income	Gender	Marital Status	Purchase Decision
35	45000	Male	Single	Yes
25	32000	Female	Married	No
45	60000	Male	Married	Yes
30	70000	Female	Single	Yes
50	80000	Male	Married	No

(a) Describe the steps involved in building a decision tree model for classification.
(b) Perform the necessary pre-processing steps for the given dataset, including handling categorical variables and splitting the dataset into training and testing sets.
(c) Construct a decision tree model using the provided dataset. Show the decision tree structure and explain how it makes predictions.

OR iii. Explain the concept of Support Vector Machines (SVMs) in machine learning. Provide a detailed explanation of the optimization objective in SVMs and how it relates to finding the optimal hyperplane for separating classes. Support your answer with relevant diagrams and mathematical formulations where necessary. **8**

Q.4 i. How does unsupervised learning contribute to the field of artificial intelligence and machine learning? Provide examples of real-world applications where unsupervised learning techniques have been successfully employed. **3**

ii. How does PCA help in simplifying complex datasets while preserving the most important information? Illustrate with a practical example how PCA can be applied to real-world datasets and discuss the potential benefits and limitations of using PCA in machine learning tasks. **7**

Marking Scheme

CS3EL15 (T) Machine Learning

Q.1	i)	C		**1**
	ii)	A		**1**
	iii)	C		**1**
	iv)	A		**1**
	v)	A		**1**
	vi)	B		**1**
	vii)	B		**1**
	viii)	C		**1**
	ix)	A		**1**
	x)	C		**1**

Q.2	i.	Understanding of Key Concepts (1 mark): Award 0.5 mark for correctly identifying the fundamental differences between supervised and unsupervised learning. Award another 0.5 mark for demonstrating an understanding of these concepts. Examples (1 mark): Award 0.5 mark for providing relevant examples for both supervised and unsupervised learning. Award another 0.5 mark if the examples effectively illustrate the application of each type of learning.	**2**
	ii.	Explanation of the k-nearest neighbors (KNN) algorithm (1 mark) Usage of KNN for classification tasks (1 mark) Discussion of the influence of the value of k on the performance of the KNN algorithm (1 mark)	**3**
	iii.	Understanding of Machine Learning in Healthcare (1 mark): Proper explanation of what machine learning is. Clear understanding of how machine learning is applied in healthcare. Examples of Applications (2 marks): Providing specific and relevant examples of machine learning applications in healthcare. Each example should be described in sufficient detail to demonstrate understanding. Discussion of Potential Impact (2 marks):	**5**
OR	iv.	Introduction to Logistic Regression (1 mark) Working Principle of Logistic Regression (1 mark) Training Process (2 marks) Optimization with Gradient Descent (1 mark)	**5**
Q.3	i.	Understanding of Techniques (1 mark): Award 0.5 marks for listing at least two techniques for addressing overfitting and underfitting. Award another 0.5 marks for providing a brief explanation or description of each technique. Examples (1 mark):	**2**
	ii.	1 mark for mentioning key steps such as selecting the best attribute for splitting, calculating impurity measures like Gini index or entropy, and recursively building the tree. 1 mark for explaining the importance of each step and how it contributes to the overall construction of the decision tree. (b) Perform the necessary pre-processing steps for the given dataset, including handling categorical variables and splitting the dataset into training and testing sets. (3 marks) (c) Construct a decision tree model using the provided dataset. Show the decision tree structure and explain how it makes predictions. (3 marks) 1 mark for correctly constructing the decision tree model using the provided dataset.	**8**

		1 mark for showing the decision tree structure visually or in a textual representation. 1 mark for explaining how the decision tree makes predictions, including how it traverses the tree based on attribute values and reaches a decision.	
OR	iii.	Understanding of Support Vector Machines (SVMs) **(2 marks)** Explanation of Optimization Objective **(3 marks)** Detailed explanation of the optimization objective, which involves finding the hyperplane that maximizes the margin between classes while minimizing classification errors. Discussion of the mathematical formulation of the optimization objective, such as the formulation of the margin and the use of support vectors. Relation to Finding Optimal Hyperplane **(2 marks)** Clear explanation of how the optimization objective relates to finding the optimal hyperplane for separating classes. Discussion of how the margin maximization helps in achieving better generalization and robustness in classification. Inclusion of Relevant Diagrams and Formulations **(1 mark)**	**8**

Q.4	i.	Explanation of Unsupervised Learning (1 mark) Contribution to AI/ML (1 mark) Real-World Examples (1 mark)	**3**
	ii.	Explanation of K-means Algorithm Steps (2 marks): Implementation of K-means Algorithm (2 marks): Determination of Final Centroids (1 mark): Cluster Assignments (1 mark): Quality of Presentation (1 mark): Dedicate 1 mark for the overall quality of presentation, including clarity, organization, and conciseness of the solution.	**7**
OR	iii.	Understanding of PCA **(1 marks)** Description of PCA and its purpose in simplifying complex datasets while retaining important information. Explanation of how PCA reduces dimensionality by transforming the original features into a new set of orthogonal components (principal components). Illustration with Practical Example **(3 marks)** Demonstrating the application of PCA to a real-world dataset. Clear explanation of how PCA is applied, including preprocessing steps such as mean centering and scaling. Interpretation of results, showing how PCA helps in visualization or analysis of the dataset. Discussion of Benefits **(2 marks)** Identification and explanation of potential benefits of using PCA in machine learning tasks. Examples could include improved model performance, faster training times, and enhanced interpretability of results. Discussion of Limitations **(1 mark)** Identification and explanation of potential limitations or drawbacks of PCA. Examples could include loss of interpretability of individual features, potential loss of information, and sensitivity to outliers.	**7**

Q.5	i.	Explanation of fundamental concept of a neural network (1 mark) Discussion of neural network's role in modern AI (1 mark) Description of basic architecture (1 mark) Discussion of layers, nodes, and activation functions (1 mark)	**4**
	ii.	Feed forward Neural Network with diagram (4 marks) Recurred Neural Network (2 marks)	**6**
OR	iii.	Award 3 mark for each library (Keras and TensorFlow) if the student provides accurate explanations of their key features and relationships. Depth of Explanation (2 marks) Award up to 2 marks based on the depth of understanding demonstrated by the student. Clarity and Structure (1 mark): Award 1 mark for clarity of expression and organization of ideas. Ensure the response is well-structured, with clear delineation of features and relationships of Keras and TensorFlow.	**6**

Deduct marks for unclear or poorly organized responses.

Q.6

i.	Understanding of Ensemble Methods: 2 marks Understanding of Bagging: 2 marks Clarity and Coherence: 1 mark	**5**
ii.	Clarity of Example (2 marks): Description of Application (2 marks Analysis of Challenges (1 mark)	**5**
iii.	Understanding of key components (2 marks) Explanation of interaction (2 marks) Insight into achieving optimal decision-making (1 mark) Demonstrate understanding: Show how these interactions lead to the agent learning an optimal policy over time.	**5**

******