Machine Learning Assignment – 8

Ans. 1) b

Ans. 2) a

Ans. 3) d

Ans. 4) c

Ans. 5) d

Ans. 6) b

Ans. 7) c

Ans. 8) a, d

Ans. 9) b, c

Ans. 10) a

Ans. 11) A.One-Hot Encoding is another popular technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. categorical variables where no such ordinal relationship exists, the integer encoding is not enough. In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value. In the “color” variable example, there are 3 categories and therefore 3 binary variables are needed. A “1” value is placed in the binary variable for the color and “0” values for the other colors

Ans. 12) Undersampling balances the dataset by reducing the size of the abundant class.This method is used when quantity of data is

sufficient.Over sampling is used when quantity of data is insufficient.It balances by increasing the size of rare samples.

Ans. 13) The key difference is that ADASYN uses density distribution,as a criterion to automatically decide the number of synthetic samplesthat must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed.

Ans. 14) Gridsearch CV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using cross validation method.After using this function we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

Ans. 15) Mean squared error-It measures the average of the squares of the errors.It is the average squared difference between estimated values and actual value.

Root mean squared error-It is the standard deviation of the residuals.Residuals are measure of how far from the regression line datapoints are.It tells you how concentrated the data is around the line of best fit.

Mean absolute error-It is a measure of errors between paired observations expressing the same phenomenon.