Problem statement

We are using adult dataset or our mini project.This dataset is created Barry Becker from 1994 Census databaset.the original dataset copy can be found at UCI Machine lernningsepertory.the dataset has 48842 instances but we will use the irst 1500 isances or our project.

This dataset can be use or survying and analyzing salaries fo people from dierent working class and with different education

Attribute used:

Age:Continous

Work class(categorical): Privatesel-emp-not-inc,sel-emp-inc,ederal-gov,local-gov,state-gov,without-pay,never-worked.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam,

Education(categorical): Bachelors, Some-college, 11th, HS-grad

Salary: >50K, <=50K.

Classiication:

Attribute used:Workclass and education

A classifier will be built by using naïve Bayesclassification algorithm. The current marketing strategy, which involves a high degree of investment per offer (driving the relatively high offer cost) has achieved a high acceptance rate in the past of seventy five percent by individuals whose salaries exceed fifty thousand US dollars. The goal of this project is therefore to create a model (or models) which can accurately identify individuals whose annual salary exceeds this amount.

Clustering:

Attribute used: age

Clustering will be done by using K-means clustering alogrithm. The clustering will be done on the bases ofage. This will help us to relate age with working class and education.

**Pattern mining:**

Attribute used: Work class, Education, marital-status, native-country.

Pattern and association rules will be generated by using Apriori algorithm.

Features

Support the business objective of maximizing return on investment and the provision of descriptions of the attributes.

Description of dataset

The dataset used in this project has forty nine thousand records and a binomial label indicating a salary of less or greater than fifty thousand US dollars, which for brevity, will be referred to as< 50K or >50K in this report. Seventy six percent of the records in the dataset have a class label of The data has been divided into a training dataset containing thirty two thousand records and a test dataset containing sixteen thousand records.

There are fourteen attributes consisting of seven polynomials, one binomial and six continuous attributes . The nominal employment class attribute describes the type of employer such as self employed or federal and occupation describes the employment type such as farming or managerial. The education attribute contains the highest level of education attained such as high school graduate or doctorate. The relationship attribute has categories such as unmarried or husband and the marital status attribute has categories such as married or separated. The final nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (which is a numerical representation of the nominal education attribute), capital gain and loss and a survey weight attribute which is a demographic score assigned to an individual based on information such as area of residence and type of employment.

Data mining algorithms used

Classification:

Naïve Bayes

In [machine learning](http://en.wikipedia.org/wiki/Machine_learning), naive Bayes classifiers are a family of simple [probabilistic classifiers](http://en.wikipedia.org/wiki/Probabilistic_classifier) based on applying [Bayes' theorem](http://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naive) [independence](http://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features.

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the [text retrieval](http://en.wikipedia.org/wiki/Information_retrieval)community in the early 1960s, and remains a popular (baseline) method for [text categorization](http://en.wikipedia.org/wiki/Text_categorization), the problem of judging documents as belonging to one category or the other (such as [spam or legitimate](http://en.wikipedia.org/wiki/Spam_filtering), sports or politics, etc.) with [word frequencies](http://en.wikipedia.org/wiki/Bag_of_words) as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including [support vector machines](http://en.wikipedia.org/wiki/Support_vector_machine). It also finds application in automatic [medical diagnosis](http://en.wikipedia.org/wiki/Medical_diagnosis).

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. [Maximum-likelihood](http://en.wikipedia.org/wiki/Maximum-likelihood_estimation) training can be done by evaluating a [closed-form expression](http://en.wikipedia.org/wiki/Closed-form_expression), which takes [linear time](http://en.wikipedia.org/wiki/Linear_time), rather than by expensive [iterative approximation](http://en.wikipedia.org/wiki/Iterative_method) as used for many other types of classifiers.

In the [statistics](http://en.wikipedia.org/wiki/Statistics) and [computer science](http://en.wikipedia.org/wiki/Computer_science) literature, Naive Bayes models are known under a variety of names, including simple Bayesand independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a [Bayesian](http://en.wikipedia.org/wiki/Bayesian_probability) method; [Russell and Norvig](http://en.wikipedia.org/wiki/Artificial_Intelligence:_A_Modern_Approach) note that "[naive Bayes] is sometimes called a Bayesian classifier, a somewhat careless usage that has prompted true Bayesians to call it the idiot Bayes model.

Clustering:

K-Means

*k*-means clustering is a method of [vector quantization](http://en.wikipedia.org/wiki/Vector_quantization), originally from signal processing, that is popular for [cluster analysis](http://en.wikipedia.org/wiki/Cluster_analysis) in [data mining](http://en.wikipedia.org/wiki/Data_mining). *k*-means clustering aims to [partition](http://en.wikipedia.org/wiki/Partition_of_a_set) *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest [mean](http://en.wikipedia.org/wiki/Mean), serving as a [prototype](http://en.wikipedia.org/wiki/Prototype) of the cluster. This results in a partitioning of the data space into [Voronoi cells](http://en.wikipedia.org/wiki/Voronoi_cell" \o "Voronoi cell).

The problem is computationally difficult ([NP-hard](http://en.wikipedia.org/wiki/NP-hard)); however, there are efficient [heuristic algorithms](http://en.wikipedia.org/wiki/Heuristic_algorithm) that are commonly employed and converge quickly to a local optimum. These are usually similar to the [expectation-maximization algorithm](http://en.wikipedia.org/wiki/Expectation-maximization_algorithm) for [mixtures](http://en.wikipedia.org/wiki/Mixture_model) of [Gaussian distributions](http://en.wikipedia.org/wiki/Gaussian_distribution) via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Pattern mining:

Apriori

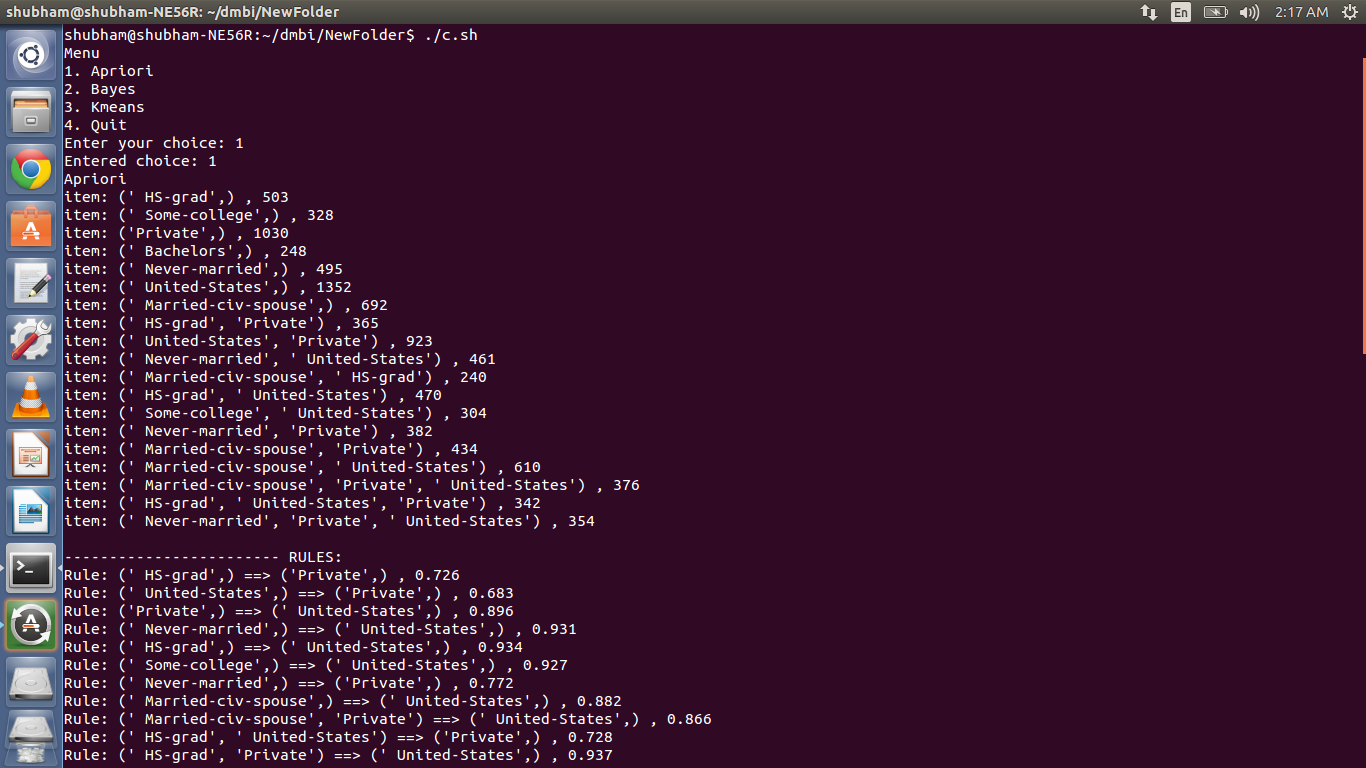
Apriori is designed to operate on [databases](http://en.wikipedia.org/wiki/Database) containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions ([Winepi](http://en.wikipedia.org/wiki/Winepi" \o "Winepi) and Minepi), or having no timestamps (DNA sequencing). Each transaction is seen as a set of items (an *itemset*). Given a threshold C, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database.

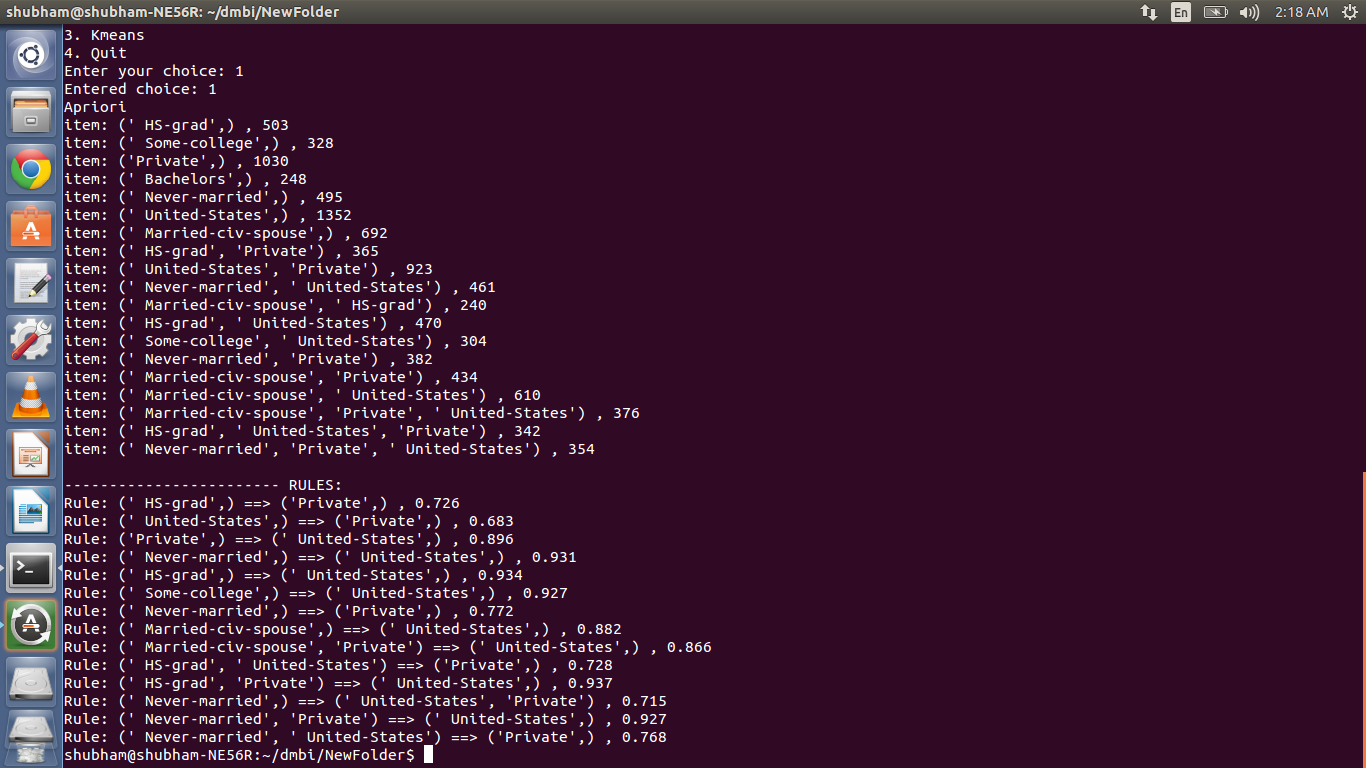
Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Apriori uses [breadth-first search](http://en.wikipedia.org/wiki/Breadth-first_search) and a [Hash tree](http://en.wikipedia.org/wiki/Hash_tree_(persistent_data_structure)) structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates which have an infrequent sub pattern. According to the [downward closure lemma](http://en.wikipedia.org/w/index.php?title=Downward_closure_lemma&action=edit&redlink=1), the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

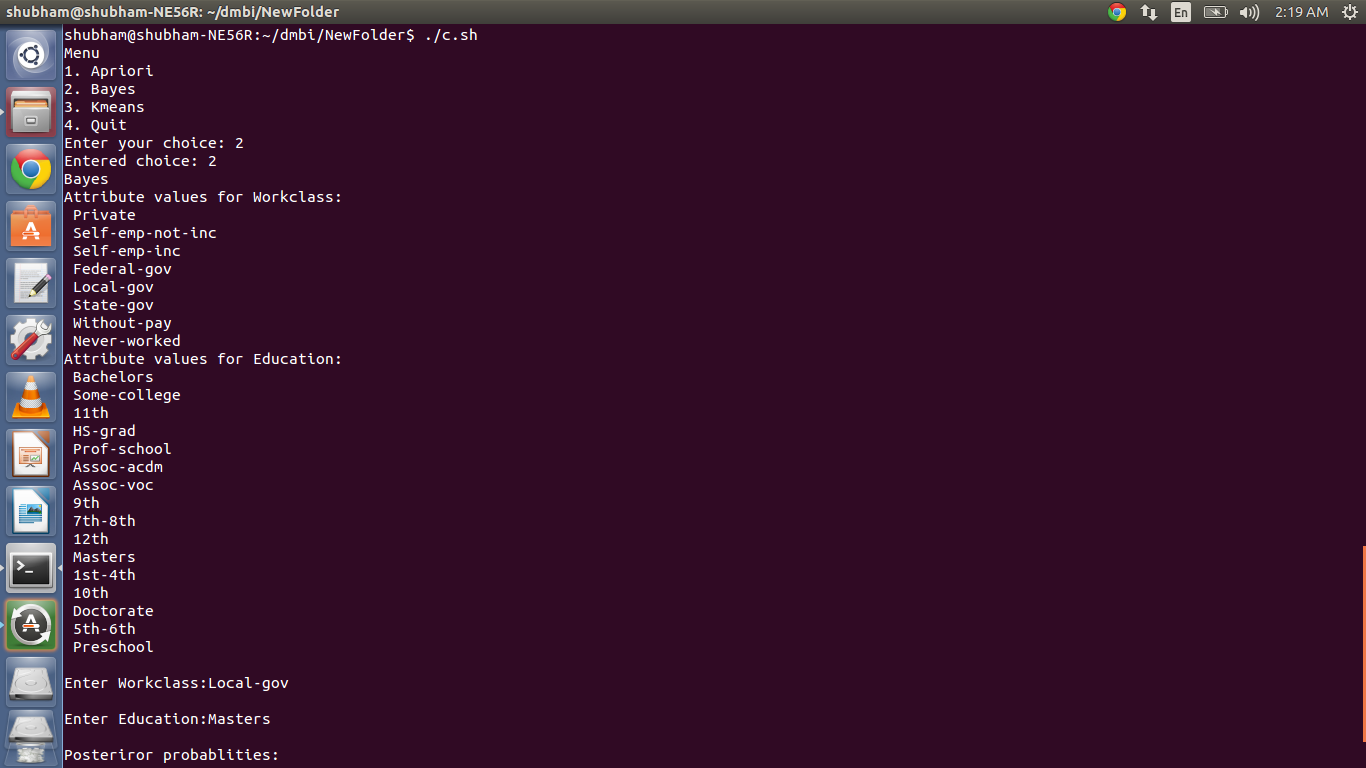
**Screenshots:**

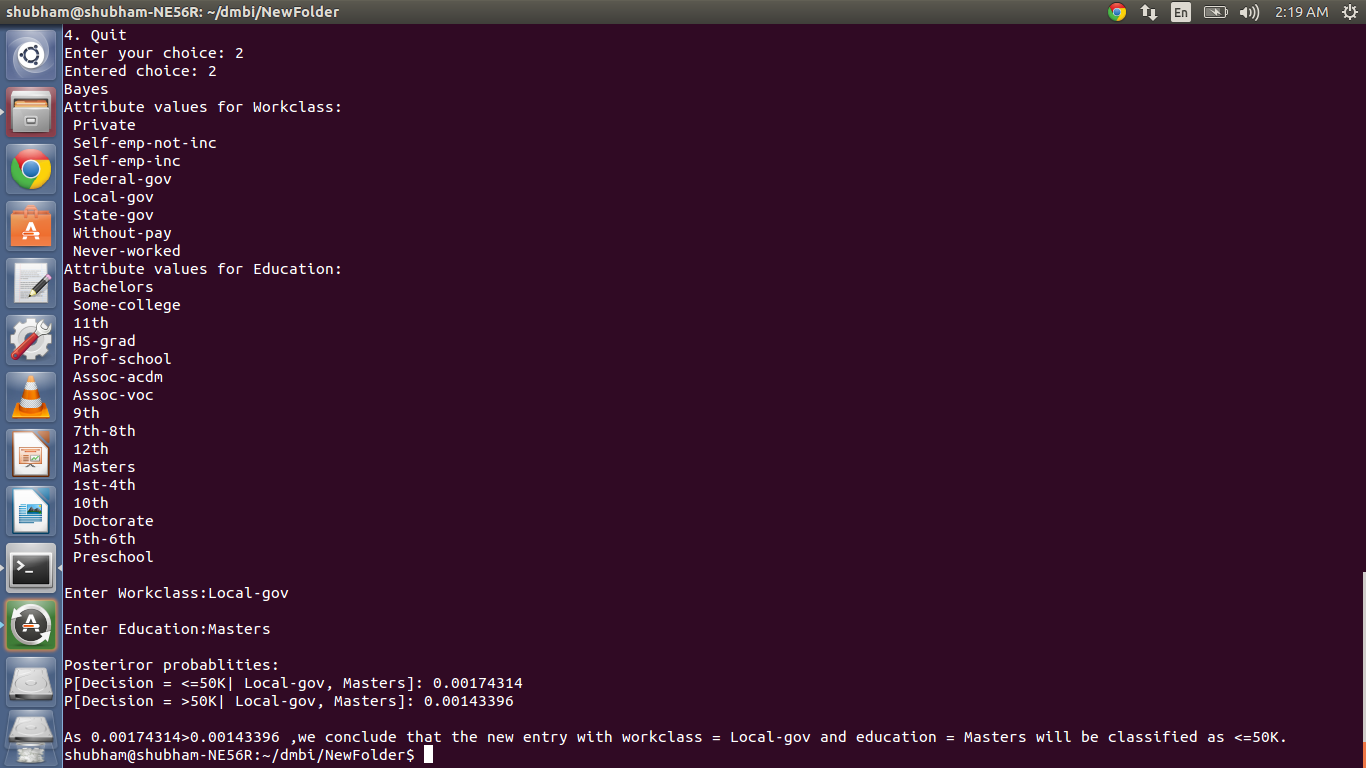
1. Apriori:





2. Bayes:





3. Kmeans:

