MACHINE LEARNING ASSIGNMENT:

1)C

2)D

3)C

4)D

5)B

6)C

7)A

8)B,C

9)B,A

10)D,A

11) The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.
However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.

- Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results.
- Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.
- Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

If a dataset has *2n / 2n+1* data points, then
Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.
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.

12) As we helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. Basic idea is to learn a set of classifiers (experts) and to allow them to vote. **Bagging** and **Boosting** are two types of **Ensemble Learning**. These two decrease the variance of a single estimate as they combine several estimates from different models. So the result may be a model with higher stability. Let's understand these two terms in a glimpse.

1. **Bagging**: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.
2. **Boosting**: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

14) **Standardization** rescales a dataset to have a mean of 0 and a standard deviation of 1. It uses the following formula to do so:

$$x_{new} = (x_i - x) / s$$

where:

- $x_i$: The $i^{th}$ value in the dataset
- $x$: The sample mean
- $s$: The sample standard deviation

**Normalization** rescales a dataset so that each value falls between 0 and 1. It uses the following formula to do so:

$$x_{new} = (x_i - x_{min}) / (x_{max} - x_{min})$$

Whether you decide to normalize or standardize your data, keep the following in mind:

- A **normalized dataset** will always have values that range between 0 and 1.
- A **standardized dataset** will have a mean of 0 and standard deviation of 1, but there is no specific upper or lower bound for the maximum and minimum values.

15) It is the process by which the machine learning models are evaluated on a separate set known as validation set or hold-out set with which the best hyper-parameters are found, so that we get the optimal model, that can be used on future data and which is capable of yielding the best possible predictions

**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.

13) *Adjusted R Squared refers to the statistical tool that helps investors measure the extent of the variable's variance, which is dependent and explained with the independent variable. It considers the impact of only those independent variables that impact the variation of the dependent variable.*

Adjusted R Squared or Modified R^2 determines the extent of the variance of the dependent variable, which the independent variable can explain. The specialty of the modified R^2 is that it does not consider the impact of all independent variables but only those which impact the variation of the dependent variable. Therefore, the value of the modified R^2 can also be negative, though it is not always negative.

The formula to calculate the adjusted R square of regression is below:

**$R^2 = \{(1 / N) * \Sigma [(xi – x) * (Yi – y)] / (\sigma x * \sigma y)\}^2$**