MACHINE LEARNING

1. B
2. B
3. C
4. C
5. D
6. B
7. C
8. A,D
9. B,D
10. A B
11.

One hot encoding is used when we have to transform categorical data to numerical data for the model to understand. We must avoid one hot encoding when the categorical data is of ordinal type. This means OHE should not be used when the values of the categorical feature have a linear relation. For eg, Outstanding →Good → Bad → Worse, in this case we can use OHE as it will provide equal weightage to all the feature and in short Outstanding and Worse would be same. This will cause a bias in the model. Thus we should avoid OHE. To counter this, we can use the ordinal encoding techniques/label encoding techniques which will preserve the relation between the values.

12.

Data Imbalance is a problem when we have a classification problem and we have very few records of one class in relative to other. The ratio between the records of classes for about 10:90 or 20:80. This would drastically favors over model towards majority class being predicting most of the time making our model bias. There are multiple techniques we could use to solve this problem.

Random Under-Sampling
Random Undersampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out.

Random Over-Sampling
Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

Synthetic Minority Over-sampling Technique
This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an

example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset.

Cluster-Based Over Sampling
K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.

Boosting-Based techniques
Boosting is an ensemble technique to combine weak learners to create a strong learner that can make accurate predictions. Boosting starts out with a base classifier / weak classifier that is prepared on the training data.

13.

SMOTE: This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset.

ADASYN : ADAptive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data. The algorithm uses Euclidean distance for KNN Algorithm.

ADASYN a improved version of Smote. What it does is same as SMOTE just with a minor improvement. After creating those sample it adds a random small values to the points thus making it more realistic. In other words instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e they are bit scattered.

14.

GridSearchCV is a function used for hyper parameter tuning of the machine learning models . GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function we get accuracy/loss for every combination of hyper parameters and we can choose the one with the best performance. It is not advices to use grid search cv on a large dataset with high number if dictionary values as training combination will crash up the systems . Iven if you chose to tune 3 parameters for which you gave 5 possible values this on a system with 5cross folds that's 5*5*5 or 625 model training will could potentially crash you system or take way long long time to train. In such cases we could try using RanddomizedSearchCv model.

15.

Some of the metrics that could be used for evaluation of regression models ae as follows:

Mean Squared Error:
The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

MSE = 1 / N * sum for i to N (y_i – yhat_i)^2
Where y_i is the i'th expected value in the dataset and yha

Root Mean Squared Error
The Root Mean Squared Error, or RMSE, is an extension of the mean squared error. it may be common to use MSE loss to train a regression predictive model, and to use RMSE to evaluate and report its performance.

The RMSE can be calculated as follows:
RMSE = sqrt(1 / N * sum for i to N (y_i – yhat_i)^2)

Mean Absolute Error MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error. MAE score is calculated as the average of the absolute error values.

The MAE can be calculated as follows:
MAE = 1 / N * sum for i to N abs(y_i – yhat_i)

R^2 Error: Coefficient of Determination or R^2 is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. R^2 is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R^2 will always be less than or equal to 1.

The R^2 is calculated using as follows:
$R^2 = 1 – (MSE(model) / MSE(baseline))$

Adjusted R^2: Adjusted R^2 depicts the same meaning as R^2 but is an improvement of it. R^2 suffers from the problem that the scores improve on increasing terms even though the model is not improving which may misguide the researcher. Adjusted R^2 is always lower than R^2 as it adjusts for the increasing predictors and only shows improvement if there is a real improvement.

The R^2 is calculated using as follows:
$R^2 = 1 – ([n-1/n-k-1])* (1 - R^2)$