#1 What is Data Science? List the differences between supervised and unsupervised learning

'''Data Science means extracting knowledge and insights from structure, semi-structured or

Supervised:
1. When the data is labelled we use supervised ML
2. We take feedback to see if we are predicting correct values or not
3. we generally get more accuracy with labelled data
4. We provide input variables along with the output variable
5. Supervised learing predicts output
6.Supervised learning needs supervision to train the model.

Unsupervised:
1. When the data is not labelled we use unsupervised ML
2. We dont take feedback in Unsupervides ML
3. we generally get less accuracy with non labelled data
4. We provide input variables only
5. Supervised learing predict patterns
6.Supervised learning doenst needs supervision i.e why it's called unsupervised

#2 What is logistic regression?

''' Logistic regression is a Supervised machine learning regression and used
for classification problems to predict probabilities of Target variables'''

#3 How will you deal with the multiclass classification problem using logistic regression?

'''Even multiclass classifiaction works as a Binary Classification in background, if there
then first it will take High vs Medium, Low then it will take Medium vs High,Low and then

#4 What is the difference between Linear Regression and Logistic Regression?

'''In Linear regression the Output variables are contious and in Logistic regression the O
Linear Line and in Log R our model is sigmoid curve. Model equations and Cost Functions ar

#5 Why is logistic regression very popular?

'''Logistic regression is a simple and more efficient method for binary and linear classif
problems output can be yes or no, Fraud or No Fraud, High or Low'''

#6 What is the formula for the logistic regression function?

P(Y) = 1/1+e^(-Y) where y is nothing but mx+c

#7 What are the assumptions made in logistic regression?

'''
The dependent variable must be categorical in nature

```
The independent variable should not have multi-collinearity
'''



#8 Why is logistic regression called regression and not classification?

'''
It is derived from linear line equation only,

y = mx + c

then we need to find the probability thats why we devide it with y-1

y/y-1, 0 when y=0 and infinity when y =1, but we want range from -inf to inf so we take lo

log(y/y-1) = mx+c

'''


#9 Explain the general intuition behind logistic regression

'''
Logistic regression is a statistical model that in its basic form uses a logistic function
variable,  although many more complex extensions exist. In regression analysis, logistic r
is estimating the parameters of a logistic model (a form of binary regression)

'''


#10 Explain the significance of the sigmoid function.

'''
Sigmoid function use to map any values into probabilities between range 0 to 1. It convert
this range the shape of this sigmoid function/Logistic Function curve is S shape. We selec
descrete values. The value above threshold tends to 1 and below tends to 0
'''


#11 How does Gradient Descent work in Logistic Regression?

'''
It will work same as Linear Regression but here only cost function will get changed, then
and start derivating it wrt m and c until we reach to global minima. Here alpha will be st
'''


#12 Why can’t we use Mean Square Error (MSE) as a cost function for logistic regression?

'''
1. If we use MSE as a Cost Function then in Gradient Descent boosting we might end up at l

2. Lets say if we have two descrete values 1 and 0 in target column, and Actual value is 1
If we find MSE then (1-0)^2 is 1 only and if we put it into the logloss then it would be i
penalise the misclassifications even for the perfect mismatch!
```

```
'''



#13 What is the Confusion Matrix?

'''
Confusion Matrix is an evaluation matrix for Classification problems. In that we get know
classified and incorrectly classified out of actual vlaues
'''


#14 What are the false positives and false negatives?

'''
False positives are Actual Negative vlaues but predicted as a positive and False Negative
predcited as positive by classifier
'''


#15 What is the true positive rate (TPR) and true negative rate (TNR)?

'''
TPR: Positive Values got correctly classified by classifier

TNR: Negative Values got correctly classified by classifier

'''


#16 What is the false-positive rate (FPR) and false-negative rate (FNR)?

'''
FPR:
Negative values got incorreclty classified by by classifier

FNR:
Positive values got incorreclty classified by by classifier

'''


#17 What are precision and recall? Why this is important in model evaluation?

'''
Only Getting high accuracy is'nt enough.

Recall is proportion of Positive Values correctly preedicted by classifier out of Total ac
TP / TP + FN
If we are making some Predictions in Medical Field then this FN values have to be low.
If someone is Cancer positive and we are predicting that person as a Negative then it woul
If we are dealing with Medical Diagnosis, then we have to consider this Recall while evalu
Maximun Recall meaning very low False Negative value

Precision is proportion of True Positive values to total values predicted as Positive by c
```

```
TP / TP + FP
When we are detecting spam mails then we have to consider this Precision in account, here
model predicting it as a Spam Mail i.e False Positive. We have to Reduce False Positive Va
Maximun Precision meaning very low False Positive value
'''



#18 What is the purpose of the precision-recall curve?

'''
We plot precision and recall for different thresholds and see threshold is suitable for ou
'''


#19 What is f1-score and Explain its importance?

'''
F1 score is nothing but harmonic mean of Precision and Recall, Formula: 2 * Precision * Re
It comes in handy when we hae to consider both Precision as Well as Recall in account i.e

'''


#20 In classification problems like logistic regression, classification accuracy alone is

'''
accuracy can be high even if FP or FN are present in an output, so accuracy standalone can
precision as well as recall to see how our model is working

'''


#21 How can you calculate accuracy using a confusion matrix?

''' TP + TN / TP + FP + TN + FN, this will give accuracy of our model '''


#22 What is Sensitivity?
'''
Sensitivity can be called as Recall or TPR(True Positive Rate), Positive values accurately
'''


#23 What is Specitivity?

'''
Specitivity is nothing but TNR(True Negative Rate), How many negative values correctly pre

'''


#24 Explain the use of ROC curves and the AUC of a ROC Curve.

'''We plot TPR vs FPR curve for different accuracies and take the threshold which is givin


#25 What is the bias-variance tradeoff in machine learning? Explain Bias and Variance.
```

'''We Generally Find Optimum model where our Variance is also less and Bias in not high. B
depends only upon Training model and Variance is difference between Training and Testing M

#26 What is Overfitting?

'''When Training Model Accuracy is high and Testing Model Accuracy is very low then that c

#27 What is Underfitting?

'''When accuracies of both models are Low then that condition is called as Underfiiting'''

#28 How do you deal with overfitting and Underfitting in machine learning?

'''
Underfitting:

Increase parameters using existing parameters

Handle Outliers

Deal with the Missing Values

Add More Datapoints

Overfitting:

Remove Correlated Features

Ensembling Methods like Bagging and Bossting

Pruning

Hyperparameter Tuning

Regularization (Ridge and Lasso)

'''

#29 Explain Bias and variance using the bulls-eye diagram

'''We Generally Find Optimum model where our Variance is also less and Bias in not high. B
depends only upon Training model and Variance is difference between Training and Testing M

#30 What are the advantages and disadvantages of Logistic Regression?

'''
Advantages:

1. Easy to implement , understand
    Logistic Function(Signmoid Function), LogLoss

```
2. Less likely to overfitted:
    if model overfitted :
        Regularization Techniques to handle Overfitting in linear model

3. Good Accuracy on simple dataset(Less number of Features):
    naive bayes  >> 10000 columns
    Text classification

4. Perform well if dataset is linearly seperable

Disadvantages:

1. When independent variables are highly correlated with each other, it may affect on
performance of model(assumption of no multicollinearity)

2. Highly sensitive to outliers

3. linearly seperable data is rarely available in real world scenarios

4. If there is no Linear Relationship between independent variable and logit odd, it may a

'''
```