Advanced topics in database systems
Methods of Knowledge Discovery in Healthcare

Project 1

Task 1

1. What is the most common disease for each age group?

   select age, max(c),diag from(select age, count(diag)as c, diag from (SELECT age,DIAGNOSIS_CODE_1 AS diag FROM data_set UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set)as x group by age,diag order by c desc)as y group by age;

age	max(c)	diag
1	1	2767
2	5	V6284
3	15	486
4	25	5856
5	14	486
6	22	486
7	22	486
8	20	5990
9	11	5849

   In the above query result, max(c) is the maximum number of times a diagnosis code is repeated for each age group and diag is the diagnosis code. For age group 1, there are more than one diagnosis codes that have max(c)=1, but the query shows only one diagnosis code.

2. What is the prevalence of the top three diseases for each age group?
   The top three diseases for each age group is shown by executing separate queries for each age group.
   For age group 1:
   select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=1 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=1)x group by diag order by c desc limit 3 ;

age	c	diag
1	1	2767
1	1	65571
1	1	5856

For age group 2:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=2 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=2)x group by diag order by c desc limit 3 ;

age	c	diag
2	5	V6284
2	4	
2	4	34590

For age group 3:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=3 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=3)x group by diag order by c desc limit 3 ;

age	c	diag
3	15	486
3	13	5849
3	13	5856

For age group 4:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=4 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=4)x group by diag order by c desc limit 3 ;

age	c	diag
4	25	5856
4	14	486
4	13	4019

For age group 5:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=5 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=5)x group by diag order by c desc limit 3 ;

age	c	diag
5	14	486
5	13	389
5	12	51881

For age group 6:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=6 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=6)x group by diag order by c desc limit 3 ;

age	c	diag
6	22	486
6	12	5849
6	12	5990

For age group 7:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=7 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=7)x group by diag order by c desc limit 3 ;

age	c	diag
7	22	486
7	14	5849
7	13	5990

For age group 8:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=8 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=8)x group by diag order by c desc limit 3 ;

age	c	diag
8	20	5990
8	14	5849
8	13	486

For age group 9:
select age,count(diag) as c,diag from (select age,DIAGNOSIS_CODE_1 AS diag FROM data_set where age=9 UNION ALL SELECT age,DIAGNOSIS_CODE_2 AS diag FROM data_set where age=9)x group by diag order by c desc limit 3 ;

age	c	diag
9	11	5849
9	8	486
9	8	5990

In the above query results, c is the count of the diagnosis codes for the age group and diag is the diagnosis code. This means for age group 9, 5849 is diagnosed 11 times.

3. What is the average death_on_discharge ratio for each admitting diagnosis code?
   SQL query: select ADMITTING_DIAGNOSIS_CODE, count(DISCHARGE_STATUS)/c as average_ratio from data_set,(select count(distinct ADMITTING_DIAGNOSIS_CODE) as c from data_set)b where DISCHARGE_STATUS='B' group by ADMITTING_DIAGNOSIS_CODE;
   Rows: 25

ADMITTING_DIAGNOSIS_CODE	average_ratio
1629	0.0022
179	0.0022
2639	0.0022
27542	0.0022
2851	0.0022
2989	0.0022
389	0.0087
40391	0.0022
4280	0.0022
431	0.0022
4329	0.0022
43491	0.0087
486	0.0022
5070	0.0022
51881	0.0152
51883	0.0022
51884	0.0022
5570	0.0022
5722	0.0022
5781	0.0022
586	0.0022
59960	0.0022
7054	0.0022
71945	0.0043
7802	0.0022

In the above query result, average_ratio is the average death_on_discharge ratio for each admitting diagnosis code.

4. What is the most common value combination for the source_of_admission and discharge_destination?

   SQL query:

   select SOURCE_OF_ADMISSION,DISCHARGE_DESTINATION,count(DISCHARGE_DESTINATION) from data_set group by SOURCE_OF_ADMISSION;

   Rows: 8

SOURCE_OF_ADMISSION	DISCHARGE_DESTINATION	count(DISCHARGE_DESTINATION)
	20	19
1	6	1146
2	6	121
4	51	121
5	3	47
6	1	18
9	1	4

   The above query result shows that for each source of admission, which discharge destination has the maximum count. Count(DISCHARGE_DESTINATION) is the count for each discharge destination in the data set.

5. Compare the in-hospital mortality of men and women.

   SQL query: select sex,count(DISCHARGE_DESTINATION) from data_set where DISCHARGE_DESTINATION=20 group by sex;

   Rows: 2

sex	count(DISCHARGE_DESTINATION)
1	18
2	29

   The above query result shows the number of male patients and number of female patients who are dead on discharge.

6. What is the most common long stay primary diagnosis (DIAGNOSIS_CODE_1) and which is the most common short stay primary diagnosis (DIAGNOSIS_CODE_1

   SQL query: select DIAGNOSIS_CODE_1, count(DIAGNOSIS_CODE_1) as one,STAY_INDICATOR from data_set where STAY_INDICATOR='S' group by DIAGNOSIS_CODE_1 order by one desc limit 1;
   Rows: 1

DIAGNOSIS_CODE_1	one	STAY_INDICATOR
389	65	S

   SQL query: select DIAGNOSIS_CODE_1, count(DIAGNOSIS_CODE_1) as one,STAY_INDICATOR from data_set where STAY_INDICATOR='L' group by DIAGNOSIS_CODE_1 order by one desc limit 1;
   Rows: 1

DIAGNOSIS_CODE_1	one	STAY_INDICATOR
V5789	18	L

   From the above query results, the most common primary diagnosis for short stay is 389 and for long stay is v5789.

7. What is the average total cost (total charges) for each length of stay?
   SQL query: select LENGTH_OF_STAY,avg(total_charges) from data_set group by LENGTH_OF_STAY ORDER BY LENGTH_OF_STAY ASC;
   Rows: 41

LENGTH_OF_STAY	avg(total_charges)
1	18305.115942028984
10	63270.4
11	73906.93548387097
110	934443
12	67934.76470588235
13	48245.0625
14	82098.15384615384
15	72132.88235294117
16	72939
17	59692.2

18	76199.625
19	92606.25
2	20836.394495412846
20	111915.125
21	273411.25
22	163407.5
23	342789
24	179778
25	118402
27	184175.5
28	151800.5
29	40754
3	26814.10843373494
31	96186
33	561217
34	127264.5
35	254410
37	228115
38	208234.33333333334
39	316960
4	27489.242857142857
41	262986
42	168752
5	35170.954545454544
51	389386
55	135805
58	217458.5
6	36867.807692307695
7	49794.030303030304
8	47305.72222222222
9	49811.333333333336

The avg(total_charges) gives the average cost for each length of stay.

8. Compare the total charges for each admission diagnosis code, for the age group 5.
   SQL query: select diag,sum(total_charges) from (select ADMITTING_DIAGNOSIS_CODE as diag,total_charges from data_set where age=5) as x group by diag;
   Rows: 25

Diag	sum(total_charges)
1589	16597
179	13284
2113	64983
2352	27830
2409	66525
24200	11699
25000	13034
25010	19864
25080	4368
2760	40939
2761	19644
27801	12259
2841	15837
28489	10533
2859	35735
28800	109804
29620	83074
29624	17667
29630	11584
29633	10237
29644	11673
30390	13200
3383	7787
34290	18230
389	192240

The sum(total_charges) gives the total charges for each admitting diagnosis code. In the above query result, diag is the admitting diagnosis code.

9. Compare the average length of stay for type of admission=1 between men and women.
SQL query: select sex,avg(LENGTH_OF_STAY) from data_set where TYPE_OF_ADMISSION=1 group by sex;

sex	avg(LENGTH_OF_STAY)
1	5.175342465753425
2	5.495967741935484

The above query result shows the average length of stay for male and female patients.

10. Find the top 3 most expensive DRG codes (i) based on the DRG price and (ii) based on the total charges.

(i) select DRG_CODE,DRG_PRICE from data_set order by DRG_PRICE desc limit 3;

DRG_CODE	DRG_PRICE
155	9994
287	9974
308	9971

The above query shows the top three expensive DRG codes based on DRG PRICE.

(ii) select DRG_CODE,TOTAL_CHARGES from data_set order by TOTAL_CHARGES desc limit 3;

DRG_CODE	TOTAL_CHARGES
244	9946
604	9935
743	9916

The above query shows the top three expensive DRG codes based on TOTAL_CHARGES.

Task 2

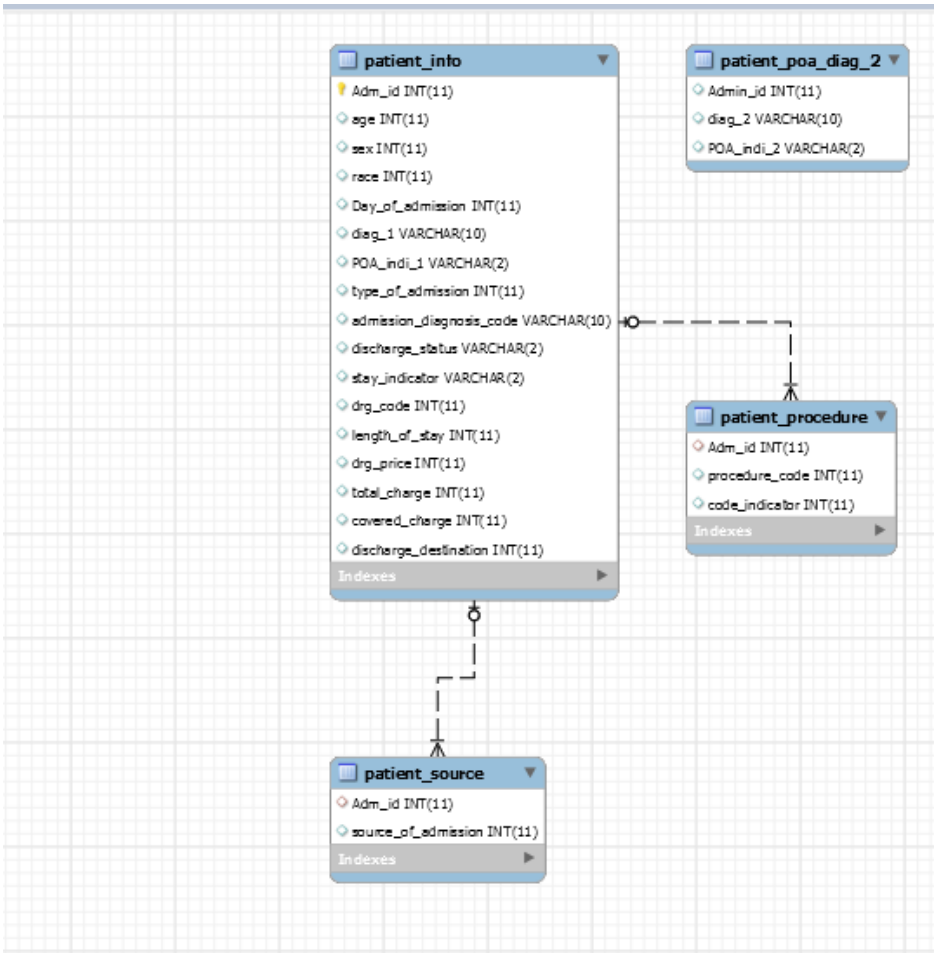
1. Observe the data and design an appropriate relational schema. Then create the schema on the DBMS system of your choice (preferably MySQL server). The schema should have the following merits:
   a. Normalization principles should be applied to avoid duplicates
   b. Appropriate data types should be defined

   SQL Code:

   1. create table patient_info(Adm_id integer primary key,age integer,sex integer,race integer,Day_of_admission integer, diag_1 varchar(10), POA_indi_1 varchar(2),type_of_admission integer,admission_diagnosis_code varchar(10),discharge_status varchar(2),stay_indicator varchar(2),drg_code integer,length_of_stay integer,drg_price integer,total_charge integer,covered_charge integer,discharge_destination integer);
   2. create table patient_POA_diag_2(Admin_id integer,diag_2 varchar(10), POA_indi_2 varchar(2));
   3. create table patient_procedure(Adm_id integer,procedure_code integer,code_indicator integer,foreign key(Adm_id) references patient_info(Adm_id));

   4. create table patient_source(Adm_id integer,source_of_admission integer,foreign key(Adm_id) references patient_info(Adm_id));

Relational Schema



The schema has four tables- patient_info, patient_poa_diag2, patient_procedure and patient_source.

2. Import the data into your newly developed schema
   Data has been imported using the php scripts attached in the zip file.
3.
   a. An appropriate query which returns a result which is identical to the given csv file. This way you are demonstrating how one can extract data from an Electronic Medical Record database, to use for data analysis.
      SQL Code : create view pro_cod_1 as select adm_id as ap1,procedure_code as procedure_code_1 from patient_procedure where code_indicator=1
      create view pro_cod_2 as select adm_id as ap2,procedure_code as procedure_code_2 from patient_procedure where code_indicator=2

      select
      a.adm_id,a.age,a.sex,a.race,a.day_of_Admission,a.stay_indicator,a.drg_code,a.length_of_stay,a.drg_price,a.total_charge,a.covered_charge,a.POA_indi_1,a.diag_1,b.POA_indi_2,b.diag_2,procedure_code_1,procedure_code_2 from patient_info a,patient_poa_diag_2 b,pro_cod_1,pro_cod_2 where a.adm_id=b.admin_id or a.adm_id=ap1 or a.adm_id=ap2;

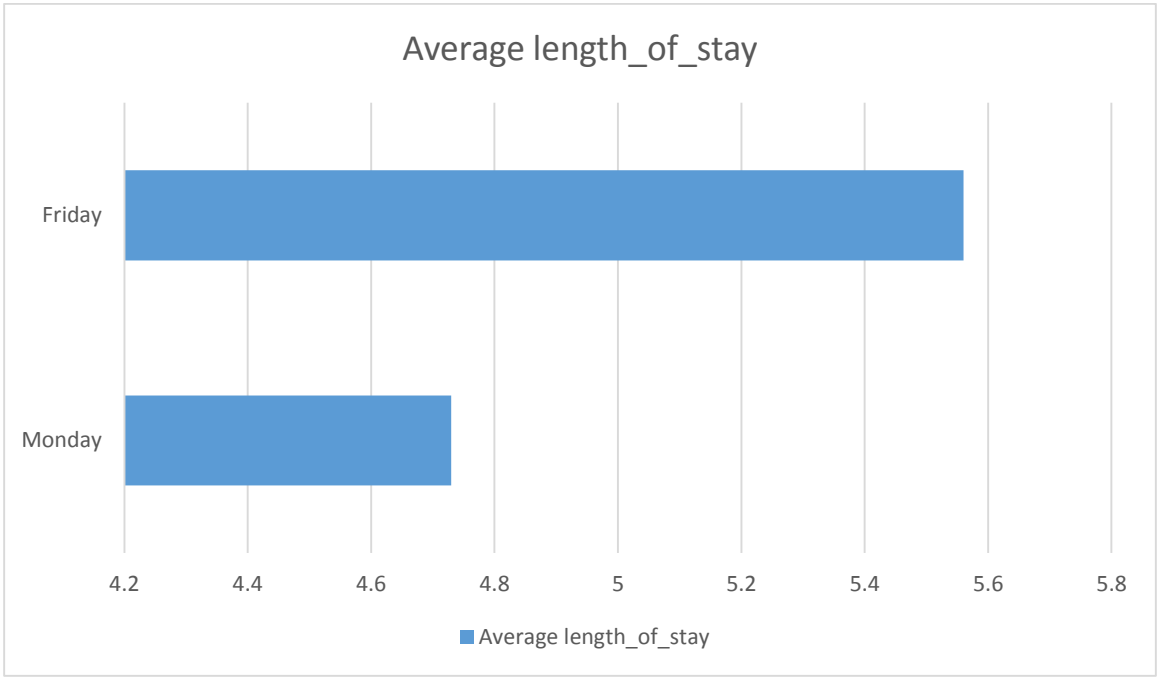
   b. Queries which return the Coverage Ratio (Coverage Ratio = COVERED_CHARGES/TOTAL_CHARGES) of patients who stayed in the hospital for a period of time longer than 5 days. How does this compare to the Coverage Ratio of patients with a Long Stay?

      SQL Query: select COVERED_CHARGES/TOTAL_CHARGES as coverage_ratio from patient_info where LENGTH_OF_STAY>5;

      select COVERED_CHARGES/TOTAL_CHARGES as coverage_ratio from patient_info where STAY_INDICATOR='L';
   c. Is there any variation in the average Length of Stay of patients admitted to the hospital in different days of the week (DAY_OF_ADMISSION)? Showcase this with an appropriate SQL query and design an appropriate graph comparing the average Length of Stay between Friday admissions (DAY_OF_ADMISSION=6) and Monday admissions (DAY_OF_ADMISSION=2).
      Answer: Yes, the average length of stay varies with the day 0f admission.

      SQL Query:select Day_of_admission,avg(LENGTH_OF_STAY) from patient_info group by Day_of_admission;
      select Day_of_admission,avg(LENGTH_OF_STAY) from patient_info where Day_of_admission='6' union
      select Day_of_admission,avg(LENGTH_OF_STAY) from patient_info where Day_of_admission='2';

Graph comparing the length of stay between Friday and Monday admissions.



d. Count all the distinct DRG Price values for each DRG code. For example for DRG code 1234 there might be the following DRG prices in dollars: $23,100 (12 occurrences), $19,000 (15 occurrences), and $15,320 (20 occurrences)**.**

SQL Query: select DRG_CODE,count(distinct DRG_PRICE) from patient_info group by DRG_CODE;

TASK 3

(a) Is clustering a supervised or an unsupervised data mining technique? Please explain your answer focusing on what differentiates unsupervised learning from the supervised techniques.

ANSWER: Clustering is unsupervised learning. In unsupervised learning, the categories are not known and appropriate categories are found along the way. Unsupervised learning is also known as descriptive data mining. Similarity is measured between the objects and unknown patterns are discovered.

Supervised learning is used when categories are known and when the parameters are small. If you have a large set of parameters and then some analysis is done to find the appropriate categories by learning process for usually a large set of data then unsupervised learning is used. Supervised is chosen when training data is available and unsupervised is used if the training data is not available.

Unsupervised is perfect in case of a health care database as the categories are not so apparent and clusters change as the parameters are added on.

(b) The appropriate number of clusters which are required to properly cluster the following admission attributes: source of admission, type of admission, age group. Use the elbow method to define the number, by evaluating the 'within cluster sum of squared errors' you get as a result in your Weka output. Draw an appropriate graph to explain your answer

Number of clusters:1
Within cluster sum of squared errors: 2177.0

Number of clusters:2
Within cluster sum of squared errors: 1873.0

Number of clusters:3
Within cluster sum of squared errors: 1697.0

Number of clusters:4
Within cluster sum of squared errors: 1532.0

Number of clusters:5
Within cluster sum of squared errors: 1411.0

Number of clusters:6
Within cluster sum of squared errors: 1257.0

Number of clusters:7
Within cluster sum of squared errors: 1143.0

Number of clusters:8
Within cluster sum of squared errors: 1124.0

Number of clusters:9

Within cluster sum of squared errors: 1097.0

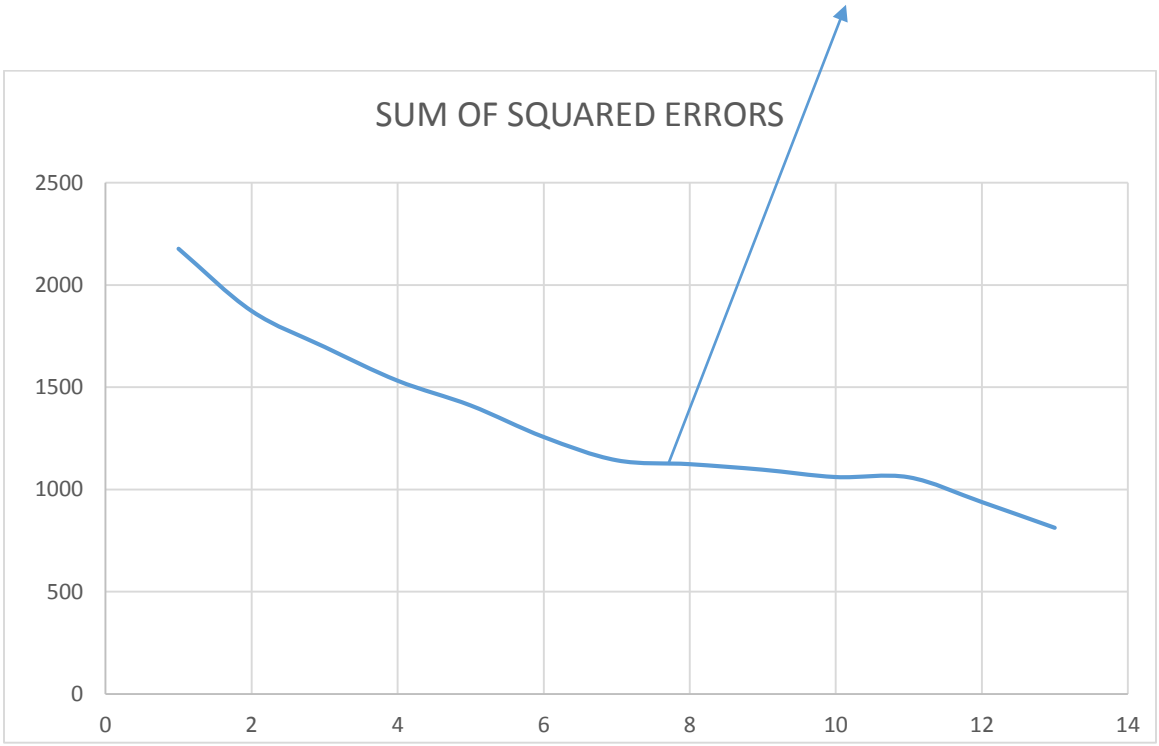
Number of clusters:10
Within cluster sum of squared errors: 1061.0

Number of clusters:11
Within cluster sum of squared errors: 1060.0

Number of clusters:12
Within cluster sum of squared errors: 939.0

Number of clusters:13
Within cluster sum of squared errors: 813.0

ELBOW POINT=AGE:7



(c)
Based on the number of clusters you specified in the above step, please calculate those clusters and explain how one may interpret the result.
Final cluster centroids:

Attribute	Cluster#	0	1	2	3	4	5	6
	Full Data (1477.0)	(452.0)	(393.0)	(167.0)	(119.0)	(121.0)	(154.0)	(71.0)
AGE	3	6	8	7	3	3	5	4
SOURCE_OF_ADMISSION	1	1	1	1	4	1	1	1
TYPE_OF_ADMISSION	1	2	1	1	3	1	1	3

Task 4

(a) Use any appropriate method to modify the class attribute values to be only of two values, either zero (DRG price less than $80,000) or one (DRG price more than $80,000) so that the problem will be binary classification. Integrate the new attribute (DRG_PRICE_BINARY) into your dataset.

SQL Code: alter table data_set add column drg_price_binary integer;

update data_set set drg_price_binary=0 where DRG_PRICE<80000;

update data_set set drg_price_binary=1 where DRG_PRICE>80000;

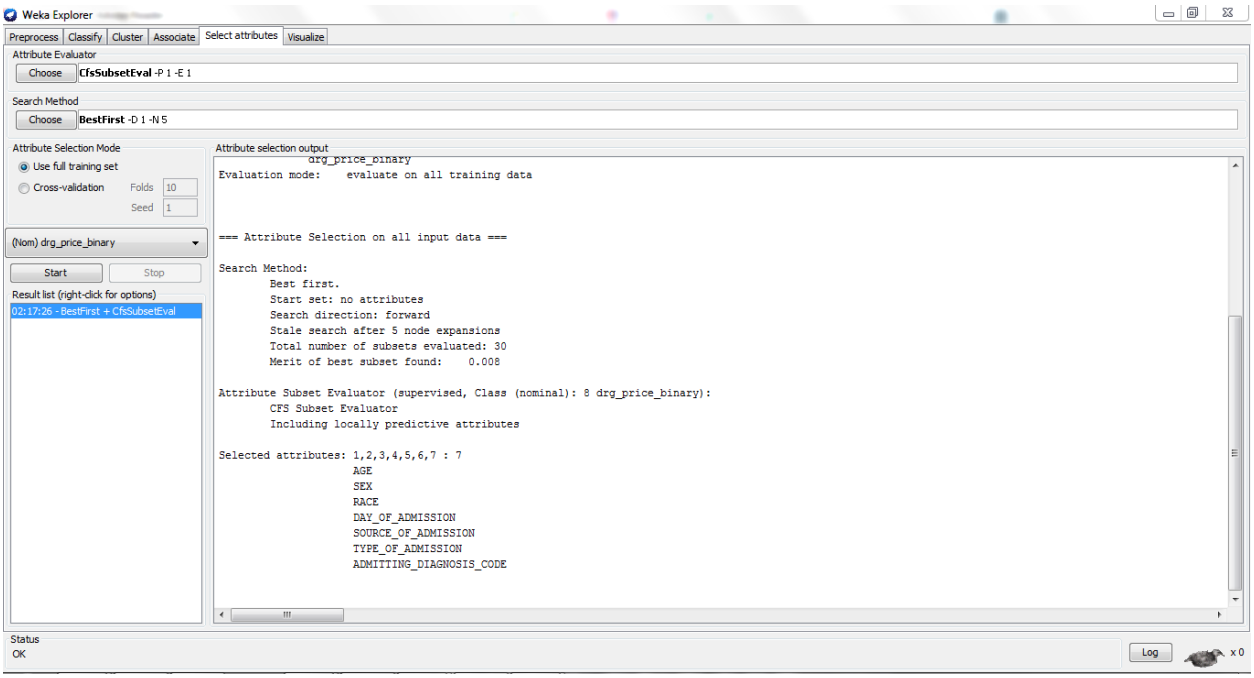
(b) Observe the available features and explain how the data are acquired in a temporal manner during the healthcare procedure in the real hospital. Specifically, define what do clinicians/administrators already know:
1. At the time when the patient enters the hospital
2. At the time when the patient is discharged from the hospital

1. When the patient enters the hospital, the clinicians will know the AGE, SEX, RACE, DAY_OF_ADMISSION, SOURCE_OF_ADMISSION, TYPE_OF_ADMISSION, ADMITTING_DIAGNOSIS_CODE.
2. When the patient is discharged from the hospital, the clinicians know DISCHARGE_STATUS, STAY_INDICATOR, LENGTH_OF_STAY, POA_DIAGNOSIS_INDICATOR_1, POA_DIAGNOSIS_INDICATOR_2, DIAGNOSIS_CODE_1, DIAGNOSIS_CODE_2, PROCEDURE_CODE_1, PROCEDURE_CODE_2, DISCHARGE_DESTINATION.

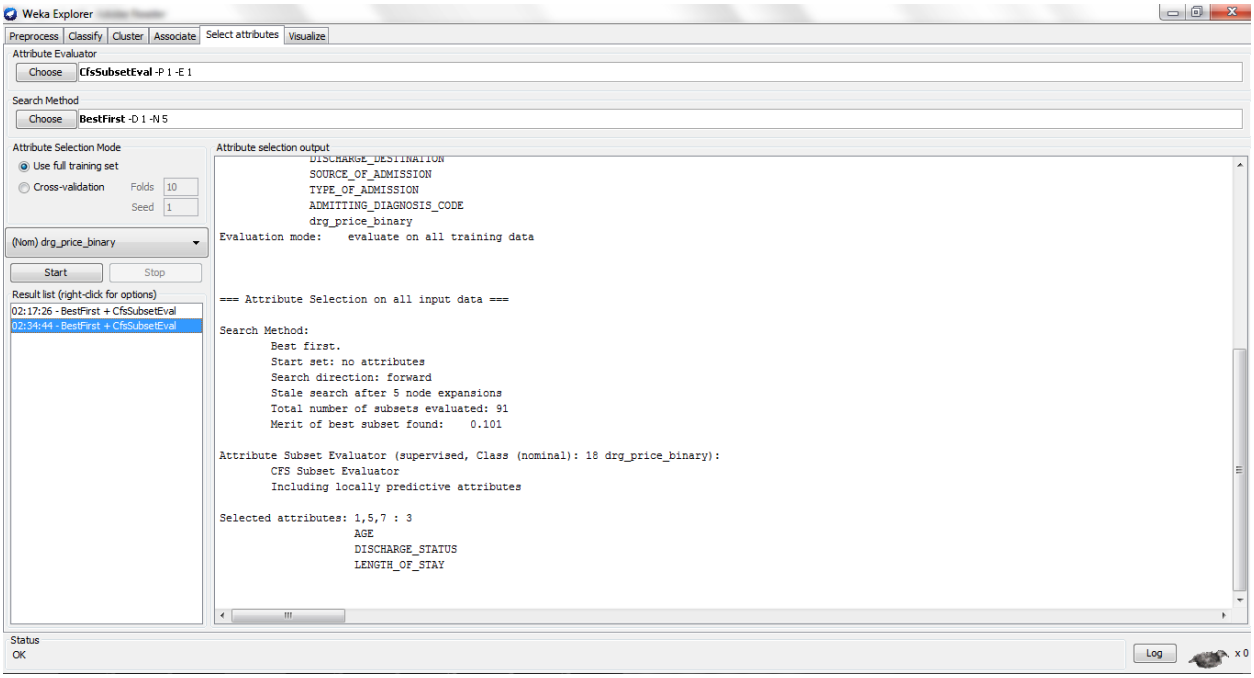
(c) Use the classifiers (i) Naïve Bayes and (ii) Logistic Regression to classify the DRG_PRICE_BINARY, for each scenario, by using the features you found to be useful during feature selection.

1. In scenario 1, we manually excluded the features 7, 10 and 11 because, the DRG_CODE, TOTAL_CHARGES and COVERED_CHARGES are not known at the time of admission.
   In scenario 2, we manually exclude 7,10 and 11 because those features are not relevant to the discharge of the patient.

2. CfsSubsetEval for scenario 1:



CfsSubsetEval for scenario 2:

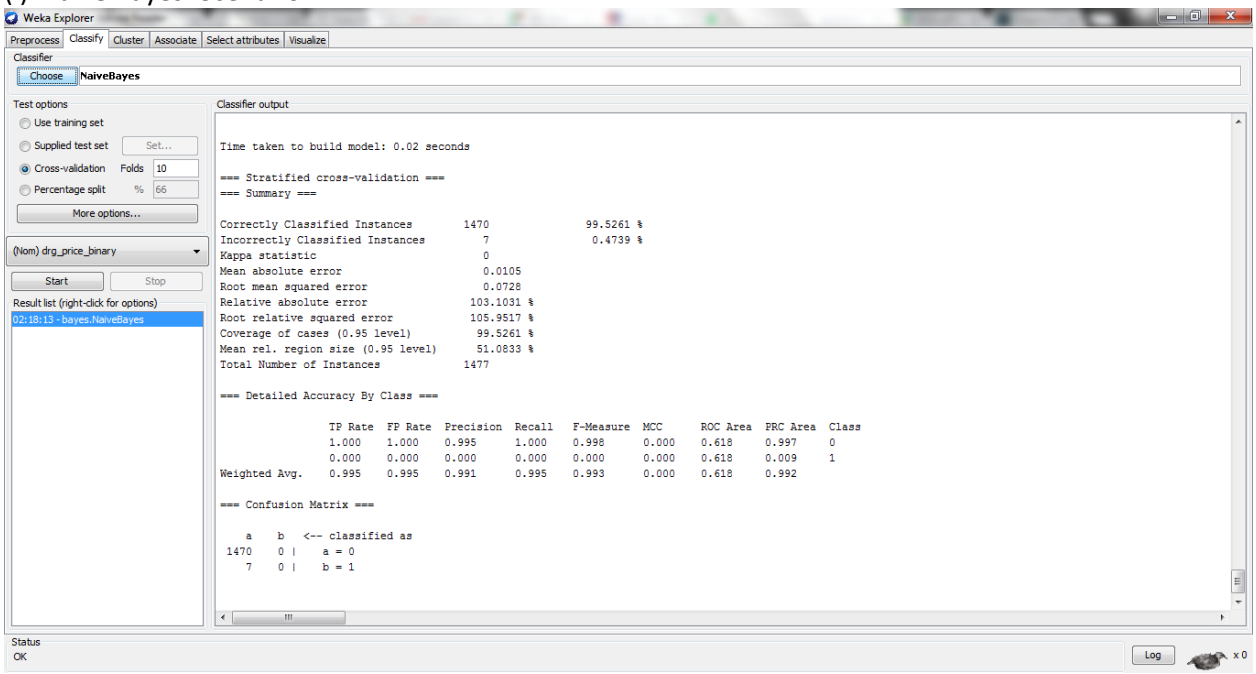


The features selected during the feature selection process using CfsSubsetEval for scenario 1 are: 1,2,3,4,5,6,7.

The features selected during the feature selection process using CfsSubsetEval for scenario 1 are: 1,5,7

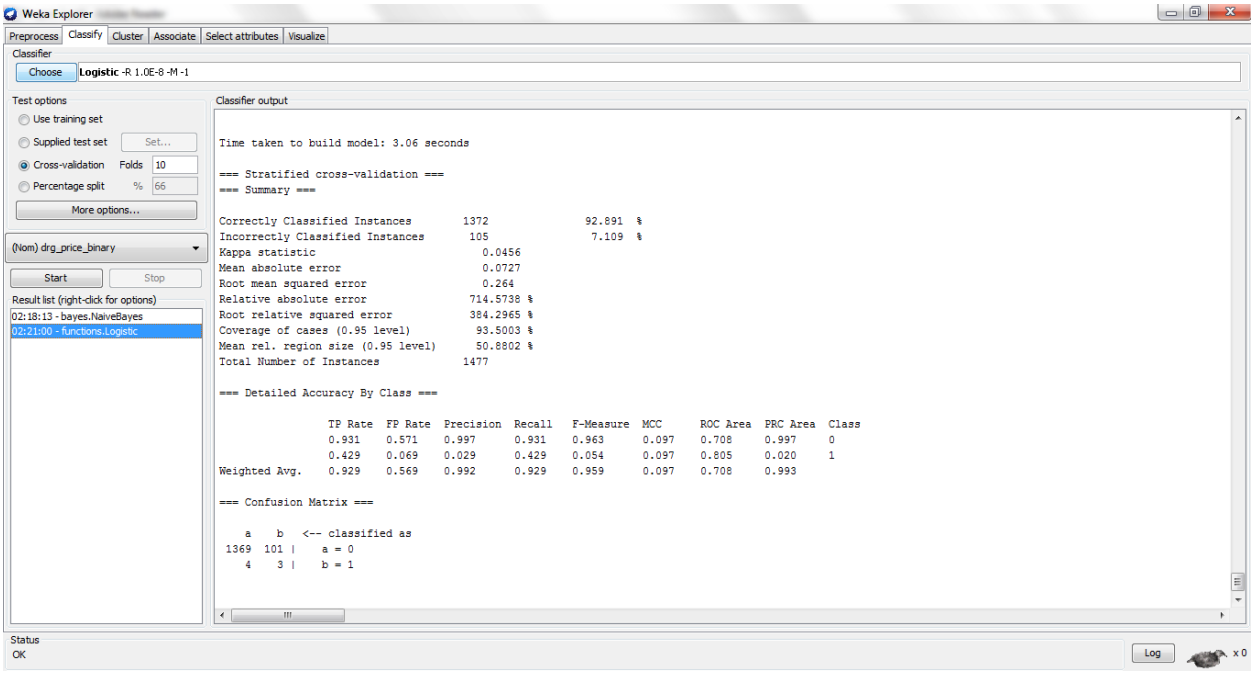
3.

(i) Naïve Bayes- Scenario 1



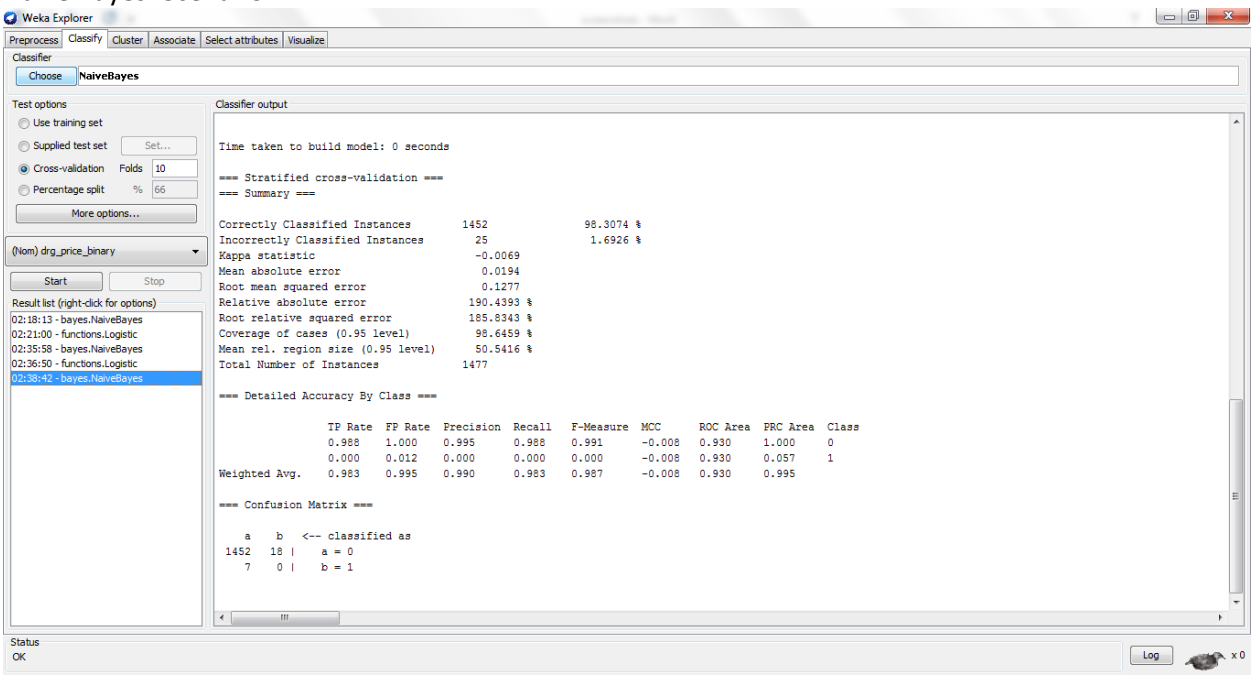
The detailed accuracy by class and the confusion matrix is shown above.

(ii) Logistic Regression- Scenario 1



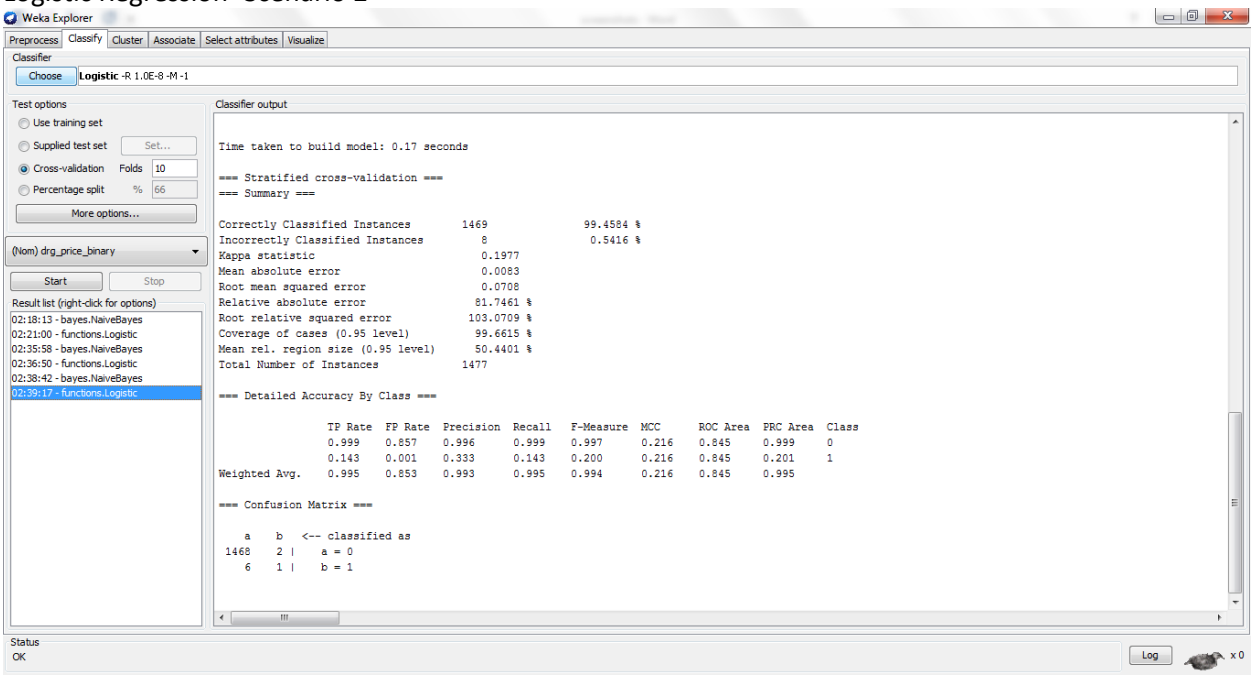
The detailed accuracy by class and the confusion matrix is shown above.

Naïve Bayes- Scenario 2



The detailed accuracy by class and the confusion matrix is shown above.

Logistic Regression- Scenario 2



The detailed accuracy by class and the confusion matrix is shown above.

4. Answer the following questions regarding the accuracy of the classification for Naïve Bayes in each scenario: ⎪ the overall accuracy ⎪ the precision for each one of our classes (DRG_PRICE_BINARY=0 and DRG_PRICE_BINARY=1) ⎪ the recall for each one of the classes (DRG_PRICE_BINARY=0 and DRG_PRICE_BINARY=1) ⎪ the f-measure for one each of the classes (DRG_PRICE_BINARY=0 and DRG_PRICE_BINARY=1)

Scenario 1

✓ overall accuracy-99.5261%

```
Correctly Classified Instances        1470               99.5261 %
Incorrectly Classified Instances         7                0.4739 %
Kappa statistic                          0
Mean absolute error                      0.0105
Root mean squared error                  0.0728
Relative absolute error                103.1031 %
Root relative squared error            105.9517 %
Coverage of cases (0.95 level)          99.5261 %
Mean rel. region size (0.95 level)      51.0833 %
Total Number of Instances             1477
```

✓ Precision for each class
(DRG_PRICE_BINARY=0)- 0.995
(DRG_PRICE_BINARY=1)- 0.000

✓ Recall for each class
(DRG_PRICE_BINARY=0)- 1.000
(DRG_PRICE_BINARY=1)- 0.000

✓ Fmeasure for each class
(DRG_PRICE_BINARY=0)- 0.998
(DRG_PRICE_BINARY=1)- 0.000

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	1.000	0.995	1.000	0.998	0.000	0.618	0.997	0
0.000	0.000	0.000	0.000	0.000	0.000	0.618	0.009	1

Scenario 2

✓ Overall Accuracy-98.3074%

Correctly Classified Instances	1452	98.3074 %
Incorrectly Classified Instances	25	1.6926 %
Kappa statistic	-0.0069	
Mean absolute error	0.0194	
Root mean squared error	0.1277	
Relative absolute error	190.4393 %	
Root relative squared error	185.8343 %	
Coverage of cases (0.95 level)	98.6459 %	
Mean rel. region size (0.95 level)	50.5416 %	
Total Number of Instances	1477	

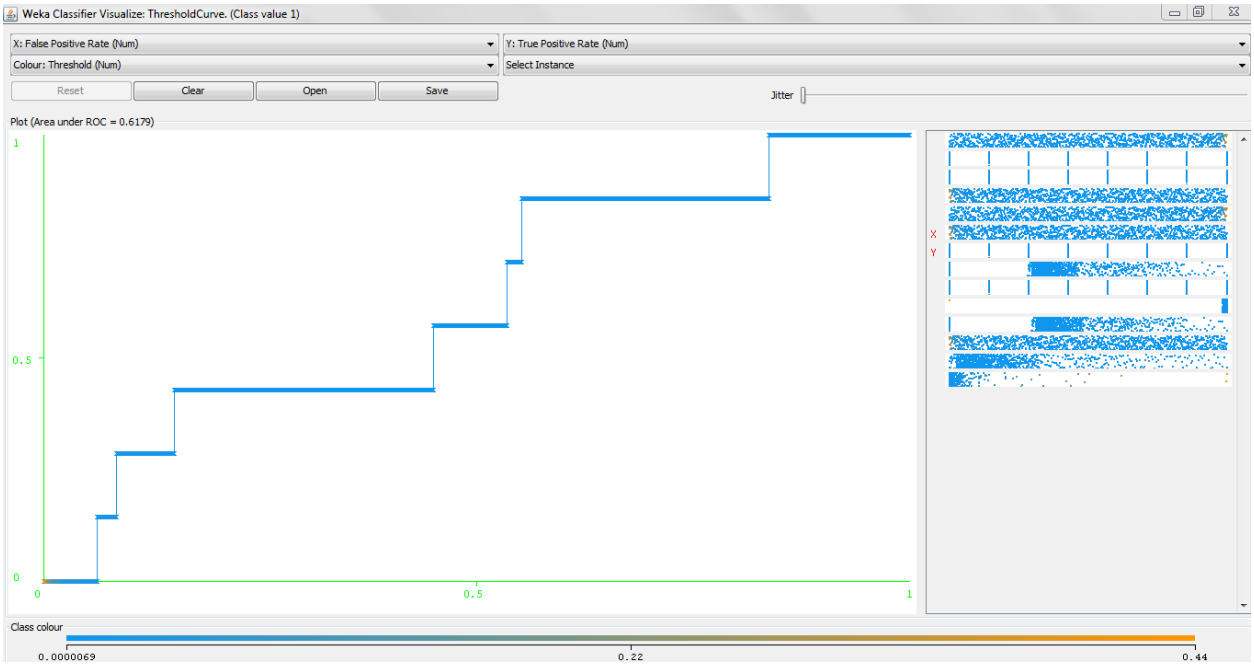
- ✓ Precision for each class
  (DRG_PRICE_BINARY=0)-0.995
  (DRG_PRICE_BINARY=1)-0.000
- ✓ Recall for each class
  (DRG_PRICE_BINARY=0)-0.988
  (DRG_PRICE_BINARY=1)-0.000
- ✓ Fmeasure for each class
  (DRG_PRICE_BINARY=0)-0.991
  (DRG_PRICE_BINARY=1)-0.000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.988	1.000	0.995	0.988	0.991	-0.008	0.930	1.000	0
	0.000	0.012	0.000	0.000	0.000	-0.008	0.930	0.057	1
Weighted Avg.	0.983	0.995	0.990	0.983	0.987	-0.008	0.930	0.995	

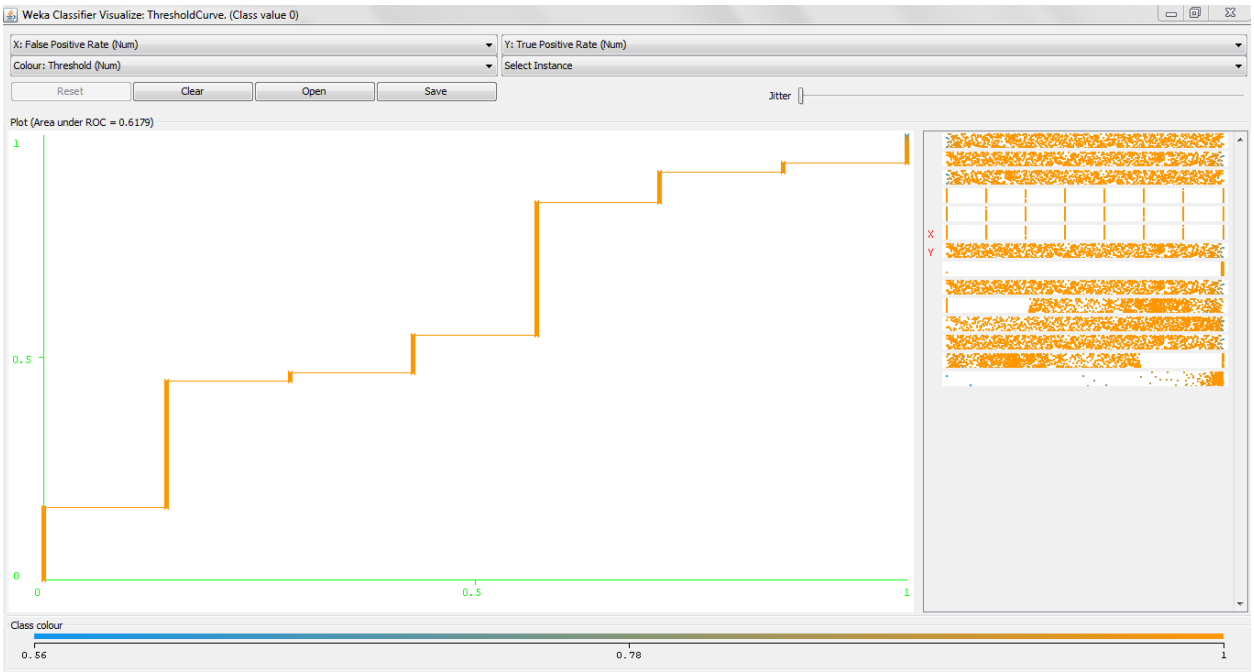
5. Design the ROC curve and calculate the ROC area of Naïve Bayes for both scenarios
   ROC curve

Scenario 1- Class 1
ROC Area: 0.6179



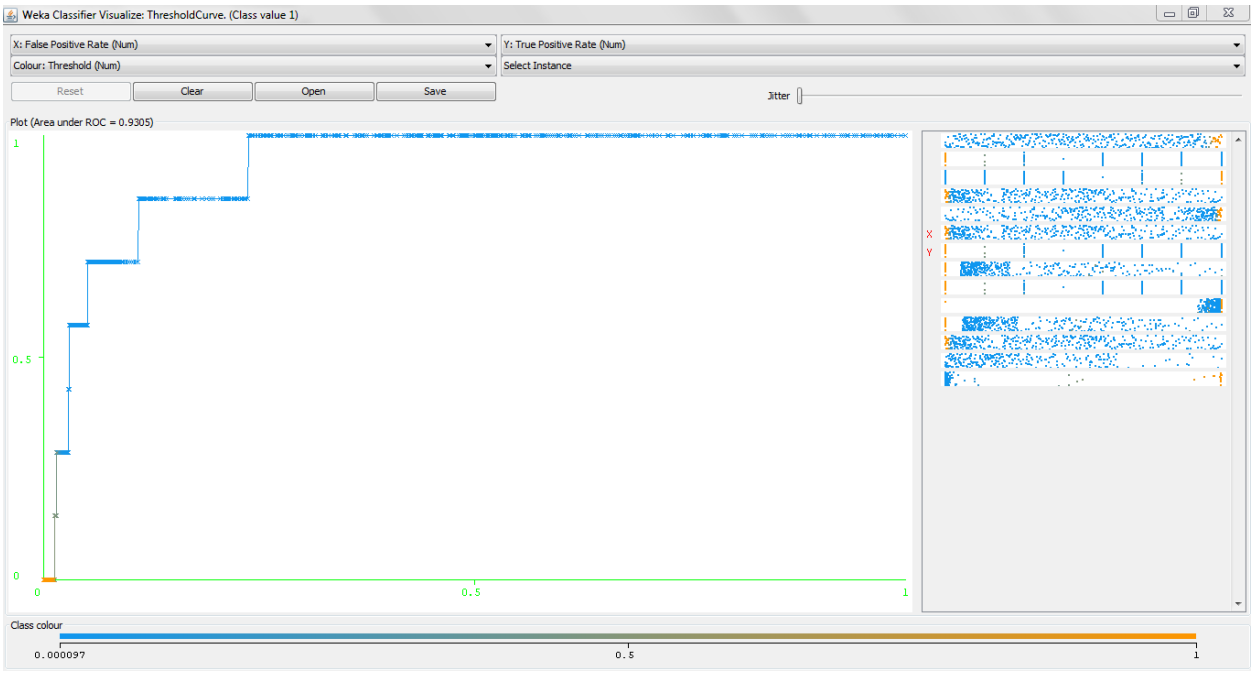
Scenario 1-Class 0

ROC Area: 0.6179



Scenario 2- Class 1

ROC Area: 0.9305



Scenario 2- Class 0

ROC Area: 0.9305