MACHINE LEARNING WORKSHEET -1

ANS.NO.1 (C)

ANS.NO.2 (A)

ANS.NO.3 (A)

ANS.NO.4 (C)

ANS.NO.5 (D)

ANS.NO.6 (B)

ANS.NO.7 (C)

ANS.NO.8 (D)

ANS.NO.9 (C)

ANS.NO.10 (B|

SUBJECTIVE ANSWER

Ans.no.11

1) One hot encoding is a process whereby variables are converted into a form that can be provided as an input to machine learning models.
2) The goal of one hot encoding is to transform data form a representation to a numeric representation.
3) One hot encoding can be performed using the Pandas library in Python. The Pandas library provided a function is called “get dummies”
4) One hot encoding creates D-dimensional vector for each instance where d is the unique number of features value in the dataset.
5) In the one hot encoding refers to variables that are made up of label values.
6) Variable could have the values “red” “blue” and “green”
7) This is where the integer encoded variable is removed and one new binary variable is added for each unique integer value in the variable.

8) This technique is used for categorical variable where order does not matter.
9) One hot encoding technique is used when features are nominal.
10) In one hot encoding for every categorical feature a new variable.

ANS.12

1) A classification dataset with skewed class proportion is called imbalance.
2) In the case of imbalanced datasets this is particular problem if the minority class has multiple concepts or clusters in the feature space.
3) Data classification is the most popular task of data mining.
4) Its problem is to correctly classify an instance with indeterminate class.
5) The used datasets are organized in the form of tables.
6) The tables' columns are called the attributes.
7) They represent the characteristics of the dataset.
8) Rule based classification algorithm have a bias toward majority classes.
9) The presents the classification problem in imbalanced dataset.
10) We present the evaluation metrics used in classification problem in imbalanced datasets.
11) In the binary imbalanced dataset, the number of instances of one higher than that of the second class.
12) The distribution of instances in imbalanced binary datasets is measured by the imbalanced.(IR)

IR=Number of majority instances

Number of minority instances

TECHNIQUE CAN BE USED TO BALANCE THE DATASET.

1) Oversampling
2) Under sampling
3) Class weight
4) Decision threshold.

1) Oversampling: = increase the number of samples of the smallest class up to the size of the biggest class.
2) Under Sampling: = decrease the number of samples of the biggest class down to the size of the smallest class.
3) Class weight: = Assign the weight to each class.
   Biggest class weight =1
   Smallest class weight= samples biggest class.

4) Decision Threshold: = if a predicted value is greater than threshold. It is set 1 otherwise it is set to 0.

QUESTION NO.13

ANS NO 13.

The difference between SMOTE AND ADASYN.

SMOTE: =first it find the n-nearest neighbour in the minority class for each of the samples in the class.

Majority class samples

Minority class sample.

Synthetic samples.

Smote stands for Synthetic Minority Over-Sampling Technique. The method was proposed in a 2002 in the Journal of artificial. Smote is an improved method of dealing with imbalanced data in classification problem.

SMOTE works by selecting examples that are close in the feature space drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Smote is a statistical technique for increasing the number of cases in your dataset in a balanced way.

SMOTE is applied prior to feeding data to these machine learning models so that the imbalance problem of the given dataset can be resolved.

ADASYN SAMPLING.

1) ADASYN is stands for ADAPTIVE OVER-SAMPLING TECHNIQUE FOR SKEWED DATASET.
2) ADASYN is to use a weighted distribution for different minority class.
3) This paper presents a novel adaptive synthetic sampling approaches for learning for imbalanced data sets.
4) Reducing the bias introduced by the class imbalance.
5) Adaptive nature of creating more data for harder-to-learn.
6) The oversampling technique implemented in I learn is adaptive synthetic sampling.
7) ADASYN is similar to smote.
8) ADASYN is a more generic framework.
9) Then higher the ratio more synthetic points are generated for that particular points. The number of synthetic observation to be created for Obs 3 is going to be double that of Obs2.
10) The ratio of majority observations in the neighbourhood.

QUESTION.NO-14

ANSWER NO.14

GridSearchCV is the process of performing hyper parameter tuning in order to determine the optimal values for a given models.

The performance of the models significantly depends on the value of hyper parameters.

We need to try all possible values to know the optimal values.

GridSearchCV is a function that comes in Scikit-learn model_selection packages.

This function helps to loop through predefined hyper parameter and fit your estimate on your training set.

Having larger datasets can be beneficial for several reasons.

1) Increase statistical power: with larger datasets its is possible to detect subtle patterns and relationships in the data that may not be apparent with smaller datasets.
2) More representative: larger datasets are more likely to be representative of the population or phenomenon being studied. This can lead to more generalizable and reliable conclusion.

QUESTION NO.15

ANSWER NO.15

REGRESSION: - REGRESSION is a type of machine learning. Which helps in finding the relationship between independent and dependent variables.

EVALUATION METRICS: - Its is necessary to obtained the accuracy on training data, but it is also important to get

genuine and approximate results on unseen data otherwise models is of no use.

If one matric is perfect there is no need to multiple metrics.

To understand the benefits and the disadvantages of evaluation metrics because different evaluation metrics fits on a different set of a dataset.

Let’s start understanding various evaluation metrics used for regression tasks.

DATASET:-

For demonstrating each evaluation metrics using the sci-kit – learn library we will use the placement dataset.

Let’s start exploring various evaluation metrics.

1) MEAN ABSOLUTE ERROR (MAE) : -

MEAN ABSOLUTE ERROR is a very simple metrics which calculates the absolute difference between actual and predicted values.

You have input data and output data and use Linear Regression which draws a best fit line.

All the errors and divided them by a total number of observation and this is MAE and we aim to gets a minimum MAE because this is loss.

- The MAE you get is in the same unit as the output variables.
- It is most robust outlier.

2) MEAN SQUARED ERROR (MSE) : -

- MSE is most used and very simple metrics with a little bit of change in mean absolute error.
- Mean squared error states that finding the squared difference between actual and predicted value.
- MSE is represents the squared distance between actual and predicted values.

3) ROOT MEAN SQUARED ERROR (RMSE)

- AS RMSE is clear by the name itself that is a simple square error.
- It is not robust to outliers as compared MAE.
- Most of the time people use RMSE as an evaluation metrics and mostly when you are working with deep learning technique the most preferred metrics is RMSE.

4) ROOT MEAN SQUARED LOG ERROR :-

- To control this situation of RMSE we take the log calculated RMSE error and resultant we get as RMSLE.
- To perform RMSLE we have to use the Numpy log function over RMSE.
- It is very simple metrics that is used by most of the data sets hosted for machine learning.

5) R SQUARED:=

1) R squared is a metrics that tells the performance of your model.
2) MAE and MSE depends on the context as we have seen whereas the R2 score is independent of context
3) R2 is a coefficient of determine or sometimes knows as goodness of fit.

$$R2 \text{ squared} = 1 - \frac{SSr}{}$$