III. PROPOSED ARCHITECTURE

In this section, we have demonstrated the proposed architecture of the system. Firstly, we have collected data utilizing self-report surveys, then the statistical analysis of the data acquired were performed. The parameters for the machine learning systems were considered as features, thus the features are numerical in this context. Furthermore, different machine learning algorithms have been applied to the synthesized features. The steps of the proposed architecture are data collection, statistical analysis, feature extraction, applying machine learning (supervised) algorithms and finally performance evaluation.

A.

Data Collection For this critical type of analysis, we have gathered data utilizing self-report survey.

Self-report surveys could be utilized to identify successful womens. We have circulated the survey among the village women of Kishoreganj,Narayanganj and Jossore. The survey has been taken using online facilities as well as printed. The statistical measurements of the collected data are mentioned in Table I. The survey was filled up by subjects who are of 18-46 +age limits. Same number of attributes are fed into the machine learning models.

Table I

|  |  |
| --- | --- |
| Number of subjects |  |
| Age Groups | 18-46+ |
| Total Collected attribute | 20 |
| No. of instances for ML |  |

B. Data pre-processing

For data pre-processing, we have to look carefully at the collected data. We found many missing data and incomplete surveys from the subjects. If there are any missing values in the gender column, then it is removed. In case of missing values in more than 2 items, the data are omitted from the dataset.

C. Feature Analysis

Feature analysis is one of the major steps in many machine learning models, so is our context. The features are generally considered to be extracted from the amount of raw data and characterize them into vectors as equation

(1). F = <f1,f2….fn…(1)

Where, fi denotes the different features and n denotes the total number of features. Sometimes the features are called as attributes of the system. Therefore, features/attributes are needed to be fetched from the raw data that we have collected. The data collection was performed on the categorical data, thus the categorical data can be used as valid attributes to the models.

Applying Machine Learning Algorithms

Applying machine learning to the selected attributes is one of the primary contributions of our research. We have analyzed the features to be used for this step and selected 20 features and 1 class label to create the annotated dataset. The annotated dataset has been used to get learned by the systems in supervised learning form. The traditional supervised learning algorithms have been used in this paper. Different machine learning classification algorithms have been utilized in this paper namely, Naïve Bayes (NB), Sequential Minimal Optimization (SMO), kNearest Neighbor (k-NN), Decision Tree (DT), Random Forest (RF), . The algorithms are from three different types of classification algorithms namely, Bayesian, function-based, instance-based, Meta classifier and tree-based classifiers.

PERFORMANCE METRICS FOR EACH CLASSIFIERS

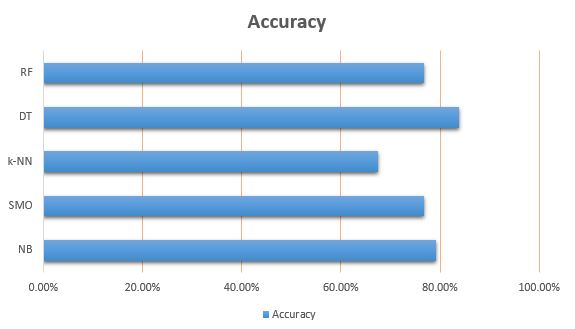
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **TPR** | **FPR** | **PR** | **RE** | **F1** |
| NB | 0.791 | 0.268 | 0.817 | 0.791 | 0.778 |
| SMO | 0.767 | 0.205 | 0.768 | 0.767 | 0.767 |
| RF | 0.767 | 0.279 | 0.789 | 0.767 | 0.754 |
| K-NN | 0.674 | 0.312 | 0.673 | 0.674 | 0.672 |
| DT | 0.837 | 0.145 | 0.845 | 0.837 | 0.831 |

[Description of Algorithms]

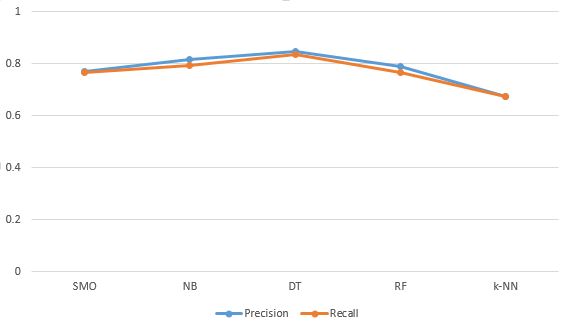
Experimental result

PERFORMANCE METRICS FOR EACH CLASSIFIERS

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **MCC** | **ROC** | **ACC** |
| NB | 0.600 | 0.774 | 79.07% |
| SMO | 0.566 | 0.810 | 76.744% |
| RF | 0.555 | 0.856 | 76.74% |
| K-NN | 0.376 | 0.717 | 67.44% |
| DT | 0.723 | 0.836 | 83.72% |



Precision-recall curve for each classifier



[Description]