Machine Learning

1. Between -1 and 1
2. PCA
3. Linear
4. Logistic Regression
5. old coefficient of ‘X’ ÷ 2.205
6. Increases
7. Random Forests reduce overfitting
8. Principal Components are calculated using supervised learning techniques and Principal Components are linear combinations of Linear Variables.
9. Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index and Identifying spam or ham emails
10. max_depth and min_samples_leaf

11. IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. **The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR** are outliers. Example: Assume the data 6, 2, 1, 5, 4, 3, 50.

12. Bagging: Bagging attempts to tackle the over-fitting issue. If the classifier is unstable (high variance), then we need to apply bagging

    Boosting: Boosting tries to reduce bias..If the classifier is steady and straightforward (high bias), then we need to apply boosting.

13. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

14. **Normalization** rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.
    Xchanged=X−XminXmax−XminXchanged=X−XminXmax−Xmin

    Standardization rescales data to have a mean (μμ) of 0 and standard deviation (σσ) of 1 (unit variance).
    Xchanged=X−μσ

15. Cross-validation is used to protect a model from overfitting, especially if the amount of data available is limited. It's also known as rotation estimation or out-of-sample testing and is mainly used in settings where the model's target is prediction.

Advantages:

- Cross-validation helps to determine a more accurate estimate of model prediction performance.

Disadvantages:

- Cross-validation is computationally very expensive as we need to train on multiple training sets.