

Data Analytics of Hospital Readmissions

Sydni Ward

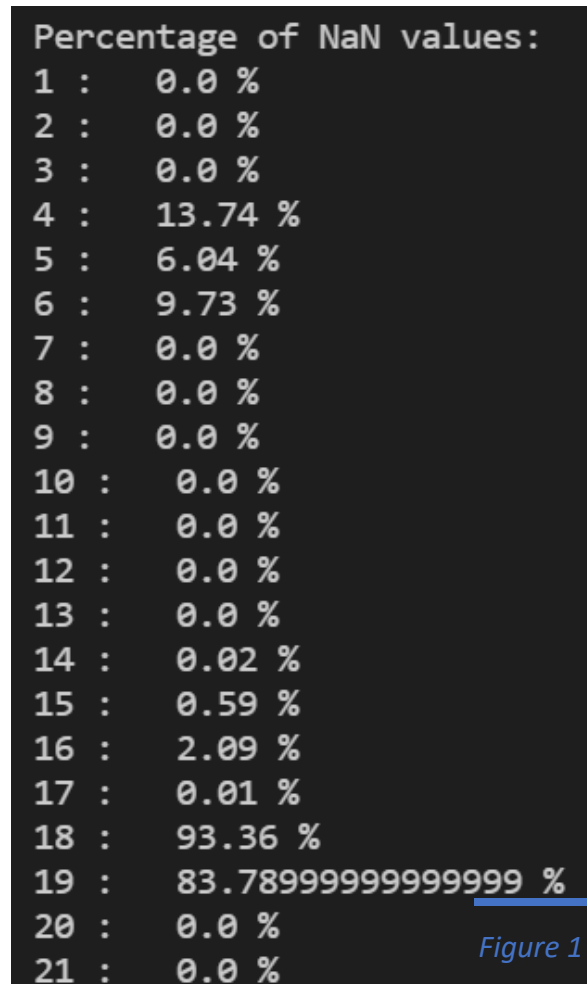
East Tennessee State University

CSCI-4047-001 Data Analytics & Visualization

Professor Ghaith Husari

## Part 1: Data Processing

I loaded the data and set the null values to a standard value, “*np.NaN*”. With a standard value it is easier to calculate the total null values in each column. I removed 2 columns: ‘*max\_glu\_serum*’, ‘*A1Cresult*’, based the high percentage of null value (see *Figure 1*) exceeding 50 percent. After I dropped the unnecessary columns, I imputed the missing values and trimmed the outliers. After weighing the pros and cons of how to impute the missing values I went with mode. Mean and Median are meant for numerical data while mode can be used for numerical and categorical data. Since the data we are filling is categorical, mode is the most appropriate. To trim the outliers, I calculated the upper and lower limit of the data for all the numerical data, I did this by getting the mean and standard deviation for each row. Adding the mean and standard deviation together for the upper limit and subtracting the mean and standard deviation for the lower limit. This removed 354 rows from the data.

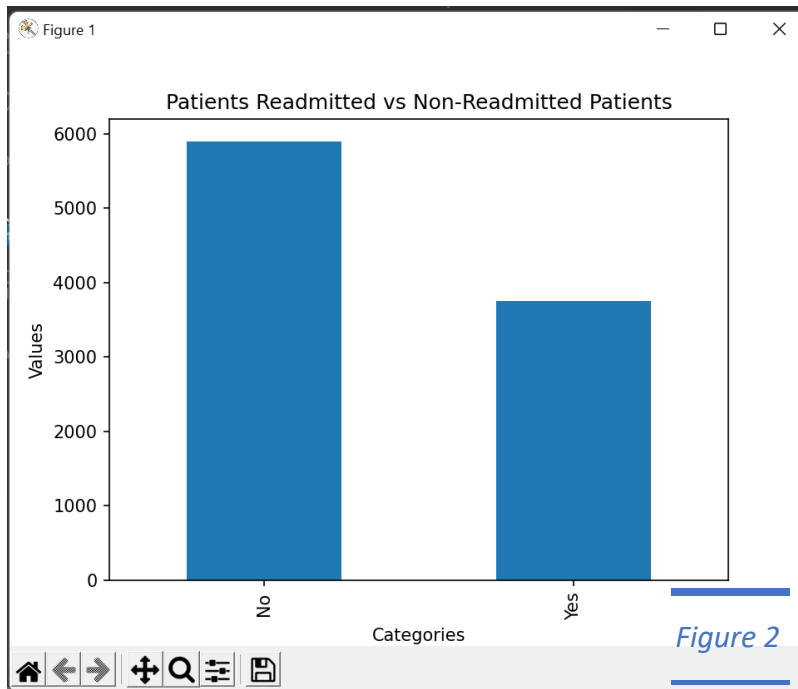


Percentage of NaN values:	
1 :	0.0 %
2 :	0.0 %
3 :	0.0 %
4 :	13.74 %
5 :	6.04 %
6 :	9.73 %
7 :	0.0 %
8 :	0.0 %
9 :	0.0 %
10 :	0.0 %
11 :	0.0 %
12 :	0.0 %
13 :	0.0 %
14 :	0.02 %
15 :	0.59 %
16 :	2.09 %
17 :	0.01 %
18 :	93.36 %
19 :	83.78999999999999 %
20 :	0.0 %
21 :	0.0 %

Figure 1

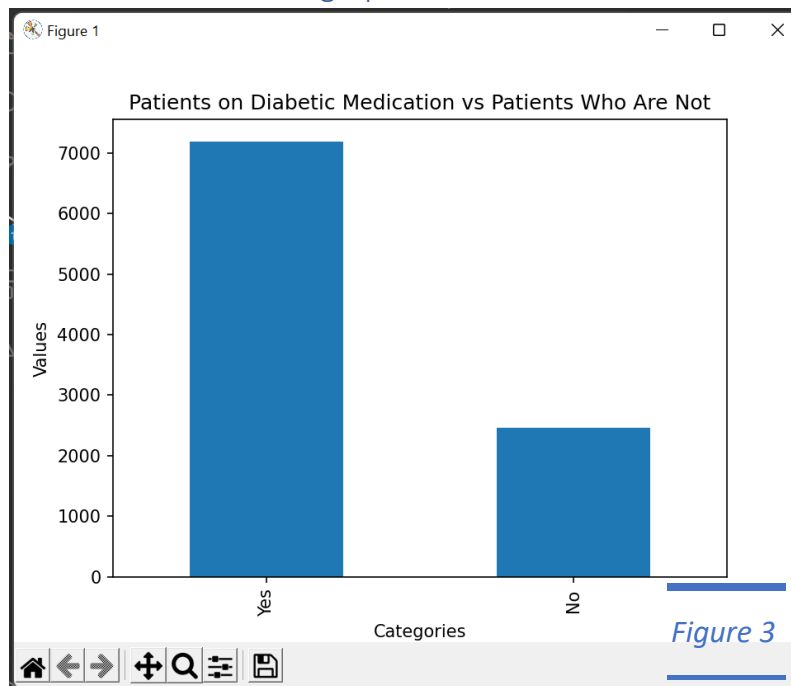
## Part 2: Exploratory Analytics

### Patients Readmitted Demographics



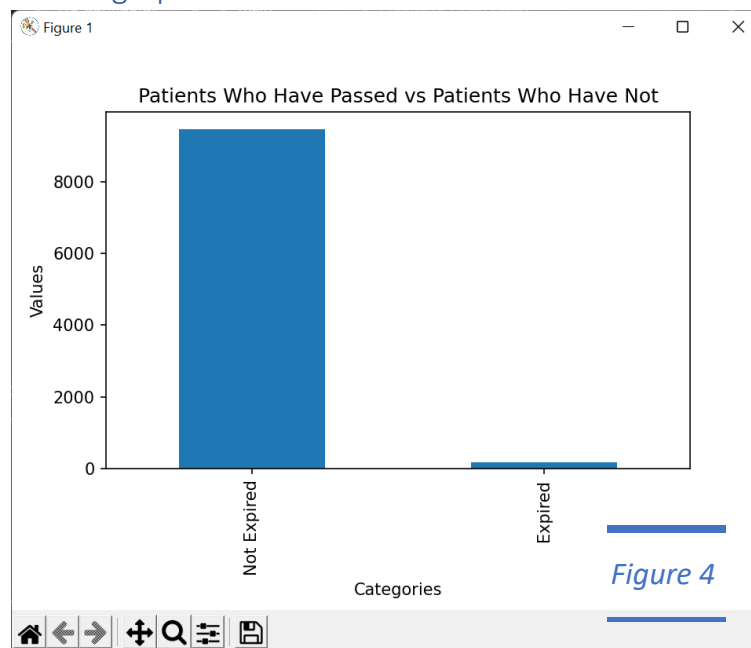
In *Figure 2*, the plot displays the comparison of patient readmissions at the hospital. The total number of records is 9,646. The majority of patients have not been readmitted at a value of 5897(61.13%). Verses the patients who were readmitted at a value of 3749 (38.86%).

## Patients on Diabetic Medication Demographics



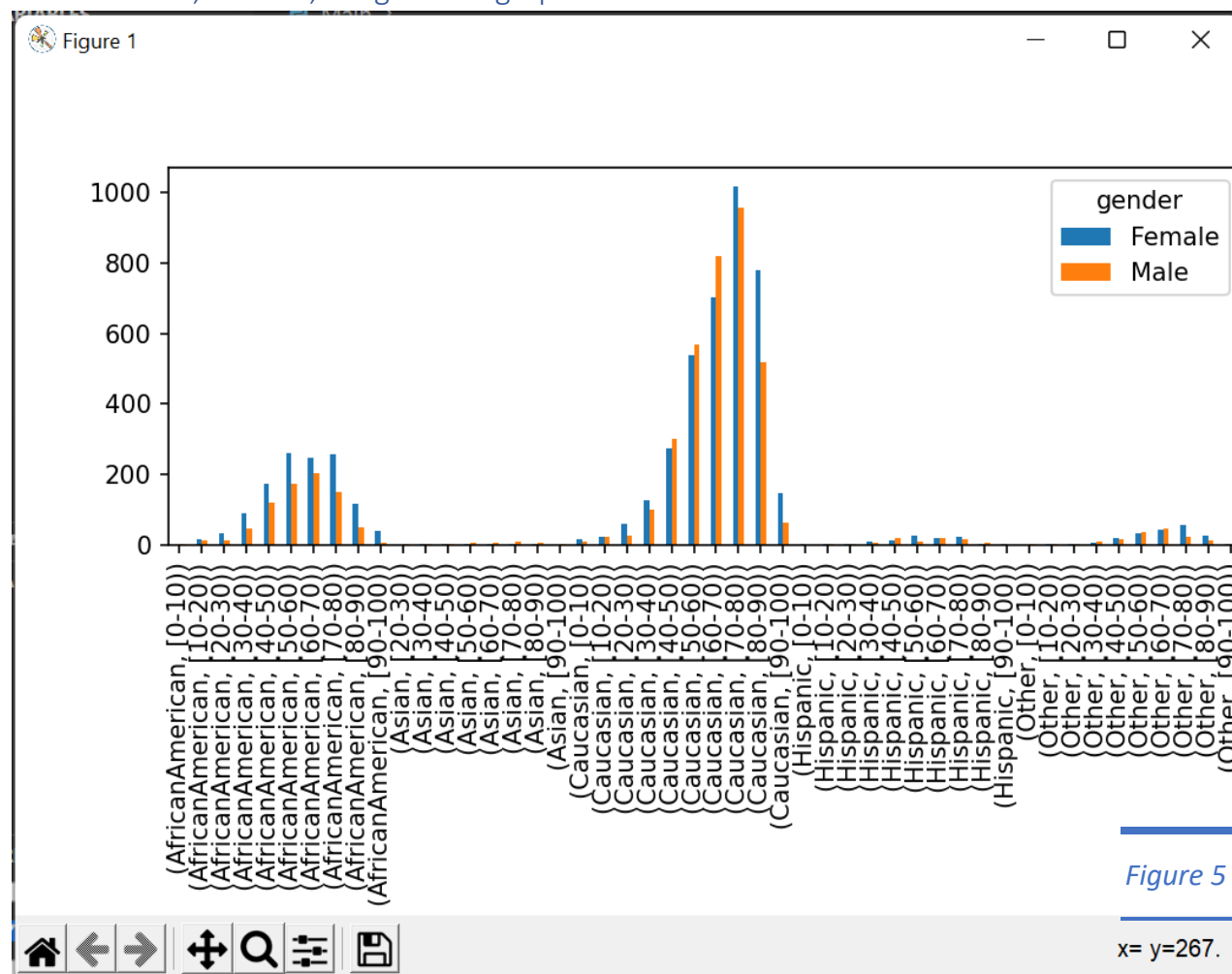
In *Figure 3*, the plot displays the patients on diabetic medication at the hospital. The total number of records is 9,646. 7,189(74.52%) patients are on diabetic medication, which is significantly higher than patients who are not on diabetic medication, at 2,457(25.47%).

## Patients Deceased Demographics



In *Figure 4*, the plot displays the patients who died verse the patients who did not die. The total number of records is 9,646. 9,460(98.07%) patients lived, with a much smaller number of patients who did not die, at 186(1.93%).

## Patients Race, Gender, & Age Demographics



Although, 6 graphs were expected, I thought looking at this data grouped together was far more interesting. Here you can visualize the rates at which each category exists. A few patterns the caught my eye are: There is a Caucasian major, especially for Caucasian over 40; The next noticeable pattern is that African American women visited constantly more than African American men.

## Part 3: Interesting Patterns

Before I start going over the patterns, here are some of the parameters to get my rules:

- min\_support=0.3
- metric="confidence"
- min\_threshold=0.3

The reason I chose a low min\_support to allow some rules to be return for me to deliver and evaluate.

## Readmitted Rules

antecedent	consequent	Ant. support	Cons. support	support	confidence	lift	leverage	conviction
Not Expired	Yes	0.9807	0.3886	0.3886	0.3963	1.0197	0.0075	1.0126

Support is the percentage of data in the set that contains the condition (Readmitted Yes). Confidence is the percentage of time that a rule proves to be true. Rules with a support or confidence less than 50 are not strong and should be discarded. I would not use this rule based on the percentage of the support and confidence.

## Not Readmitted Rules

antecedents	consequents	Ant. support	Cons. support	support	confidence	lift	leverage	conviction
Caucasian	No	0.7333	0.6113	0.4346	0.5927	0.9695	-0.0137	0.9542
Female	No	0.5409	0.6113	0.3273	0.6050	0.9897	-0.0034	0.9840
Emergency	No	0.6261	0.6113	0.3761	0.6008	0.9827	-0.0066	0.9735
Not Expired	No	0.9807	0.6113	0.5921	0.6037	0.9875	-0.0075	0.9807
Emergency Room	No	0.5889	0.6113	0.3429	0.5823	0.9525	-0.0171	0.9305
Not Expired & Caucasian	No	0.7185	0.6113	0.4199	0.5843	0.9558	-0.0194	0.9350
Not Expired & Female	No	0.5310	0.6113	0.3173	0.5976	0.9776	-0.0073	0.9659
Emergency & Not Expired	No	0.6119	0.6113	0.3619	0.5915	0.9675	-0.0121	0.9514
Emergency & Emergency Room	No	0.5246	0.6113	0.3090	0.5891	0.9637	-0.0117	0.9459

Remembering to filter out all rules with a support or confidence less than 50%; only one rule left. Rule: If the patient is not expired, then a patient will likely not be readmitted. But notice that the lift is 0.9875. Lift is to filter out misleading supports and confidence. If lift is less than 1, then that rule is misleading and is not reliable.

## Expired Rules

The is so little data for Expired Rule, I could not return any rules for expired patients.

## Interesting Reliable Patterns

index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
14	Not Expired	Female	0.980717396	0.540949616	0.530997	0.541437632	1.000902	0.0004786	1.00106423
15	Female	Not Expired	0.540949616	0.980717396	0.530997	0.981602146	1.000902	0.0004786	1.04808988
24	Emergency	Emergency Room	0.626062617	0.588948787	0.52457	0.837887067	1.422682	0.155851	2.53558557
25	Emergency Room	Emergency	0.588948787	0.626062617	0.52457	0.890688259	1.422682	0.155851	3.42083458
90	Emergency & Not Expired	Emergency Room	0.611859838	0.588948787	0.512233	0.837173839	1.421471	0.1518789	2.52447894
92	Not Expired & Emergency Room	Emergency	0.575782708	0.626062617	0.512233	0.889629096	1.420991	0.151757	3.3880069
93	Emergency	Not Expired & Emergency Room	0.626062617	0.575782708	0.512233	0.818181818	1.420991	0.151757	2.33319511
95	Emergency Room	Emergency & Not Expired	0.588948787	0.611859838	0.512233	0.869741243	1.421471	0.1518789	2.97976251

Although not required, I wanted to show the rule associations that are reliable, based on the criteria I discussed above. Notice that some of the consequents and antecedents are reversed, but the confidence is not the same. This is because one might be more likely to predict correctly, based on one consequent than another. For example, the rule If female then Not Expired, has a lower confidence than, if not expired then female. I believe this is true because Not Expired occurs more frequently than female and around 53% of not expired patients are female.

## Part 4: Predictive Analytics

To do predictive analytics, the data must be split into training and testing data. I chose a 80:20 ratio. For the learning algorithm, I decided to do two and compare results: Decision Tree and Gaussian Naïve Bayes.

### Decision Tree:

#### Confusion Matrix:

```
[[ 894 280]
 [ 517 239]]
```

#### Accuracy:

	precision	recall	f1-score	support
0	0.63	0.76	0.69	1174
1	0.46	0.32	0.37	756
accuracy			0.59	1930
macro avg	0.55	0.54	0.53	1930
weighted avg	0.57	0.59	0.57	1930

### Gaussian Naive Bayes:

#### Confusion Matrix:

```
[[ 88 1086]
 [ 27 729]]
```

#### Accuracy:

	precision	recall	f1-score	support
0	0.77	0.07	0.14	1174
1	0.40	0.96	0.57	756
accuracy			0.42	1930
macro avg	0.58	0.52	0.35	1930
weighted avg	0.62	0.42	0.31	1930

Recall is the % of true positives. So, if the recall of the Naïve Bayes algorithms is 0.07, means only 0.07 are labeled at positive. Precision is the percentage that the classifier labeled as positive. Meaning of the 0.07 recalled there is a 77% chance that it's correctly labeled. When comparing the two learning algorithms, Decision Tree is the more reliable and consistent in returning true positives at 894. I would choose the decision tree as my learning algorithm.