

**Data Analytics**

**Assignment - 3**

Dharmesh Harsha (ESD15I018)

Akshay Kumar (CED15I031)

Vivek Khatua (MFD15I014)

### **Description:**

The motivation of the project aim to predict the age group and the number of cigarettes smoked by the particular age group. The following data mining techniques were used to arrive at the model to predict the above objective.

### **Data Mining Techniques:**

* 1. Association Rule Mining
     1. Apriori
     2. FP Growth
  2. Classification
     1. Decision Tree
     2. Naive Bayes
     3. Gaussian Naive Bayes
  3. Clustering
     1. K Means
     2. DBSCAN
     3. Agglomerative

### **About :**

### DataSet Source :

* + 1. Youth Tobacco Survey and Random Fille

### **Association Rule Mining:**

**Association rule learning** is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. This rule-based approach also generates new rules as it analyzes more data. The ultimate goal, assuming a large enough dataset, is to help a machine mimic the human brain’s feature extraction and abstract association capabilities from new uncategorized data.

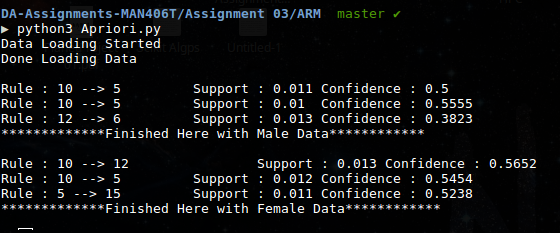
**The following rule mining algorithms predict the hour at which one would probably smoke a cigarette and also the hour at which he will smoke.**

1. **Apriori:**
   1. **Working:**

Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space.We will first describe this property, and then show an example illustrating its use.

Apriori property: All nonempty subsets of a frequent itemset must also be frequent.

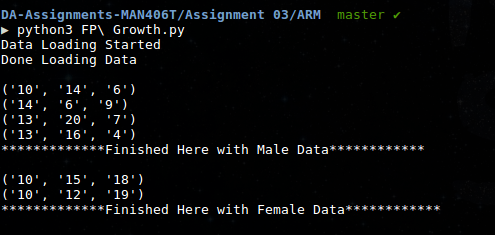
* 1. **Results:**

****

1. **FP growth:**
   1. **Working:**

FP-growth, which adopts a divide-and-conquer strategy as follows. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree,which retains the itemset association information. It then divides the compressed database into a set of conditional databases (a special kind of projected database), each associated with one frequent item or “pattern fragment,” and mines each such database separately.

* 1. **Results:**

****

### **5. Classification:**

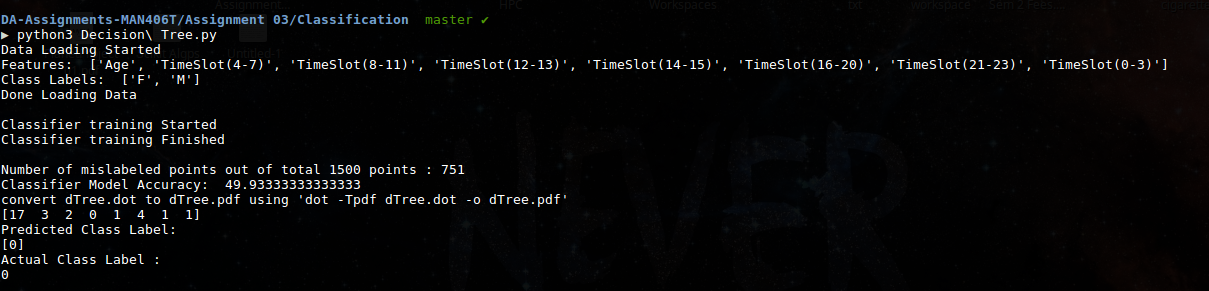
Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

**The objective of using the following algorithms are to predict the gender of the smoker given the hours the smoker prefers or say likes to smoke a cigarette.**

1. **Decision Tree:**
   1. **Working:**

Decision tree algorithm generates a tree of decisions based on the attribute list, The multidimensional training data set used for the decision tree generation is partitioned based on the attribute inferring and delivering highest information gain from the attribute list. The obtained class for the decision tree are majorly categorical in nature.

* 1. **Result:**

****

1. **Naive Bayes:**
   1. **Working:**

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

* 1. **Result:**

****

### **6. Clustering:**

**Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

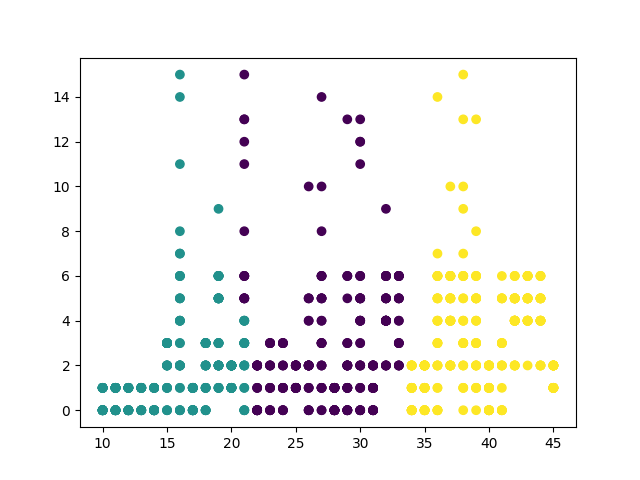
**The objective of the deploying the following algorithm is to estimate the age group in which chain smoking as a habit is prevalent.**

1. **K-Means:**
   1. **Working:**

*k*-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

This clustering approach clusters the data set based on distance between the cluster means and the data points. The K here means the number of cluster groups desired.

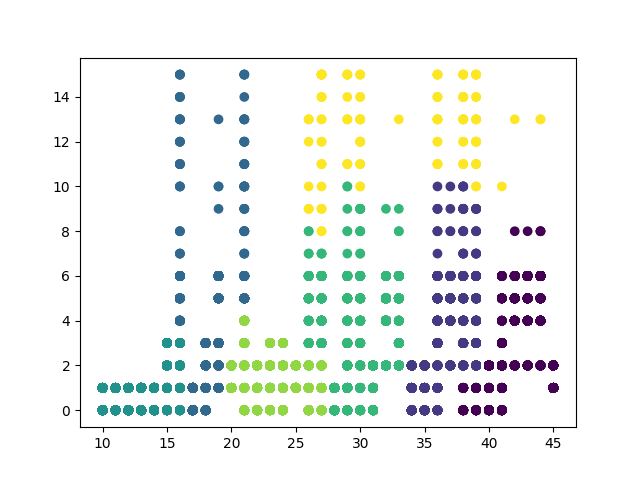
* 1. **Results:**

****

1. **Agglomerative:**
   1. **Working:**

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Agglomerative is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

* 1. **Results:**

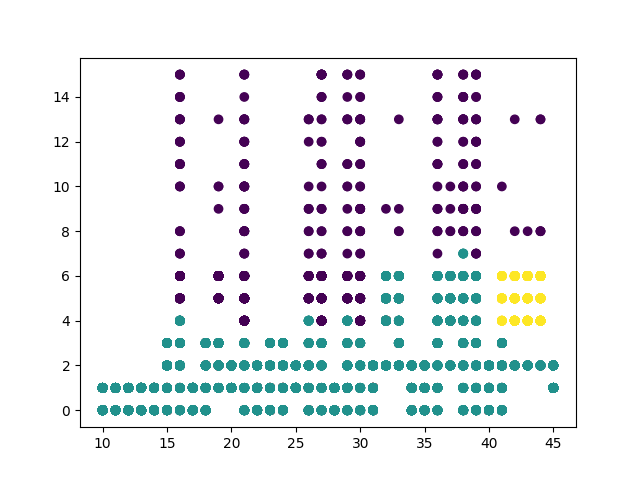
****

1. **DB scan:**
   1. **Working:**

DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a dense region (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster.

If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.DBSCAN can be used with any distance function (as well as similarity functions or other predicates). The distance function (dist) can therefore be seen as an additional parameter.

* 1. **Results:**

****