12/11/2017

PREDICTIVE ANALYSIS ON UCI CRIME DATASET

PROJECT REPORT - INSY 5339 – PRINCIPLES OF BUSINESS DATA MINING

Group 4
Gayathri Etraj Janaki
Madhu Balan Subramaniyan Dhanabalan
Kowshik Paramesh Venkatesh
Anudeep Kunchala
Sanjana Mutalikdesai

TABLE OF CONTENTS

1 DATASET INTRODUCTION……………………………………………………..2

1.1 DATASET INFORMATION……………………………………………………..2

1.2 DATASET AND ATTRIBUTES…………………………………………………2

1.3 DISTRIBUTION OF THE GOAL VARIABLE…………………………………3

2 DATA PREPARATION……………………………………………………………3

2.1 DATA CLEANING TOOLS……………………………………………………....4

2.2 DATA CLEANING………………………………………………………………..4

2.3 DATA DISCRETIZATION………………………………………………………..5

2.4 DATA SKEW……………………………………………………………………...6

2.5 FALSE PREDICTORS…………………………………………………………….6

2.6 RESOLVING MISSING VALUES………………………………………………..7

3 SELECTION OF ATTRIBUTES……………………………………………………..7

4 CLASSIFIER SELECTION…………………………………………………………...8

5 EXPERIMENTAL DESIGN…………………………………………………………12

5.1 FOUR CELL EXPERIMENTAL DESIGN…………………………………….. .12

5.2 RESULTS FROM EACH CLASSIFIER…………………………………………13

6 ANALYSIS BASED ON THE CLASSIFIER RESULTS…………………………..21

6.1 ROC (RECEIVER OPERATOR CURVE) ………………………………………21

6.2 SINGLE ROC CURVES…………………………………………………………21

6.3 GENERATING MULTIPLE ROC CURVES……………………………………23

7 CONCLUSION………………………………………………………………………..25

8 REFERENCES……………………………………………………………………… 26

1 DATA SET INTRODUCTION:

Our dataset combines the socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

1.1 DATA SET INFORMATION:

Variables which contributed for learning and which can be tested after allocating weights to them were only included in this dataset. Attributes which had plausible connection to crime and also predict the Per Capita Violent Crimes counted to about 122. Variables related to the community such as the percentage of urban population, median family income and percent of family members working in the police department and drug units.

Per capita violent crime variable was calculated using population and sum of crime variables considered violent crimes in United States. Crimes which are considered as violent are murder, rape, robbery and assault. In some states there was some controversy regarding rape thus resulting in missing values, thereby resulting in incorrect per capita violent crime. Mostly these communities were omitted belonged to the midwestern USA.

The data present here are normalized original values ranging between 0.00-1.00 using Unsupervised, equal-interval binning method. Attributes retain their distribution and skew characteristics. Most of the communities are small resulting in attributes having mean values as less as 0.06. The attributes present here such as the “mean people per household” values are normalized versions of their original values.

The normalization preserves rough ratios of values within an attribute, i.e. values which are more than 3 SD above the mean are normalized to 1.00 and all those values having 3SD below the mean are normalized to 0.00.

But the normalization does not preserve relationships between values between attributes as it would not be meaningful. For example, it would not meaningful to compare values of whitePerCap and blackPerCap for a community.

Some of the dataset was contributed by the LEMAS survey. This was done by the team consisting of at-least 100 police officials plus a random sample of smaller departments. But it has mostly missing values
for many communities. Communities that were not found in both census and crime datasets were omitted.

1.2 DATASET AND ATTRIBUTES:

Dataset Characteristic	: Multivariate
Original number of attributes	: 122
Number of instances	: 1994
Goal Variable	: Violent Crimes per Population

1.3 DISTRIBUTION OF THE GOAL VARIABLE:

Range	Frequency
0.000-0.067	484
0.067-0.133	420
0.133-0.200	284
0.200-0.267	177
0.267-0.333	142
0.333-0.400	113
0.400-0.467	59
0.467-0.533	76
0.533-0.600	57
0.600-0.667	38
0.667-0.733	37
0.733-0.800	20
0.800-0.867	23
0.867-0.933	14
0.933-1.000 50	50

2 DATA PREPARATION:

Data Cleaning can be defined as a process in which the amendment and removal of data from a database. The basic idea is to correct or delete incorrect, improperly formatted, incomplete or duplicate data. This process helps maintaining the quality of data for analysis purpose. This is the crucial step in Data Analysis field, as the most part of analysis depends on the data quality.

The general framework of data cleaning:

- Define the error types to determine them
- Explore the data to identify errors
- Correct the errors for improved data quality
- Document error types and identified errors
- Plan on reduction potential future errors.

There is abundance of data everywhere, but not all of it is “fit for use” by the users. To make it fit for use, data cleansing is needed, which will enhance the data quality. Sources of data errors are misspelling due to data entry, outliers in the data, missing or null values, other invalid data etc. Data cleaning is very time-consuming process, but it is very crucial part before model building. Although data cleaning process can be a time consuming and a tedious process, but it is very important that the errors in the data be corrected and that the changes made are traced. To avoid losing information, we should be working on data cleaning on a copy of original data, but not the original data itself.

2.1 DATA CLEANING TOOLS

Microsoft Excel & Weka were used for removing & splitting the attributes and for selecting the appropriate class variable.

Land Area	Pop Dens	Pct UsePub Trans	PolicCars	Polic OperBudg	Lemas PctPolic OnPatr	Lemas GangUnit Deploy	Lemas PctOffic DrugUn	PolicBudg PerPop	Violent Crimes PerPop
0.12	0.26	0.2	0.06	0.04	0.9	0.5	0.32	0.14	0.2
0.02	0.12	0.45	?	?	?	?	0	?	0.67
0.01	0.21	0.02	?	?	?	?	0	?	0.43
0.02	0.39	0.28	?	?	?	?	0	?	0.12
0.04	0.09	0.02	?	?	?	?	0	?	0.03
0.01	0.58	0.1	?	?	?	?	0	?	0.14
0.05	0.08	0.06	?	?	?	?	0	?	0.03
0.01	0.33	0	?	?	?	?	0	?	0.55
0.04	0.17	0.04	?	?	?	?	0	?	0.53
0	0.47	0.11	?	?	?	?	0	?	0.15
0.02	1	1	?	?	?	?	0	?	0.24
0.01	0.63	1	?	?	?	?	0	?	0.08
0.03	0.18	0.59	?	?	?	?	0	?	0.06
0.08	0.04	0	?	?	?	?	0	?	0.09
0.02	0.4	0.15	?	?	?	?	0	?	0.21
0.04	0.15	0.04	?	?	?	?	0	?	0.3
0.06	0.39	0.84	0.06	0.06	0.91	0.5	0.88	0.26	0.49
0.03	0.09	0.21	?	?	?	?	0	?	0.07
0.03	0.2	0.07	?	?	?	?	0	?	0.15

2.2 DATA CLEANING

Dataset before cleaning

This is a glimpse of our initial dataset with enough missing values and Target attribute ViolentCrimePerPop not being normalized.

Data cleaning is one of the vital part in data mining since unwanted data might affect the target attribute prediction to greater extent and the final model would be sloppy. The UCI crime dataset had 127 attributes in total. The LEMAS data in the dataset had 84.5% of the missing values and were removed. The data transformation to these missing values did not help with the accuracy in Violent Crime prediction, hence these attributes were removed.

Sl.No	Reason	Attributes Removed
1	>90% missing values	Community, County
2	85-90% missing values	zzxxLemasGangUnitDeploy, LemasPctOfficDrugUn, LemasPctPolicOnPatr, LemasSwFTFieldOps, LemasSwFTFieldPerPop, LemasSwFTPerPop LemasSwornFT, LemasTotalReq, LemasTotReqPerPop
3	80-85% missing values	NumKindsDrugsSeiz, OfficAssgnDrugUnits, PctPolicAsian PctPolicBlack, PctPolicHisp, PctPolicMinor PctPolicWhite, PolicAveOTWorked, PolicBudgPerPop PolicCars, PolicOperBudg, PolicReqPerOffic

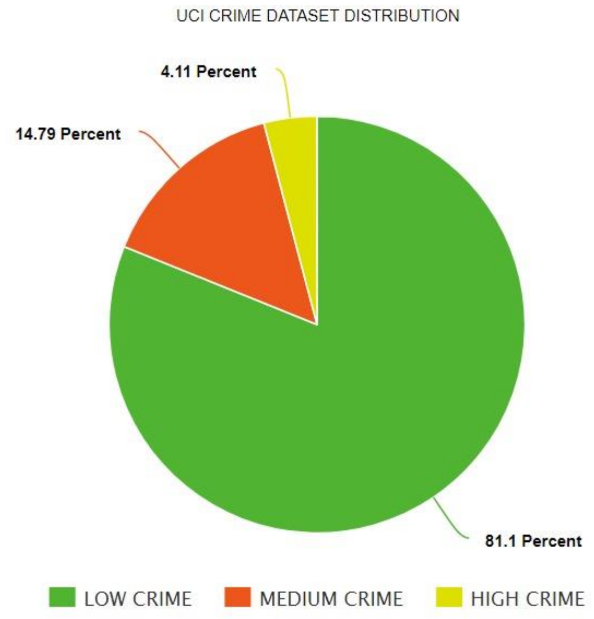
2.3 DATA DISCRETIZATION

Discretization refers to binning or grouping the continuous values of an attribute into a discrete category which in turn limits the number of possible categories for the target variable and aids in efficient model building. The class attribute ViolentCrimePerPop is discretized using WEKA and class attribute was discretized into 3 different bins as follows.

RANGE	COUNT	CLASS CATEGORY
‘(-inf – 0.33333]’	1372	LOW
‘(0.33333 – 0.666667]’	234	MEDIUM
‘(0.66667 – inf)’	69	HIGH

Based on this discretization by WEKA, we have categorized the crime severity as

- LOW crime
- MEDIUM crime
- HIGH crime



2.4 DATA SKEW

Upon discretization, the target variable ViolentCrimePerPop is categorized to Low, Medium and High. The Crime dataset is 81.91 percent skewed to Low Crime since around 80 percent of the population is susceptible to this category. Due to this imbalance in the distribution of class attribute we tried using the SMOTE technique to overcome the data skewness. But even after oversampling using SMOTE, we only got a marginal increase in the size of MEDIUM and High-class category. This did not improve the accuracy of the dataset either. This concludes that even though the data is skewed, it is the actual representation of our class attribute.

2.5 FALSE PREDICTORS

False predictors are certain attribute characteristics in the dataset which seem to give tremendous contribution to the target variable prediction by increasing the accuracy. In contrast, these predictions are not viable indicators of a successful model.

Removing the False contributors:

1. When using OneR to which results the best attribute that fits the model in the target prediction, we found that "AsianPerCapita" gave a good prediction with 92 percent Accuracy.
2. This is not the case with other classifiers - Naive Bayes and J48 which seemed to produce better clarity on the accuracy which was around 75 to 80 percent.
3. With the appropriate Domain knowledge in this area we suspected this attribute to be false predictors and our training set results showed that "AsianPeCapita" is turn producing the false results.

2.6 RESOLVING MISSING VALUES

We removed the LEMAS Data since more than 80% missing values.

3. SELECTION OF ATTRIBUTES:

Further analysis was done using the "Select Attributes" in Weka. We choose the given Attribute Evaluator,

Search Method combinations:

Using the following attribute evaluators with particular search method combinations we were able to rank order the attributes and select the best attributes which have a close association in predicting the target variable "ViolentCrimePerPop".

Attribute Evaluator	Search Method
BestFirst	CfsSubsetEval
Ranker	ChiSquaredAttribute
GreedyStepwise	ClassifierSubset
BestFirst	ClassifierSubset
GeneticSearch	ClassifierSubset
GeneticSearch	CostSensitivity
Ranker	FilteredAttributeEval
GreedyStepwise	FilteredSubset
Ranker	GainRatioAttribute
Ranker	ONeRAttributeEval
Ranker	PrincipalComponent
Ranker	OneRAttributeEval

4 CLASSIFIER SELECTIONS:

We have predominantly used the following classifiers to predict the target variable

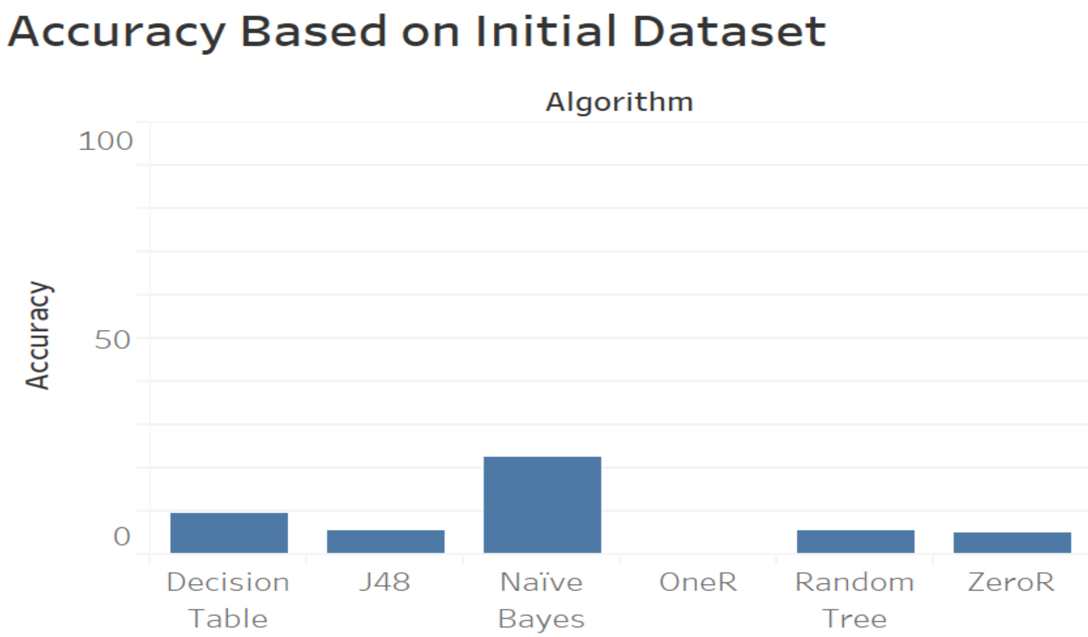
- OneR
- NaiveBayes
- J48
- ZeroR
- Decision Table
- Random Tree

These classifiers tend to have close association to the structure of our dataset and the models built from these classifiers tends to show better results when compared to the rest of the classifiers in WEKA

Based on the initial 127 attributes we were able to achieve a maximum accuracy of 22% with Naive Bayes without the target attribute being transformed into a Nominal attribute.

Data testing & analyzing: Accuracy analysis with 127 Attributes

ALGORITHM	ACCURACY (% SPLIT)	ACCURACY (TRAINING SET)
OneR	0.0015	0.0022
NaiveBayes	0.0561	0.2272
Decision Table	0.0782	0.0933
ZeroR	0.0575	0.0522
J48	0.0523	0.056
Random Tree	0.0546	0.0575

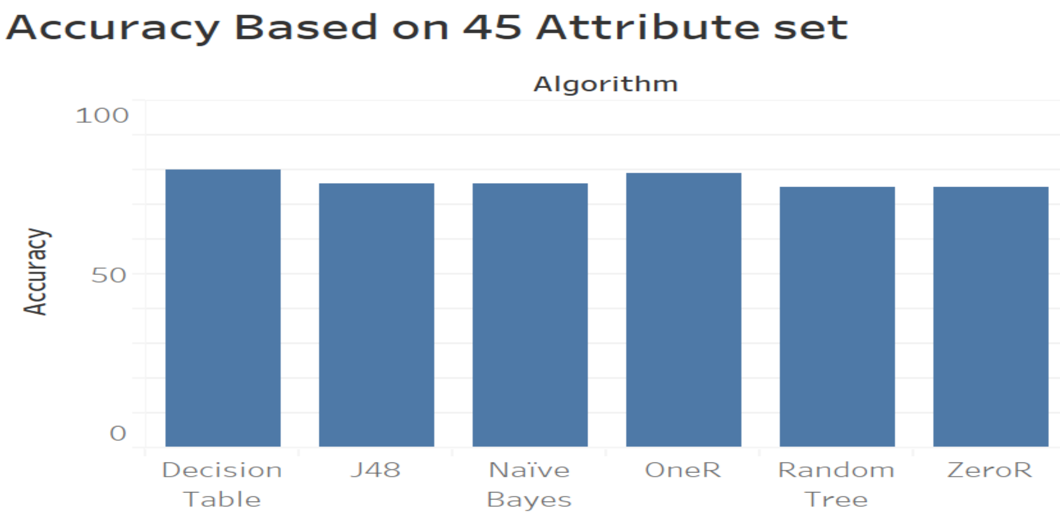


Going forward, we have reduced the no of attributes to 45 by applying WEKA's attribute selection with appropriate combination of attribute and search method evaluators as described in the table above.

Data testing & analyzing: Accuracy analysis with 45 Attributes

ALGORITHM	ACCURACY (% SPLIT)	ACCURACY (TRAINING SET)
OneR	0.7876	0.7885
NaiveBayes	0.7582	0.7844
Decision Table	0.7979	0.8295
ZeroR	0.7478	0.7558
J48	0.7581	0.7869
Random Tree	0.7492	1

These accuracy predictions were finalized by running the dataset 10 times in each desired classifier and computing the average on the runs. Since the Classifier models behave tends to show better clarity on the prediction only when running multiple times, it was appropriate to follow this to train an efficient Model.

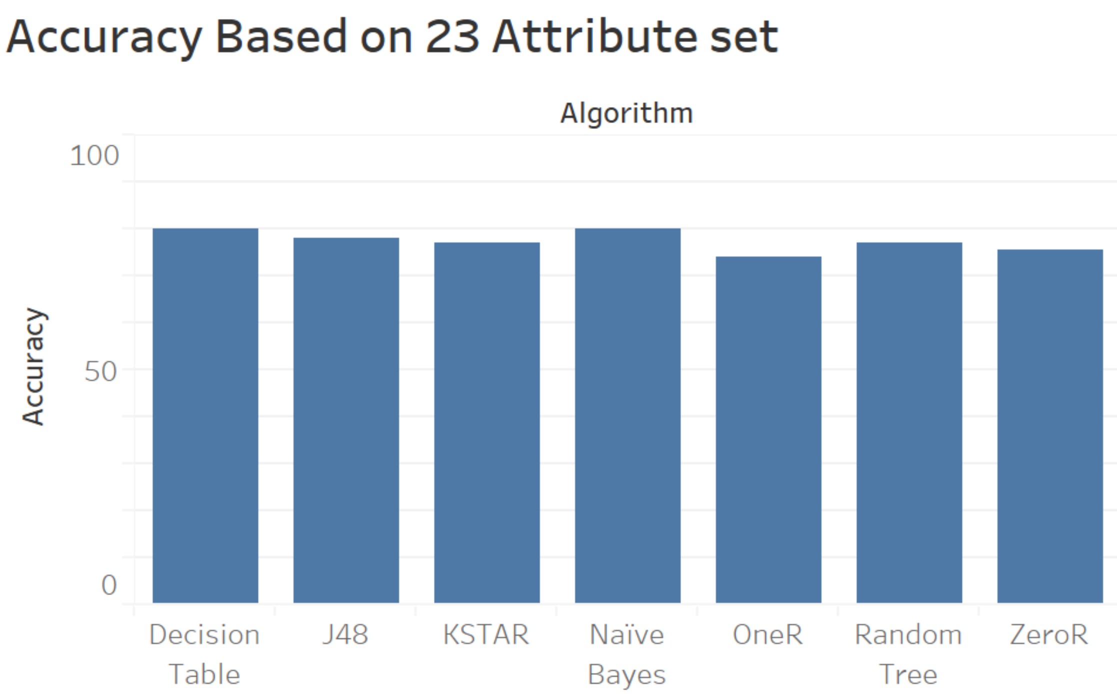


Based on Rank ordering the attributes from the combination of different search evaluations, the following are the final set of 23 attributes which were efficient enough to predict the class attribute:

State	PctKids2Par
CommunityNameString	PctYoungKids2Par
RacePctBlack	PctTeen2Par
RacePctWhite	NumIlleg
NumbUrban	PctIlleg
PctWInvInc	PctLargHouseFam
PctWPubAsst	NumInShelters
BlackPerCap	NumStreet
NumUnderPov	PopDens
PctUnemployed	Crime Severity
PctFam2Par	MalePctNevMarr
	FemalePctDiv

Data testing & analyzing: Accuracy analysis with 23 Attributes

ALGORITHM	ACCURACY (% SPLIT)
OneR	0.7423
NaiveBayes	0.7979
Decision Table	0.794
ZeroR	0.7568
J48	0.778
Random Tree	0.7699
KSTAR	0.7714



We were able to come up with a final set of 23 attributes which gave us better prediction results by 2 percent when compared to the previous Dataset which had 45 set attributes. The following three attributes were added out of Domain knowledge, which significantly improved our target prediction.

- PctUnemployed - Percentage of Population Unemployed
- NumbUrban - number of people living in areas classified as urban
- PctWPubAsst - percentage of households with public assistance income

ATTRIBUTE REDUCTION:

- The main purpose of attribute reduction is to build an efficient model that performs well in real world situation i.e. Performance of the model in the new records which the training set had never trained for.
- Reduction of attributes from 45 to 23 helps the model in predicting with better accuracy and clarity
- Complexity of the model reduced, so that prediction becomes much faster in real world.
- Simpler model always tends to yield better results. This is achieved by not Overfitting the model to predict exact training scenarios and approaching with a broader perspective.

CLASSIFIER SELECTION:

- Based on the final attribute set, we have concluded the following three algorithms to be accurate enough to predict the CLASS attribute.
    - NAIVEBAYES
    - DECISION TREE
    - J48
- We computed the Receiver Operating characteristic (ROC) graph plotted against TruePositive vs FalsePositive for the 7 Algorithms and found that the above mentioned three algorithms had enough accuracy and also better Area under Curve.
- This confirmed our findings that NaiveBayes , J48 and Decision Tree Models are accurate enough to predict the CRIME CATEGORY

5 EXPERIMENTAL DESIGN

5.1 FOUR CELL EXPERIMENTAL DESIGN

- **Two Factor Design:**
  Our project's experimental design has 2 factors**:**
    1. Attributes with or without noise (Factor-1)
    2. Percentage Split (Factor-2)
- **Four Criteria of the Design:**
  These factors are further divided into 4 criteria with one factor varying and keeping other constant, and vice versa. This is illustrated more clearly in the table below.

We selected the following classifiers for our Experimental Design:
a. Naïve Bayes
b. Decision Table
c. J48

	% Split – 66%	**% Split – 80%**
Without Noise	**C1**	**C2**
With 10% Noise	**C3**	**C4**

Four Cell Experimental Design:

- It consists of 4 Conditions:
  - C1: Percentage Split 66% without Noise.
  - C2: Percentage Split 80% without Noise.
  - C3: Percentage Split 66% with 10% Noise.
  - C4: Percentage Split 80% with 10% Noise.

Total number of experiment runs = Number of criteria * Number of Classifiers * 10
= 4 * 3 * 10 = 120 runs

5.2 RESULTS FOR EACH CLASSIFIER

The table below describes the 12 possible combinations of our 4 criteria with the 3 selected classifiers. We ran each of these combinations 10 times and averaged their accuracy and variance:

E1= Performance of Naïve Bayes when, Attributes without noise + Percentage Split of 66%:34%
E2= Performance of Naïve Bayes when, Attributes without noise + Percentage Split of 80%:20%
E3= Performance of Naïve Bayes when, Attributes with noise + Percentage Split of 66%:34%
E4= Performance of Naïve Bayes when, Attributes with noise + Percentage Split of 80%:20%

E1= Performance of J48 when, Attributes without noise + Percentage Split of 66%:34%
E2= Performance of J48 when, Attributes without noise + Percentage Split of 80%:20%
E3= Performance of J48 when, Attributes with noise + Percentage Split of 66%:34%
E4= Performance of J48 when, Attributes with noise + Percentage Split of 80%:20%

E1= Performance of Decision Table when, Attributes without noise + Percentage Split of 66%:34%
E2= Performance of Decision Table when, Attributes without noise + Percentage Split of 80%:20%
E3= Performance of Decision Table when, Attributes with noise + Percentage Split of 66%:34%
E4= Performance of Decision Table when, Attributes with noise + Percentage Split of 80%:20%

Above Algorithms used to run the sets of experiments were Naïve Bayes, J48 and Decision Table.

1. **Naïve Bayes:** In Naïve Bayes, we ran four experiments, E1 to E4. They were as follows:

E1 – Without Noise with 80-20 split.
E2 – Without Noise with 66-34 split.
E3 – With Noise with 80-20 split.
E4 – With Noise with 66-34 split.

E1 - Without Noise with 80-20 split.

SEED	CLASSIFIER	PERCENTAGE SPLIT	ACCURACY
1	NavieBayes	80	76.19
2	NavieBayes	80	80.7
3	NavieBayes	80	80.45
4	NavieBayes	80	79.94
5	NavieBayes	80	81.45
6	NavieBayes	80	77.94
7	NavieBayes	80	82.706
8	NavieBayes	80	77.694
9	NavieBayes	80	80.2
10	NavieBayes	80	81.7
		AVERAGE	**79.867**
		VARIANCE	**4.113586889**

E2 - Without Noise with 66-34 split

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY
1	NavieBayes	66	78.1711
2	NavieBayes	66	77.1386
3	NavieBayes	66	77.8761
4	NavieBayes	66	81.7109
5	NavieBayes	66	83.0383
6	NavieBayes	66	79.351
7	NavieBayes	66	80.6785
8	NavieBayes	66	78.4661
9	NavieBayes	66	80.6785
10	NavieBayes	66	81.5634
		Average	**79.897**
		VARIANCE	**3.796072303**

E3: With Noise 80-20 Split

SEED	CLASSIFIER	PERCENTAGE SPLIT	ACCURACY WITH 10% NOISE
1	NavieBayes	80	69.423
2	NavieBayes	80	74.93
3	NavieBayes	80	73.93
4	NavieBayes	80	73.18
5	NavieBayes	80	73.43
6	NavieBayes	80	69.42
7	NavieBayes	80	76.44
8	NavieBayes	80	69.42
9	NavieBayes	80	73.68
10	NavieBayes	80	74.682
		Average	**72.8535**
		VARIANCE	**5.81453305**

E4: With Noise 66- 34 Split

SEED	CLASSIFIE R	PERCENTAGE SPLIT	%ACCURACY WITH 10% NOISE
1	NavieBayes	66	70.7965
2	NavieBayes	66	70.649
3	NavieBayes	66	70.5015
4	NavieBayes	66	75.3687
5	NavieBayes	66	75.5162
6	NavieBayes	66	70.944
7	NavieBayes	66	73.0088
8	NavieBayes	66	71.9764
9	NavieBayes	66	73.8938
10	NavieBayes	66	73.0088
		Average	**71.56637**
		VARIANCE	**3.276069414**

2. **Decision Table**: In Decision Table, we ran four experiments, E1 to E4. They were as follows:

E1 – Without Noise with 80-20 split.
E2 – Without Noise with 66-34 split.
E3 – With Noise with 80-20 split.
E4 – With Noise with 66-34 split.

E1 – Without Noise with 80-20 split.

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY
1	DecisionTable	80	79.44
2	DecisionTable	80	81.45
3	DecisionTable	80	78.44
4	DecisionTable	80	80.45
5	DecisionTable	80	78.69
6	DecisionTable	80	81.704
7	DecisionTable	80	81.45
8	DecisionTable	80	80.7
9	DecisionTable	80	79.62
10	DecisionTable	80	80.2
		AVERAGE	**80.2144**
		VARIANCE	1.089922036

E2 – Without Noise with 66-34 split.

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY
1	DecisionTable	66	77.7286
2	DecisionTable	66	78.7611
3	DecisionTable	66	78.0236
4	DecisionTable	66	81.4159
5	DecisionTable	66	82.8909
6	DecisionTable	66	80.0885
7	DecisionTable	66	78.6136
8	DecisionTable	66	80.0885
9	DecisionTable	66	77.2861
10	DecisionTable	66	79.646
		AVERAGE	**79.4542**
		VARIANCE	2.492054477

E3 – With Noise with 80-20 split.

SEED	CLASSIFIER	PERCENTAGE SPLIT	ACCURACY WITH 10% NOISE
1	DecisionTable	80	70.92
2	DecisionTable	80	73.68
3	DecisionTable	80	72.68
4	DecisionTable	80	71.17
5	DecisionTable	80	74.18
6	DecisionTable	80	69.67
7	DecisionTable	80	76.69
8	DecisionTable	80	73.684
9	DecisionTable	80	76.44
10	DecisionTable	80	72.68
		AVERAGE	**73.179**
		VARIANCE	4.219955686

E4 – With Noise with 66-34 split

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY WITH 10% NOISE
1	DecisionTable	66	70.2065
2	DecisionTable	66	72.5664
3	DecisionTable	66	70.649
4	DecisionTable	66	74.9263
5	DecisionTable	66	75.5162
6	DecisionTable	66	72.5664
7	DecisionTable	66	71.5339
8	DecisionTable	66	73.5988
9	DecisionTable	66	76.2537
10	DecisionTable	66	68.8791
		AVERAGE	**72.66963**
		VARIANCE	4.793973007

3.J48: In J48, we ran four experiments, E1 to E4. They were as follows:

E1 – Without Noise with 80-20 split.
E2 – Without Noise with 66-34 split.
E3 – With Noise with 80-20 split.
E4 – With Noise with 66-34 split.

E1 – Without Noise with 80-20 split.

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY
1	J48	80	79.44
2	J48	80	81.45
3	J48	80	78.44
4	J48	80	80.45
5	J48	80	78.69
6	J48	80	81.704
7	J48	80	81.45
8	J48	80	80.7
9	J48	80	79.62
10	J48	80	80.2
		AVERAGE	**78.081**
		VARIANCE	1.2123

E2 – Without Noise with 66-34 split.

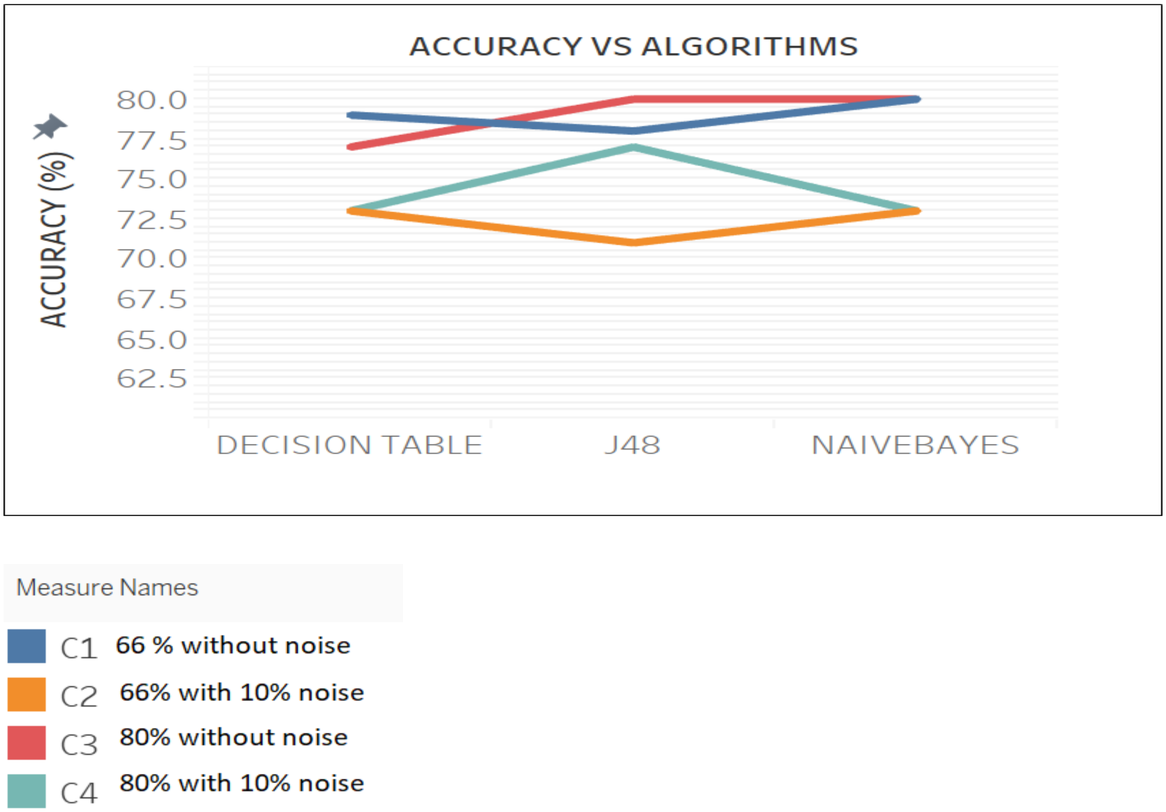
SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY
1	J48	66	75.2065
2	J48	66	78.5664
3	J48	66	76.649
4	J48	66	78.9263
5	J48	66	77.5162
6	J48	66	77.5664
7	J48	66	75.5339
8	J48	66	74.5988
9	J48	66	79.2537
10	J48	66	76.8791
		AVERAGE	**77.3954**
		VARIANCE	1.2312763

E3 – With Noise with 80-20 split.

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY WITH 10% NOISE
1	J48	80	70.92
2	J48	80	68.68
3	J48	80	72.68
4	J48	80	71.17
5	J48	80	70.18
6	J48	80	69.67
7	J48	80	70.69
8	J48	80	68.684
9	J48	80	71.44
10	J48	80	72.68
		AVERAGE	**70.685**
		VARIANCE	6.71443486

E4 – With Noise with 66-34 split

SEED	CLASSIFIER	PERCENTAGE SPLIT	%ACCURACY WITH 10% NOISE
1	J48	66	70.2065
2	J48	66	72.5664
3	J48	66	70.649
4	J48	66	68.9263
5	J48	66	70.5162
6	J48	66	69.5664
7	J48	66	71.5339
8	J48	66	70.5988
9	J48	66	72.2537
10	J48	66	68.8791
		AVERAGE	**70.383**
		VARIANCE	5.7321



Accuracy

Average Summary	E1	E2	E3	E4
Decision Table	80.2144	79.4542	73.179	72.66963
Naive Bayes	79.867	79.897	72.566	71.566
J48	78.081	77.394	70.685	70.383

Based on this we can conclude that both Decision Table and Naive performed really well in E1 and E2 without noise and approximately there is an 8 percent decrease in the accuracy when 10 percent noise is introduced. J48 performed well having close to 78 percent in Accuracy.

Variance

Average Summary	E1	E2	E3	E4
Decision Table	1.0899	2.492	4.2199	4.7139
Naive Bayes	4.1135	3.7960	5.8145	3.2760
J48	1.2123	1.2312	6.7144	5.7321

Here both Decision Table and J48 had really low Variance without noise and performed well in predicting the target variable when compared to Naive Bayes. As expected there is increase in variance in all the three cases when 10 percent noise is introduced.

6 ANALYSES BASED ON THE CLASSIFIER RESULTS:

We used ROC technique for analysis and interpreting results.

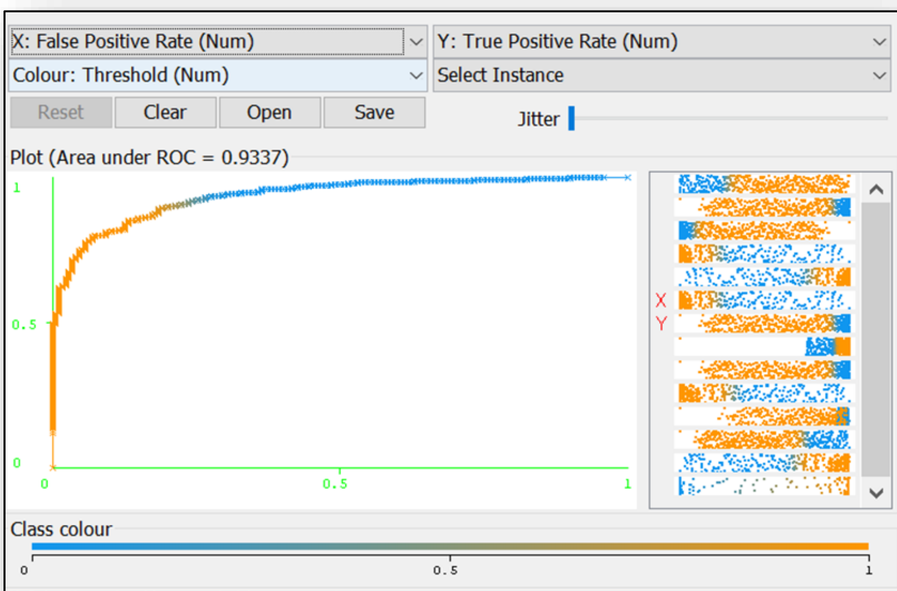
6.1 ROC (RECEIVER OPERATOR CURVE):

A Receiver Operator Characteristics (ROC) Curve is a graphical representation that estimates the performance of a binary classifier as its threshold is varied. At various threshold settings, The ROC curve is created by mapping the true positive rate (TPR) against the false positive rate (FPR).

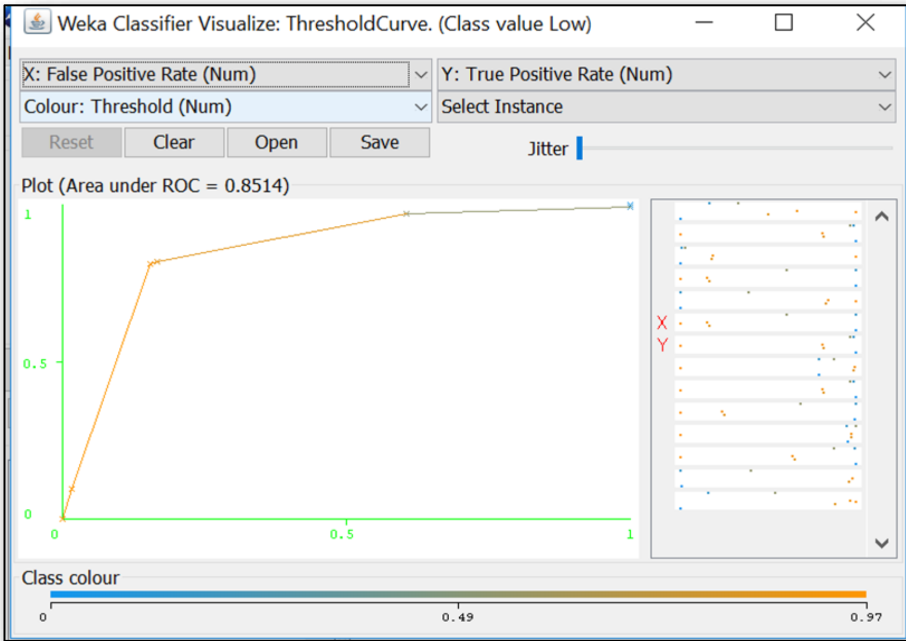
Our accuracy is found as a result with which our classifiers can predict the true positives and true negatives. This method of determining the classifier and factor overall efficiency is by 'how much area is covered under the ROC curve. Higher the area, better the model. So, If the Area under the ROC Curve is large, the model that has been built is better and efficient enough to predict the CLASS attribute.

6.2 SINGLE ROC CURVES:

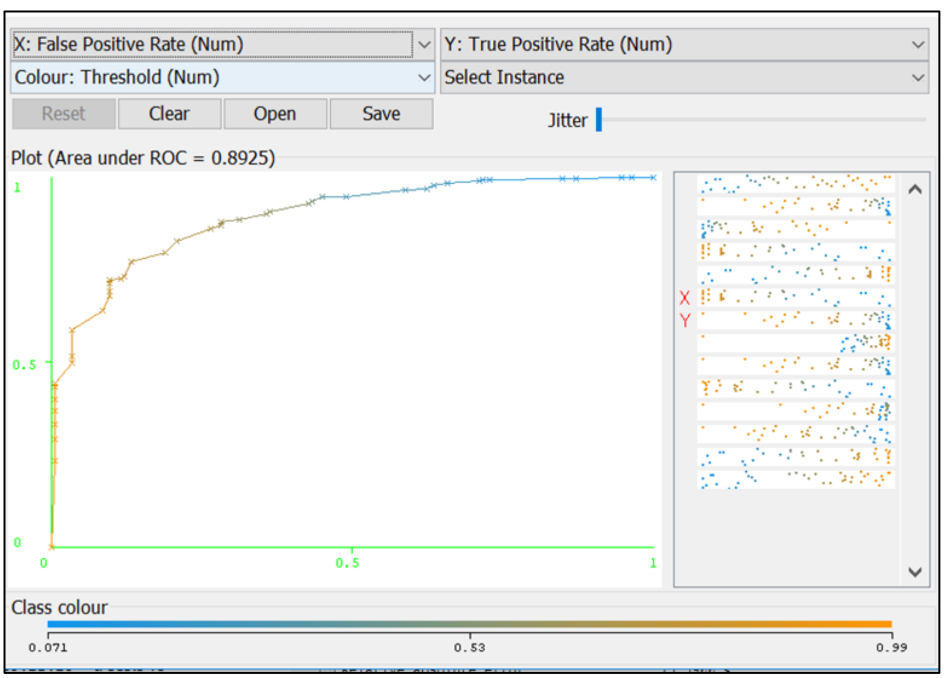
DECISION TABLE – LOW VS OTHER CLASSES



J48 – LOW VS OTHER CLASSES

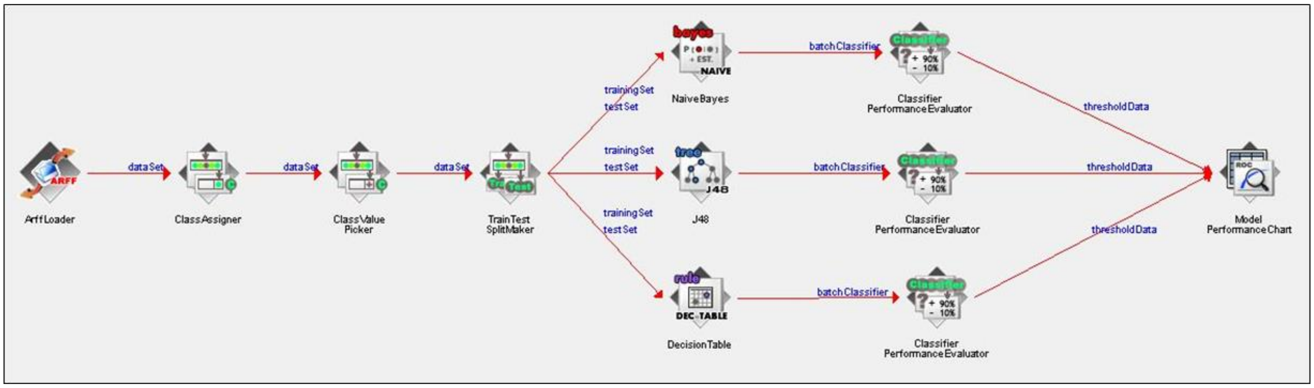


NAIVE BAYES – LOW VS OTHER CLASSES

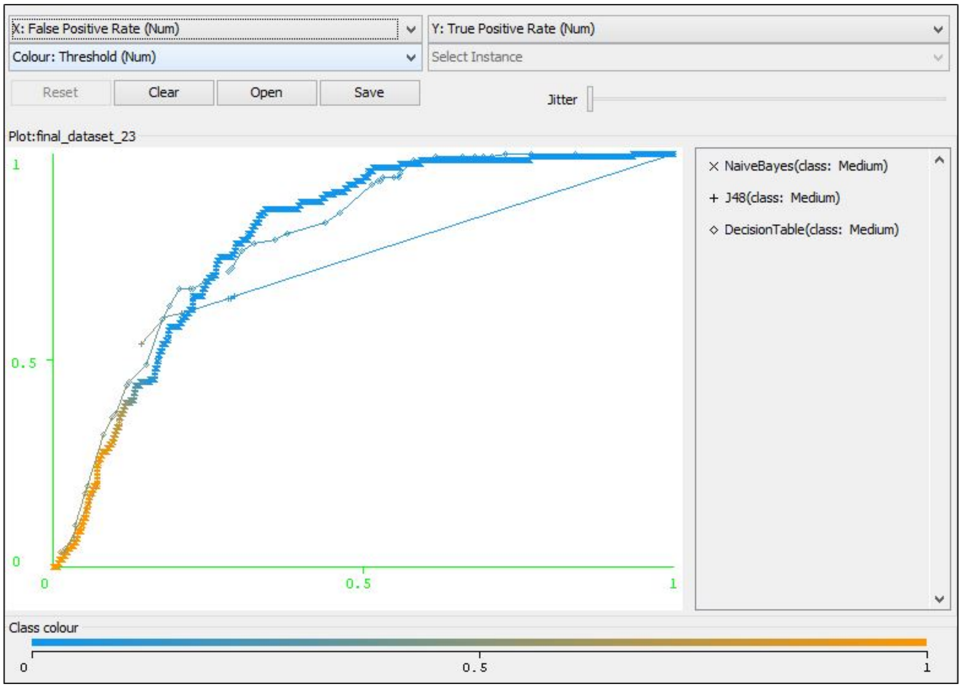


6.3 GENERATING MULTIPLE ROC CURVES:

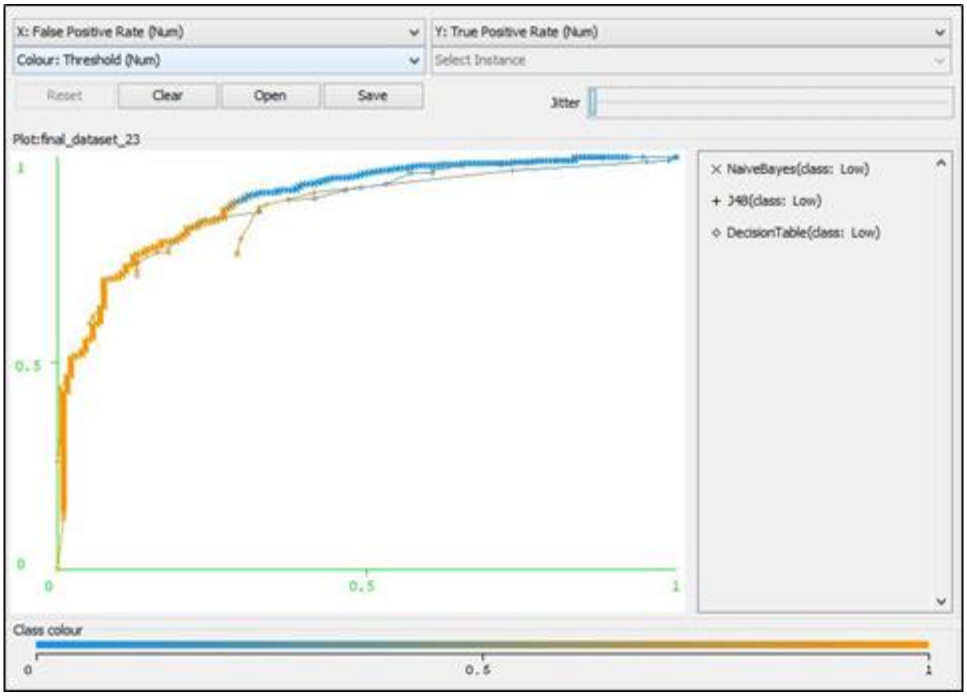
A multiple ROC Curve model is designed using the "Knowledge Flow" feature in Weka to illustrate Multiple ROC curves for one factor comparison with others. The knowledge flow layout of the same is represented below:



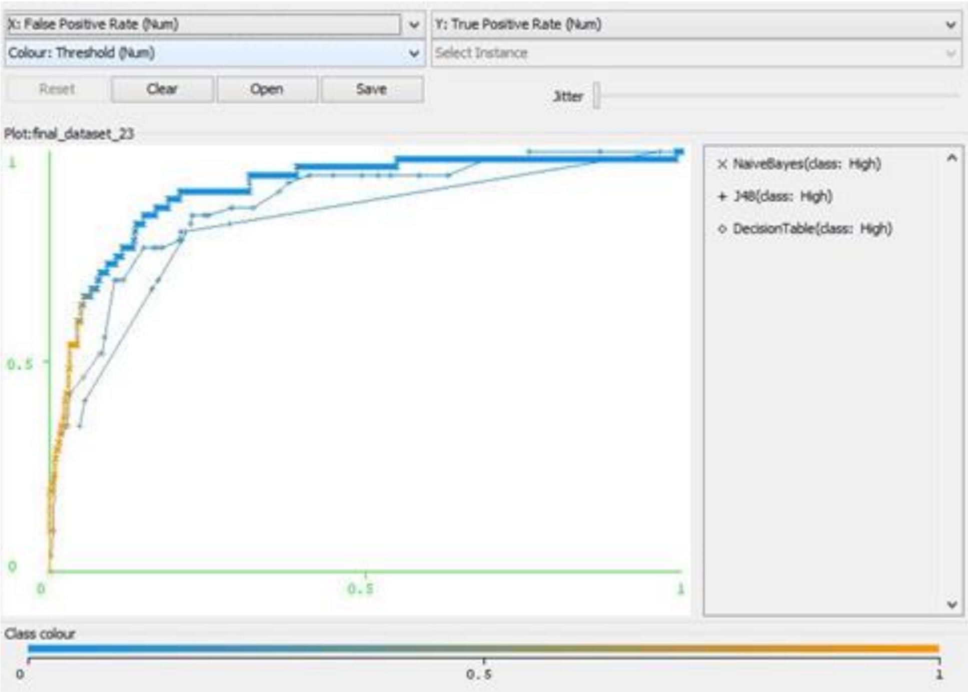
MEDIUM CLASS VS 3 CLASSIFIERS



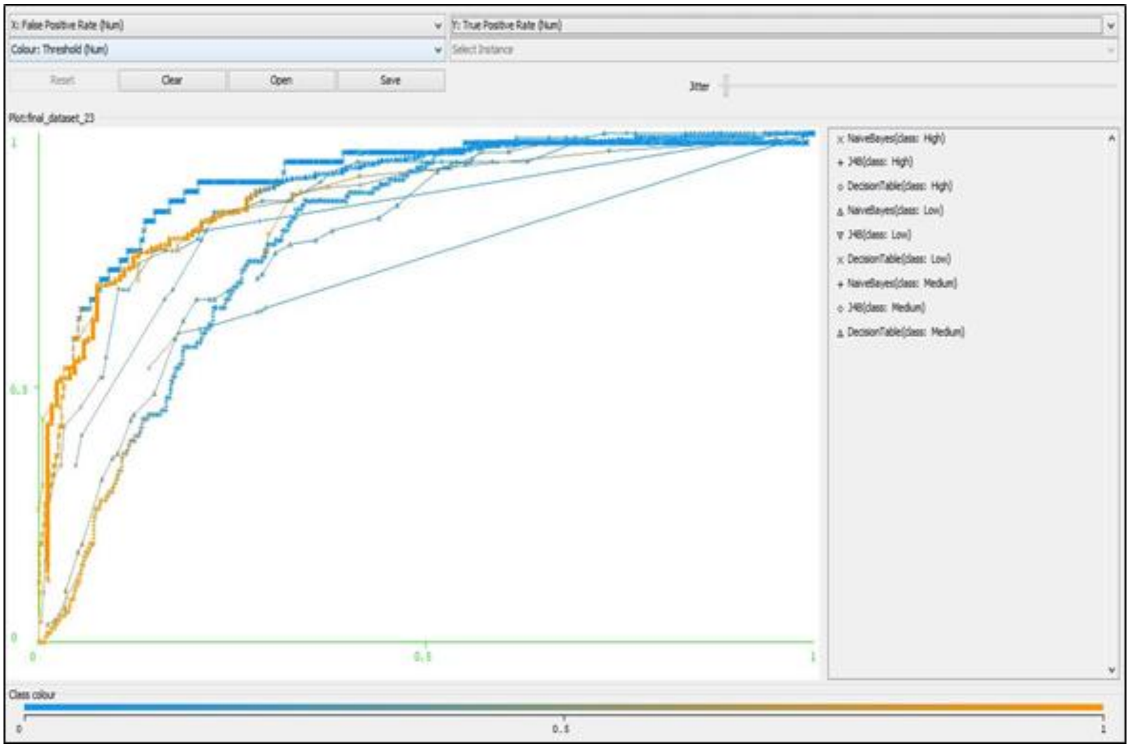
LOW CLASS VS 3 CLASSIFIERS



HIGH CLASS VS 3 CLASSIFIERS



3 CLASSES VS 3 CLASSIFIERS



7. CONCLUSION:

Area under ROC:

- We can say from the ROC curves that the curve plotted with the NaïveBayes algorithm tends to be more efficient since it has a large area under the curve compared to the curves of J48 and Decision Table algorithms.

Accuracy:

- Based on the accuracy analysis we were able to get the best possible accuracy in Naïve Bayes – 79.79%.

Experimental Design:

The results from the Experimental Design showed us that when the 10% Noise is introduced to our dataset, there was an approximate dip of 7-8% in the accuracy.

Overall:

Overall decision table performs better in terms of area under the curve but Naive Bayes has marginally higher accuracy when compared Decision Table. But the difference between these two classifiers in both the methods were very minimal. Hence, we can conclude that both Decision Table and Naive Bayes predicts fairly better than other algorithms when all factors are weighed in.

8. REFERENCES:

- A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection - by Masita
  http://www.naun.org/main/NAUN/computers/2014/a022007-096.pdf

- A Study on Classification Algorithms for Crime Records – by K. B. Sundhara Kumar , N. Bhalaji
  https://link.springer.com/chapter/10.1007/978-981-10-3433-6_104