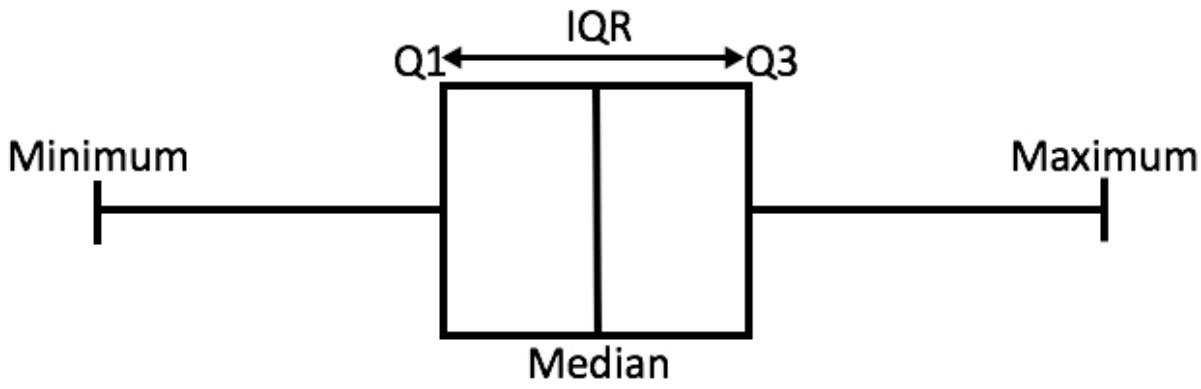
MACHINE LEARNING ASSIGNMENT 4 SOLUTION

1. C) between -1 and 1
2. C) Recursive feature elimination
3. A) linear
4. A) Logistic Regression
5. D) Cannot be determined
6. B) increases
7. C) Random Forests are easy to interpret
8. C) Principal Components are linear combinations of Linear Variables.
9. B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
   C) Identifying spam or ham emails
10. A) max_depth
    D) min_samples_leaf

11. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

To explain IQR Method easily, let's start with a box plot.



A box plot tells us, more or less, about the distribution of the data. It gives a sense of how much the data is actually spread about, what's its range, and about its skewness. As you might have noticed in the figure, that a box plot enables us to draw inference from it for an ordered data, i.e., it tells us about the various metrics of a data arranged in ascending order.

minimum is the minimum value in the dataset,

and *maximum* is the maximum value in the dataset.

So the difference between the two tells us about the range of dataset.

The *median* is the median (or centre point), also called second quartile, of the data (resulting from the fact that the data is ordered).

Q1 is the first quartile of the data, i.e., to say 25% of the data lies between *minimum* and *Q1*.

Q3 is the third quartile of the data, i.e., to say 75% of the data lies between *minimum* and *Q3*.

The difference between *Q3* and *Q1* is called the Inter-Quartile Range or IQR.

```
IQR = Q3 - Q1
```

12.Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

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.

R^2 = Explained variation / Total Variation

R-Squared is also called coefficient of determination. It lies between 0% and 100%. A r-squared value of 100% means the model explains all the variation of the target variable. And a value of 0% measures zero predictive power of the model. Higher R-squared value, better the model.

14. In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.

15. Cross-validation is a technique that allows us to utilize our training data better for training and evaluating the model. For example, while using cross-validation, you effectively use complete data for training the model. Cross-validation also helps in finding the best hyperparameter for the model.

Advantages:

Checking Model Generalization: Cross-validation gives the idea about how the model will generalize to an unknown dataset
Checking Model Performance: Cross-validation helps to determine a more accurate estimate of model prediction performance

Disadvantages:

Higher Training Time: with cross-validation, we need to train the model on multiple training sets.
Expensive Computation: Cross-validation is computationally very expensive as we need to train on multiple training sets.