Machine learning assignment 4

Answer-1 C) between -1 and 1

Answer-2 A) Lasso Regularisation

Answer-3 A) linear

Answer-4 A) Logistic Regression

Answer-5 B) same as old coefficient of 'X'

Answer-6 B) increases

Answer-7 A) Random Forests reduce overfitting

Answer-8
B) Principal Components are calculated using unsupervised learning techniques

Answer-9
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.
A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

Answer-10 D) min_samples_leaf

Answer 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 observations.

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

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.

Answer-12

S.NO	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.

S.NO	Bagging	Boosting
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.

Answer-13

Adjusted R^2 is **a corrected goodness-of-fit (model accuracy) measure for linear models**. It identifies the percentage of variance in the target field that is explained by the input or inputs. R^2 tends to optimistically estimate the fit of the linear regression.

Adjusted R^2 might decrease if a specific effect does not improve the model. Adjusted R squared is calculated by **dividing the residual mean square error by the total mean square error** (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

Answer-14

Normalisation	**Standardisation**
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers
It is applied when we are not sure about	It is used when the data is Gaussian or

the data distribution	normally distributed
It is also known as Scaling Normalization	It is also known as Z-Score

Answer-15

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Advantages of cross-validation: **More accurate estimate of out-of-sample accuracy**. More “efficient” use of data as every observation is used for both training and testing.

The disadvantage of this method is that **the training algorithm has to be rerun from scratch k times**, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.