Readmission Risk Analysis

Introduction

In this report, predictive analysis model to detect premature discharge of patients from the hospital is described in detail.

Health Institute loose millions due to readmission of patients discharged prematurely. A model is developed to predict these patients. This model will be used by doctors to identify such patients and reduce the risk of premature discharge.

A dataset of 10000 patient shared by hospital was used to train and test the model.

Methodology

Data is provided from the Hospital; hence data collection was not performed.

Firstly, the collected data was preprocessed to remove unnecessary details and make data uniform.

Data was then visualized in a form of pie chat to analyze the skewness of the data.

Apriori algorithm was then used to discover interesting patterns in the data and to establish relationship between different variables.

Provided data is then split into test and train data and three predictive algorithm: Decision Tree, Gaussian Naïve Bayes and Logistic Regression is used for prediction and their effectiveness is further analyzed.

Findings

Data Preprocessing:

Data preprocessing was done in two phases:

In first phase, we tried to clean the data by removing the columns that had more than 50% of the missing data. For this, we first replaced all values corresponding to missing data like “?”, “None”, “Not Available”, “Not Mapped” by “NaN”. Then for each column we calculated the percentage of “NaN”, and we dropped the column for which missing records was more than 50%.

Output:

```
Total complete data for race 9779
Percentage of complete data = 97.78999999999999
Total complete data for gender 10000
Percentage of complete data = 100.0
Total complete data for age 10000
Percentage of complete data = 100.0
Total complete data for admission_type_id 8626
Percentage of complete data = 86.26
Total complete data for discharge_disposition_id 9396
Percentage of complete data = 93.96
Total complete data for admission_source_id 9027
Percentage of complete data = 90.27
Total complete data for time_in_hospital 10000
Percentage of complete data = 100.0
Total complete data for num_lab_procedures 10000
Percentage of complete data = 100.0
Total complete data for num_procedures 10000
Percentage of complete data = 100.0
Total complete data for num_medications 10000
Percentage of complete data = 100.0
Total complete data for number_outpatient 10000
Percentage of complete data = 100.0
Total complete data for number_emergency 10000
Percentage of complete data = 100.0
Total complete data for number_inpatient 10000
Percentage of complete data = 100.0
Total complete data for diag_1 9998
Percentage of complete data = 99.98
Total complete data for diag_2 9941
Percentage of complete data = 99.41
Total complete data for diag_3 9791
Percentage of complete data = 97.91
Total complete data for number_diagnoses 9999
```

```
Percentage of complete data = 99.99

Total complete data for max_glu_serum 664

Percentage of complete data = 6.64

Total complete data for A1Cresult 1621

Percentage of complete data = 16.21

Total complete data for diabetesMed 10000

Percentage of complete data = 100.0

Total complete data for readmitted 10000

Percentage of complete data = 100.0

Columns with more than 50% missing data : ['max_glu_serum', 'A1Cresult']

max_glu_serum removed from data frame.

A1Cresult removed from data frame.
```

As we can see two columns, namely “max_glu_serum” and “A1Cresult” had less that 50% of complete data hence they were removed.

In second phase, we imputed the missing values.

From above output we can see that following columns have incomplete data:

"admission_type_id", "discharge_disposition_id", "admission_source_id", "diag_1", "diag_2", "diag_3", "number_diagnoses"

We used the **mode** of each of the above column to impute the missing values.

As there was no fixed pattern for the values in these columns and a fixed relationship between these values and the nearby values was difficult to establish, mode was used to impute. Precedence is given to value that repeats the most as it was generally the case in the data. Also, as the values were both numerical and string mode was easier to implement as well.

Exploratory Analysis

Visualization

Exploratory analysis is done to visualize some basic patterns present on the data. Here we find how much the data is skewed. We find the percentage of certain columns and use matplotlib to plot pie charts to visualize how much data is skewed.

Output:

discharge_disposition_id

```
Discharged to home                                              66.60
Discharged/transferred to SNF                                   11.90
Discharged/transferred to home with home health service         11.60
Expired                                                          1.95
```

Discharged/transferred to another short term hospital	1.85
Discharged/transferred to another rehab fac including rehab units of a hospital.	1.71
Discharged/transferred to another type of inpatient care institution	1.55
Discharged/transferred to ICF	0.99
Discharged/transferred to a long term care hospital.	0.49
Left AMA	0.45
Hospice / home	0.33
Hospice / medical facility	0.28
Discharged/transferred to home under care of Home IV provider	0.14
Discharged/transferred/referred to a psychiatric hospital of a psychiatric distinct part unit of a hospital	0.13
Discharged/transferred within this institution to Medicare approved swing bed	0.03

Readmitted

No	60.35
Yes	39.65

diabetesMed

Yes	74.78
No	25.22

race

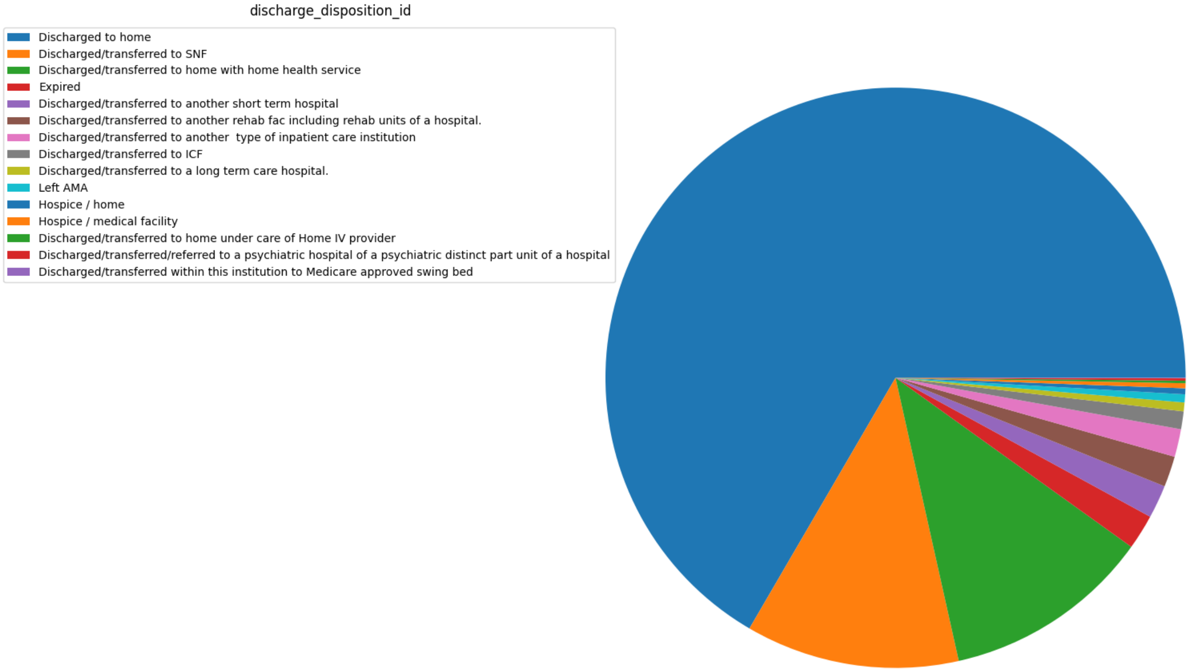
Caucasian	75.253093
AfricanAmerican	21.096227
Hispanic	1.850905
Other	1.237345
Asian	0.562430

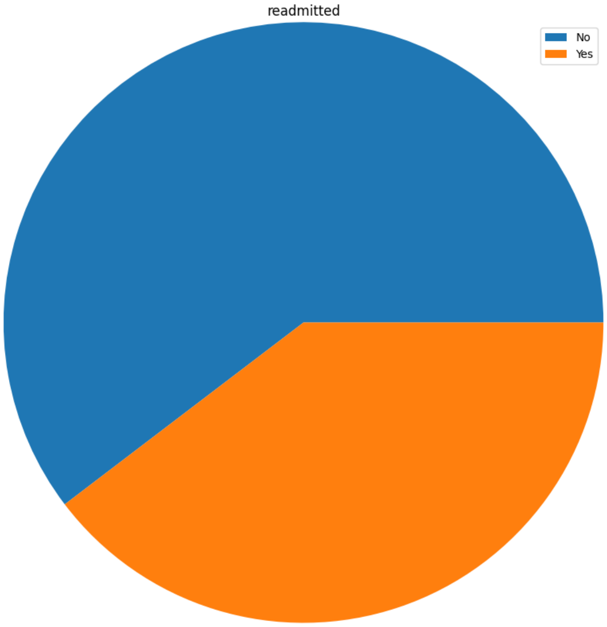
gender

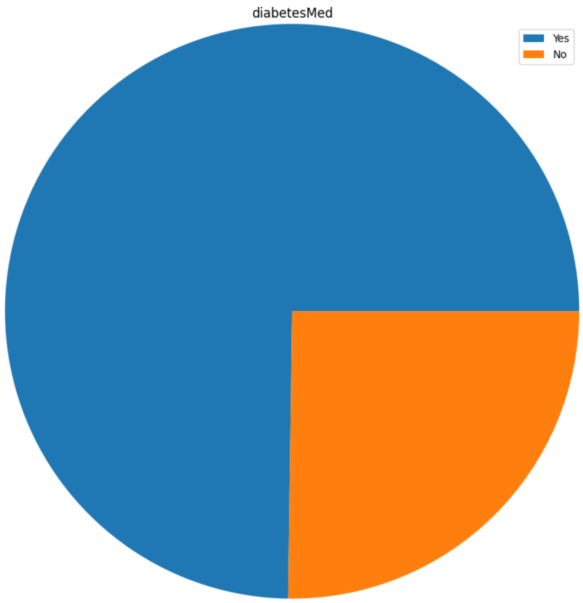
Female	53.98
Male	46.02

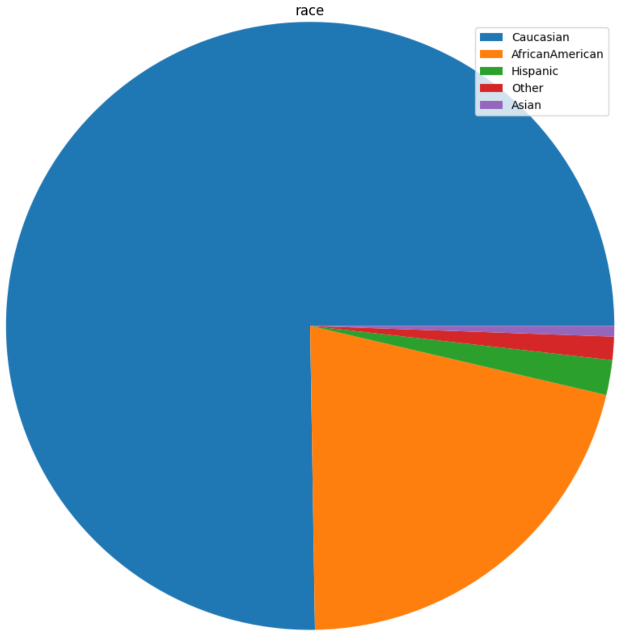
age

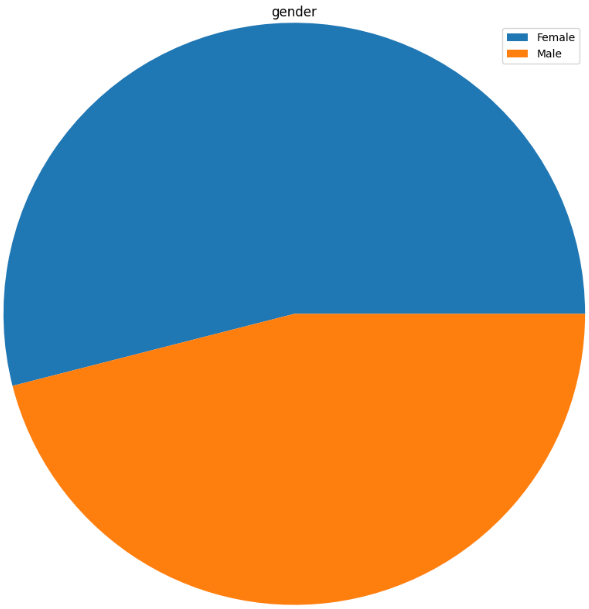
[70-80)	25.95
[60-70)	21.87
[50-60)	17.22
[80-90)	15.77
[40-50)	9.78
[30-40)	4.17
[90-100)	2.68
[20-30)	1.43
[10-20)	0.82
[0-10)	0.31

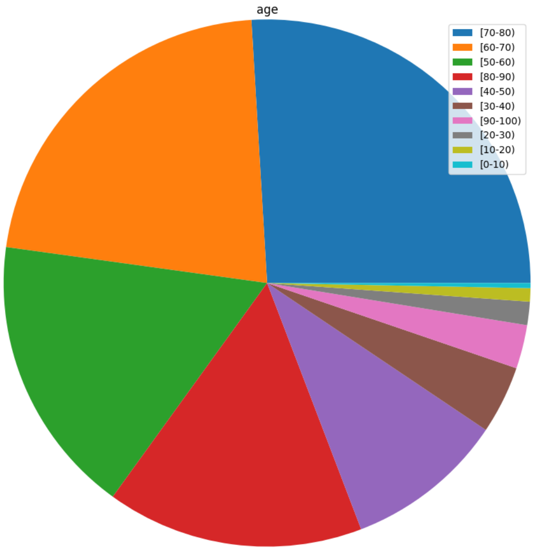

discharge_disposition_id
Discharged to home
Discharged/transferred to SNF
Discharged/transferred to home with home health service
Expired
Discharged/transferred to another short term hospital
Discharged/transferred to another rehab fac including rehab units of a hospital.
Discharged/transferred to another type of inpatient care institution
Discharged/transferred to ICF
Discharged/transferred to a long term care hospital.
Left AMA
Hospice / home
Hospice / medical facility
Discharged/transferred to home under care of Home IV provider
Discharged/transferred/referred to a psychiatric hospital of a psychiatric distinct part unit of a hospital
Discharged/transferred within this institution to Medicare approved swing bed


readmitted
No
Yes


diabetesMed
🟦 Yes
🟧 No


race
🟦 Caucasian
🟧 AfricanAmerican
🟩 Hispanic
🟥 Other
🟪 Asian


gender
Female
Male

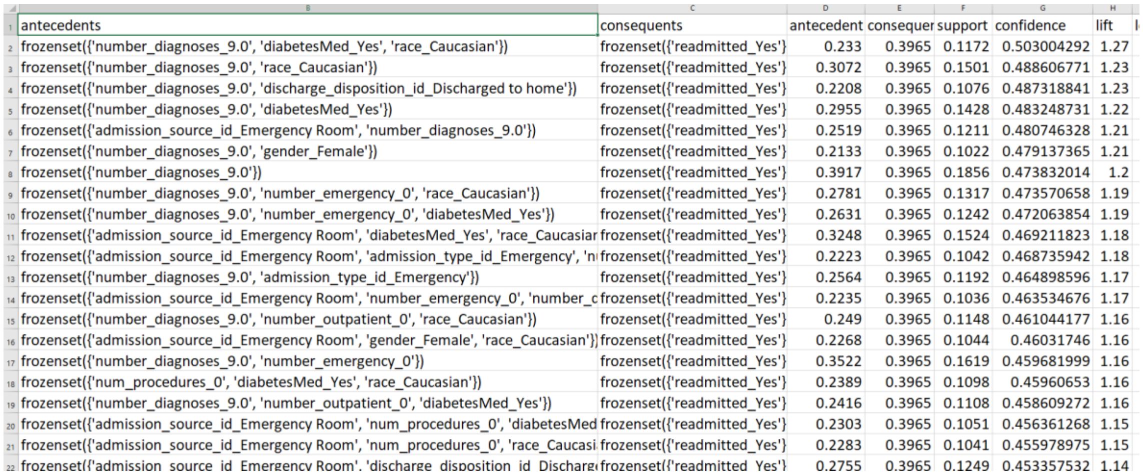

age
[70-80)
[60-70)
[50-60)
[80-90)
[40-50)
[30-40)
[90-100)
[20-30)
[10-20)
[0-10)

Pattern Mining

Apriori algorithm and association rule is used find the relation between different attributes and person being readmitted or not, and rules that cause patient to pass away.

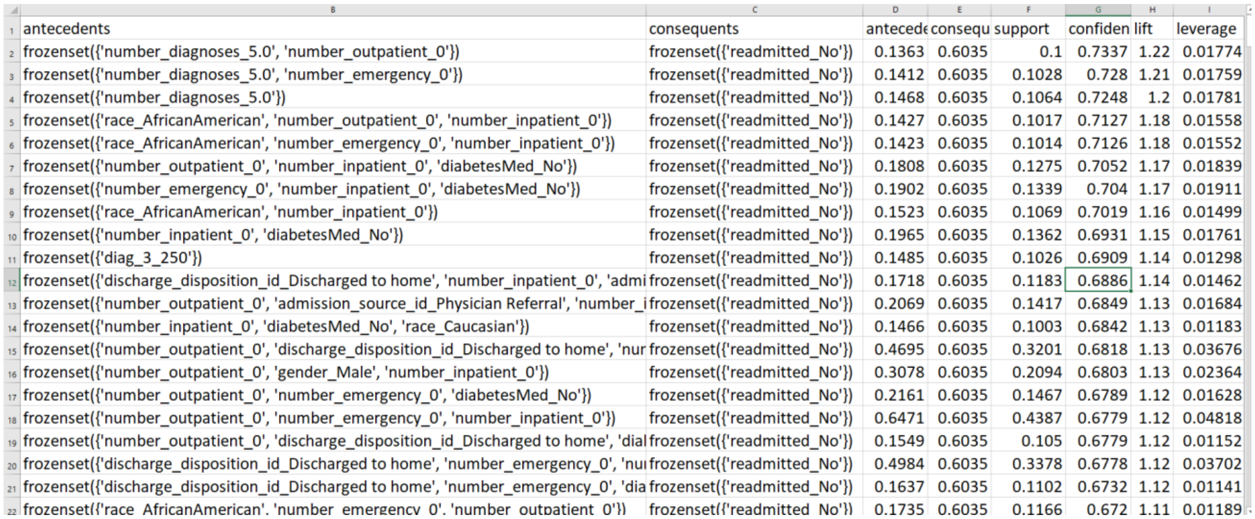
Full result provided in readmitted_yes_rules.csv file alongside report.

Output and discussion:



antecedents	consequents	antecedent	consequer	support	confidence	lift	l
frozenset({'number_diagnoses_9.0', 'diabetesMed_Yes', 'race_Caucasian'})	frozenset({'readmitted_Yes'}	0.233	0.3965	0.1172	0.503004292	1.27	
frozenset({'number_diagnoses_9.0', 'race_Caucasian'})	frozenset({'readmitted_Yes'}	0.3072	0.3965	0.1501	0.488606771	1.23	
frozenset({'number_diagnoses_9.0', 'discharge_disposition_id_Discharged to home'})	frozenset({'readmitted_Yes'}	0.2208	0.3965	0.1076	0.487318841	1.23	
frozenset({'number_diagnoses_9.0', 'diabetesMed_Yes'})	frozenset({'readmitted_Yes'}	0.2955	0.3965	0.1428	0.483248731	1.22	
frozenset({'admission_source_id_Emergency Room', 'number_diagnoses_9.0'})	frozenset({'readmitted_Yes'}	0.2519	0.3965	0.1211	0.480746328	1.21	
frozenset({'number_diagnoses_9.0', 'gender_Female'})	frozenset({'readmitted_Yes'}	0.2133	0.3965	0.1022	0.479137365	1.21	
frozenset({'number_diagnoses_9.0'})	frozenset({'readmitted_Yes'}	0.3917	0.3965	0.1856	0.473832014	1.2	
frozenset({'number_diagnoses_9.0', 'number_emergency_0', 'race_Caucasian'})	frozenset({'readmitted_Yes'}	0.2781	0.3965	0.1317	0.473570658	1.19	
frozenset({'number_diagnoses_9.0', 'number_emergency_0', 'diabetesMed_Yes'})	frozenset({'readmitted_Yes'}	0.2631	0.3965	0.1242	0.472063854	1.19	
frozenset({'admission_source_id_Emergency Room', 'diabetesMed_Yes', 'race_Caucasiar	frozenset({'readmitted_Yes'}	0.3248	0.3965	0.1524	0.469211823	1.18	
frozenset({'admission_source_id_Emergency Room', 'admission_type_id_Emergency', 'n	frozenset({'readmitted_Yes'}	0.2223	0.3965	0.1042	0.468735942	1.18	
frozenset({'number_diagnoses_9.0', 'admission_type_id_Emergency'})	frozenset({'readmitted_Yes'}	0.2564	0.3965	0.1192	0.464898596	1.17	
frozenset({'admission_source_id_Emergency Room', 'number_emergency_0', 'number_c	frozenset({'readmitted_Yes'}	0.2235	0.3965	0.1036	0.463534676	1.17	
frozenset({'number_diagnoses_9.0', 'number_outpatient_0', 'race_Caucasian'})	frozenset({'readmitted_Yes'}	0.249	0.3965	0.1148	0.461044177	1.16	
frozenset({'admission_source_id_Emergency Room', 'gender_Female', 'race_Caucasian'})	frozenset({'readmitted_Yes'}	0.2268	0.3965	0.1044	0.46031746	1.16	
frozenset({'number_diagnoses_9.0', 'number_emergency_0'})	frozenset({'readmitted_Yes'}	0.3522	0.3965	0.1619	0.459681999	1.16	
frozenset({'num_procedures_0', 'diabetesMed_Yes', 'race_Caucasian'})	frozenset({'readmitted_Yes'}	0.2389	0.3965	0.1098	0.45960653	1.16	
frozenset({'number_diagnoses_9.0', 'number_outpatient_0', 'diabetesMed_Yes'})	frozenset({'readmitted_Yes'}	0.2416	0.3965	0.1108	0.458609272	1.16	
frozenset({'admission_source_id_Emergency Room', 'num_procedures_0', 'diabetesMed	frozenset({'readmitted_Yes'}	0.2303	0.3965	0.1051	0.456361268	1.15	
frozenset({'admission_source_id_Emergency Room', 'num_procedures_0', 'race_Caucasi	frozenset({'readmitted_Yes'}	0.2283	0.3965	0.1041	0.455978975	1.15	
frozenset({'admission_source_id_Emergency Room', 'discharge_disposition_id_Discharg	frozenset({'readmitted_Yes'}	0.2755	0.3965	0.1249	0.453357532	1.14	

Above we can see readmission is highly probably where the number of diagnoses is 9, and patient is taking diabetes medication and if the patient is Caucasian.



antecedents	consequents	antecede	consequ	support	confiden	lift	leverage
frozenset({'number_diagnoses_5.0', 'number_outpatient_0'})	frozenset({'readmitted_No'})	0.1363	0.6035	0.1	0.7337	1.22	0.01774
frozenset({'number_diagnoses_5.0', 'number_emergency_0'})	frozenset({'readmitted_No'})	0.1412	0.6035	0.1028	0.728	1.21	0.01759
frozenset({'number_diagnoses_5.0'})	frozenset({'readmitted_No'})	0.1468	0.6035	0.1064	0.7248	1.2	0.01781
frozenset({'race_AfricanAmerican', 'number_outpatient_0', 'number_inpatient_0'})	frozenset({'readmitted_No'})	0.1427	0.6035	0.1017	0.7127	1.18	0.01558
frozenset({'race_AfricanAmerican', 'number_emergency_0', 'number_inpatient_0'})	frozenset({'readmitted_No'})	0.1423	0.6035	0.1014	0.7126	1.18	0.01552
frozenset({'number_outpatient_0', 'number_inpatient_0', 'diabetesMed_No'})	frozenset({'readmitted_No'})	0.1808	0.6035	0.1275	0.7052	1.17	0.01839
frozenset({'number_emergency_0', 'number_inpatient_0', 'diabetesMed_No'})	frozenset({'readmitted_No'})	0.1902	0.6035	0.1339	0.704	1.17	0.01911
frozenset({'race_AfricanAmerican', 'number_inpatient_0'})	frozenset({'readmitted_No'})	0.1523	0.6035	0.1069	0.7019	1.16	0.01499
frozenset({'number_inpatient_0', 'diabetesMed_No'})	frozenset({'readmitted_No'})	0.1965	0.6035	0.1362	0.6931	1.15	0.01761
frozenset({'diag_3_250'})	frozenset({'readmitted_No'})	0.1485	0.6035	0.1026	0.6909	1.14	0.01298
frozenset({'discharge_disposition_id_Discharged to home', 'number_inpatient_0', 'admi	frozenset({'readmitted_No'})	0.1718	0.6035	0.1183	0.6886	1.14	0.01462
frozenset({'number_outpatient_0', 'admission_source_id_Physician Referral', 'number_i	frozenset({'readmitted_No'})	0.2069	0.6035	0.1417	0.6849	1.13	0.01684
frozenset({'number_inpatient_0', 'diabetesMed_No', 'race_Caucasian'})	frozenset({'readmitted_No'})	0.1466	0.6035	0.1003	0.6842	1.13	0.01183
frozenset({'number_outpatient_0', 'discharge_disposition_id_Discharged to home', 'nur	frozenset({'readmitted_No'})	0.4695	0.6035	0.3201	0.6818	1.13	0.03676
frozenset({'number_outpatient_0', 'gender_Male', 'number_inpatient_0'})	frozenset({'readmitted_No'})	0.3078	0.6035	0.2094	0.6803	1.13	0.02364
frozenset({'number_outpatient_0', 'number_emergency_0', 'diabetesMed_No'})	frozenset({'readmitted_No'})	0.2161	0.6035	0.1467	0.6789	1.12	0.01628
frozenset({'number_outpatient_0', 'number_emergency_0', 'number_inpatient_0'})	frozenset({'readmitted_No'})	0.6471	0.6035	0.4387	0.6779	1.12	0.04818
frozenset({'number_outpatient_0', 'discharge_disposition_id_Discharged to home', 'dial	frozenset({'readmitted_No'})	0.1549	0.6035	0.105	0.6779	1.12	0.01152
frozenset({'discharge_disposition_id_Discharged to home', 'number_emergency_0', 'nui	frozenset({'readmitted_No'})	0.4984	0.6035	0.3378	0.6778	1.12	0.03702
frozenset({'discharge_disposition_id_Discharged to home', 'number_emergency_0', 'dia	frozenset({'readmitted_No'})	0.1637	0.6035	0.1102	0.6732	1.12	0.01141
frozenset({'race_AfricanAmerican', 'number_emergency_0', 'number_outpatient_0'})	frozenset({'readmitted_No'})	0.1735	0.6035	0.1166	0.672	1.11	0.01189

In case of patient not being readmitted, most of them have number of diagnoses as 5 and number of outpatient visits of the patient in the year preceding the encounter is 0.

The result above is sorted according to Lift.

We use lift to evaluate our model. Lift is the probability of occurrence of given case over all average case. The value of lift is between 0 and infinity. If the lift is greater than 1 means occurrence of the scenario is greater than average. Therefore, higher the value, the relationship is strong.

For the rules that cause the patient to pass away, rules could not be generated with very low threshold of min_support of 0.01 and min_confidence of 0.05. As these criteria already low, any rule generated below these values is considered random is not included.

Predictive Analysis:

In our models, we used 80% of the data for the training, whereas 20% of the data for testing.

We used the following predictive models:

a) Decision Tree
b) Gaussian Naïve Bayes
c) Logistic Regression

Analysis of Models:

a) Decision Tree:

	precision	recall
0	0.65	0.67
1	0.47	0.44
average	0.58	0.58

Confusion matrix for Decision Tree Classifier:

	Predicted 0	Predicted 1
Actual 0	808(TP)	396(FN)
Actual 1	443(FP)	535(TN)

b) Gaussian Naïve Bayes

	precision	recall
0	0.78	0.11
1	0.41	0.95
average	0.63	0.44

Confusion matrix for Gaussian Naïve Bayes:

	Predicted 0	Predicted 1
Actual 0	103(TP)	1074(FN)
Actual 1	37(FP)	759(TN)

c) Logistic Regression

	precision	recall
0	0.68	0.79
1	0.58	0.42
average	0.64	0.65

a) Confusion matrix for Logistic Regression:

	Predicted 0	Predicted 1
Actual 0	955(TP)	249(FN)
Actual 1	459(FP)	337(TN)

Precision gives us how much of the result our model predicted true, is true. As we can see for logistic regression performs better than both decision tree and naïve bayes. However, precision of 0.64 is just a fair result.

Recall gives the percentage of positive records, model predicted true. For this also logistic regression performs better than both decision tree and naïve bayes. Also, 0.65 is just fair.

True positive are records which are true and flagged as true by the model.

False positive are records which are false but flagged true by the model.

False negative are records which are true, but model flagged as false.

True negative are the records which are false and flagged as false by the record.

High TP and less FP means that model is precise.

Similarly, high TP and small FN means model has high recall.

From confusion matrix also we can conclude that the Logistic Regression is a better model than both decision tree and naïve bayes. Also, as FN is very high on naïve bayes, it is considerably bad model.