MACHINE LEARNING WORKSHEET-4

In Q1 to Q7, only one option is correct, Choose the correct option:

Ques1. The value of correlation coefficient will always be:
C) between -1 and 1

Ques2. Which of the following cannot be used for dimensionality reduction?
C) Recursive feature elimination

Ques3. Which of the following is not a kernel in Support Vector Machines?
A) linear

Ques4. Amongst the following, which one is least suitable for a dataset having non-linear decision
boundaries?
A) Logistic Regression

Ques5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be? (1 kilogram = 2.205 pounds)
B) same as old coefficient of 'X'

Ques6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
B) increases

Ques7. Which of the following is not an advantage of using random forest instead of decision trees?
C)Random Forests are easy to interpret

In Q8 to Q10, more than one options are correct, Choose all the correct options:

Ques8. Which of the following are correct about Principal Components?
D) All of the above

Ques9. Which of the following are applications of clustering?
A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Ques10. Which of the following is(are) hyper parameters of a decision tree?
A) max_depth
C) n_estimators
D) min_samples_leaf

Q10 to Q15 are subjective answer type questions, Answer them briefly.

Ques11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.
Ans: Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations. Outliers impact the prediction in a dataset and often these needs to be treated before predicting a model. One of the ways to deal Outliers in model building in machine learning is by importing zscore.

Interquartile range or IQR is used to measure variability by dividing a data set into quartiles.The data is sorted in ascending order and split into 3 equal parts- Q1, Q2, Q3.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

Ques12. What is the primary difference between bagging and boosting algorithms?
Ans: Bagging and Boosting are two types of Ensemble Learning. These two decrease the variance of a single estimate as they combine several estimates from different models.
Bagging attempts to tackle the over-fitting issue. Boosting tries to reduce bias. Each model in bagging is trained parallelly and indeendently where in a final prediction is created from the prediction of every models.Boosting is an interative process which trains all the models together and gets a certain prediction, a second model is then built ries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

Ques13. What is adjusted R2 in linear regression. How is it calculated?
Ans: The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

Ques14. What is the difference between standardisation and normalisation?
Ans: Normalization is a part of data processing and cleansing techniques. The main goal of normalization is to make the data homogenous over all records and fields. It helps in creating a linkage between the entry data which in turn helps in cleaning and improving data quality.

Normalization of data is a type of Feature scaling and is only required when the data distribution is unknown or the data doesn't have Gaussian Distribution.
Standardization is the process of placing dissimilar features on the same scale. Standardized data in other words can be defined as rescaling the attributes in such a way that their mean is 0 and standard deviation becomes 1. It is usually applied when the data has a bell curve i.e. it has gaussian distribution.

Ques15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
Ans: Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models.

Advantage: It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

Disadvantage: Cross Validation is computationally very expensive in terms of processing power required.