* What is Typecasting?

Typescasting is a way of changing an object from one data type to the other

* What are different types of machine learning algorithms?

1. Supervised
2. Unsupervised
3. Reinforcement learning

* What is supervised learning?

Supervised learning is a machine learning algorithm of inferring a functions from labeled training data. The training data consists of a set of training examples

* Give examples of Supervised learning?

1. Support Vector Machines
2. Regression
3. Naïve Bayes
4. Decision Trees
5. K-nearest neighbour algorithm and neutral networks

* What is unsupervised learning?

Unsupervised learning is also a type of machine learning algorithm used to find patterns on the set of data given. In this, we don’t have any dependent variable or label to predict.

* Give examples of unsupervised learning?

1. K-means clustering
2. Principal component analysis
3. Hierarchical clustering

* What is ‘Naïve’ in a Naïve Bayes?

The Naïve Bayes method is a supervised learning algorithm, it is naïve since it makes assumptions by applying bayes’ theorem that all attributes are independent of each other.

* What is PCA? When do you use it?

Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets, at the same time minimizing information loss.

PCA technique is particularly useful in processing data where multi-colinearity exists between the features/variables. PCA can be used when the dimensions of the input features are high (e.g. a lot of variables). PCA can be also used for denoising and data compression.

* What is Multicollinearity?

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.

* Problems with having Multicollinearity?

Mulicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable.

* What is VIF in machine learning when will we use it?

The variance inflation factor(VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

* Solutions for multicollinearity

1. Drop the variables causing the problem.
2. Re-code the form of the independent variables. Also we can say feature engineering

For example, if x1 and x2 are collinear, you might try using x1 and the ratio x2/x1 instead.

1. Ridge and Lasso regression – This is an alternative estimation procedure to ordinary least squares. Penalizes for the duplicate information and shrinks or drops to zero the parameters of a regression model.
2. Increase in sample size may sometimes solve the problem of multicollinearity.

* How do I combine two data frames?

Using the concatenation function

* What is the main difference between classification regression and clustering techniques?

Regression and Classification are types of supervised learning algorithms while clustering is a type of unsupervised algorithm.

* What is the main difference between regression and classification techniques?

When the output variable is continuous, then it is a regression problem whereas when it contains discrete values, it is a classification problem.

* What is the primary purpose of histogram?
* What is the primary purpose of boxplot?
* Name two technique for outlier treatment?
* What is difference between wins and trim?