PROJECT REPORT: HOSPITAL CHARGES FOR INPATIENTS

INSY 5339 – PRINCIPLES OF BUSINESS DATA MINING

GROUP 9 : Aditya Gorde, Kartik Chavan, Pratik Shetty, Arpeet Hoskote, Shrish Murthy.

TABLE OF CONTENTS

1. DATA BACKGROUND

The data provided here include hospital-specific charges for the more than 3,000 U.S. hospitals that receive Medicare Inpatient Prospective Payment System (IPPS) payments for discharges, paid under Medicare based on a rate per discharge using the Medicare Severity Diagnosis Related Group (MS-DRG) for Fiscal Year (FY) 2014. These MS-DRGs represent more than 7 million discharges or 75 percent of total Medicare IPPS discharges. Variation of hospital charges in the various hospitals in the US for the top 100 diagnoses. This dataset will show how prices for the same diagnosis and the same treatment and in the same city can vary differently across different providers. It might help to find a better hospital for your treatment. You can also analyze to detect fraud among providers.

1.1 DATASET & ATTRIBUTES

The data set that was provided contained ***13 attributes*** and about ***163065 instances***.

These attributes are:

Attribute Name	Descriptions
DRG Definition	The code and description identifying the MS-DRG. MS-DRGs are a classification system that groups similar clinical conditions (diagnoses) and the procedures furnished by the hospital during the stay.
Provider Id	The CMS Certification Number (CCN) assigned to the Medicare certified hospital facility.
Provider Name	The name of the provider.
Provider Street Address	The provider's street address.
Provider City	The city where the provider is located.
Provider State	The state where the provider is located.
Provider Zip Code	The provider's zip code.
Provider HRR	The Hospital Referral Region (HRR) where the provider is located.

Total Discharges	The number of discharges billed by the provider for inpatient hospital services.
Average Covered Charges	The provider's average charge for services covered by Medicare for all discharges in the MS-DRG. These will vary from hospital to hospital because of differences in hospital charge structures.
Average Total Payments	The average total payments to all providers for the MS-DRG including the MSDRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also, included in average total payments are co-payment and deductible amounts that the patient is responsible for and any additional payments by third parties for coordination of benefits
Average Medicare Payments	The average amount that Medicare pays to the provider for Medicare's share of the MS-DRG. Average Medicare payment amounts include the MS-DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Medicare payments Do Not include beneficiary co-payments and deductible amounts nor any additional payments from third parties for coordination of benefits.

2. DATA CLEANING PROCESS

Data Cleaning can be defined as a process in which the amendment and removal of data from a database takes place. The data removed is generally incorrect, incomplete, improperly formatted or duplicated. Data cleaning can also be defined as a process which is carried out to determine inaccurate or unreasonable data. This process can then also be used to improve the data quality through correction of the detected errors. The process of Data Cleaning can broadly be divided in to a general framework:

- Define and determine error types
- Search and identify error instances
- Correct the errors
- Document error instances and error types

- Modify data entry procedures to reduce future errors.

Need for Data Cleaning is simply to improve the quality of the data and to bring the data in a form that can be deemed as "fit for use" by the users. It is very common for any dataset to have a field error rate of 1-5%. A good understanding of the error can also lead to an improvement in the data quality and active quality control. 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. The corrections should always be done in a separate field and not the original data so that the original data can be retrieved at any point in time. Data Cleaning is required in Single data collections such as files and databases when it faces data quality problems which can be due to misspellings due to data entry, missing information or other invalid data also the need for data cleaning increases significantly when Multi data sources need to be integrated like data warehouses, federated database systems or global web-based information systems.

2.1 DATA CLEANING TOOLS

We used Microsoft Excel and Weka for splitting the attributes, removing the attributes and selecting the appropriate class variable.

2.2 DATA CLEANING

Dataset before cleaning

DRG Definiti	Provider I	Provider N	Provider S	Provider City	Provider S	Provider Zip Co	Hospital Referral Region Description	Total Disc	Average Covered Charges	Average Total Payments	Average Medicare Payment
039 - EXTRAC	10001	SOUTHEA	1108 ROSS	DOTHAN	AL	36301	AL - Dothan	91	$32,963.07	$5,777.24	$4,763.73
039 - EXTRAC	10005	MARSHAL	2505 U S H	BOAZ	AL	35957	AL - Birmingham	14	$15,131.85	$5,787.57	$4,976.71
039 - EXTRAC	10006	ELIZA COF	205 MARE	FLORENCE	AL	35631	AL - Birmingham	24	$37,560.37	$5,434.95	$4,453.79
039 - EXTRAC	10011	ST VINCEN	50 MEDIC/	BIRMINGHAM	AL	35235	AL - Birmingham	25	$13,998.28	$5,417.56	$4,129.16
039 - EXTRAC	10016	SHELBY BA	1000 FIRST	ALABASTER	AL	35007	AL - Birmingham	18	$31,633.27	$5,658.33	$4,851.44
039 - EXTRAC	10023	BAPTIST M	2105 EAST	MONTGOMER	AL	36116	AL - Montgomery	67	$16,920.79	$6,653.80	$5,374.14
039 - EXTRAC	10029	EAST ALA	2000 PEPP	OPELIKA	AL	36801	AL - Birmingham	51	$11,977.13	$5,834.74	$4,761.41
039 - EXTRAC	10033	UNIVERSI	619 SOUTH	BIRMINGHAM	AL	35233	AL - Birmingham	32	$35,841.09	$8,031.12	$5,858.50
039 - EXTRAC	10039	HUNTSVIL	101 SIVLEY	HUNTSVILLE	AL	35801	AL - Huntsville	135	$28,523.39	$6,113.38	$5,228.40
039 - EXTRAC	10040	GADSDEN	1007 GOO	GADSDEN	AL	35903	AL - Birmingham	34	$75,233.38	$5,541.05	$4,386.94
039 - EXTRAC	10046	RIVERVIEW	600 SOUTH	GADSDEN	AL	35901	AL - Birmingham	14	$67,327.92	$5,461.57	$4,493.57
039 - EXTRAC	10055	FLOWERS	4370 WES	DOTHAN	AL	36305	AL - Dothan	45	$39,607.28	$5,356.28	$4,408.20
039 - EXTRAC	10056	ST VINCEN	810 ST VIN	BIRMINGHAM	AL	35205	AL - Birmingham	43	$22,862.23	$5,374.65	$4,186.02
039 - EXTRAC	10078	NORTHEA	400 EAST 1	ANNISTON	AL	36207	AL - Birmingham	21	$31,110.85	$5,366.23	$4,376.23
039 - EXTRAC	10083	SOUTH BA	1613 NOR	FOLEY	AL	36535	AL - Mobile	15	$25,411.33	$5,282.93	$4,383.73
039 - EXTRAC	10085	DECATUR	1201 7TH S	DECATUR	AL	35609	AL - Huntsville	27	$9,234.51	$5,676.55	$4,509.11
039 - EXTRAC	10090	PROVIDEN	6801 AIRP	MOBILE	AL	36608	AL - Mobile	27	$15,895.85	$5,930.11	$3,972.85
039 - EXTRAC	10092	D C H REG	809 UNIVE	TUSCALOOSA	AL	35401	AL - Tuscaloosa	31	$19,721.16	$6,192.54	$5,179.38
039 - EXTRAC	10100	THOMAS H	750 MORP	FAIRHOPE	AL	36532	AL - Mobile	18	$10,710.88	$4,968.00	$3,898.88
039 - EXTRAC	10103	BAPTIST M	701 PRINC	BIRMINGHAM	AL	35211	AL - Birmingham	33	$51,343.75	$5,996.00	$4,962.45
039 - EXTRAC	10104	TRINITY M	800 MONT	BIRMINGHAM	AL	35213	AL - Birmingham	29	$55,219.31	$5,710.31	$4,471.68
039 - EXTRAC	10113	MOBILE IN	5 MOBILE	MOBILE	AL	36652	AL - Mobile	66	$14,948.15	$5,550.90	$4,219.90

Our dataset was made of a list of large number of diseases but for mining to be conducted on the dataset we had to select a disease because we were going to take the hospitals that were there in the dataset and check in which of the three ranges provided will the cost for treatment of the diseases will fall. This would not have been possible for more than one disease. Before the data cleaning process was started we had about 21 lakh records.

INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION was the disease that was selected since it had enough records for mining to be conducted on. After selecting the disease, we were left with approximately 63 thousand records.

2.3 ATTRIBUTE CLEANING

Our Data Cleaning process was mainly based on the attributes. Below is a list of the attributes that the cleaning process was done upon.

- **INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION** was removed because each record was identified based on the Comorbidity and Complexity values.
- **Hospital Referral Region Description**- Hospital Referral Region Description was a single attribute which contained both Referral State and Referral City. We split it into Referral State and Referral City and made it into two separate attributes to check if both independently influenced the output.
- **Average Medicare Payments**- The values that were present in Average Medicare Payments were very different from the rest of the data and it was not providing any information. Therefore, it had to be removed since its values were not in coordination with the rest of the dataset.
- **Average Covered Charges, Average Total Payments**- The selection of our class variable had to be done from Average Covered Charges or Average Total Payments. Hence, in order to select the class variable, we made 3 different files and ran algorithms like Zero R, One R, J48, Naïve Bayes algorithms on them to get the file which had the highest accuracy

Files that were made had class variables as:

- Average Total Payments
- Average Covered Charges
- Average Total Payments + Average Covered Charges

The results showed that accuracy of Average Total Payments was the highest. Therefore, the file with class variable as Average Total Payments was selected and the remaining two files were discarded.

2.4 DERIVED ATTIBUTES

Comorbidity and Complexity (Severity) of INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION was added. It is the extent or the severity of the disease. This was added because the disease INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION had three different levels of Comorbidity and Complexity they were:

a) **MCC- Major Comorbidity and Complexity**: This is the highest level of Comorbidity and Complexity. It is the most severe and serious condition of INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION.

b) **CC- Comorbidity and Complexity**: This is the medium level of Comorbidity and Complexity. It is the not as severe condition of INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION as MCC.

c) **W/O CC- Without Comorbidity and Complexity**: This signifies that there is no Comorbidity and Complexity but INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION is present.

2.5 DATASET AFTER CLEANING

Comorbidity	Provider Id	Provider Name	Provider Street Address	Provider City	Provider State	Provider Zip Code	Referral State	Referral City	Total Discharges	Average Total Payments
MCC	10001	SOUTHEAST ALABAMA MEDICAL CEI	1108 ROSS CLARK CIRCLE	DOTHAN	AL	36301	AL	Dothan	84	10,260.21
MCC	10005	MARSHALL MEDICAL CENTER SOUTH	2505 U S HIGHWAY 431 NORTH	BOAZ	AL	35957	AL	Birmingham	13	10,562.23
MCC	10006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	FLORENCE	AL	35631	AL	Birmingham	32	10,439.46
MCC	10010	MARSHALL MEDICAL CENTER NORTH	8000 ALABAMA HIGHWAY 69	GUNTERSVILLE	AL	35976	AL	Huntsville	21	9,116.66
MCC	10011	ST VINCENT S EAST	50 MEDICAL PARK EAST DRIVE	BIRMINGHAM	AL	35235	AL	Birmingham	13	10,174.69
MCC	10016	SHELBY BAPTIST MEDICAL CENTER	1000 FIRST STREET NORTH	ALABASTER	AL	35007	AL	Birmingham	21	10,256.90
MCC	10019	HELEN KELLER MEMORIAL HOSPITAL	1300 SOUTH MONTGOMERY A	SHEFFIELD	AL	35660	AL	Birmingham	13	10,071.23
MCC	10023	BAPTIST MEDICAL CENTER SOUTH	2105 EAST SOUTH BOULEVARD	MONTGOMER	AL	36116	AL	Montgomery	48	12,314.04
MCC	10024	JACKSON HOSPITAL & CLINIC INC	1725 PINE STREET	MONTGOMER	AL	36106	AL	Montgomery	36	10,746.80
MCC	10025	GEORGE H. LANIER MEMORIAL HOS	4800 48TH ST	VALLEY	AL	36854	AL	Birmingham	11	10,442.54
MCC	10029	EAST ALABAMA MEDICAL CENTER A	2000 PEPPERELL PARKWAY	OPELIKA	AL	36801	AL	Birmingham	34	10,584.79
MCC	10033	UNIVERSITY OF ALABAMA HOSPITA	619 SOUTH 19TH STREET	BIRMINGHAM	AL	35233	AL	Birmingham	119	14,728.40
MCC	10035	CULLMAN REGIONAL MEDICAL CENT	1912 ALABAMA HIGHWAY 157	CULLMAN	AL	35058	AL	Birmingham	24	10,313.79
MCC	10039	HUNTSVILLE HOSPITAL	101 SIVLEY RD	HUNTSVILLE	AL	35801	AL	Huntsville	178	11,501.20
MCC	10040	GADSDEN REGIONAL MEDICAL CENT	1007 GOODYEAR AVENUE	GADSDEN	AL	35903	AL	Birmingham	18	9,908.50
MCC	10045	FAYETTE MEDICAL CENTER	1653 TEMPLE AVENUE NORTH	FAYETTE	AL	35555	AL	Tuscaloosa	11	10,979.18
MCC	10046	RIVERVIEW REGIONAL MEDICAL CEN	600 SOUTH THIRD STREET	GADSDEN	AL	35901	AL	Birmingham	16	9,374.81
MCC	10055	FLOWERS HOSPITAL	4370 WEST MAIN STREET	DOTHAN	AL	36305	AL	Dothan	26	13,016.23
MCC	10056	ST VINCENT S BIRMINGHAM	810 ST VINCENT S DRIVE	BIRMINGHAM	AL	35205	AL	Birmingham	42	10,162.76
MCC	10059	LAWRENCE MEDICAL CENTER	202 HOSPITAL STREET	MOULTON	AL	35650	AL	Birmingham	12	10,921.91
MCC	10078	NORTHEAST ALABAMA REGIONAL M	400 EAST 10TH STREET	ANNISTON	AL	36207	AL	Birmingham	30	10,349.56
MCC	10085	DECATUR GENERAL HOSPITAL	1201 7TH STREET SE	DECATUR	AL	35609	AL	Huntsville	25	9,477.32

3. CLASSIFIER SELECTION

After data cleansing, & deriving attributes, we tried several algorithms. Few important are mentioned below-

- Zero R-33.345% (Benchmark Accuracy)
- Naive Bayes (Bayes)
- J48 (Tree)
- OneR (Rule)
- Stacking (Meta)
- IBk (Lazy)

Above algorithms worked as mentioned below-

a) Zero R-33.345% (Benchmark Accuracy):

As Zero R forms, only one rule with maximum frequency in the class attribute; medium (middle range) was chosen for classification. In our case, medium (middle range) had an extra instance compared to high and low, and gave the accuracy of 33.35%, which we made our benchmark.

b) Naive Bayes (Bayes)

Naïve Bayes which belongs to Bayes family of classifier that takes probabilistic approach for classifying.

c) J48 (Tree)

J48 tree classifier which works on Decision tree learning. The tree it created for our dataset, contained only one attribute i.e.; if disease was MCC then it would classify high, for CC it would be classify as medium range, & for w/o CC and MCC it would result in Low.

d) OneR (Rule)

OneR uses attribute with minimum error or variance for classification. This also used same attribute as in J48. Hence it gave same accuracy as J48 when tested without noise.

e) Stacking (Meta)

Stacking usually combines several algorithms to come up with a better algorithm. But we observed that used ZeroR & hence gave us accuracy equal to ZeroR which was our benchmark.

f) IBk

IBk uses k nearest neighbor classifier, that gave us accuracy about 22.58%.

g) Decision Table

Class for building and using a simple decision table majority classifier

3.1 FOUR CELL EXPERIMENT DESIGN

- **Two Factor Design**: Our experiment design contained of two factors:
  1. Factor 1 (F1) → **Attributes with or without noise**
  2. Factor 2 (F2) → **Percentage Split**

- **Four Criteria of the Design**: The two factors are to be divided up into 4 criteria by keeping one factor constant and varying the other factor between two values and vice versa. This is illustrated more clearly in the table blow.

Percentage Split	Without Noise	With Noise
20%/80%	C1	C3
80%/20%	C2	C4

a. **F11, C1**= Attributes without Noise + Percentage Split of 20%/80%
b. **F12, C2**= Attributes without Noise + Percentage Split of 80%/20%
c. **F21, C3**= Attributes with Noise + Percentage Split of 20%/80%
d. **F22, C4**= Attributes with Noise + Percentage Split of 80%/20%

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

4. EXPERIMENT RESULTS

4.1 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 20%:80%
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 20%:80%
E4= Performance of Naïve Bayes when, Attributes with noise + Percentage Split of 80%:20%
E5= Performance of J48 when, Attributes without noise + Percentage Split of 20%:80%
E6= Performance of J48 when, Attributes without noise + Percentage Split of 80%:20%
E7= Performance of J48 when, Attributes with noise + Percentage Split of 20%:80%
E8= Performance of J48 when, Attributes with noise + Percentage Split of 80%:20%
E9= Performance of OneR when, Attributes without noise + Percentage Split of 20%:80%
E10= Performance of OneR when, Attributes without noise + Percentage Split of 80%:20%
E11= Performance of OneR when, Attributes with noise + Percentage Split of 20%:80%
E12= Performance of OneR when, Attributes with noise + Percentage Split of 80%:20%

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

1. **Naïve Bayes:**

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

E1 – Without Noise with 20-80 split.

E2 – Without Noise with 80-20 split.

E3 – With Noise with 20-80 split.

E4 – With Noise with 80-20 split.

	E1(NB 20:80)	
trails	seeds	O/P
1	1	66.4269
2	2	66.6667
3	3	66.9283
4	4	64.3558
5	5	65.1624
6	6	66.6013
7	7	65.7946
8	8	66.0998
9	9	64.6392
10	10	66.4923
	AVG	**65.91673**
	VARIANCE	**0.814660965**

	E2(NB 80:20)	
trails	seeds	O/P
1	1	56.5824
2	2	55.9721
3	3	55.9721
4	4	57.9773
5	5	56.6696
6	6	57.6286
7	7	56.8439
8	8	56.7568
9	9	57.803
10	10	61.9878
	AVG	**57.41936**
	VARIANCE	**3.066089687**

	E3(NB 20:80)	
trials	seeds	O/P
1	1	57.2705
2	2	56.2677
3	3	56.2677
4	4	53.8478
5	5	55.8099
6	6	54.5891
7	7	54.9597
8	8	55.5483
9	9	54.8943
10	10	55.7881
	AVG	**55.52431**
	VARIANCE	**0.967462201**

	E4(NB 80:20)	
trials	seeds	O/P
1	1	43.6792
2	2	47.5153
3	3	46.0331
4	4	45.51
5	5	45.51
6	6	45.0741
7	7	42.197
8	8	45.1613
9	9	47.0793
10	10	48.1255
	AVG	**45.58848**
	VARIANCE	**3.130690164**

We got the highest accuracy in E1, i.e. with 20-80 split without noise.

Similarly, we ran eight more experiments, four for J48 algorithm, and the remaining four using OneR.

2. J48

E5 – Without Noise with 20-80 split.

E6 – Without Noise with 80-20 split.

E7 – With Noise with 20-80 split.

E8 – With Noise with 80-20 split.

	E5(J48-20:80)	
trials	seeds	O/P
1	1	78.5263
2	2	78.3301
3	3	78.4173
4	4	78.8097
5	5	78.7225
6	6	78.6135
7	7	78.8315
8	8	78.6571
9	9	78.2865
10	10	78.5481
	AVG	**78.57426**
	VARIANCE	**0.035780292**

	E6(J48-80:20)	
trials	seeds	O/P
1	1	78.2912
2	2	77.5937
3	3	78.204
4	4	78.9887
5	5	77.9425
6	6	79.3374
7	7	79.2502
8	8	79.2502
9	9	80.4708
10	10	77.245
	AVG	**78.65737**
	VARIANCE	**0.948943376**

	E7(J48-20:80)	
trials	seeds	O/P
1	1	71.8334
2	2	71.5936
3	3	71.4192
4	4	71.8988
5	5	71.8552
6	6	71.6808
7	7	71.8552
8	8	71.986
9	9	71.4628
10	10	71.7244
	Avg	**71.73094**
	VARIANCE	**0.03612352**

	E8(J48-80:20)	
trials	seeds	O/P
1	1	72.7986
2	2	70.8806
3	3	70.1831
4	4	71.7524
5	5	71.1421
6	6	70.8806
7	7	72.1883
8	8	72.1011
9	9	73.6704
10	10	70.4446
	AVG	**71.60418**
	VARIANCE	**1.214527462**

We got highest accuracy in E6, i.e. with 80-20 split without noise.

3. OneR

E9 – Without Noise with 20-80 split.

E10 – Without Noise with 80-20 split.

E11 – With Noise with 20-80 split.

E12 – With Noise with 80-20 split.

E9(OneR 20:80)

trials	seeds	O/P
1	1	28.6898
2	2	27.7305
3	3	28.5154
4	4	28.3192
5	5	29.213
6	6	28.9514
7	7	28.9078
8	8	27.9922
9	9	28.0358
10	10	28.6898
	AVERAGE	**28.50449**
	VARIANCE	**0.227630757**

E10(OneR 80:20)

trials	seeds	O/P
1	1	78.2912
2	2	77.5937
3	3	78.204
4	4	78.9887
5	5	77.9425
6	6	79.3374
7	7	79.2502
8	8	79.2502
9	9	80.4708
10	10	77.245
	AVERAGE	**78.65737**
	VARIANCE	**0.948943376**

E11 OneR(20:80)

TRIALS	SEEDS	O/P
1	1	30.1068
2	2	29.2784
3	3	29.6272
4	4	29.1476
5	5	30.8698
6	6	29.867
7	7	29.7798
8	8	29.5618
9	9	29.3438
10	10	29.976
	Average	**29.75582**
	Variance	**0.249706937**

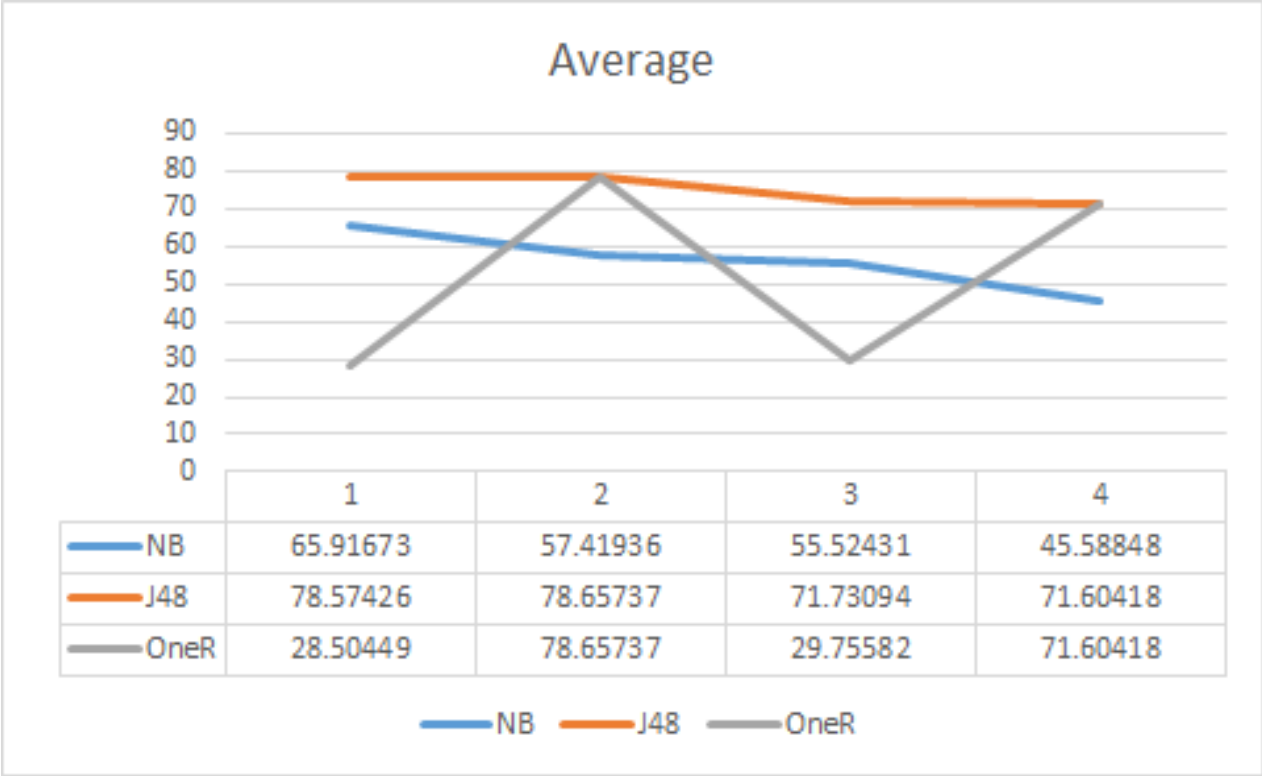
E12 OneR(80:20)

trial	seeds	O/P
1	1	72.7986
2	2	70.8806
3	3	70.1831
4	4	71.7524
5	5	71.1421
6	6	70.8806
7	7	72.1883
8	8	72.1011
9	9	73.6704
10	10	70.4446
	Average	**71.60418**
	Variance	**1.214527462**

We got highest accuracy in E10, i.e. with 80-20 split without noise.

4.2 SUMMARY OF RESULTS

- **ACCURACY GRAPH**



	1	2	3	4
NB	65.91673	57.41936	55.52431	45.58848
J48	78.57426	78.65737	71.73094	71.60418
OneR	28.50449	78.65737	29.75582	71.60418

As the graph above shows, the average accuracy was the highest for J48 overall.

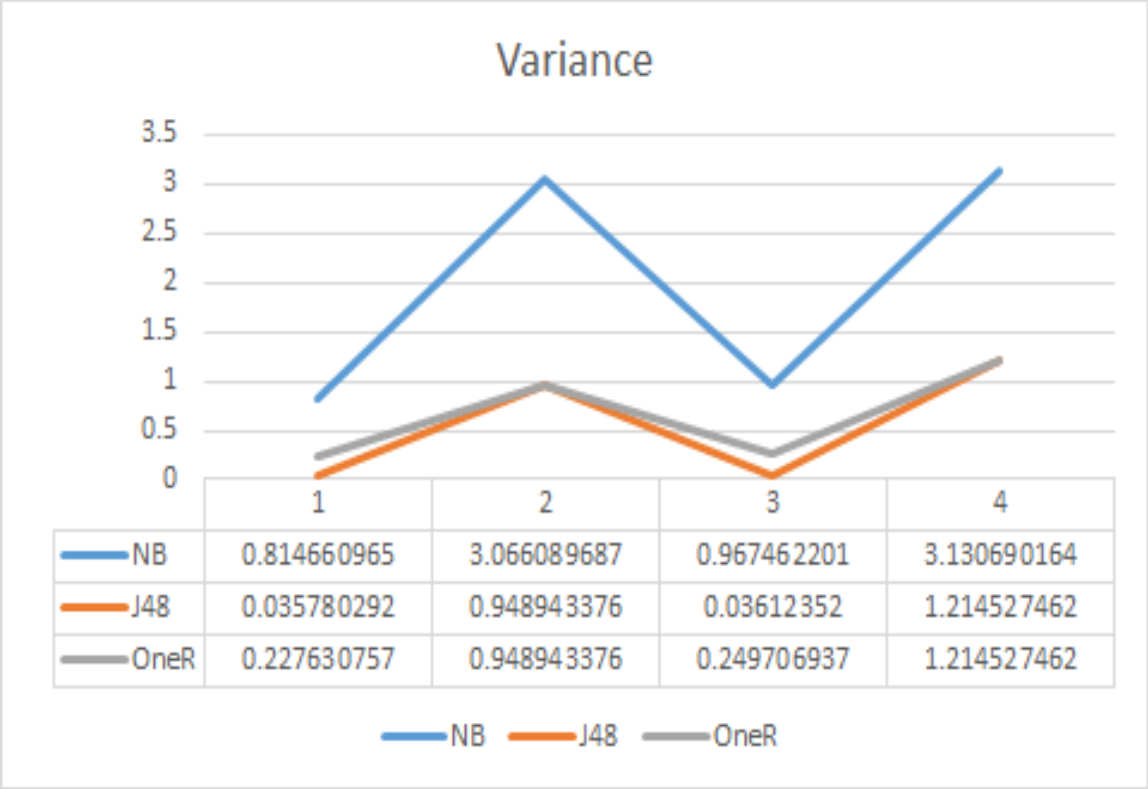
Interpretation of Accuracy

OneR uses simple rules generation technique to evaluate the dataset. The result of the experiment for attributes with noise and without noise shows that accuracy of OneR fluctuates a lot.

Naives Bayes on the other hand make use of probabilistic approach to analyse the dataset. With increase in noise the accuracy decreases by considerable amount.

J48 tree classifier which works on Decision tree learning. The tree it created for our dataset, contained only one attribute i.e.; if disease was MCC then it would classify high, for CC it would be classify as medium range, & for w/o CC and MCC it would result in Low. Due to this, even after addition of noise there is no significant effect on the accuracy.

- **VARIANCE**



As the graph above shows, the average variance was the lowest for J48 overall.

Interpretation of Variance

OneR: With noise the variance increase significantly

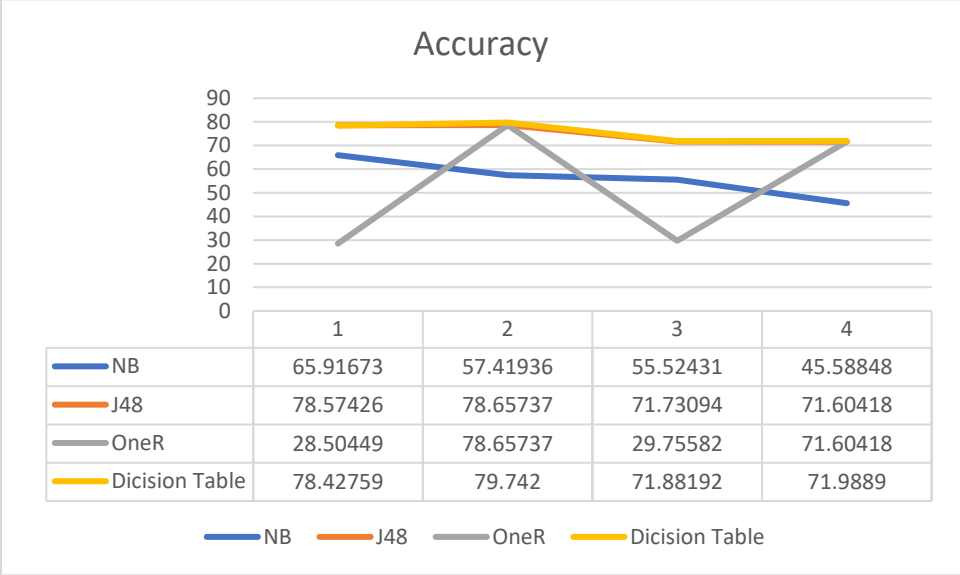
Naives Bayes: With increase in noise the variance increases considerably

J48: With increase in noise the variance increases but in amount.

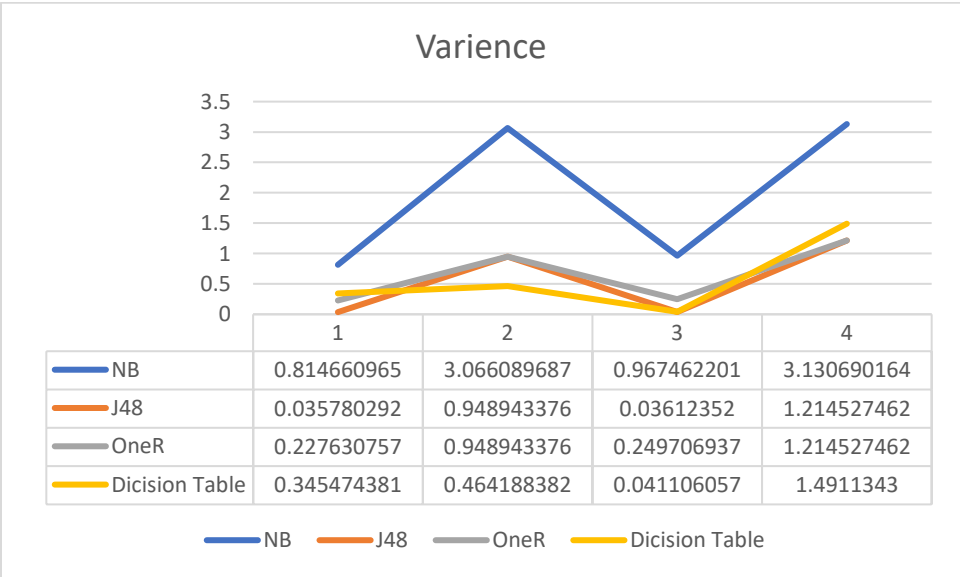
5. ANALYSIS & CONCLUSION

5.1 SUGGESSTION

On later period, while conducting experiment we recognize with the help of Dr. Sikora that the results of OneR are too fluctuating and needs to be reconsidered. Decision tree classifier generates rules for classifying each attribute. While using Decision tree as a classifier the results of the experiments were as shown below.



	1	2	3	4
NB	65.91673	57.41936	55.52431	45.58848
J48	78.57426	78.65737	71.73094	71.60418
OneR	28.50449	78.65737	29.75582	71.60418
Dicision Table	78.42759	79.742	71.88192	71.9889



	1	2	3	4
NB	0.814660965	3.066089687	0.967462201	3.130690164
J48	0.035780292	0.948943376	0.03612352	1.214527462
OneR	0.227630757	0.948943376	0.249706937	1.214527462
Dicision Table	0.345474381	0.464188382	0.041106057	1.4911343

If we compare the accuracy in above chart, J48 and Decision tree give approximately similar results, but variance charts indicates that J48 still gives better results than decision tree.

- **10 Fold Cross Validation Results by removing 1st attribute**

  As in most of the runs our Comorbidity attribute was playing much important role, we tried to run removing that attribute. The results found were as below-

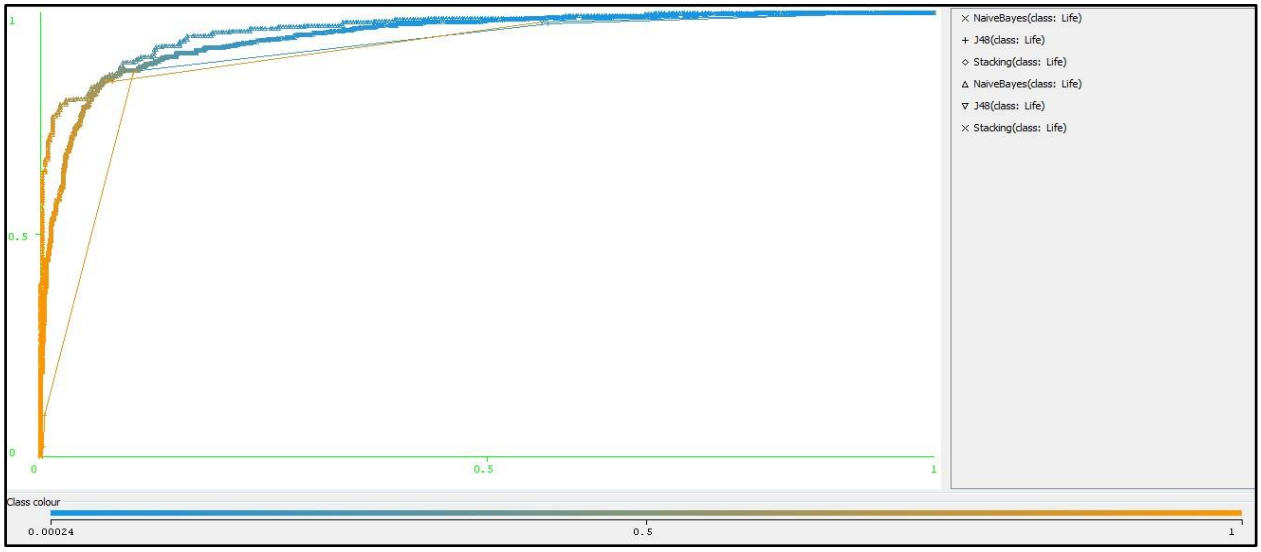
	Accuracy
ZeroR	33.3101
OneR	15.0506
Naïve Bayes	26.4737
J48	34.1472
Dicision Table	42.2393

5.1 RECEIVER OPERATING CHARACTERISTIC

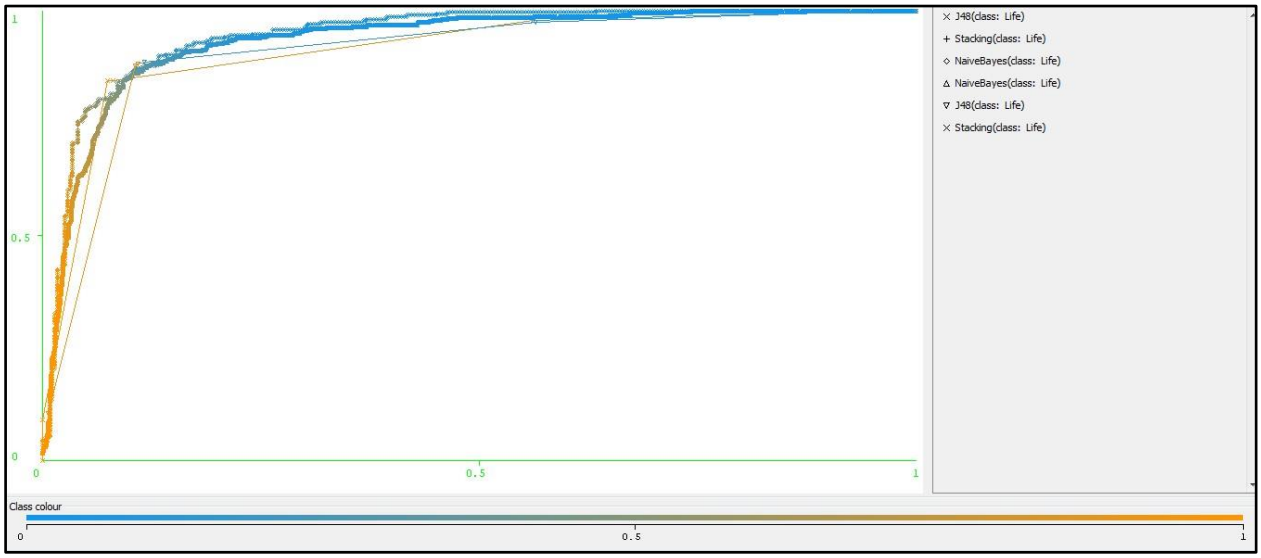
Developed in 1950s for signal detection theory to analyze noisy signal. ROC characterize the trade-off between positive hits and false alarms. It plots True positive (TP) on Y- axis and false positive on X-axis. The ROC graph for Inpatient datasets is as follows:

We generated two multiple ROC curves using this feature above:

1. ROC Curve for **Attributes without noise** (both 20%/80% and 80%/20% split)



2. ROC Curve for **Attributes with noise** (both 20%/80% and 80%/20% split)



A quick analysis of these ROC curves helps us infer that the efficiency of '**All Attributes**' is higher than selected attributes because the *Area under the ROC curve is larger* for all attributes.

5.2 CLASSIFIER ANALYSIS

As per the results of the experiments conducted, we can say that

- We can say that, J48 has the highest accuracy when there is no noise.
- J48 has the highest accuracy and lowest variance. It is very good for prediction and it does not over fit for large data sets.
- Naïve Bayes does not work to the expected level, when there is addition of noise.
- OneR failing in creation of rules with addition of noise resulting in fluctuating accuracy.

5.3 CONCLUSION

With the added noise, percentage split, discretize data and average accuracy, variance graphs, we recommend the following for our data set.

- When there is no noise and without data discretion, all the classifiers performed well.
- Noise is the deciding factor of accuracy percentage in our data set.
- J48 is the best classifier with low variance and high accuracy.

6. REFERENCES

- http://**betterevaluation**.org/sites/default/files/data_cleaning.pdf
- **Google Refine** [Online] / auth. Authors Multiple // Google. - Google Refine, 7 1, 2013. - https://code.google.com/p/google-refine/.
- http://www.**gbif**.org/system/files_force/gbif_resource/resource-80528/Principles%20and%20Methods%20of%20Data%20Cleaning%20-%20ENGLISH.pdf?download=1
- **Making Sense of Data** [Online] / auth. Authors Multiple // Google . - Google, 3 1, 2014. - https://datasense.withgoogle.com/course.
- **Multiple ROC Curves** [Online] / auth. Shams Rushdi // Youtube. - 10 28, 2013. - https://www.youtube.com/watch?v=rZHw3gGe7DA.