Tab 1

HACK-A-STAT-25

Teamname: Hackverse

Members:
Ms. Aayushi Fariya
Ms. Sayali Mahurkar
Ms. Sneha Maheshwari

PROBLEM STATEMENT:

You are provided with a dataset of 10 years (1999-2008) representing 130 records of hospitals of subjects with diabetes who underwent laboratory, medication and stayed up to 14 days. The goal is to determine the readmission of the subject within 30 days of discharge.

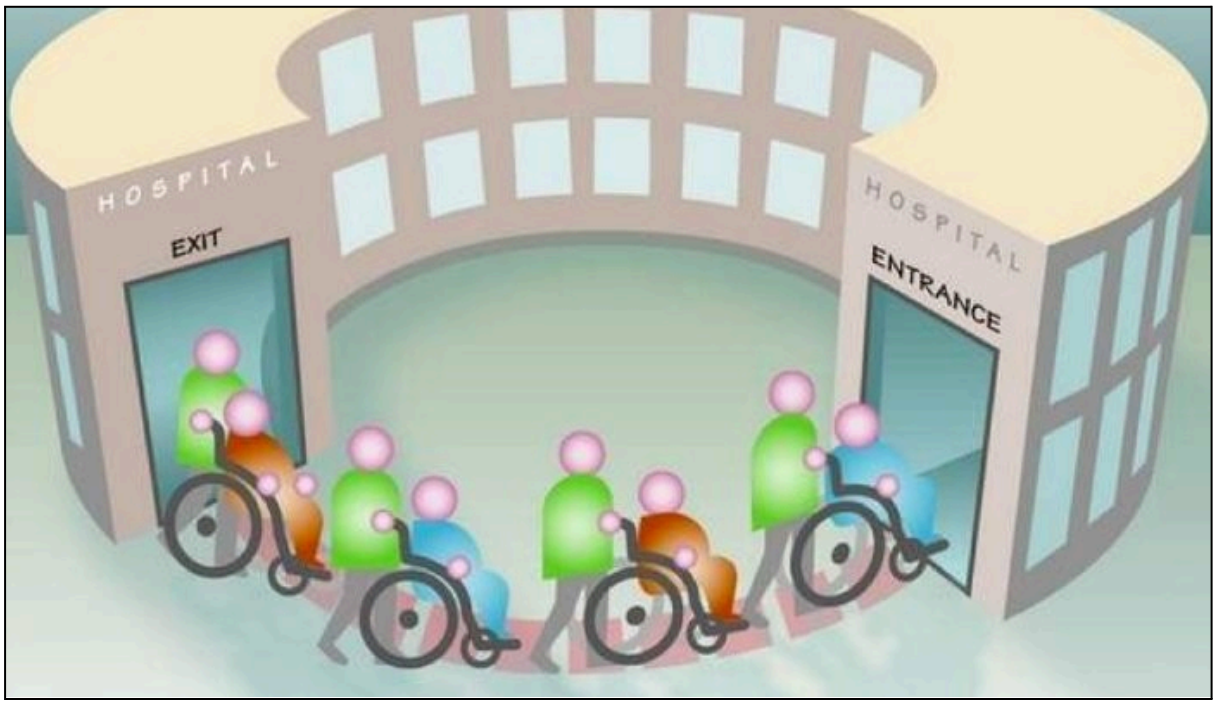


TABLE OF CONTENTS

OBJECTIVES

1. Conducting an initial EDA between explanatory variable to understand several correlations.
2. Imputing missing data using predictive model such as KNN Imputation instead of simple statistical imputation.
3. Analysing the impact of variable-HbA1c on dependent variable-hospitals readmission.
4. Seeing how diabetes medication changes under different scenarios of HbA1c measurements.
5. Testing for multicollinearity.

6. Performing model validation.

INTRODUCTION

Diabetes is a chronic disease , which affects a person's health and quality of life. The one who gets diabetes has to go through many complications, regular check-ups, get hospitalsised or in some cases there may be readmission. The goal of our project is to check whether there will be a readmission in a hospitals within 30 days of discharge. The provided dataset of 10 years (1999-2008) represents hospitals records of patients with diabetes who underwent laboratory medication and stayed up to 14 days.

Since readmission rates are very crucial in the healthcare sector as it reflects badly on the treatment provided by them and also indicates the post-discharge care to prevent the further complications .

The insights from this project will help the healthcare people in improving their treatments and in reducing unnecessary readmissions.

DATA DESCRIPTION

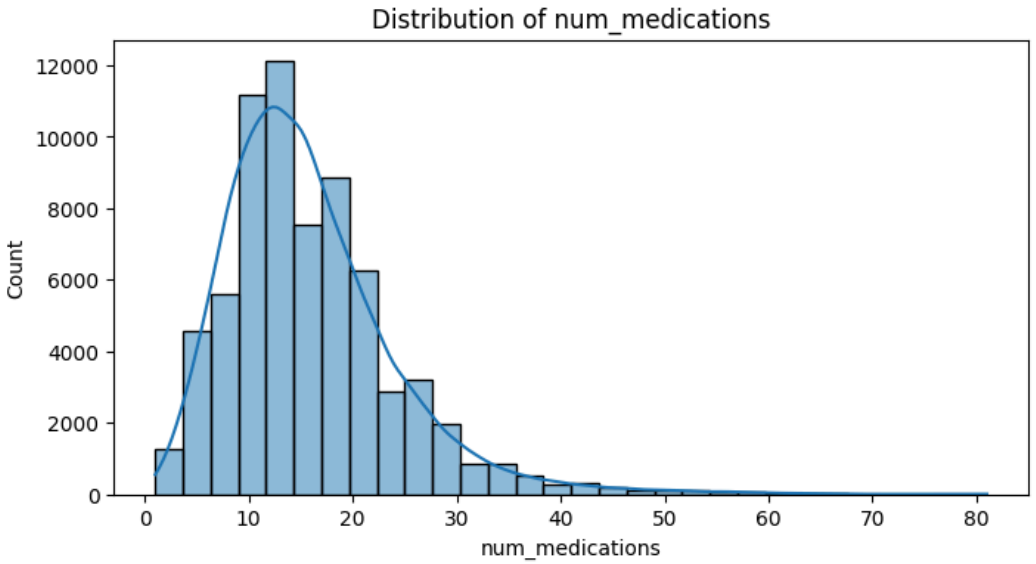
Following are the unique values in the dataset:

Feature Name	Unique Values
race	['Caucasian', 'AfricanAmerican', nan, 'Other', 'Asian', 'Hispanic']
gender	['Female', 'Male', 'Unknown/Invalid']
age	['[0-10)', '[10-20)', '[20-30)', '[30-40)', '[40-50)', '[50-60)', '[60-70)', '[70-80)', '[80-90)', '[90-100)']
admission_type_id	[6, 1, 2, 3, 4, 5, 8, 7]
discharge_disposition_id	[25, 1, 3, 6, 2, 5, 10, 7, 4, 18, 8, 12, 22, 17, 23, 9, 16, 15, 28, 24, 27]
admission_source_id	[1, 7, 2, 4, 20, 6, 5, 3, 17, 8, 9, 14, 10, 22, 11, 25, 13]
medical_specialty	['Pediatrics-Endocrinology', nan, 'InternalMedicine', 'Pediatrics', 'Otolaryngology', 'Surgery-Colon&Rectal', ….....
max_glu_serum	['NO', '>300', 'Norm', '>200']
A1Cresult	['NO', '>7', '>8', 'Norm']

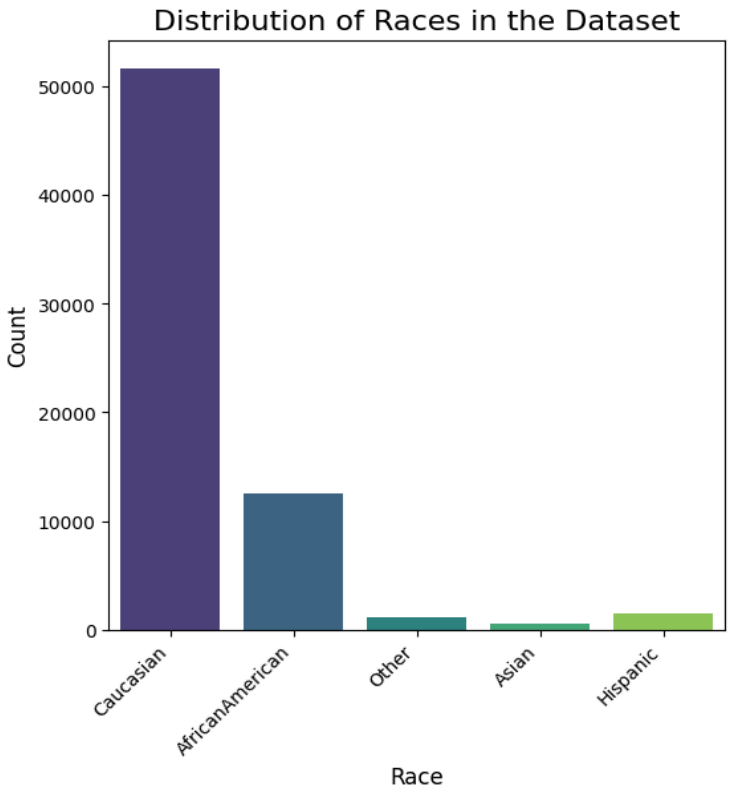
metformin	['No', 'Steady', 'Up', 'Down']
repaglinide	['No', 'Steady', 'Up', 'Down']
nateglinide	['No', 'Steady', 'Up', 'Down']
chlorpropamide	['No', 'Steady', 'Down', 'Up']
glimepiride	['No', 'Steady', 'Down', 'Up']
acetohexamide	['No', 'Steady']
glipizide	['No', 'Steady', 'Up', 'Down']
glyburide	['No', 'Steady', 'Up', 'Down']
tolbutamide	['No', 'Steady']
pioglitazone	['No', 'Steady', 'Up', 'Down']
rosiglitazone	['No', 'Steady', 'Up', 'Down']
acarbose	['No', 'Steady', 'Up', 'Down']
miglitol	['No', 'Steady', 'Up', 'Down']
troglitazone	['No', 'Steady']
tolazamide	['No', 'Steady']
examide	['No']
citoglipton	['No']
insulin	['No', 'Up', 'Steady', 'Down']
glyburide.metformin	['No', 'Steady', 'Down', 'Up']
glipizide.metformin	['No', 'Steady']
glimepiride.pioglitazone	['No']
metformin.rosiglitazone	['No', 'Steady']
metformin.pioglitazone	['No', 'Steady']
change	['No', 'Ch']
diabetesMed	['No', 'Yes']
readmitted	['NO', '>30', '<30']
diag_1_name	['diabetes', 'Other', 'neoplasms', 'circulatory diseases', 'respiratory diseases', 'injury', 'musculoskeletal diseases', 'digestive diseases',

	'genitournary']
diag_2_name	['Other', 'diabetes', 'neoplasms', 'circulatory diseases', 'respiratory diseases', 'injury', 'musculoskeletal diseases', 'digestive diseases', 'genitournary']
diag_3_name	['Other', 'circulatory diseases', 'diabetes', 'respiratory diseases', 'injury', 'neoplasms', 'genitournary', 'musculoskeletal diseases', 'digestive diseases']

EXPLORATORY DATA ANALYSIS

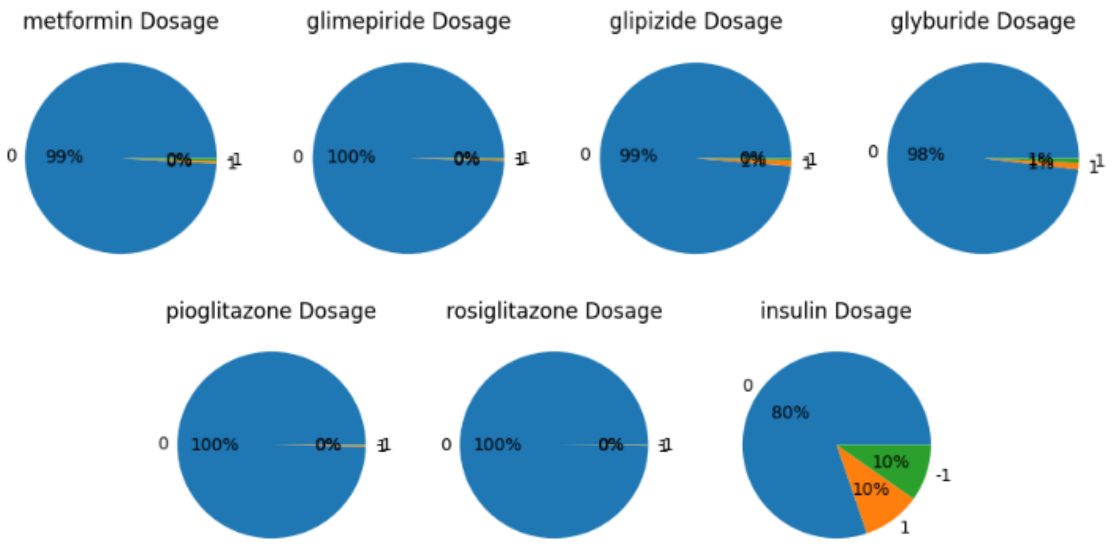


- By plotting the frequency distribution of the variable num_medications, the average number of medications given to a patient are 15.8025 which is approximately close to high. The minimum number of medications is 1 and the maximum number is 81. Also the data is highly positively skewed, this is due to the presence of outliers. As the mean is approximately 16, it suggests that the hospitals might be making the subjects ingest one too many medications! Similar is the case of num_lab_procedures where the mean being 43 lab procedures is too high.



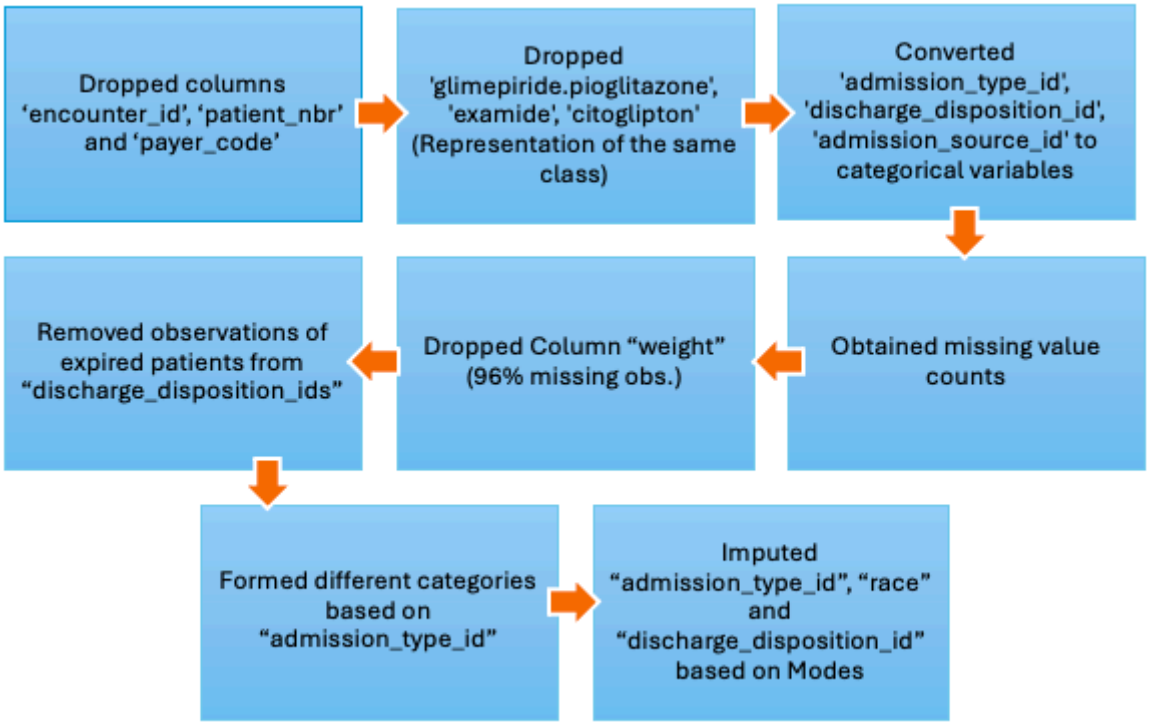
- From the above bar plot, it can be seen that the count of Caucasian ethnicity is the highest in the given dataset. This is most likely due to the geographical location of the hospitals (the hospitals is located in the US) where the population of Caucasians is the highest.

Drugs Dosage Changes During Encounters



- In the above pie charts “0” depicts: Steady and No, “-1” depicts: Down and “1” stands for Up. This informs us that while there was almost a Steady and No change in the drug dosages during the hospitals visits of the patients, but 10% of the population’s insulin dosage was increased and another 10% 's was decreased.

DATA CLEANING



- The column Index I'd has been dropped to reduce the dimensions from 51 to 50.
- We also dropped the column "Weight" because it has 96% MCAR values which cannot be imputed. Also the variable in itself does not provide any important information about subject readmission.
- After thoroughly examining the dataset , we found that there were no duplicate values identified.

MISSING VALUES

It is to be noted that the values "None" in the variables "max_glu_serum" and "A1Cresult" mean that no test was administered and are not NAN values as would be read by pandas. Thus we replaced these "None" values by "No" in the dataset before importing in the python script.

Most null values lie in Medical Speciality (46%)

```
readmitted
NO     0.732679
>30    0.219852
<30    0.047469
Name: proportion, dtype: float64
readmitted
NO     0.770629
>30    0.183989
<30    0.045382
Name: proportion, dtype: float64
```

(Proportion of Missing values vs Non-missing values in case of Medical Speciality)

As proportions are more or less similar therefore Medical Speciality values are **MCAR.**

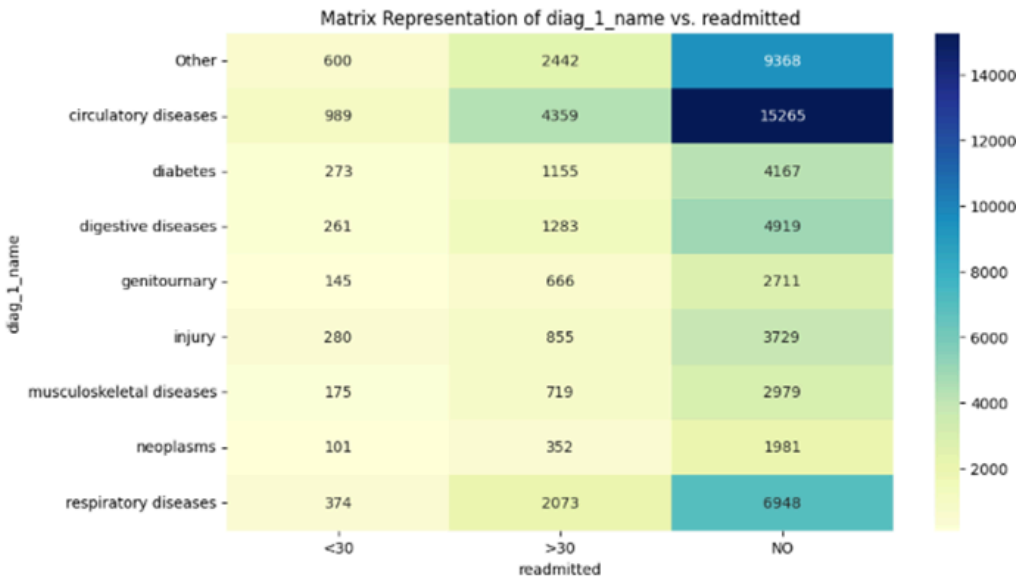
OUTLIER HANDLING

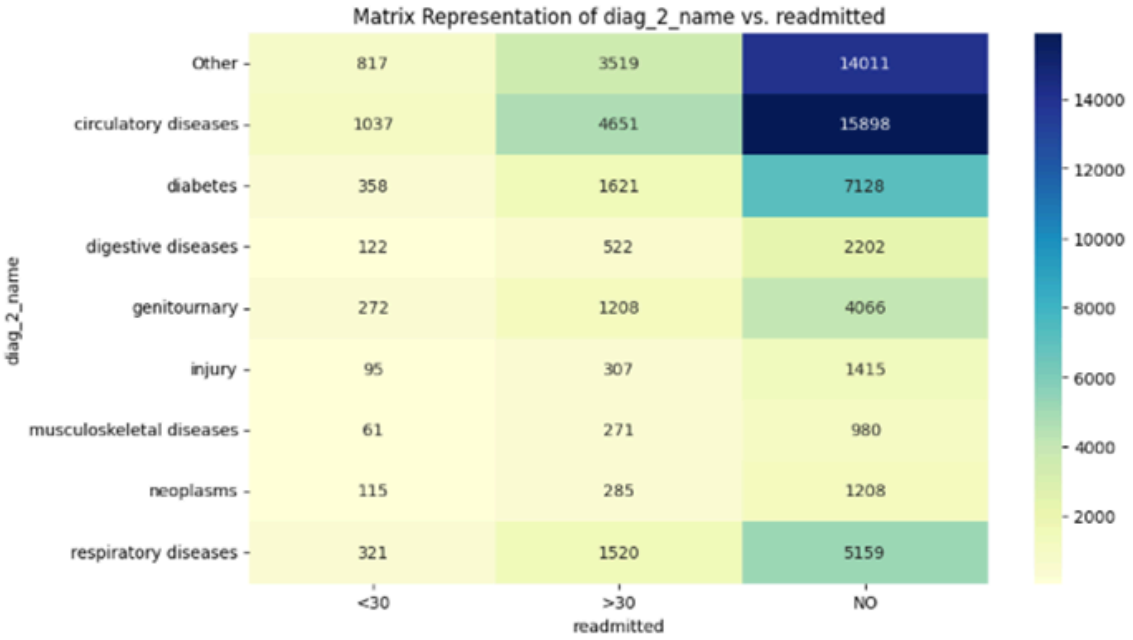
```
Column: time_in_hospitals
Percentage of outliers: 2.08%
--------------------
Column: num_lab_procedures
Percentage of outliers: 0.15%
--------------------
Column: num_procedures
Percentage of outliers: 5.22%
--------------------
Column: num_medications
Percentage of outliers: 2.63%
--------------------
Column: number_outpatient
Percentage of outliers: 14.59%
--------------------
Column: number_emergency
Percentage of outliers: 8.92%
--------------------
Column: number_inpatient
Percentage of outliers: 4.25%
--------------------
Column: number_diagnoses
Percentage of outliers: 0.33%
--------------------
```

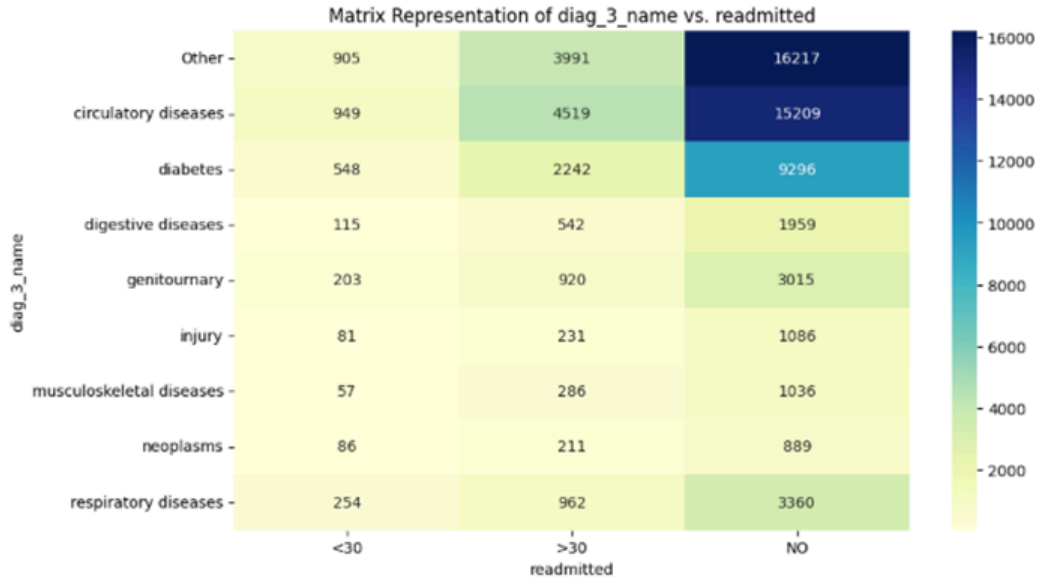
Since the count of outliers is below 15% in case of each covariate, we leave the outliers as is.

POST EDA

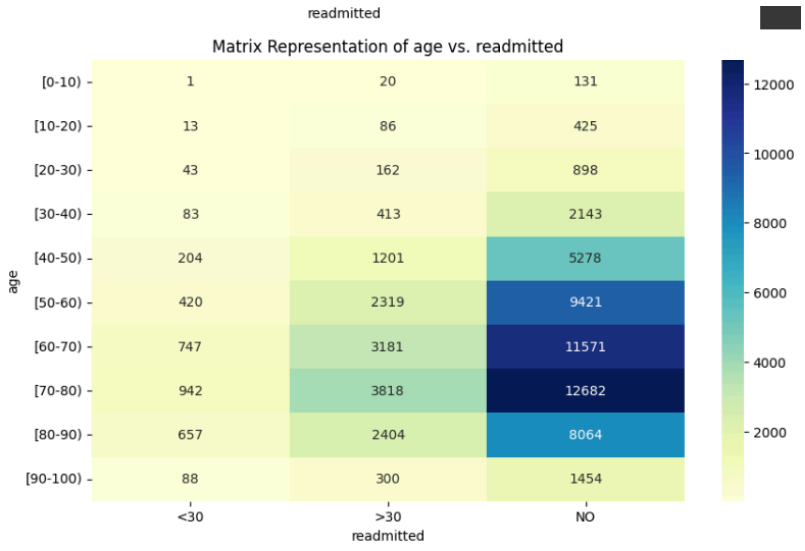
- The circulatory diseases dominate as primary , secondary and additional secondary diagnoses in diabetic patients , which highlights the need for managing cardiovascular risks. Diabetics also appears frequently across all three levels of diagnosis, underscoring its role in complications and readmissions.



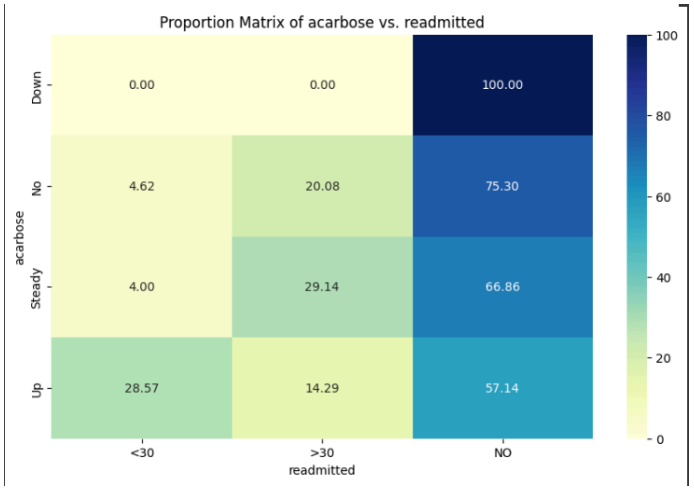


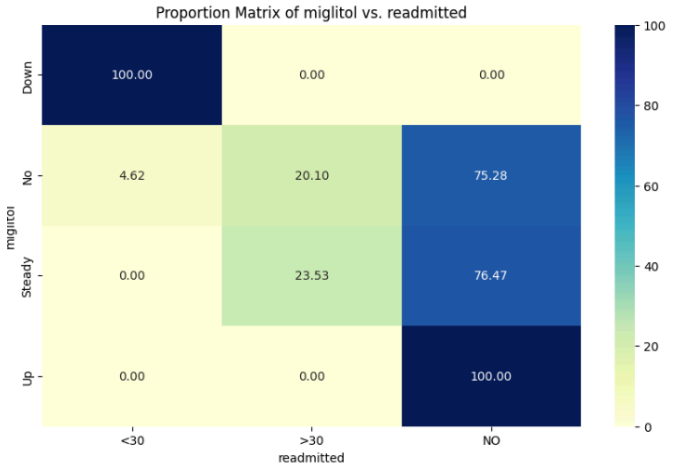


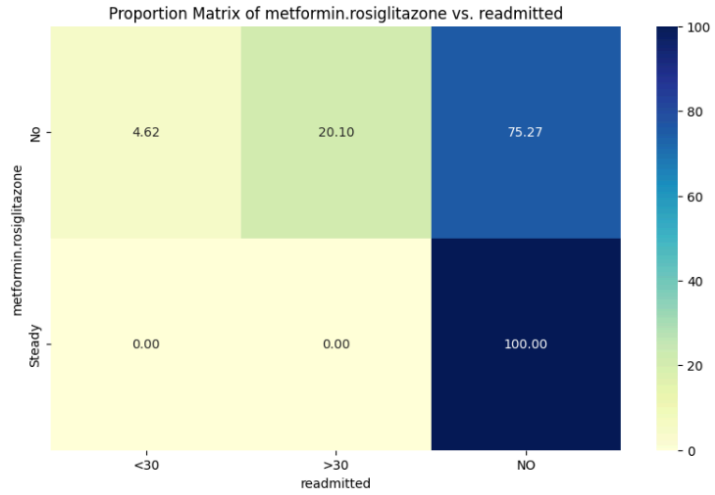
- Based on the above heatmap, it can be said that middle-age groups show higher readmissions compared to other age groups.



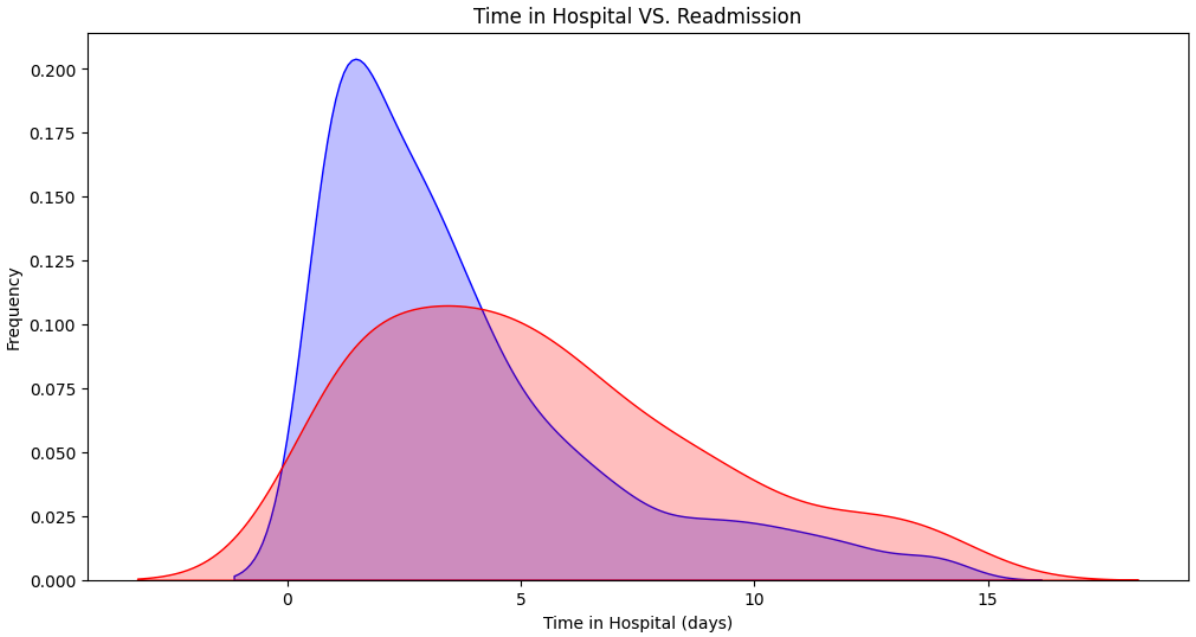
- If Arcabose dose is reduced for the subject then there is a 100% probability that they will not be readmitted within 30 days.
- If Miglitol dosage is reduced for the subject then there is a 100% probability that they will be readmitted within 30 days.
- If Metformin.rosiglitazone dosage is steady for the subject then there is a 100% probability that they will not be readmitted within 30 days.
- Similarly, If Metformin.pioglitazone dosage is reduced for the subject then there is a 100% probability that they will be not be readmitted within 30 days.
- Finally, if the dosage of Glyburide.Metformin is reduced then there is a 100% probability that the subject will not be readmitted within 30 days.



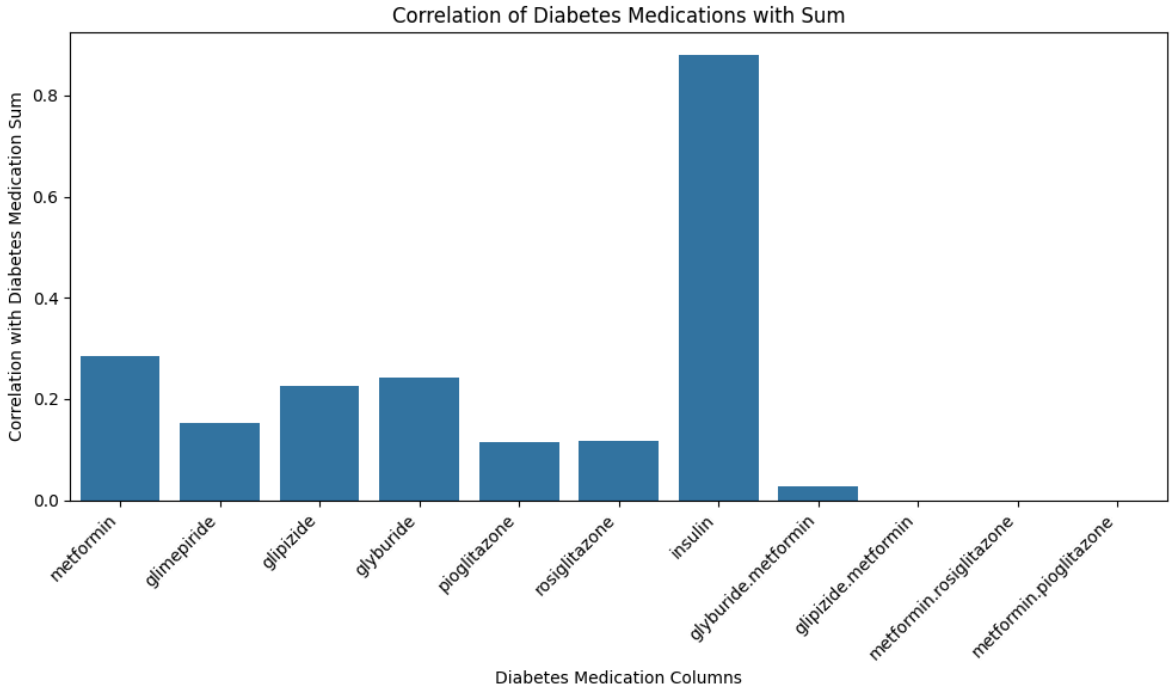




- The blue curve indicates non readmitted patients which peaks approximately at 2 days implies shorter hospitals stays i.e effective diabetic management and less complications . Whereas the red curve which is the admitted patients peaks approximately at 4 days hence it suggests unresolved issues and more diabetic complications.



- We created a new variable DM which is the sum of individual effects of the drugs that were not dropped. Here, the classes in each drug were encoded as 0 for No or Steady dosage, 1 for Up and (-1) for down. Then, we plotted interaction effect value against each medication and found that insulin levels are highly correlated with the new interaction effect DM. This is potentially because insulin is the most prescibed drug to treat diabetes.



METHODOLOGY

- After conducting an in-depth exploratory analysis, the next step was to impute or remove the missing values in the dataset.

```
weight                      95.988087
medical_specialty           48.069222
payer_code                  42.264598
admission_type_id           10.405472
discharge_disposition_id     4.304511
race                         2.663043
           dtype: float64
```

It is seen from table – that there are missing values in columns Weight , Medical_speciality , Admission_type_id, Race and discharge_desposition_id, since this percentage is less than or close to 10% in case of admission_type_id,race and discharge_desposition_id and are also Missing completely at random (MCAR) therefore we can use simple statistical imputation techniques such as mode to impute these missing values.

In case of weight, the missing values are nearly 96% this means that the column weight is redundant and can be removed from the further course of this analysis.

However, in case of Medical Speciality, the missing value percentage is approx. 48.062 %, since this percentage is less than 50, we cannot omit this variable. Also, considering that the values are MCAR, we will look into imputing these values by fitting predictive model, in particular, **Random Forest. Why?** Random Forest can naturally handle categorical variables (Medical Speciality), not necessarily requiring encoding. While algorithms such as KNN struggle with high-dimensional or

large datasets and models like logistic regression require assumptions such as normality, linearity, which are not holding for Medical Speciality. Thus, Random Forest is the best predictive model to deal with missing Medical Speciaity data.

- To study how diabetes medication changes (variable “change”) were being done under different scenarios of HbA1c measurements. I.e. to study the correlations between Hb41Ac and Change in diabetes medication, we plotted the confusion matrix which was found using Chi-Square test of association.

- Understanding the impact of blood glucose level (HbA1c) and readmission decision was a crucial part of this study. We applied various predictive ML models such as, Logistic Regression, Random Forest and XG Boost because of their ability to handle imbalanced and non-linear data.

ANALYSIS AND RESULTS

- We first applied the Chi-square test for feature selection. Then using Principal Component Analysis we did dimensionality reduction. Finally, we applied Random Forest Classifier to impute missing values in Medical Speciality.

Chi-sq selected values

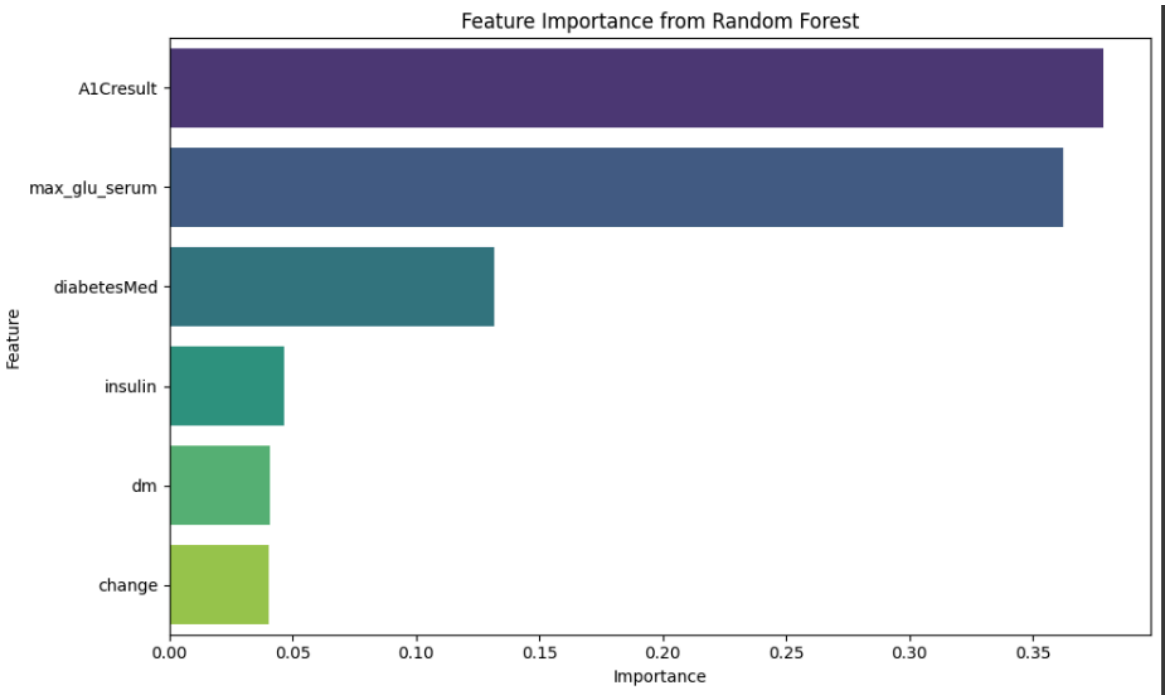
```
Index(['race', 'gender', 'age', 'admission_type_id',
'discharge_disposition_id', 'admission_source_id',
'time_in_hospitals','num_lab_procedures', 'num_procedures',
'num_medications','number_outpatient', 'number_emergency',
'number_inpatient','number_diagnoses', 'A1Cresult', 'change',
'readmitted', 'diag_1_name',  'diag_2_name',
'diag_3_name'],dtype='object')
```

After imputing Missing values

```
                    medical_specialty
InternalMedicine                29805
Other                           15311
Emergency/Trauma                 8683
Family/GeneralPractice           7573
Cardiology                       7486
   Name: count, dtype: int64
```

After imputing missing values we have to select features impacting the target variable. We check for correlation, but there is no specific high or low relation in variables wrt the target variable on which feature selection can be done.

Since, diagnosis_1 are to be considered, the other two are dropped off. Then the following feature selection graph is creating using randomForest amongst the shown variables only, as they are domain related.



Amongst the related variables, these were checked to know which has more importance and then the last 3 were removed.

```
Feature  Importance
4      A1Cresult    0.378761
5  max_glu_serum    0.362309
2    diabetesMed    0.131573
0        insulin    0.046622
1             dm    0.040658
3         change    0.04007
```

Race and gender showing similar proportions to not being admitted and the highest too are thus removed.

On the finally feature selected variables that are : `'age', 'admission_type_id', 'discharge_disposition_id','admission_source_id', 'time_in_hospitals', 'medical_specialty','max_glu_serum', 'A1Cresult', 'medical_complexity', 'num_visits'`

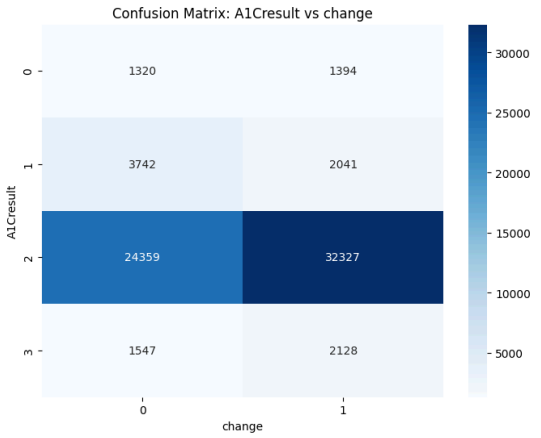
Which are 10 amongst 50 variables filtered at diag_1_name = diabetes.

The following models were evaluated with the given score metrics -

Model	Class	Precision	Recall	Accuracy	F1_score	support
Logistic Regression		-	-	0.94	-	-
Randomforest Classifier	0 1	0.95 0.08	0.91 0.15	0.93	0.96 0.06	13151 621
XGboost	0 1	0.95 0.12	0.99 0.04	0.67	0.97 0.06	1059 55

If Overall accuracy is the goal : XGBoost is slightly better, but poor recall may not be a good choice

- The confusion matrix for HbA1c measurements and Diabetes Medication Change was obtained as:



- Here “0”,”1”,”2” and “3” A1Cresult values are No test, Normal, >7 and >8 respectively and “0” and “1” Change values are No Change was made to medication and change was made to diabetes medication respectively.

    1. Almost equal numbers of patients (1320 against 1394) had changes made to their medication, despite no A1C test. This suggests decisions were based on factors other than the test.
    2. A significant number of patients with their A1Cresult>7 had their medications adjusted but it is not expected as medically, if a person is diabetic but their blood glucose lies between 7 and 8, then usually medications are not adjusted much. This could suggest that they might also be diagnosed with/suffering with other medical conditions as well.

3. 1547 people received no change in their diabetes medication as compared to the 2128 that did, even though the hb4A1C level >8, this could be a serious issue and further investigation should be done on it.

- To examine if there is multicollinearity in the dataset, we plotted a correlation matrix between the variables, but it was not very informative. Then we moved to Variance Inflation Factor(VIF) and got the following results:

	Feature	VIF
0	age	6.342085
1	admission_type_id	2.345049
2	discharge_disposition_id	1.367123
3	admission_source_id	9.466665
4	time_in_hospitals	4.232566
5	medical_specialty	5.009134
6	max_glu_serum	25.997039
7	insulin	5.393138
8	diabetesMed	5.298389
9	diag_1_name	2.197467
10	dm	6.275119
11	medical_complexity	10.470701
12	num_visits	1.277937
13	A1Cresult_1	2.594557
14	A1Cresult_2	16.768115
15	A1Cresult_3	2.019038
16	change_1	4.109236

But some variables here have multicolinearity as they are >10. To tackle this multicollinearity, we removed the variables which have VIF values >10 and thus removed multicollinearity as well.

Updated VIF Data:

	Feature	VIF
0	age	5.458128
1	admission_type_id	2.299173
2	discharge_disposition_id	1.359219
3	admission_source_id	8.277598
4	time_in_hospitals	3.318171
5	medical_specialty	4.625067
6	insulin	5.344505
7	diabetesMed	4.349381
8	diag_1_name	2.127703
9	dm	6.202635
10	num_visits	1.261067
11	A1Cresult_1	1.145115
12	A1Cresult_3	1.065335
13	change_1	3.205285

No features with VIF > 10 found. Exitin

Final - post removing multicollinearity variables

Model	Class	Precision	Recall	Accuracy	F1_score	support
Logistic Regression	0 1	0.95 0.12	1.00 -	0.95	0.98 -	13151 621
Balanced Randomforest Classifier	0 1	0.97 0.08	0.74 0.45	0.95	0.84 0.13	13151 621
XGboost	0 1	0.96 0.38	1.00 0.01	0.95	0.98 0.02	13151 621

Based on the metrics obtained, **Balanced Randomforest Classifier** is the most balanced model, with moderate precision, recall, and F1-score. If high accuracy is a priority, **Logistic Regression** might be more suitable.

INTERPRETATION

The application of predictive modeling techniques in particular the random forest classifier was an appropriate choice of model since the data comprised of a lot of observations and missing data. this approach helped us in imputing missing data.

Our analysis was able to capture the relationship between HbA1C and the change in diabetes medication. This could be crucial information when routine check-ups over multiple encounters are done and to optimize patient care.

After our initial modelling we discovered the imbalance in the dataset and thus it was important to address the presence of missing values and as well as conducting feature engineering to obtain relevant features. Once the dataset was pre processed and scrutinized well, and muticollinearity was dealt with, we fitted popular ML models such as Balanced Random Forest, XG Boost and Logistic Regression and saw that a Balanced Random Forest is the best choice out of all the predictive models.

Also, It was a concerning investigation to find the striking rate of number of medicines dosed to the patients. It might be important for the hospitals administration to look into the same.

CONCLUSIONS

By understanding the factors that contribute to readmissions, healthcare providers can better plan resource allocation so that issues such as “not prescribing medication, even if the HbA1C levels are high do not arise”. Improved allocation could boost patient satisfaction scores and would improve overall quality of healthcare.

The EDA done on different drug results can be helpful for pharma companies to assess the quality of drug and compare different drugs together to better understand which drug combinations are suitable for diagnosing which medical illnesses.

LIMITATIONS AND FUTURE SCOPE

The data has high dimensions and is highly imbalanced. Newer methods to tackle feature imbalance can be used apart from using SMOTE for class imbalance data.

Patient segmentation on demographics can be done to better understand it from a hospital as well as a pharmacological view.