3) A chemist wants to find some interesting patterns in which patients are behaving upon administering the drug

Ans: The problem statement clearly infers finding results based on identifying pattern behavior how drug reacts to each patient. The end goal is to make decisions whether drug reacted to patient as excepted or not. For this problem statement decision tree model help in finding accurate results.

**Decision trees** is often similar to people’s decision process and is easy to interpret

4) How will you select suitable machine learning algorithm for a problem statement

Ans:

**Machine Learning algorithms** are classified two types:

1. supervised
2. un supervised

Supervised learning algorithms make predictions based on a set of examples. For instance, historical stock prices can be used to hazard guesses at future prices. Each example used for training is labeled with the value of interest—in this case the stock price. A supervised learning algorithm looks for patterns in those value labels. It can use any information that might be relevant—the day of the week, the season, the company's financial data, the type of industry, the presence of disruptive geopolitical events—and each algorithm looks for different types of patterns. After the algorithm has found the best pattern it can, it uses that pattern to make predictions for unlabeled testing data—tomorrow's prices.

In unsupervised learning, data points have no labels associated with them. Instead, the goal of an unsupervised learning algorithm is to organize the data in some way or to describe its structure. This can mean grouping it into clusters or finding different ways of looking at complex data so that it appears simpler or more organized.

**Choosing a Model :**

Understand the problem statement, check whether the problem can be resolved using machine learning algorithm. Identify target variable if the target variable is continuous and linearly dependent then choose Linear regression (Here all the independent variables should linearly depend on dependent variables).If the outcome variable is classification then choose logistic regression.i.e if we are checking for customer churn in customer dataset the outcome will be yes/No. If the dependent variable is the outcome of decision then choose decision tree i.e if the outcome variable has multiple groups and categorical variable prefer decision tree. If data sets consists un labeled data clustering techniques can help identifying structure and behavior of data.

**Linear regression and Linear classifier.**Despite an apparent simplicity, they are very useful on a huge amount of features where better algorithms suffer from overfitting.

**Logistic regression** is the simplest non-linear classifier with a linear combination of parameters and nonlinear function (sigmoid) for binary classification.

**Decision trees** is often similar to people’s decision process and is easy to interpret. But they are most often used in compositions such as Random forest or Gradient boosting.

**K-means** is more primal, but a very easy to understand algorithm, that can be perfect as a baseline in a variety of problems.

**PCA** is a great choice to reduce dimensionality of your feature space with minimum loss of information.

5) Define one problem statement for Education industry?

Ans:

Predicting performance of a student in annual exam based on his previous performance in quaterly,half yearly , pre annual and assignments marks.

Data set should consists of student data includes no\_of\_days,school\_attended,no\_of\_classes\_attended,subjects\_interested,assignment\_marks,qmarks,hmarks..etc