**CS 634 Data Mining Final Term Project**

**Submitted To:**

**Yasser Abduallah**

**Submitted By:**

**Devansh Sharma**

**Department of Computer Science New Jersey Institute of Technology**

**Option 1:** I have selected option 1 and the below mentioned software tool and algorithms are used.

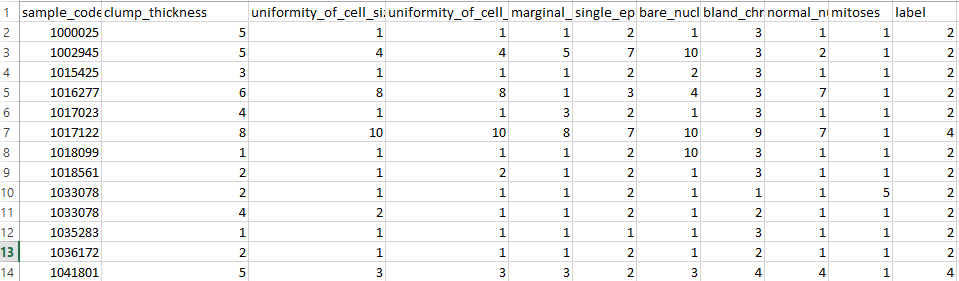
**Algorithms:** Vector Machines, Decision Trees, Algorithm (Naïve Bayes) , Random Forest

**Source of dataset**: <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)>

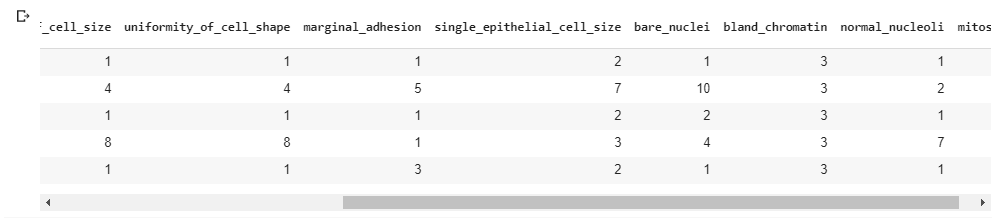
**Software Tool:** <https://colab.research.google.com/>

**Data Set:**

The data set used in final term project is named as “Breast cancer Wisconsin dataset” this data set is modification of original Wisconsin data set. The original dataset have 699 instances and 32 attributes and 2 classes of malignant and benign cancer.



**Feature Sample**



**Algorithms**

In my final data mining project the algorithms I have used are

1. **Naïve Bayes**
2. **Decision Trees**
3. **SVM**
4. **Random Forest**

**Naive Bayes**

Naive Bayes is a machine learning algorithm for classification problems. It is based on Thomas Bayes’s probability theorem. It is primarily used for text classification which involves high-dimensional training datasets. A few examples are spam filtration, sentimental analysis, and classifying news articles. Naive Bayes algorithm. Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. In short, it is a probabilistic classifier. The Naive Bayes algorithm is called “naive” because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features. The “Bayes” part refers to the statistician and philosopher Thomas Bayes and the theorem was named after him, Baye’s theorem, which is the base for Naive Bayes algorithm. Naive Bayes algorithm is Baye’s theorem or alternatively known as Baye’s rule or Baye’s law. It gives us a method to calculate the conditional probability, i.e. the probability of an event based on previous knowledge available on the events. More formally, Baye’s theorem is stated as the following equation.

**P (A|B) = P(B|A)P(A)/P(B)**

**Decision Tree**

It is based on Hunt’s algorithm. Hunt’s algorithm grows a decision tree in a recursive fashion by partitioning the training records into successively purer subsets. J48 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses gain ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

**SVM**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they’re able to categorize new text.

**Experimental Tool**

The tool that have been used for experimental purposes is Google Colab in which python version 3.7 has been used to analyze and train data set results. Four classifiers naïve Bayes, decision tree, random forest and SVM have been used for prediction of breast cancer and also their respective accuracy, sensitivity and specificity has been analyzed and detailed comparison have been also done. Moreover, some other statistical measures has been also used i.e. Precision, accuracy and recall on the basis of 10 fold cross validation.

**Results and Discussions**

The breast cancer database consists of nine conditional attributes. The decisional attribute takes the values 0 or 1. As presented in Figure 1, the distributions of almost all values of attributes are even. In case of almost all the attributes, the number of instances in which the attributes take the lowest values is the greatest. All conditional attributes are multi-valued. The results of the comparison of the algorithms are presented in Table 1. The table shows the ranking of the algorithm in case of each of the performance measures and databases.

**Table 1: Comparison of Different classifiers on the selected classifiers**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **True Positive Rate** | **True Negative Rate** | **False Positive Rate** |
| **RF** | 0.94 | 0.94 | 0.94 | 0.94 | 0.91 | 0.95 | 0.044 |
| **NB** | 0.97 | 0.96 | 0.98 | 0.97 | 1.0 | 0.95 | 0.044 |
| **DF** | 0.97 | 0.98 | 0.96 | 0.97 | 0.79 | 0.89 | 0.10 |
| **SVM** | 0.97 | 0.96 | 0.97 | 0.97 | 0.79 | 0.90 | 0.10 |