INSY 5339 – PRINCIPLES OF BUSINESS DATA MINING

Project Report
Predicting the Readmission of Diabetic Patients

Group 11

Anish Grandhi

Dhineshvikram Krishnamurthy

Gokul R Pandiaraj

Harish Kumar Vasikaran

Pranay K Bairneni

1 Contents

2 Data Background

1.1 What is the dataset about?

The dataset contains patient information data from 130 hospitals in the United States (1998 – 2008). Details of patient's hospital encounters, medical information such as diagnostics that the patients have received, insurance company paying for the fees, patient being readmitted etc. Patients often get readmitted to the hospitals for further treatments.

1.2 Details of the dataset

- The number of instances in the dataset – 101766
- Total number of attributes – 55
- The attribute names along with the type, descriptions, values, and percentage of missing values are as follows –

Attribute Name	Type	Description and values	% values missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, new-born, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%

Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.	0%

Note – The above information regarding the details of the attributes were taken from https://www.hindawi.com/journals/bmri/2014/781670/

1.3 Criteria of Attribute

The attributes were recorded based on 5 major criteria –

i. It is an inpatient encounter (a hospital admission).
ii. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
iii. Length of stay – Minimum of 1 Day and Maximum of 14 Days.
iv. Laboratory tests that were performed.
v. Medications that were administered during the time patient was in the hospital.
vi. The data contains other attributes such as patient number, age, admission type, time in hospital, diagnosis, number of medications the patient took, medical specialty of the doctor, number of lab test performed, inpatient, and emergency visits in the year before the hospitalization, race, gender, etc.

1.4 Class attribute

Our class attribute is the attribute "Readmitted". It initially contained 3 values – No, <30, >30. 'No' indicates that the patient was not readmitted. '<30', '>30' tells us whether the patient was admitted less than or greater than 30 days of the last hospital encounter.

3 Data Cleaning

Human beings are prone to cause errors. To reduce errors in a dataset, data cleaning is an important process. Data Cleaning is the process of finding and taking away or modifying inaccurate or irrelevant data from the dataset. Data Cleaning will make the data consistent, will increase the accuracy and performance of prediction.

3.1 Data cleaning tools

The tools and software we have used for cleaning our data are –

a) Microsoft Excel
b) Visual Basic scripts
c) Weka

Microsoft Excel was used to remove same patient's records after they were recorded in the dataset the first time. This would help the classifier to not treat the same patient who have the same set of attributes as unique instances.

Visual Basic scripts were created and run to help eliminate

Weka software showed the histogram of the values of every attribute. Attribute having more than 98% of the same value were removed. Usually the attribute would take only one value differently and the rest same. Such attributes which have the same value 98% of the time contribute less to the model.

3.2 Reducing redundancy

In the original data set, there are two attributes Encounter ID which is the unique ID given when the patient got admitted and Patient ID which is the unique ID for individual patient throughout their healthcare cycle. But each Patient ID corresponded to multiple Encounter ID. To reduce redundancy and nullify the dependency of one patient ID over the other, the instance where patients have multiple Encounter ID has been removed by keeping only the first encounter.

ORIGINAL DATA	
ENCOUNTER ID	PATIENT ID
E ID 1	PID 1
E ID 2	PID 1
E ID 3	PID 1
E ID 4	PID 2
E ID 5	PID 2

PROCESSED DATA	
ENCOUNTER ID	PATIENT ID
E ID 1	PID 1
E ID 4	PID 2

Figure 1

3.3 Irrelevant attributes

Several attributes that were recorded in the original dataset have been removed because they proved to be irrelevant. The irrelevant attributes are –

Encounter and Patient ID – These identification codes were used to uniquely identify the patients and their encounters to the hospitals. Unique ID values are to be removed because they do not provide any information to the classifier. They could mislead the classifier in some cases.

3.4 Conversions and transformations

Few attributes have been converted from numeric to nominal. The number of distinct values that each attribute took are given below –

- Admission Type ID – 8 distinct values
- Discharge Disposition ID - 26 distinct values
- Admission Source ID - 17 distinct values
- Time in hospital - 14 distinct values

The class attribute was reduced to 2 values from 3 values. The original attribute took the values "No", "<30", ">30". After transformation, the class attribute "Readmitted" took the values "No" and "Yes".

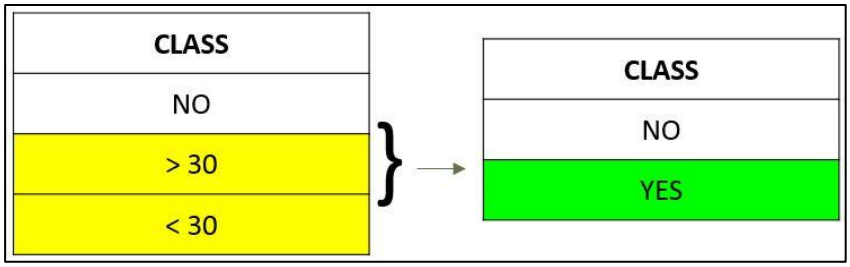

Figure 2

3.5 Missing values

Different classifiers differently treat missing values. To make sure that all the classifiers are given the same data as input, the attributes with a high percentage of missing values have been removed. The attributes with the missing values percentage are as follows –

S. No	Attribute	Percentage missing values
1.	Weight	98%
2.	Medical Speciality	53%
3.	Payer Code	52%

Table 1

Apart from that, for the attribute Discharge Disposition ID which is the ID corresponds to how the patients got discharge, those instances which corresponded to Death or Hospice has been remove as they might be considered as error terms if we included it in the analysis.

3.6 Skewed Data

Most of the attributes were skewed. The attributes usually took a single value 90% of the time and the other value around 10%. This was one of the primary reasons that different classifiers could not show an improvement in the accuracy. Figure 3. shows the case where in the case of Examide all the instances belong to a single category and in the Acetohexamide only one instance belongs to the other category. The skewness is found in 10 attributes in total which are **acetohexamide, tolbutamide, troglitaeone, tolazamide, examide, ciroglipton, glipzide, glimepiride, metformine, pioglitozone**

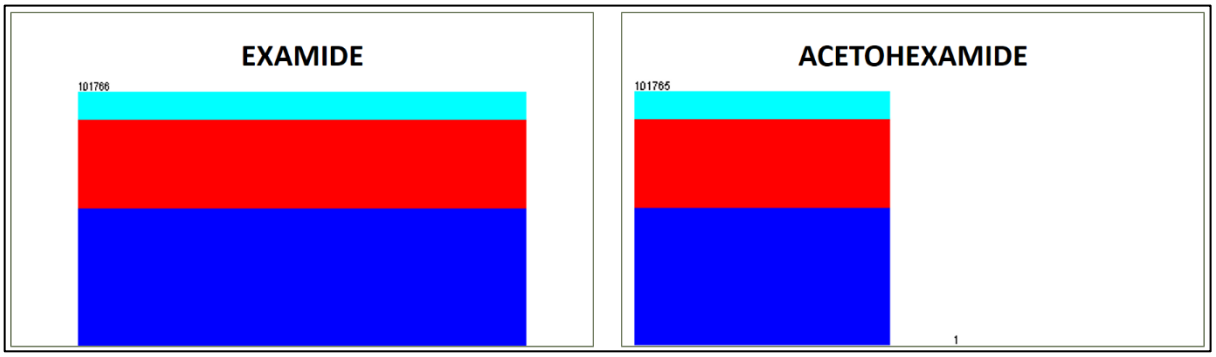

Figure 3

Initially the Diabetic data set had 101766 instances with 55 attributes. After processing the data, they have been reduced to 69673 instances with 35 attributes.

3.7 Results of Data cleaning

Before Cleaning

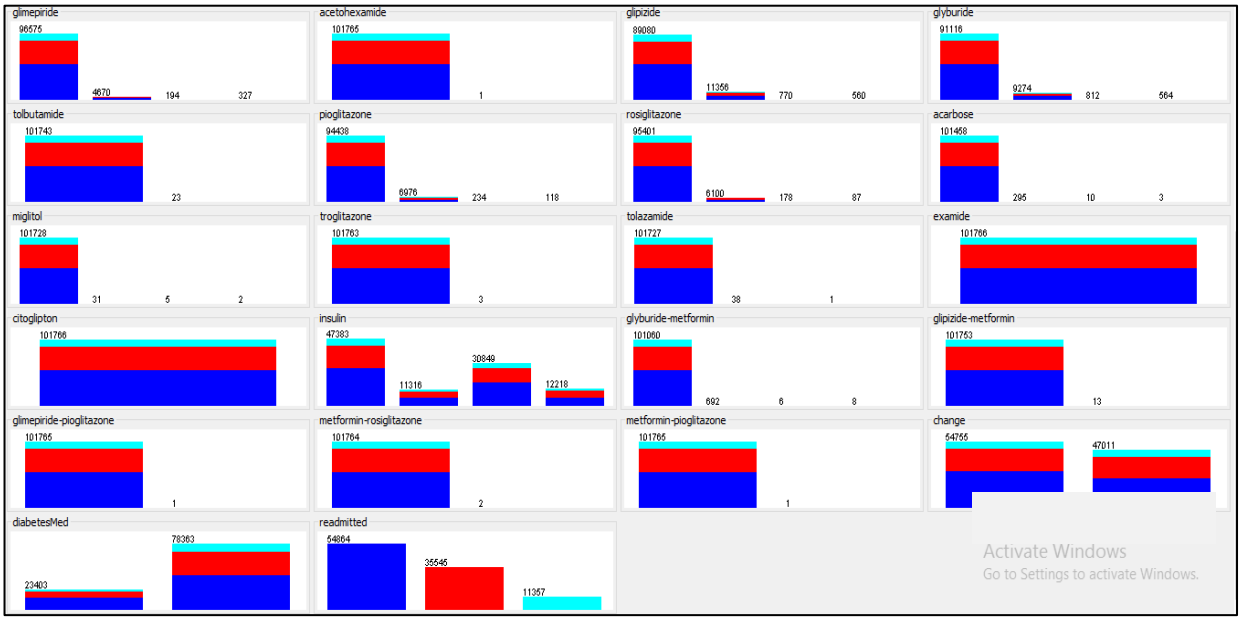

Figure 4.1 Before Cleaning Data

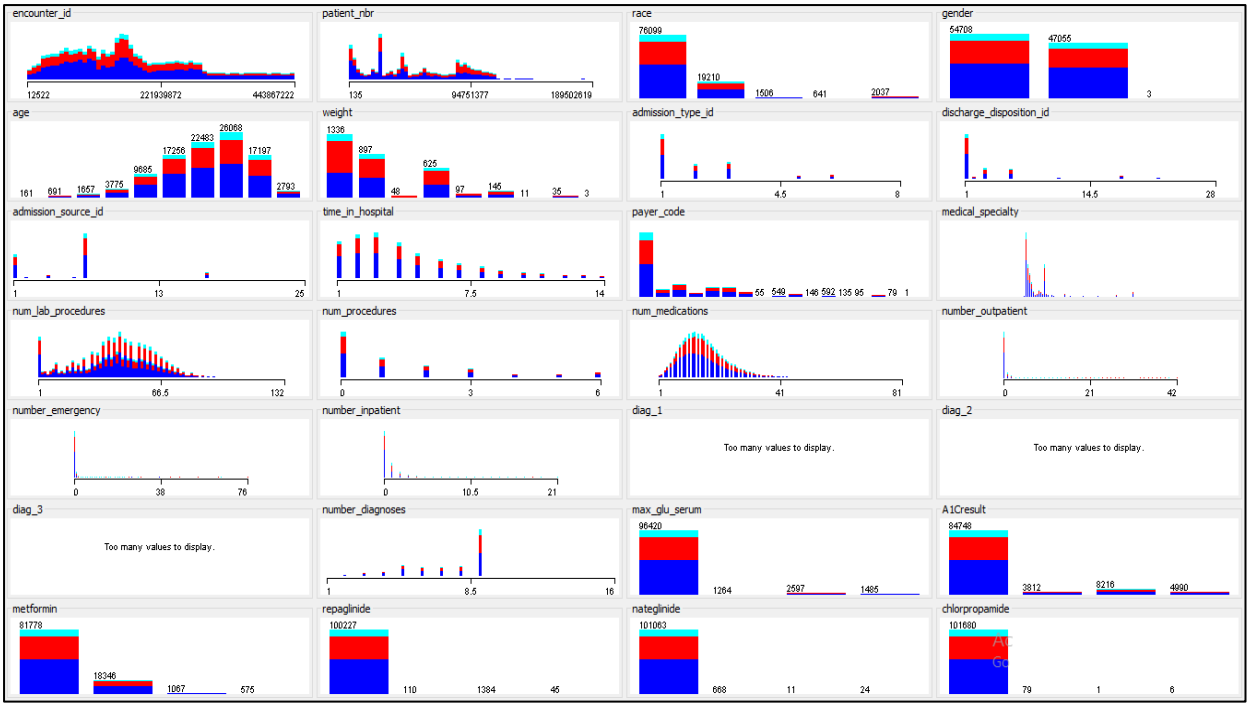

Figure 4.2 Before Cleaning data

After Cleaning Data

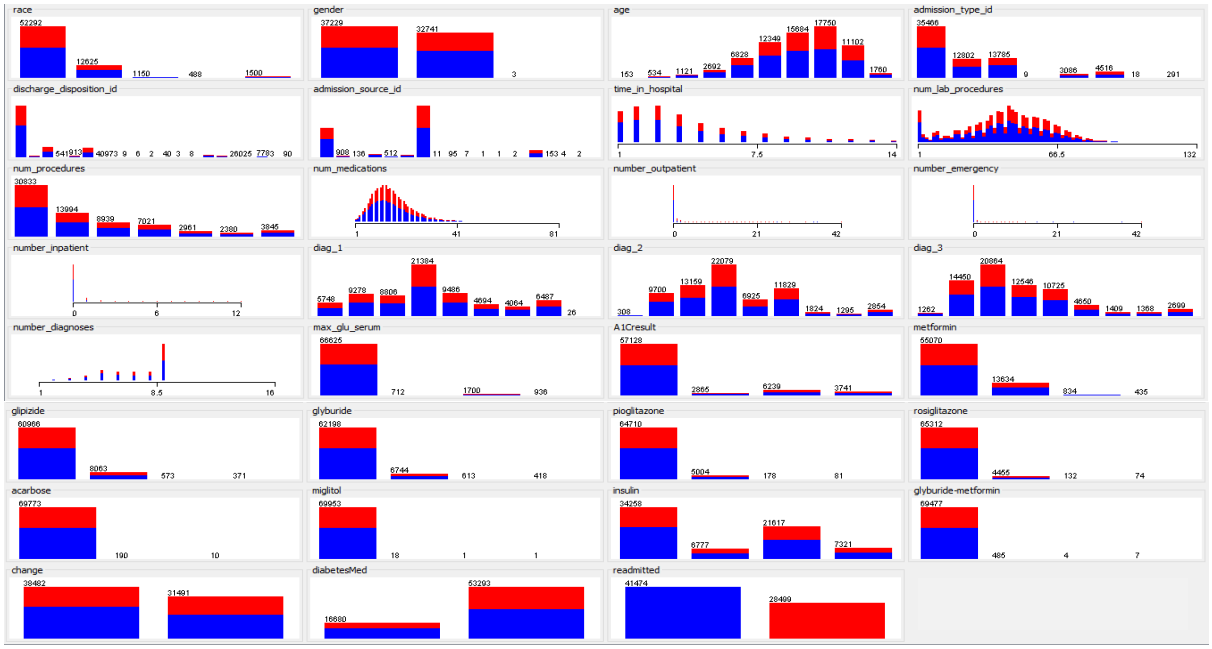

Figure 5 After Cleaning Data

4 Experiment Design

4.1 Classifier Selection

Selection of classifiers was made by keeping in mind, the two important ideas

a. High prediction accuracy

b. High stability of the model with reasonable prediction accuracy

We selected the following classifiers: -

a) **J48** (tree) - Decision tree builds classification or regression models in the form of a tree structure. It breaks down the dataset into smaller subsets and associated decision tree is incrementally developed. J48 classifier can build high accuracy models.
b) **IBK** (lazy algorithm) - lazy algorithm memorizes the training dataset and don't learn any discriminative function from training dataset. The input is classified by taking a majority vote of the k (where k is some user specified constant) closest training records across all d attributes. Each time for prediction, the nearest neighbour (K-NN) in training dataset is searched.
c) **Decision table** – Decision tables are a precise way that model complex rules and their corresponding actions. Decision tables such as if-else, switch statements associate conditions with actions to perform in a better way.

4.2 Four Cell Experiment Design

Our experiment design contained of two factors

- Factor 1 (F1) -> Noise
- Factor 2 (F2) -> Percentage split

The two factors are divided into four criteria by keeping one factor constant and varying the other factor between the two values.

	0% Noise	20% Noise
Percentage split (80%/20%)	C1	C3
Percentage split (20%/80%)	C2	C4

Table 2

a) F11, C1 = 0% Noise with 80/20 percentage split.
b) F12, C2 = 0% Noise with 20/80 percentage split.
c) F21, C3 = 20% Noise with 80/20 percentage split.
d) F22, C4 = 20% Noise with 20/80 percentage split.

Note: To make our training and test data truly representative as the data might lose its properties due to sampling and while running the classifiers, we are doing ten runs for each criterion, each classifier with a distinct value of seed.

The number of criteria is 4.

The number of algorithms are 3.

Therefore, the number of experiment runs are = 4*3*10 = 120 runs

5 Experiment Results

The following describes 4 possible combinations of each algorithm. We have used 3 algorithms which gives us 12 different experiments. For each experiment, the test is performed for 10 times and their accuracy is calculated.

E1= Performance of IBK for Percentage Split of 80%:20% without noise
E2= Performance of IBK for Percentage Split of 20%:80% without noise
E3= Performance of IBK for Percentage Split of 80%:20% with 20% noise
E4= Performance of IBK for Percentage Split of 20%:80% with 20% noise
E5= Performance of J48 for Percentage Split of 80%:20% without noise
E6= Performance of J48 for Percentage Split of 20%:80% without noise
E7= Performance of J48 for Percentage Split of 80%:20% with 20% noise
E8= Performance of J48 for Percentage Split of 20%:80% with 20% noise
E9= Performance of Decision Table for Percentage Split of 80%:20% without noise
E10= Performance of Decision Table for Percentage Split of 20%:80% without noise
E11= Performance of Decision Table for Percentage Split of 80%:20% with 20% noise
E12= Performance of Decision Table for Percentage Split of 20%:80% with 20% noise

Table 3

5.1 Results for each classifier

The results for each Classifier is given below,

Table for E1		
Trail	Seed	Accuracy
1	5	60.2501
2	10	60.8789
3	15	60.6645
4	20	61.0861
5	25	61.0504
6	30	61.0718
7	35	61.0146
8	40	60.5002
9	45	59.8071
10	50	60.2715
Average		60.65952
Standard Deviation		0.440852

Table 4.1

Table for E2		
Trail	Seed	Accuracy
1	5	60.2701
2	10	60.1879
3	15	60.2004
4	20	60.2505
5	25	60.1165
6	30	60.1772
7	35	59.9575
8	40	60.2969
9	45	59.9789
10	50	60.1665
Average		60.16024
Standard Deviation		0.11426

Table 4.2

Table for E3		
Trail	Seed	Accuracy
1	5	57.1633
2	10	56.6702
3	15	56.9203
4	20	57.3062
5	25	57.1347
6	30	57.7278
7	35	57.1561
8	40	57.099
9	45	56.4487
10	50	57.1561
Average		57.07824
Standard Deviation		0.347153

Table 4.3

Table for E4		
Trail	Seed	Accuracy
1	5	56.7259
2	10	56.6133
3	15	56.6633
4	20	56.599
5	25	56.8027
6	30	56.5615
7	35	56.5383
8	40	56.8098
9	45	56.7366
10	50	56.6526
Average		56.6703
Standard Deviation		0.095704

Table 4.4

Table for E5		
Trail	Seed	Accuracy
1	5	61.129
2	10	61.129
3	15	60.8574
4	20	61.129
5	25	61.0718
6	30	61.3934
7	35	61.3862
8	40	60.9718
9	45	60.4359
10	50	60.5216
Average		61.00251
Standard Deviation		0.320959

Table 4.5

Table for E6		
Trail	Seed	Accuracy
1	5	60.6667
2	10	60.3701
3	15	60.5309
4	20	59.8771
5	25	60.2844
6	30	60.0754
7	35	60.9132
8	40	59.9861
9	45	60.2647
10	50	59.9843
Average		60.29529
Standard Deviation		0.332911

Table 4.6

Table for E7		
Trail	Seed	Accuracy
1	5	55.8771
2	10	55.8771
3	15	56.7345
4	20	56.4059
5	25	56.9561
6	30	56.806
7	35	56.1986
8	40	56.5559
9	45	56.1986
10	50	56.4559
Average		56.40657
Standard Deviation		0.371331

Table 4.7

Table for E8		
Trail	Seed	Accuracy
1	5	56.0631
2	10	55.3485
3	15	56.081
4	20	55.7237
5	25	55.7808
6	30	55.6576
7	35	55.3253
8	40	55.3485
9	45	55.6379
10	50	55.9666
Average		55.6933
Standard Deviation		0.288267

Table 4.8

Table for E9		
Trail	Seed	Accuracy
1	5	62.0079
2	10	62.0293
3	15	62.0222
4	20	62.308
5	25	62.1436
6	30	62.2794
7	35	61.9864
8	40	61.9364
9	45	61.7792
10	50	62.0507
Average		62.0543
Standard Deviation		0.15671

Table 4.9

Table for E10		
Trail	Seed	Accuracy
1	5	61.8314
2	10	61.8761
3	15	61.9136
4	20	61.8189
5	25	61.576
6	30	61.7314
7	35	61.7242
8	40	61.8082
9	45	61.3562
10	50	61.4277
Average		61.7064
Standard Deviation		0.19109

Table 4.10

Table for E11		
Trail	Seed	Accuracy
1	5	58.4352
2	10	58.478
3	15	57.7421
4	20	57.8564
5	25	58.871
6	30	59.0425
7	35	58.4637
8	40	58.2136
9	45	58.4066
10	50	58.0779
Average		58.3587
Standard Deviation		0.40768

Table 4.11

Table for E12		
Trail	Seed	Accuracy
1	5	58.4712
2	10	58.1496
3	15	58.0585
4	20	58.2139
5	25	58.1425
6	30	58.1711
7	35	58.2747
8	40	58.2818
9	45	58.0746
10	50	58.0889
Average		58.1927
Standard Deviation		0.12483

Table 4.12

5.2 Summary of Results

The benchmark algorithm used is ZeroR. It's accuracy levels are given below.

	C1	C2	C3	C4
Accuracy	59.26%	59.32%	56.18%	56.54%

Table 5

From the above results, the accuracy of ZeroR algorithm is around 59% without noise and around 56% with 20% noise irrespective of the split. So, our selected classifiers should have a greater accuracy than ZeroR.

Summary of Results- Accuracy:

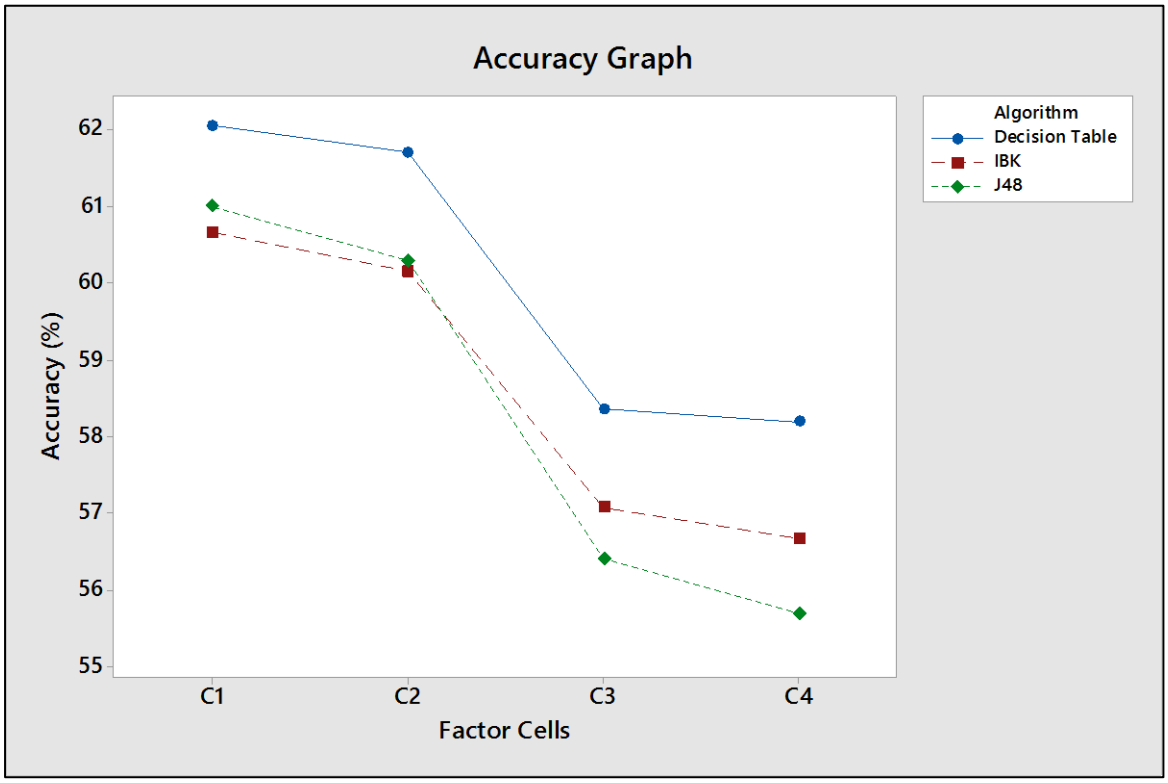

Figure 6

Algorithms	C1	C2	C3	C4
Decision Table	**62.05431**	**61.70637**	**58.3587**	**58.19268**
J48	**61.00251**	**60.29529**	**56.40657**	**55.6933**
IBK	**60.65952**	**60.16024**	**57.07824**	**56.6703**

Table 7

At results and graph obtained above, the J48 classifier has a higher accuracy when compared to IBK for C1 and C2. However, when 20% noise is added in C3 and C4, the accuracy of J48 is decreases but becomes lesser than IBK. This means J48 is does not do good with noise. Decision Table on the other hand has the highest accuracy irrespective of percentage split and noise added. The decision table performs consistently and significantly better than the other two classifiers.

Summary of results- Standard Deviation:

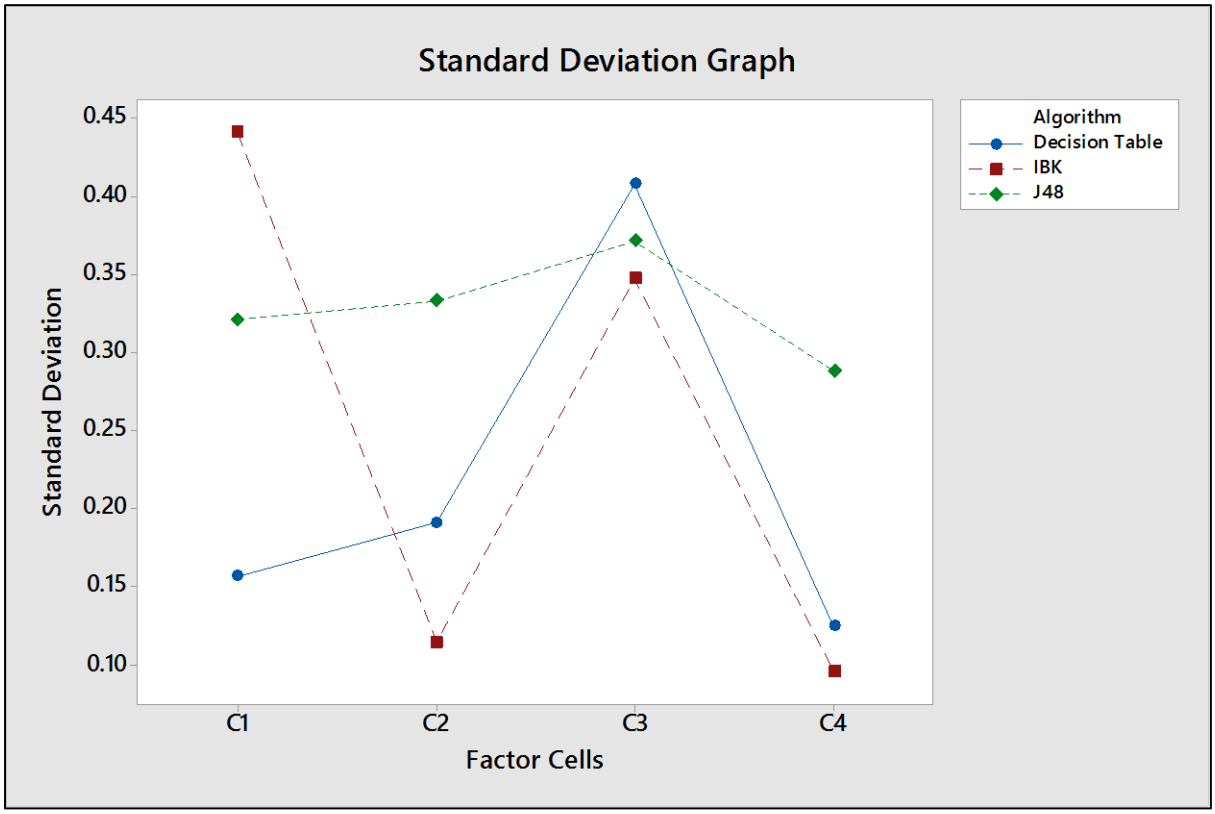

Figure 7

	C1	C2	C3	C4
Decision Table	0.156709	0.191089	0.407682	0.124833
J48	0.320959	0.332911	0.371331	0.288267
IBK	0.440852	0.11426	0.347153	0.095704

Table 8

In terms of accuracy J48 didn't perform well, but in terms of standard deviation J48 performs consistently when compared to IBK and Decision Table. For these two classifiers, there is a drastic decrease in standard deviation when the percentage split is changed from 80%/20% split to 20%/80% split. So, we can conclude that J48 had much more stable results than Decision Table and IBK classifiers.

6 Analysis and Conclusion

6.1 ROC Curve

Receiver Operating Characteristic curve is used to analyse the performance of a classifier. The ROC curve is plotted against portion of true positives from the total actual positives and portion of false positives from total actual negatives.

The Multiple ROC curve is obtain using "Knowledge Flow" feature in weka as Explorer has some limitations. The designed model is showed in the figure below.

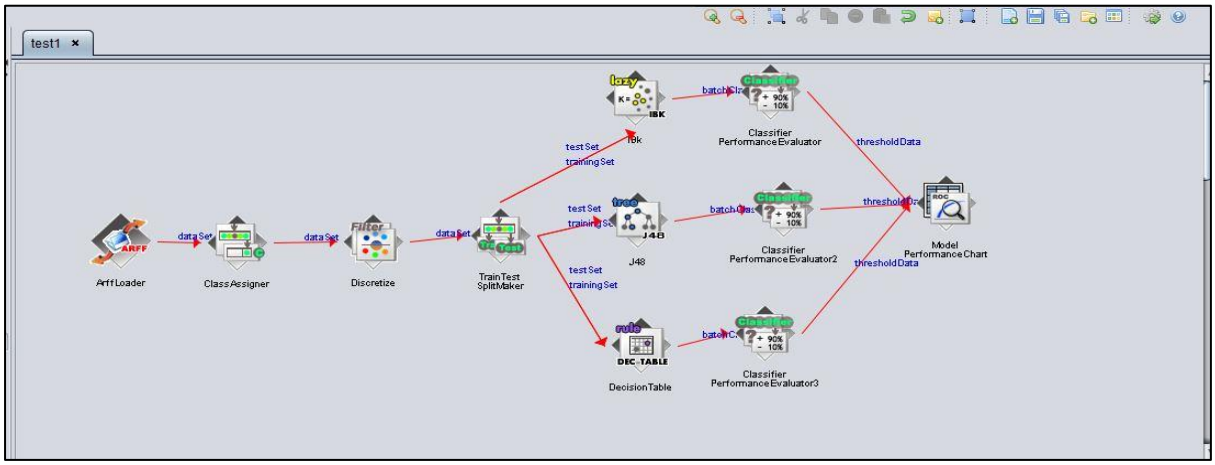

Figure 8

Four ROC curves have been generated and are as follows.

a. Multiple ROC curve of with added noise data (80-20 split)

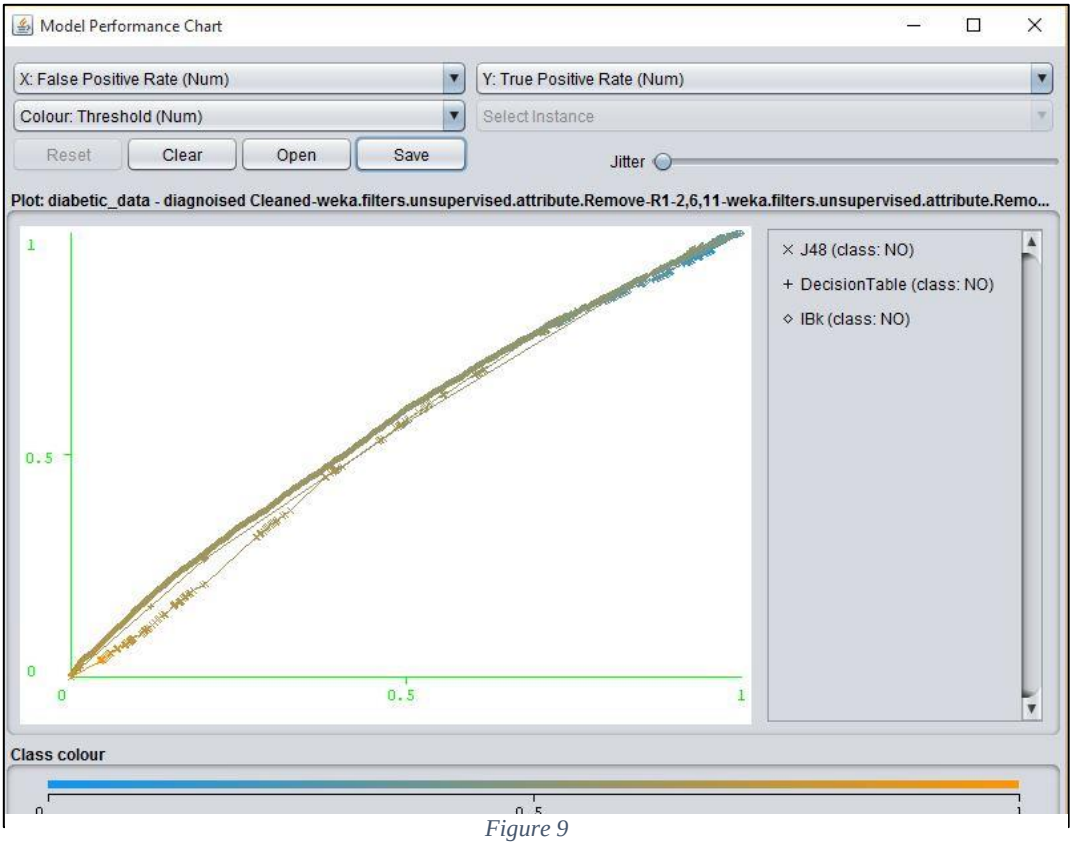

Figure 9

b. Multiple ROC curve of with added noise data (20-80 split)

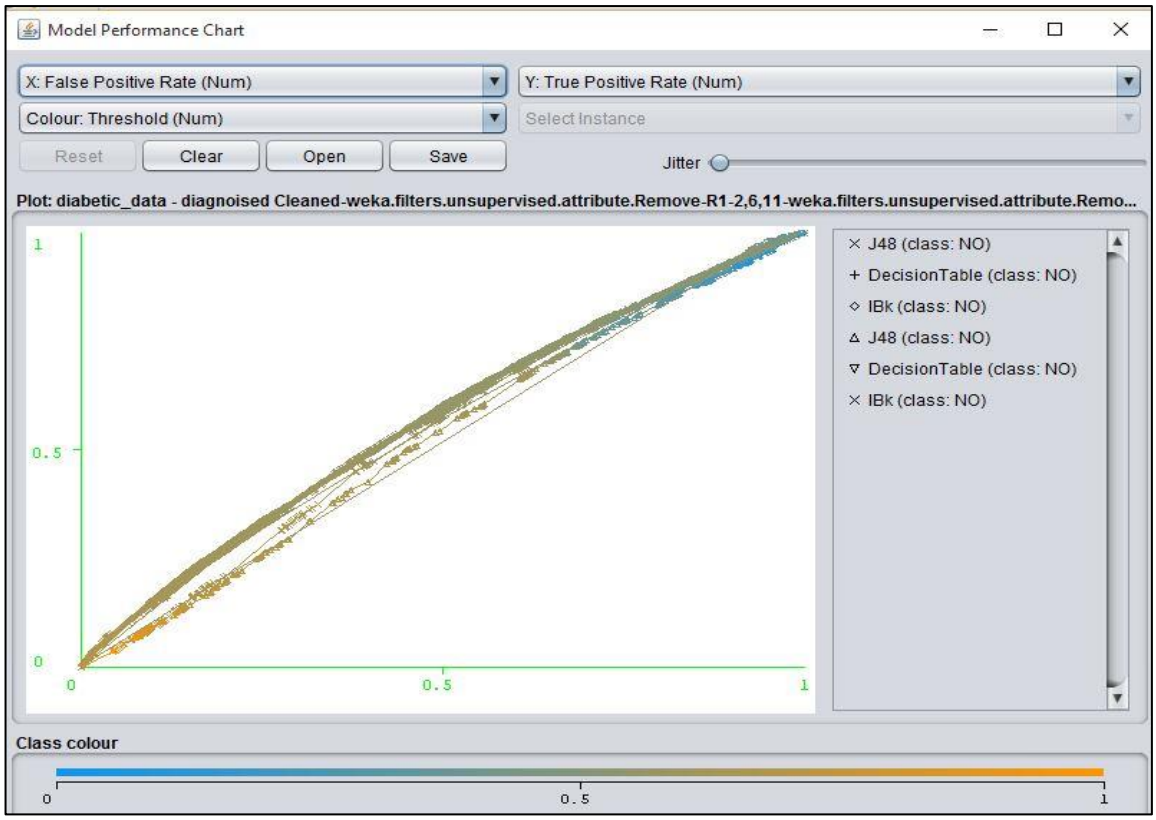

Figure 10

c. Multiple ROC curve of Without noise data (80-20 split)

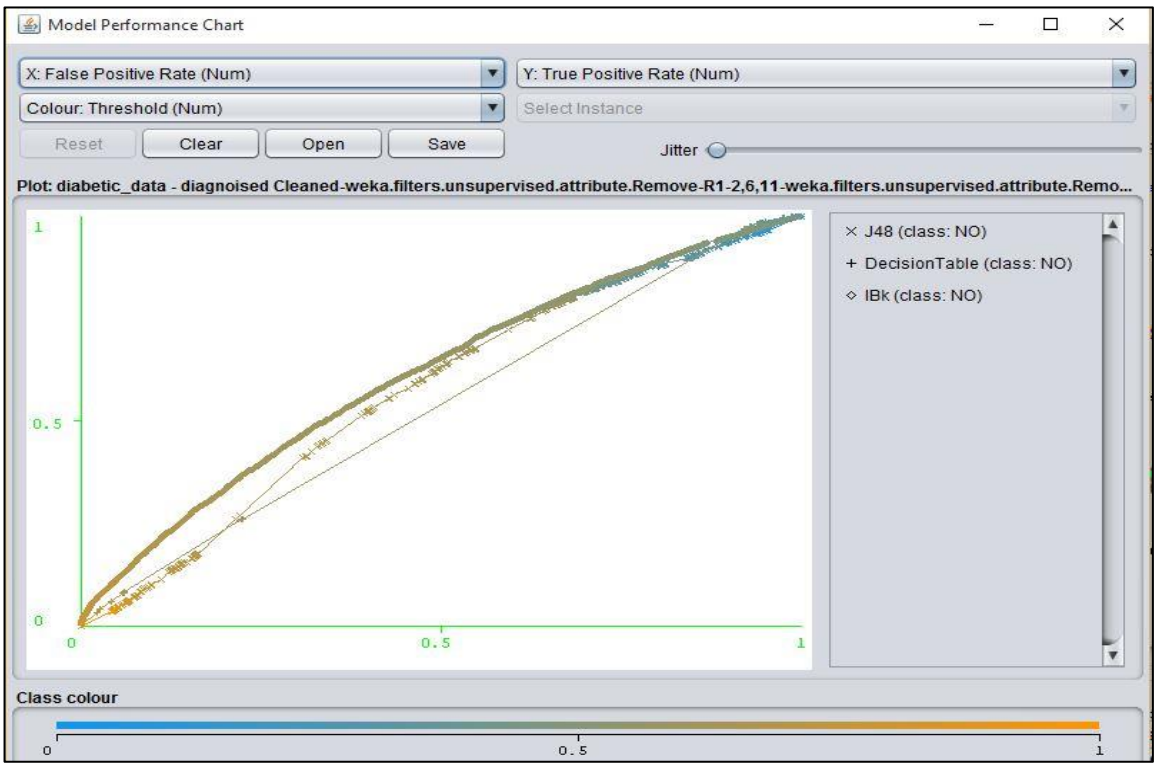

Figure 11

d. Multiple ROC curve of without noise data (20-80 split)

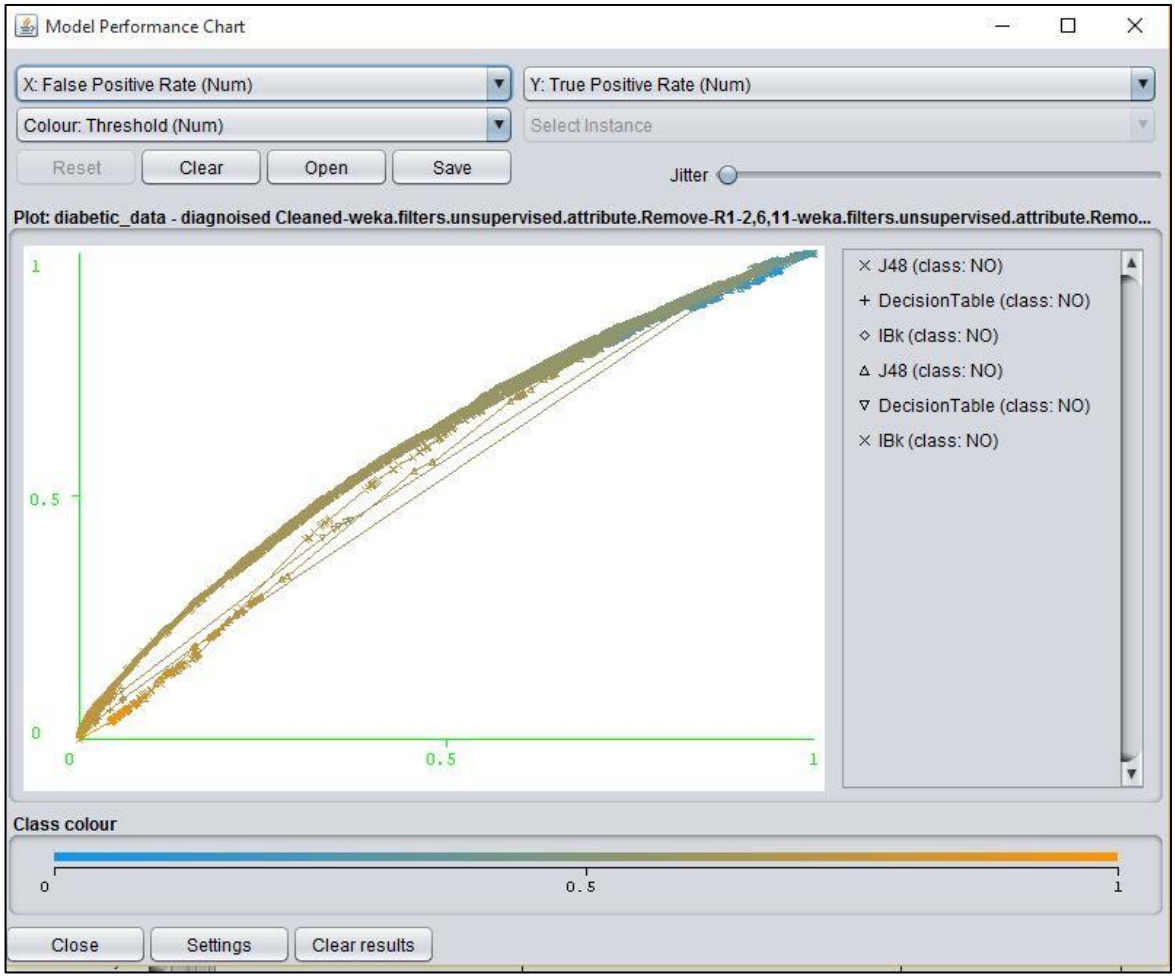

Figure 11

Hence Multiple ROC curve is used in analysing the efficiency of different classifier and best is selected based on higher area under the ROC curve for all the attributes.

6.2 Classifier Analysis

The highest and lowest standard deviation is given below

Classifier Name	Highest Accuracy	Lowest Standard Deviation
Decision Table	62.05431	0.124833
J48	61.00251	0.288267
IBK	60.65952	0.11426

Figure 9

Here decision table has the highest accuracy and IBK has the least standard deviation. However, the difference in standard deviation between decision table and IBK is very small. So, taking both accuracy and standard deviation into account we can say that Decision Table as the best classifier model as it has the highest accuracy among the three classifiers and considerably small standard deviation.

6.3 Attribute Analysis

There were some attributes which played a significantly important role in building the classifier model and in the output prediction. They are

- **number_inpatient**: Duration of the patients stay in hospital
- **number_emergency:** Number of emergency visits made by the patient in the preceding year of the encounter.
- **number_diagnoses:** Number of diagnoses entered in the system for that patient

6.4 Conclusion

- We can conclude that Decision Table is the best classifier algorithm for the given data set. Some of the reasons are:
  - High accuracy compared to others algorithm.
  - Less complicated than decision trees.
  - Decision Table has mutually exclusive and exhaustive characteristics.
- The 80%/20% split has more accuracy.
- Adding noise decreased the accuracy significantly.

7 References

- Strack, B., Deshazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014, 1-11. doi:10.1155/2014/781670