**Application of Machine learning**

* Image recognition
* Speech recognition
* Traffic prediction
* Self- driving cars
* Email Spam and Malware Filtering
* Virtual Personal Assistant
* Online Fraud Detection

**Association Rules mining and “Apriori Principle”**

* Association Rule Mining, as the name suggest, association rules are simple if/Then statements that help discover relationship between seemingly independent relational database or other data repositories.
* Most machine learning algorithms works with numeric datasets and hence tend to be mathematical. However, association rule mining is suitable for non-numeric categorical data and requires just a little bit more than simple counting.
* Association Rule Mining in is an Unsupervised Non-linear algorithm to uncover how the items are associated with each other. In it, frequent Mining shows which items appear together in a transaction or relation. It is majorly used by retailers, grocery stores, an online marketplace that has a large transactional database. The same way when any online social media, marketplace, and e-commerce websites know what you buy next using recommendations engines. The recommendations you get on item or variable, while you check out the order is because of Association rule mining boarded on past customer data.
* There are three common ways to measure association:

1. Support
2. Confidence
3. Lift

**Support**

* it says how popular an item is, as measured in the proportion of transactions in which an item set appears **Support (A => B) = P (A U B)**

**Confidence**

* it says how likely item Y is

**Application of Association Rule mining**

* Market basket Analysis
* Medical Diagnosis
* Census Data
* Protein Sequence

**Apriori Principle**

* Read each item in the transaction.
* Calculate the support of every item.
* If support is less then minimum support, discard the item. Else, insert into frequent itemset.
* Calculate confidence for each non-empty subset.
* If confidence is less than minimum confidence, discard the subset. Else, it into strong rules

**Strengths and the weaknesses of K-means clustering.**

**Strength**

* Simple – easy to understand and to implement.
* Efficient – Time complexity
* K -means is the most popular clustering algorithm.

**Weakness**

* The algorithm is only applicable if the mean is defined.
* For categorical data, k-mode – the centroid is represented by most frequent values.
* The user needs to specify K.
* The algorithm is sensitive to **outliers.**
* Outliers are data points that are very far away from other data points.

**Advantages and Disadvantages of Principal Component Analysis (PCA)**

**Advantages**

* Removes Correlated Features
  + In a real-world scenario, this is very common that you get thousands of features in your dataset. You cannot run your algorithm on all the features as it will be reduce the performance of your algorithm and it will not be easy to visualize that many features in any kind of graph. So, you must reduce the number of features in your dataset.
  + You need to find out the correlation among the features (Correlated variables). Finding correlation manually in thousands of features is nearly impossible, frustrating, and time-consuming. PCA does this for you efficiently.
  + After implementing the PCA on your dataset, all the principal components are independent of one another. There is no correlation among them.
* Improves Algorithm Performance
  + With so many features, the performance of your algorithm will drastically degrade. PCA is a very common way to speed up your Machine Learning algorithm by getting rid of correlated variables which do not contribute in any decision making. The training time of the algorithms reduces significantly with less number of features.
  + So, if the input dimensions are too high, then using PCA to speed up the algorithm is a reasonable choice.
* Reduces Overfitting
  + Overfitting mainly occurs when there are too many variables in the dataset. So, PCA helps in overcoming the overfitting issue by reducing the number of features.
* Improves Visualization
  + It is very hard to visualize and understand the data in high dimensions. PCA transform a high dimensional data (2 dimension) so that it can be visualized easily.

**Disadvantages**

* Independent variables become less interpretable:
  + After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.
* Data standardization is must before PCA:
  + You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.   
      
    For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.  
      
    Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied.  
      
    PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use **StandardScaler**from **Scikit Learn** to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.
* Information Loss:
  + Although Principal Components try to cover maximum variance among the features in a dataset, if we do not select the number of Principal Components with care, it may miss some information as compared to the original list of features.