Applications of Machine Learning Assignment 1 ECE5464

By Srilalith Nampally

# Part I:

## Question 3: Data\_Type Sheet

|  |  |  |
| --- | --- | --- |
| Field | Role | Type |
| encounter\_id | ID | Categorical |
| patient\_nbr | ID | Categorical |
| race | feature | Categorical |
| gender | feature | Categorical |
| age | feature | Ordinal |
| weight | feature | Ordinal |
| admission\_type\_id | ID | Categorical |
| discharge\_disposition\_id | ID | Categorical |
| admission\_source\_id | ID | Categorical |
| time\_in\_hospital | feature | intervel |
| payer\_code | feature | categorical |
| medical\_specialty | feature | categorical |
| num\_lab\_procedures | feature | numeric |
| num\_procedures | feature | numeric |
| num\_medications | feature | numeric |
| number\_outpatient | feature | numeric |
| number\_emergency | feature | numeric |
| number\_inpatient | feature | numeric |
| diag\_1 | feature | categorical |
| diag\_2 | feature | categorical |
| diag\_3 | feature | categorical |
| number\_diagnoses | feature | numeric |
| max\_glu\_serum | feature | Ordinal |
| A1Cresult | feature | Ordinal |
| metformin | feature | Ordinal |
| repaglinide | feature | Ordinal |
| nateglinide | feature | Ordinal |
| chlorpropamide | feature | Ordinal |
| glimepiride | feature | Ordinal |
| acetohexamide | feature | Binary |
| glipizide | feature | Ordinal |
| glyburide | feature | Ordinal |
| tolbutamide | feature | Binary |
| pioglitazone | feature | Ordinal |
| rosiglitazone | feature | Ordinal |
| acarbose | feature | Ordinal |
| miglitol | feature | Ordinal |
| troglitazone | feature | Binary |
| tolazamide | feature | Ordinal |
| examide | feature | Binary |
| citoglipton | feature | Binary |
| insulin | feature | Ordinal |
| glyburide-metformin | feature | Ordinal |
| glipizide-metformin | feature | Binary |
| glimepiride-pioglitazone | feature | Binary |
| metformin-rosiglitazone | feature | Binary |
| metformin-pioglitazone | feature | Binary |
| change | feature | Binary |
| diabetesMed | feature | Binary |
| readmitted | feature | Ordinal |

## Question 4: Impossible Values

* I identified the impossibly large values by observing the unnaturally large variance in the data\_statistics sheet.
* When I checked the column name, it was “number of medications”.
* Logically speaking I think there is no person that will ever be on 9999999 medications.
* As such I replaced 5 values in the column, at rows 333, 818, 1079, 86289.

## Question 5: Data Statistics Sheet

Due to the size of the number of features in the Dataset, for the purpose of readability, in this submission the table of size (14x51) has been split into 8 parts

### Table Part 1:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Statistic | encounter\_id | patient\_nbr | race | gender | age | weight |
| Mean | N/A | N/A | N/A | N/A | N/A | N/A |
| Min | N/A | N/A | N/A | N/A | N/A | N/A |
| Max | N/A | N/A | N/A | N/A | N/A | N/A |
| Range | N/A | N/A | N/A | N/A | N/A | N/A |
| Median | 152388294 | 97438842 | Caucasian | #N/A | [60-70) | ? |
| Mode | #N/A | 88785891 | Caucasian | #N/A | [70-80) | ? |
| Variance | N/A | N/A | N/A | N/A | N/A | N/A |
| Std Deviation | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 1 | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 2 | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 3 | N/A | N/A | N/A | N/A | N/A | N/A |
| # Missing | 0 | 0 | 0 | 2 | 0 | 0 |
| nRows | 101767 | 101767 | 101767 | 101764 | 101767 | 101767 |

### Table Part 2 and 3

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | admission\_type\_id | | discharge\_disposition\_id | | admission\_source\_id | | time\_in\_hospital | | payer\_code | |
| Mean | | N/A | | N/A | | N/A | | 4.395986872 | | N/A |
| Min | | N/A | | N/A | | N/A | | 1 | | N/A |
| Max | | N/A | | N/A | | N/A | | 14 | | N/A |
| Range | | N/A | | N/A | | N/A | | 13 | | N/A |
| Median | | 1 | | 1 | | 7 | | 4 | | MC |
| Mode | | 1 | | 1 | | 7 | | 3 | | ? |
| Variance | | N/A | | N/A | | N/A | | 8.910868383 | | N/A |
| Std Deviation | | N/A | | N/A | | N/A | | 2.985107767 | | N/A |
| Quartile 1 | | N/A | | N/A | | N/A | | 2 | | N/A |
| Quartile 2 | | N/A | | N/A | | N/A | | 4 | | N/A |
| Quartile 3 | | N/A | | N/A | | N/A | | 6 | | N/A |
| # Missing | | 0 | | 0 | | 0 | | 0 | | 0 |
| nRows | | 101767 | | 101767 | | 101767 | | 101767 | | 101767 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Statistic | medical\_specialty | num\_lab\_procedures | num\_procedures | num\_medications | number\_outpatient |
| Mean | N/A | 43.09564098 | 1.339657698 | 16.0212766 | 0.369357153 |
| Min | N/A | 1 | 0 | 1 | 0 |
| Max | N/A | 132 | 6 | 81 | 42 |
| Range | N/A | 131 | 6 | 80 | 42 |
| Median | InternalMedicine | 44 | 1 | 15 | 0 |
| Mode |  | 1 | 0 | 13 | 0 |
| Variance | N/A | 387.0805299 | 2.909375424 | 66.04661179 | 1.605960825 |
| Std Deviation | N/A | 19.67436225 | 1.705689135 | 8.126906656 | 1.267265097 |
| Quartile 1 | N/A | 31 | 0 | 10 | 0 |
| Quartile 2 | N/A | 44 | 1 | 15 | 0 |
| Quartile 3 | N/A | 57 | 2 | 20 | 0 |
| # Missing | 0 | 0 | 42 | 10 | 0 |
| nRows | 101767 | 101767 | 101724 | 101756 | 101767 |

### Table Part 4 and 5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | number\_emergency | number\_inpatient | diag\_1 | diag\_2 | diag\_3 | number\_diagnoses | max\_glu\_serum |
| Mean | 0.197836 | 0.635566 | N/A | N/A | N/A | 7.422607 | #DIV/0! |
| Min | 0 | 0 | N/A | N/A | N/A | 1 | 0 |
| Max | 76 | 21 | N/A | N/A | N/A | 16 | 0 |
| Range | 76 | 21 | N/A | N/A | N/A | 15 | 0 |
| Median | 0 | 0 | 440 | 425 | 403 | 8 | None |
| Mode | 0 | 0 | 428 | 276 | 250 | 9 | None |
| Variance | 0.865779 | 1.594824 | N/A | N/A | N/A | 3.73881 | N/A |
| Std Deviation | 0.930472 | 1.262863 | N/A | N/A | N/A | 1.9336 | N/A |
| Quartile 1 | 0 | 0 | N/A | N/A | N/A | 6 | N/A |
| Quartile 2 | 0 | 0 | N/A | N/A | N/A | 8 | N/A |
| Quartile 3 | 0 | 1 | N/A | N/A | N/A | 9 | N/A |
| # Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| nRows | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | A1Cresult | metformin | repaglinide | nateglinide | chlorpropamide | glimepiride | acetohexamide | glipizide |
| Mean | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Min | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Max | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Range | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Median | None | No | No | No | No | No | No | No |
| Mode | None | No | No | No | No | No | No | No |
| Variance | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Std Deviation | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 1 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 2 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| # Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| nRows | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 |

### Table part 6 and 7

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | glyburide | tolbutamide | pioglitazone | rosiglitazone | acarbose | miglitol | troglitazone | tolazamide | examide |
| Mean | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Min | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Max | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Range | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Median | No | No | No | No | No | No | No | No | No |
| Mode | No | No | No | No | No | No | No | No | No |
| Variance | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Std Deviation | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 1 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 2 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| # Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| nRows | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Statistic | citoglipton | insulin | glyburide-metformin | glipizide-metformin | glimepiride-pioglitazone | metformin-rosiglitazone |
| Mean | N/A | N/A | N/A | N/A | N/A | N/A |
| Min | N/A | N/A | N/A | N/A | N/A | N/A |
| Max | N/A | N/A | N/A | N/A | N/A | N/A |
| Range | N/A | N/A | N/A | N/A | N/A | N/A |
| Median | No | Up | No | No | No | No |
| Mode | No | No | No | No | No | No |
| Variance | N/A | N/A | N/A | N/A | N/A | N/A |
| Std Deviation | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 1 | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 2 | N/A | N/A | N/A | N/A | N/A | N/A |
| Quartile 3 | N/A | N/A | N/A | N/A | N/A | N/A |
| # Missing | 0 | 0 | 0 | 0 | 0 | 0 |
| nRows | 101767 | 101767 | 101767 | 101767 | 101767 | 101767 |

### Table part 8

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | metformin-pioglitazone | change | diabetesMed | readmitted |
| Mean | N/A | N/A | N/A | N/A |
| Min | N/A | N/A | N/A | N/A |
| Max | N/A | N/A | N/A | N/A |
| Range | N/A | N/A | N/A | N/A |
| Median | No | No | Yes | NO |
| Mode | No | No | Yes | NO |
| Variance | N/A | N/A | N/A | N/A |
| Std Deviation | N/A | N/A | N/A | N/A |
| Quartile 1 | N/A | N/A | N/A | N/A |
| Quartile 2 | N/A | N/A | N/A | N/A |
| Quartile 3 | N/A | N/A | N/A | N/A |
| # Missing | 0 | 0 | 0 | 0 |
| nRows | 101767 | 101767 | 101767 | 101767 |

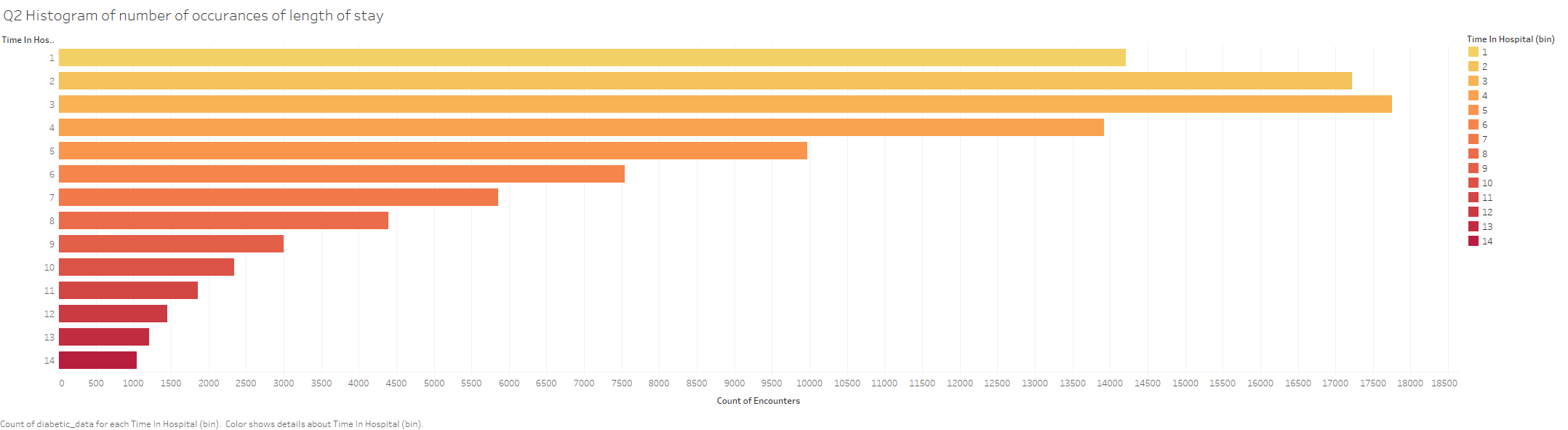
# Part II:

## Question 7:

### A The 3 patients who spent the most time in the Hospital are: 84428613, 3481272 and 88227540. The number of encounters they had is 180, 173 and 145 respectively.

## A graph with numbers and a line Description automatically generated

### B Histogram for Number of occurrences of lengths of stay:



### C Percentage breakdown of different A1C results among each racial group:

A graph of different colored bars

Description automatically generated with medium confidence

### D

Percentage of Patients re-admitted in <30 days = 12.35%

Percentage of Patients re-admitted in >30 days = 33.75%

A graph with different colored squares

Description automatically generated

### E

Percentage of Encounters which resulted in readmission in < 30 days = 11.16%

Percentage of Encounters which resulted in readmission in > 30 days = 34.93%

A graph with different colored bars

Description automatically generated with medium confidence

### F Average length of time in Hospital for people based on their A1C result when segregated by age group:

A screenshot of a graph

Description automatically generated

# Part III:

## Question 8:

### Output of the program which reads the excel workbook and prints the size of the resulting dataframe

‘

/usr/local/bin/python3.11/Users/srilalithnampally/Classes/AppML\_Assignments/Assignment\_1/read\_data.py

(101766, 50)

Process finished with exit code 0

’

## Question 9:

### Data\_Statistics.py

import pandas as pd

from StatsReport import StatsReport

excel\_data = "diabetic\_data.xlsx"

diabetes\_df = pd.read\_excel(excel\_data)

# print(diabetes\_df.shape)

labels = diabetes\_df.columns

report = StatsReport()

# print(diabetes\_df['age'].dtype)

for i in labels:

thisCol = diabetes\_df[i]

report.addCol(i, thisCol)

print(report.to\_String())

toExcel\_file = "data\_report.xlsx"

print(f'\nSuccessfully generated a data statistics report\nTranscribing data into {toExcel\_file} ....')

report.write\_to\_file(toExcel\_file)

### StatsReport.py

import pandas as pd

class StatsReport:

def \_\_init\_\_(self):

self.data\_stats\_df = pd.DataFrame()

self.data\_stats\_df['Data\_Stats'] = ['type', 'cardinality', 'mean', 'median', 'n\_at\_median', 'mode', 'n\_at\_mode',

'stddev', 'min', 'max', 'n\_rows', 'n\_zero', 'n\_ques', 'n\_missing']

pass

def addCol(self, colName, column):

try:

mean\_value = column.mean()

median\_value = column.median()

std\_dev\_value = column.std()

min\_value = column.min()

max\_value = column.max()

except TypeError:

mean\_value = "N/A"

median\_value = "N/A"

std\_dev\_value = "N/A"

min\_value = "N/A"

max\_value = "N/A"

mode\_value = column.mode().iloc[0] if not column.mode().empty else "N/A"

n\_at\_mode = (column == mode\_value).sum()

n\_at\_median = (column == median\_value).sum()

n\_zeros = (column == 0).sum()

n\_ques = (column == "?").sum()

n\_missing = column.isna().sum()

self.data\_stats\_df[colName] = [column.dtype, column.nunique(), mean\_value, median\_value, n\_at\_median,

mode\_value, n\_at\_mode, std\_dev\_value, min\_value, max\_value, len(column), n\_zeros,

n\_ques, n\_missing]

def to\_String(self):

return self.data\_stats\_df.to\_string()

def write\_to\_file(self, filepath):

self.data\_stats\_df.to\_excel(filepath, index=False, engine='openpyxl')