MACHINE LEARNING

Ans 1: Between (-1,1)

Ans 2: PCA

Ans 3: Linear

Ans 4: Decision Tree Classifier

Ans 5: Same as old coefficient of “X”

Ans 6: Increases

Ans 7: Random Forests reduce overfitting

Ans 8: Principal Components are calculated using Unsupervised learning techniques

Ans 9: Identifying developed, developing, and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index.

Ans 10: min\_sample\_leaf

Ans 11: 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 observation. Outliers are straggles- extremely high or extremely low values, in a data set that can throw off your stats. Inter Quartile Range (IQR) is an outlier detection method. we can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any value that falls outside of this fence is considered an outlier. To build this fence we take, 1.5 times the IQR and then subtract the value from Q1 and add the value to Q3. This gives us the minimum and maximum fence posts that we compare with each observation. Any observations that are more than 1.5 IQR below (Q1) or more than 1.5 IQR above (Q3) are considered as outliers. This is a method that uses to identify outliers by default.

Ans 12

|  |  |
| --- | --- |
| Bagging | Boosting |
| The simplest way of combining predictions that belong to the same type | A way of combining predictions that belong to the different types |
| Each model is built Independently | New models are influenced y the performance of the previously built model |
| Bagging tries to solve the over-fitting problem | Boosting tries to reduce bias. |
| In Bagging, classifiers are trained parallelly | In Boosting, classifiers are trained Sequentially |
| Each model receives equal weight | Models are weighted according to their performance |
| Decrease Variance not Bias | Decrease Bias not Variance |
| For Ex: Random Forest uusesBagging | For Ex: AdaBoost uses Boosting |

Ans 13: Adjusted R2  give an idea of how many data points fall within the line of the regression equation. The adjusted R2 tells the percentage of variation explained by only independent variables that actually affect the dependent variable. Its value can be calculated based on the value of R2, the number of independent variable in a model. The R2 increases, even if the independent variable is insignificant It never declines.

Ans 14:

|  |  |
| --- | --- |
| Normalization | Standardization |
| It is used when features are of different scales | It is used when we want to ensure zero mean and until the standard deviation |
| It is affected by outliers | It is much less affected by outliers |
| MinMaxSclare is used for normalization | StandardScaler is used for Standardization |
| It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian |
| It is also known as Scaling Normalization | It is also known as Z-score Normalization |
| Scales values between (0,1) or (-1,1) | It is not bounded to a certain range |
| The minimum and Maximum values of features are used for scaling | Mean and Standard Deviation is used for scaling |

Ans 15: Cross-validation is a technique for evaluating Machine Learning models by training several Machine Learning models on subsets of the available input data and evaluating them on the complementary subset of the data. In Machine Learning, we couldn’t fit the model on the training data and can’t say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up to much noise. For this, we use the cross-validation technique. Use Cross-validation to detect overfitting,failing to generalize a pattern.

**Advantages of Cross-Validation**  
  
**1. Reduces Overfitting:** In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.  
  
  
**2. Hyperparameter Tuning:** Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.  
  
  
  
**Disadvantages of Cross-Validation**  
  
**1. Increases Training Time:** Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.   
  
For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.  
  
**2. Needs Expensive Computation:** Cross Validation is computationally very expensive in terms of processing power required.