Methods:

We divided our tasks in two steps. First of all, we predicted whether a person drinks coffee or not based on waking up time, day planning and CGPA. And later, we found out correlation between various types of habits which will be discussed on.

Preprocessing

Feature Selection

Classifier

Clustering

Output Best Category

Accuracy

**Steps to solution:**

* Preprocessing
* Classifier
* Clustering
* Accuracy

Preprocessing:

* Marge all data
* Feature selection
* Label the data

Classifier:

* Gaussian Naive Bayes

### Bernouli Naive Bayes

### SVM with RBF kernel

### SVM

### Random Forest

### KNN

### Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of [feature](https://en.wikipedia.org/wiki/Feature_vector) values, where the class labels are drawn from some finite set. There is not a single [algorithm](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Independence_(probability_theory)) of the value of any other feature, given the class variable.

### In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), support-vector machines (SVMs, also support-vector networks[[1]](https://en.wikipedia.org/wiki/Support-vector_machine#cite_note-CorinnaCortes-1)) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis).

### K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition

### Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

### Result:

### We First applied classifier to find accuracy and then applied clustering for finding best category.

### For problem 1 we applied different Classifier .there accuracy is below:

|  |  |
| --- | --- |
| Classifier | Accuracy |
| DT | 56.363636 |
| GN | 51.363636 |
| NB | 59.772727 |
| SVC | 62.348485 |
| LSVC | 54.469697 |
| RFC | 58.636364 |
| KNN | 51.969697 |

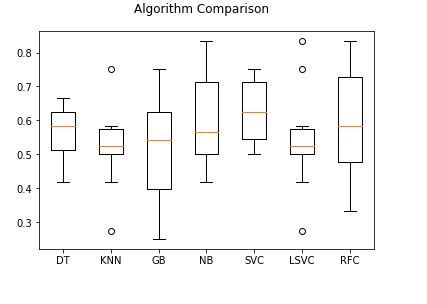


Fig: Boxplot Comparison Accuracy(problem 1)

For problem 2 we applied different Classifier .there accuracy is below:

|  |  |
| --- | --- |
| Classifier | Accuracy |
| DT | 73.712121 |
|  |  |
| GN | 62.727273 |
| NB | 75.227273 |
| SVC | 77.727273 |
| LSVC | 68.636364 |
| RFC | 69.242424 |
| KNN | 70.227273 |

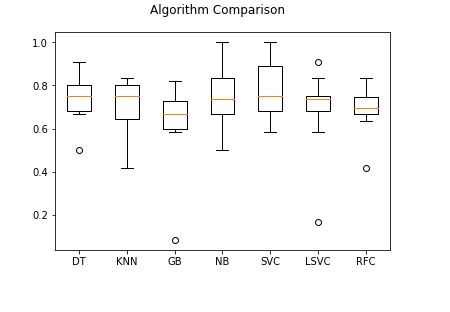


Fig: Boxplot Comparison Accuracy(problem 2)