* Explain the steps in making a decision tree

1. Take the entire data set as input
2. Calculate entropy of the target variable, as well as the predictor attributes
3. Calculate your information gain of all attributes (we gain information on sorting different objects from each other)
4. Choose the attribute with the highest information gain as the root node
5. Repeat the same procedure on every branch until the decision node of each branch is finalized.

* How can you avoid overfitting your model?

Overfitting refers to a model that is only set for a very small amount of data and ignores the bigger picture. There are three main methods to avoid overfitting:

1. Keep the model simple-take fewer variables into account, thereby removing some of the noise in the training data
2. Use cross-validation techniques, such as k folds cross-validation
3. Use regularization techniques, such as LASSO, that penalize certain model parameters if they’re likely to cause overfitting.

* What are dimensionality reduction and its benefits?

The Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there’s no point in storing a value in two different units (meters and inches).

* How should you maintain a deployed model?

The steps to maintain a deployed model are:

**Monitor:**

Constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it’s doing what it’s supposed to do.

**Evaluate:**

Evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

**Compare:**

The new models are compared to each other to determine which model performs the best.

**Rebuild:**

The best performing model is re-built on the current state of data.

* How can you select k for k-means?

We use the elbow method to select k for k-means clustering. The idea of the elbow method is to run k-means clustering on the data set where ‘k’ is the number of clusters.

Within the sum of squares(WSS), it is defined as the sum of the squared distance between each number of the cluster and it’s centroid.

* How can outlier values be treated?

You can drop outliers only if it is a garbage value.

Example: height of an adult=abc ft. this cannot be true, as the height cannot be a string value. In this case, outliers can be removed.

If the outliers have extreme values, they can be removed. For example, if all the data points are clustered between zero to 10, but one point lies at 100, then we can remove this point.

If you cannot drop outliers, you can try the following:

1. Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.
2. Try normalizing the data, this way, the extreme data points are pulled to a similar range.
3. You can use algorithms that are less affected by outliers; an example would be random forests.

* You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn’t you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection, and can greatly improve a patient’s prognosis.

Hence, to evaluate model performance, we should use sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.

* We want to predict the probability of death from heart disease based on three risk factors: age, gender and blood cholesterol level. What is the most appropriate algorithm for this case?

Logistic Regression

* What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that’s easy to analyze.

* What is logistic regression?

Logistic regression is also known as the logit model. It is a technique used to forecast the binary outcome from a linear combination of predictor variables.

* What are recommender systems?

Recommender systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product.

* Explain cross-validation?

Cross-validation is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is to forecast and one wants to estimate how accurately a model will accomplish in practice.

* What is collaborative filtering?

Most recommender systems use this filtering process to find patterns and information by collaborating perspectives, numerous data sources, and several agents.

* What are the drawbacks of the linear model?

1. The assumption of linearity of the errors
2. It can’t be used for count outcomes or binary outcomes.
3. There are overfitting problems that it can’t solve.

* How regularly must an algorithm be updated?

You will have to update an algorithm when:

1. You want the model to evolve as data streams through infrastructure.
2. The underlying data source is changing.
3. There is a case of non-stationarity.

* Why is resampling done?

Resampling is done in any of these cases:

1. Estimating the accuracy of sample statistics by using subsets of accessible data, or drawing randomly with replacement from a set of data ponts.
2. Substituting labels on data points when performing significance tests.
3. Validating models by using random subsets (Bootstrapping, cross-validation).

* What is Selection bias?

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.

* What are the types of biases that can occur during sampling?

1. Selection bias
2. Undercoverage bias
3. Survivorship bias

* What is survivorship bias?

Survivorship bias is the logical error of focusing on aspects that support surviving a process and casually overlooking those did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.