1. **Assumptions of Linear Regression**

a. Linear Relationship

b. Multivariate Normalist : Residuals should be normally distributed

c. No or little multicollinearity

d. Homoscedacity

2**. Difference between Accuracy and precision**

a. ACcurate is Correct (or Close to real value)

b. PRecise is Repeating (or Repeatable)

3. **Compare the performance of an organization with other organization and check which performs better.**

a. Z-Score

(x-mean)/STD

4. **Power of Hypothesis test and why it is important**

a. The probability of not committing a Type II error is called the power of a hypothesis test.

5. **Difference between KNN and K-Means**

a. **KNN** is a supervised classification algorithm. The prediction of a test sample is based on the similarity of its features to its neighbors. The similarity is computed based on a measure such as Euclidean Distance. K here refers to the number of neighbors with whom similarity is being compared.

b**. K-Means** is an unsupervised clustering technique.

It is a process of defining clusters or groups, around pre-defined centroids based on similarity of each data point to the other. K here refers to number of centroids around which each cluster will be defined.

6. **Random Forest :**

It is a supervised tree based machine learning algorithm that uses an ensemble of decision trees and random selection of features to reduce the error in prediction.

In Layman terms:

Let's say you have received 3 job offers and you want to make a decision on which job to accept. There are a number of features on which your decision is based such as: Brand name, employee satisfaction, compensation, growth, travel, culture etc.

You can reach out to 10 of your linkedin contacts and quiz them about the 3 companies. Everytime you choose a random set of 3 features and ask a contact to recomment you one out of the three companies based on the 3 features. The company which gets the most recommendations will be your final choice.

7. **K fold cross validation and why it is done:**

a.It is a method to avoid overftting.

b. Final accuracy = Average of each round.

8. **Difference between Precision and Recall:**

a. **Recall** : Recall is the ratio of a number of events you can correctly recall to a number of all correct events.

b. If can recall all the events but it's not so precise.

c. **Precision** : Precision is the ratio of a number of events you can correctly recall to a number of all events you recall.(mix of correct and wrong recalls).

9. **True Positive, False Positive, True Negative, False Negative :**

a. **True Positive** : If the alarm goes on in case of a fire

i. Fire is positive and prediction made by the system is true.

b. **False Positive** : If alarm goes on, and there is no fire

i. System predicted fire to be positive which is a wrong prediction, hence the prediction is fake.

c. **False Negative** : If alarm does not go on but there was a fire,

i. System predicted fire to be negative which was false since there was fire

d**. True Negative** : If alarm does not go on and there was no fire

i. The fire is negative and this prediction was true.

10. **Inductive and Deductive Learning :**

a. Inductive Learning : Observation -> Conclusion

b. Deductive Learning : Conclusion -> Observation

11. **ROC Curve :**

a. ROC Curve is a fundamental tool for diagnostic test evaluation and is a plot of the true positive rate(Sensitivity) against the false positive rate(Specificity) for the different possible cut-off points of a diagnostic test.

Performance paramter for a Model and is used in the cases of binary classification

i. It shows the tradeoff between sensitivity and specificity(any increase in sensitivity will be accompanied by a decrease in specificity).

ii. The closer the curve follows the left hand border and then the top border of the ROC space, the more accurate the test.

iii. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

iv. The slope of the tangent line at cutpoint gives the likelihood ratio(LR) for that value of the test.

v. The area under the curve is a measure of text accuracy.

12. **Difference b/w Type I and Type II error:**

a. Type I : Type I error is a false positive. It is claiming something has happened when it hasn't.

b. Type II : Type II error is a false negative. It is claiming nothing when in fact something has happened.

13. **Which is important : model accuracy or model performance :**

a. Model accuracy is only a subset of model performance.

14. **Difference between Gini Impurity and Entropy in a Decision Tree:**

a. **Impurity** : How disclassified your classes are within the tree.

b. WHen you make a splits, how your classes are being split.

i. There two are the metrics for deciding how to split a tree.

ii. **Gini** measurement is the probability of a random sample being classified correctly if you randomly pick a label according to the distribution in the branch.

iii. **Entropy** is a measurement of lack of information. You calculate the information gain by making a split, which is the difference in entropies. This measures tells how you reduce the uncertainity about the label.

15. **Difference between Entropy and Information Gain :**

a. Entropy is an indicator of how messy your data is. It keeps on decreasing as you reach closer to the leaf node.

b. The information gain is based on the decrease in entropy after a dataset is split on an attribute. It keeps on increasing as you reach closer to the leaf node.

16. **How to ensure that the model is not overfitted:**

i. Collect more data

ii. Use **ensembling** methods that "average" the models

iii. Choose simpler models(Logistic regression).

iv. **Regularization**(L1 and L2 which penalized the model and tries to balance the data)

17. **Explain Ensemble learning technique:**

a. Ensemble is nothing but different model combined together and each model tries to understand the data in a different manner.

b. Each model will try to capture different pattern in the data.

c. Each model is a week learner and when combined together gives a strong model which is better predictor model.

d. In regression we take the average of the values and create the new value and in classification we take the value with the majority of votes.

d. So, training large number of models and combine the predictions of those model and generate a conclusion out of it.

Two types : Bagging and Boosting

i. **Bagging** : Different samples will be trained on a single algorithm and then combine the output of the algorithm

ii. **Boosting** : Model is trying the misclassification which is there in the previous model and tries to learn from it and make better prediction for the next one.

iii. **Stacking** : Is combines multiple classification and regression model via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training data, then the meta-model is trained on the outputs of the base level models as features.

--Meta-classifiers are XGBoost, Neural Network, and Adaboost.

Base level often consists of different learning algorithms and therefore stacking are often heterogeneous.

18. **Bagging and Boosting :**

a. **Similarities :**

i. Both are ensemble methods to get N learns from 1 learner.

ii. Both generate several training data sets by random sampling.

iii. Both make the final decision by taking the average of N learners.

iv. Both are good at reducing variance and provide higher scalability.

b. **Difference :**

i. While they are built independently for Bagging, Boosting tries to add new models that do well where previous models fail.

ii. Only Boosting determines weight for the data to tip the scales in favour of the most difficult cases.

iii. Is an equally average for Bagging and a weighted average for Boosting more weight in those with better performance on training data.

iv. Only Boosting tries to reduce bias. On the other hand, Bagging may solve the problem of over-fitting, while boosting can increase it.

19. **Screen for Outliers and what to do if you find one:**

a. **How to find the outlier:**

i. Box Plot : Extreme Value Analysis
ii. Probalistic and Statistical Model
iii. Linear Models
iv. Proximity-based Models : K-means clustering
v. Information Theoretic Models
vi. High Dimensional Outlier Detection

b. **How to handle :**

i. Dropping the outliers
ii. Cap the data using percentile
iii. Impute based on rule

20. **Collinearity and Multicolinearity:**

a. **Collinearity :** Occurs when two predictor variables(e.g., x1 and x2) in a multiple regression have some correlation.

b. **Multicollinearity :** Occurs when more than two predictor variables(e.g., X1, X2 and X3) are inter-correlated.

21. **Eigenvectors and Eigenvalues :**

eigenvector of a square matrix A is a nonzero vector x such that for some number lambda('a') we have following :

$$Ax = ax$$

where a is an eigenvalue and x is an eigen vector.

22. **A/B Testing :**

a. Statistical hypothesis testing for randomized experiment with two variables A and B.

b. **Goal** : Identify any changes to the web page to maximize or increase the outcome of an interest.

c. **Example** : Identifying the click-through rate for a banner ad

23. **Cluster Sampling :**

a. It is a process of randomly selecting intact groups within the defined population, sharing similar characteristics.

b. Cluster sample is a probability sample where each sampling unit is a collection or cluster of elements.

24. **How does the tree decide on which variable to split at the root node and its succeeding child nodes:**

a. Calculate Gini for sub-nodes, using formula sum of probability for success and failure(p^2+q^2).

b. Calculate Gini for split using weighted Gini score of each node of that split.

c. Entropy is the measure of impurity or randomness in the data, (for binary class):

Entropy = -plogp - qlogq

d. Here p and q is probability of sucess and failure respectively in that node.

e. Entropy is zero when a node is homogeneous and is maximum when both the classes are present in the node at 50% - 50%. All we want is to achieve lower entropy.

25. **Handle duplicate values in a dataset.**

```
dataset.drop_duplicated()
dataset.duplicated()
```

26. **Confidence Interval will increase with the introduction to the outliers.**

As confidence interval depends on the standard deviation of the data. SO increase in outliers will increase the standard deviation.
Standard error of the mean is the standard deviation by the square root of the number of values :

Bias-Variance Tradeoff :
Bias means "how well the model fits the data".
It is the error due to an overly simplistic assumptions in the learning algorithm.
A high bias error means an under-performing model, which keeps on missing important trends.
Variance means the magnitude of the change in the model based on the changes in the data - in case of an overfitting model.
Meaning the predictive model being using is highly complex.
Getting more data can help in case of model with a very high variance.
How to debug algorithm to check high bias or high variance.

To fix high variance :

a. Get more training examples
b. Try some smaller sets of features.
c. Try increasing the Regularization parameter(lambda)

To fix high bias :

a. Try getting additional features.
b. Try adding polynomial features.
c. Try decreasing the Regularization partameter(lambda).