**ASSIGNMENT 3**

**DS2306**

**TOLULOPE OLAJOBI**

QUESTION 1.

d) Collinearity. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can help reduce collinearity in the dataset by transforming the original variables into a new set of uncorrelated variables.

QUESTION 2.

b) Random Forest. Random Forest is an ensemble learning algorithm based on the idea of bagging. It combines multiple decision trees to make predictions by averaging the results of individual trees.

QUESTION 3.

c) Decision Trees are prone to overfit. Decision trees have a tendency to overfit the training data, meaning they can create overly complex trees that perform well on the training data but generalize poorly to new data.

QUESTION 4.

c) Training data. The machine learning algorithms build a model based on training data, which is a labeled dataset used to train the model and learn the relationships between the input variables and the target variable.

QUESTION 5.

c) Anomaly detection. Anomaly detection is a machine learning technique that helps in detecting outliers in data by identifying patterns that deviate significantly from the norm or expected behavior.

QUESTION 6.

c) Case based. "Case based" is not a numerical function in machine learning. The other options (a) Support Vector, (b) Regression, and (d) Classification, are valid representations of different types of machine learning algorithms.

QUESTION 7.

d) Both a and b. The analysis of machine learning algorithms requires both statistical learning theory and computational learning theory. Statistical learning theory provides the theoretical foundation for understanding the properties and performance of algorithms, while computational learning theory focuses on the computational aspects of learning algorithms.

QUESTION 8.

c) Both a and b. The k-nearest neighbor algorithm faces the "curse of dimensionality," where the algorithm's performance degrades as the number of dimensions (features) increases. Additionally, it needs to calculate the distance of the test case for all training cases, which can be computationally expensive for large datasets.

QUESTION 9.

b) 2. Radial basis function neural networks typically have two types of layers: the input layer and the hidden layer. The hidden layer contains the radial basis functions that process the input data.

QUESTION 10.

a) PCA (Principal Component Analysis). PCA is an unsupervised dimensionality reduction technique and not a supervised learning algorithm. The other options (b) Naïve Bayes, (c) Linear regression, and (d) KMeans are supervised learning algorithms.

QUESTION 11.

c) Neither feature nor number of groups is known. Unsupervised learning refers to machine learning tasks where there is no explicit information about the groups or categories in the data. The goal is to discover patterns, structures, or relationships in the data without prior knowledge of the outcomes or labels.

QUESTION 12.

b) SVG. SVG is not a machine learning algorithm. The other options (a) SVM, (c) Random Forest Algorithm, are valid machine learning algorithms used for classification and regression tasks.

QUESTION 13.

b) Underfitting. Underfitting occurs when a model fails to capture the underlying trend or patterns in the input data. It usually happens when the model is too simple or lacks complexity to represent the relationships adequately.

QUESTION 14.

a) Reinforcement learning. Real-Time decisions, Game AI, Learning Tasks, Skill acquisition, and Robot Navigation are applications of reinforcement learning, where an agent learns to make decisions or take actions based on feedback from the environment.

QUESTION 15.

b) Mean squared error. The average squared difference between the predicted output of a classifier and the actual output is known as the mean squared error (MSE). It is a common metric used to evaluate the performance of regression models.

QUESTION 16.

a) Linear, binary. Logistic regression is a linear regression technique used to model data with a binary outcome (two classes). It estimates the probability of the binary outcome based on linear combinations of the input features.

QUESTION 17.

a) supervised learning. Classifying reviews of a new Netflix series using labeled reviews as training data falls under supervised learning, where the model learns from labeled examples to make predictions on new, unseen data.

QUESTION 18.

C) both a and b. Euclidean distance and Manhattan distance are both powerful distance metrics used by geometric models. Euclidean distance measures the straight-line distance between two points, while Manhattan distance measures the sum of the absolute differences between corresponding coordinates.

QUESTION 19.

d) none of these. The given options do not represent techniques for reducing dimensions of a dataset. Common techniques for dimensionality reduction include PCA, LDA, and feature selection methods such as variance thresholding or recursive feature elimination.

QUESTION 20.

c) input attribute. Both supervised learning and unsupervised clustering require input attributes (features) to train the model or group the data. The output attribute (target variable) is only required for supervised learning.

QUESTION 21.

a) SVM allows very low error in classification. Hard margin in SVM refers to allowing very low error in classification, aiming to find a decision boundary that separates the classes with maximum margin.

QUESTION 22.

b) Only 2. Increasing the depth of the trees in a Random Forest can lead to overfitting, as the trees become more complex and capture noise in the training data. Increasing the number of trees generally improves the model's performance.

QUESTION 23.

a) -(6/10 log(6/10) + 4/10 log(4/10)). The entropy of the target variable can be calculated using the formula -(p \* log₂(p) + q \* log₂(q)), where p and q represent the proportions of the two classes (0 and 1) in the target variable. In this case, p = 6/10 and q = 4/10.

QUESTION 24.

a) weights are regularized with the l1 norm. Lasso regression is a linear regression technique where the weights are regularized with the l1 norm, promoting sparsity by shrinking some coefficients to exactly zero. This allows for feature selection and can simplify the model.

QUESTION 25.

a) Perceptron and logistic regression. In Perceptron and logistic regression, the decision boundary is sensitive to individual data points, including those far away from the decision boundary. Adding a new labeled data point far from the decision boundary can change the learned decision boundary.

QUESTION 26.

d) Either 2 or 3. Instead of deleting both collinear variables (option 1), it is common to remove one of the variables (option 2) to mitigate multicollinearity. However, removing correlated variables may result in information loss, so penalized regression models like ridge or lasso regression (option 3) can be used to retain important information.

QUESTION 27.

b) Increase by 5 pounds. According to the given least squares line equation y = 120 + 5x, the weight (y) should increase by 5 pounds when the height (x) is increased by one inch.

QUESTION 28.

d) Minimize the squared distance from the points. The line described by the linear regression equation (OLS) minimizes the squared distance (residuals) between the predicted values and the actual values, aiming to find the best-fitting line.

QUESTION 29.

b) As the value of one attribute increases, the value of the second attribute also increases. A correlation coefficient of 0.85 indicates a strong positive linear relationship between the two attributes. As the value of one attribute increases, the value of the second attribute tends to increase as well.

QUESTION 30.

b) Convolutional Neural Network (CNN). Convolutional Neural Networks are specifically designed for image-related tasks, such as image identification or recognition. They excel at extracting hierarchical features from images and are widely used in computer vision tasks.