**Answer 1 –** B)

**Answer 2 –** A) depth of tree, overfitting can occur when the depth of tree is increased.

**Answer 3 –** C) as using random under sampler we will lose data.

**Answer 4 –** C)

**Answer 5 –** D)

**Answer 6 –** B) time complexity of SVM is order of n cube.

**Answer 7 –** C)

**Answer 8 –** A) and D)

**Answer 9 –** B)

**Answer 10 –** A) high variance means over fitting

**Answer 11 –** When the data is ordinal and number of categories is high then one hot encoding is a not good approach, that time we can use label encoder. It is suitable for ordinal data and data with many categories.

**Answer 12 –** In case of highly imbalanced data we can use the following techniques:  
 1. Under Sampling – This technique balances the data set by reducing the size of abundant  
 class. This technique should be used only when we have enough data. As this technique  
 causes reduction of data.  
  
 2. Over Sampling - Oversampling is used when the quantity of data is   
 insufficient. It tries to balance the dataset by increasing the size of  
 rare samples.  
  
 3. K-fold cross validation - The k-fold cross-validation procedure involves splitting the training dataset into k folds. The first k-1 folds are used to train a model, and the holdout kth fold is used as the test set. This process is repeated and each of the folds is given an opportunity to be used as the holdout test set.  
  
 4. Ensemble technique – Ensemble techniques uses resampling of data.   
  
 5. Class weight – when training any model, we can set the class weight. By this we can assign the class weights inversely proportion to their respective frequencies.

Answer 13 –

**Answer 14 –** The grid search cv is used for finding the best hyper parameter for our model. It tries all the values for the parameters that we give to our model and then find the best parameter. It is not good for heavy dataset, as gridsearch cv takes too much time if the dataset is high and also if the parameters are more.

**Answer 15 –** Evaluation metrics used for regression models are :  
1 – Mean squared error - The **mean squared error** (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the [mean](https://www.statisticshowto.com/mean/)squared error as you’re finding the average of a set of errors. The lower the MSE, the better the forecast.  
  
2 – Mean absolute error - **Absolute Error is the amount of error in your measurements. It** is the difference between the measured value and “true” value. For example, if a scale states 90 pounds but you know your true weight is 89 pounds, then the scale has an absolute error of 90 lbs – 89 lbs = 1 lbs.  
  
3 – Root mean squared error – It is the root of mean squared error.  
  
4 – R-squared - This is the most important evaluation metric for regression problem.  
R2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data.