Data Science Project Report

Data Science Enthusiasts

Aditya Jain (2020554), Aryan Vohra (2020557), Rupin Oberoi (2020571), Sarthak Maini (2020576)

Data Description

- The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals. Each row pertains to the medical records of individuals diagnosed with diabetes. These patients underwent laboratory tests, received prescribed medications, and were hospitalized for a duration of up to 14 days.
- It includes 47 features representing patient and hospital outcomes.
- It contains 10 numerical and 37 categorical features.
- Some of these attributes contain missing values like race, weight, payer_code, medical_speciality and diagnosis.
- Few columns are unique to each row, like the encounter_id, patient_nbr .
- Columns admission_type_id, discharge_disposition_id, admission_source_id are the categorical columns, each entity mapping to a special category.

admission_type_id	discharge_disposition_id	number_emergency	number_outpatient	number_inpatient	diag_1	diag_2	diag_3	max_glu_serum
1	1	0	0	0	276	250.01	255	None
6	1	0	0	0	491	250.03	999	>300
6	3	0	0	0	332	294	425	>200

Some of the most important features of the dataset are:

- Discharge_disposition_id - It is an integer identifier corresponding to 29 unique values, such as being discharged to home, having expired, or not being available, for instance.
- Diagnosis - The patient's diagnosis is encompassed by 3 categorical attributes, namely diag_1, diag_2, and diag_3. They represent the different types of diagnosis undergone by the patient. diag_1 represents the primary diagnosis and has 848 distinct values, diag_2 represents the secondary diagnosis and has 923 distinct values and diag_3 represents the additional secondary diagnosis and has 954 distinct values.
- Max_glu_serum - It contains categorical values representing the range of the glucose test results like >200, >300, normal, and none if not measured.
- Admission_type_id - It is an integer mapping to different types of admissions of the patients admitted in these 130 hospitals. It contains 9 distinct values like 1 → emergency, 4 → newborn, 7 → trauma center etc.
- Number of outpatients, inpatients and emergencies - These features contain integer values representing the count of the corresponding attribute visits in the year preceding the encounter. Here outpatient refers to a patient who does not stay in the hospital for treatment overnight, whereas inpatient refers to one who lives in the hospital under treatment.

Existing Analysis

The following URLs were obtained on the first page for the search query ‘analyze diabetes 130-us hospitals for years 1999-2008’:

- https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
- https://www.kaggle.com/datasets/brandao/diabetes
- https://saurabhraj5162.medium.com/diabetes-130-us-hospitals-for-years-1999-2008-hospital-readmission-823ff48272f9
  Data cleaning:
  - ‘weight’ column dropped due to high proportion of missing values
  - Imputing missing values in race according to the proportion of each race.
  - For age range, replace with median age.
- https://medium.com/analytics-vidhya/diabetes-130-us-hospitals-for-years-1999-2008-e18d69beea4d
  Data cleaning:
  - For age range, replace with median age.
  - Dropping duplicate records
  - Merging some categories into fewer ones, such as codes corresponding to similar descriptions for admission sources were merged into one.
  - Most common value used in place of ‘?’ for missing values
  - Creating new features
    a) Health_index = (1 / (number_emergency + number_inpatient + number_outpatient))
    b) severity_of_disease = (time_in_hospital + num_procedures + num_medications + num_lab_procedures + number_of_diagnoses)
    c) Number of changes:- (changes in proportion of medicines for patients)
  - Removing least significant features chi-squared test, and Spearman correlation coefficient
- https://fairlearn.org/main/user_guide/datasets/diabetes_hospital_data.html
- https://github.com/Mohith-Kota/Diabetic-patients-Readmission-in-Hospitals-Prediction
  Data cleaning:
  - ‘weight’ column dropped due to high proportion of missing values
  - Disregarding the entries of patients who might not return to the hospital. The discharge reason of the patient has been described in the discharge_disposition_id. The entries of patients whose discharge_disposition_id is expired, hospice/home, hospice/medical facility, expired at home were removed.
  - Features ‘encounter_id’ and ‘patient_nbr’ are removed as they represent the unique row and do not contribute to the model building
  - One hot encoding of categorical features
- https://github.com/VolodymyrGavrysh/My_RoadMap_Data_Science/blob/master/data_analysis/Diabetes%20130-US%20hospitals%20for%20years%201999-2008%20Data.ipynb
  Data cleaning:
  Replace ‘?’ with Nan
- https://www.researchgate.net/publication/342095039(Paper not publicly available)
- Data Science Project.ipynb

Problem Statement

Challenges Faced: The main challenge encountered while data preparation was that of feature selection. Since there were very few attributes in the dataset which had a significant correlation with the target variable, therefore additional analysis with techniques like Recursive Feature Elimination needed to be performed. In addition to this, many features in the dataset had a significantly high correlation to each other, thereby, they were combined into a single feature. Additional analysis had to be done in this regard as well, like observing the Variance Inflation Factor of those attributes to check for multicollinearity.

Other New Techniques Explored : We applied the technique of Near Zero-Variance variable removal by which one of those attributes which had close to zero variance between them was removed. It was done because it would offer little information in the prediction of the target feature. Also, different data balancing techniques were explored, and the best performance was given by ROSE (Random Over Sampling Examples), which was then applied to the unbalanced attributes.

Hypotheses anticipated:

1) Independence of race and readmission: Since both are categorical in nature, therefore this can be verified by the Chi-square test for independence between features.

2) High dependence between the features - time_in_hospital and number of unique medications (num_medications). This is anticipated because if a patient is given a variety of medications, it is an indication that the patient is not reacting to the treatment positively they are more likely to stay in the hospital longer.

3) High dependence between the features - num_procedures and num_medications.This is anticipated because more distinct procedures are expected to involve a higher number of unique medications as well.

4) We also expect that older people are more likely to be readmitted as compared to younger people because there might be increased chances of side effects and complications due to a weaker immune system.

5) We also anticipate that older people will also have a higher time in the hospital because the healing process might be slower.

Features to learn from data

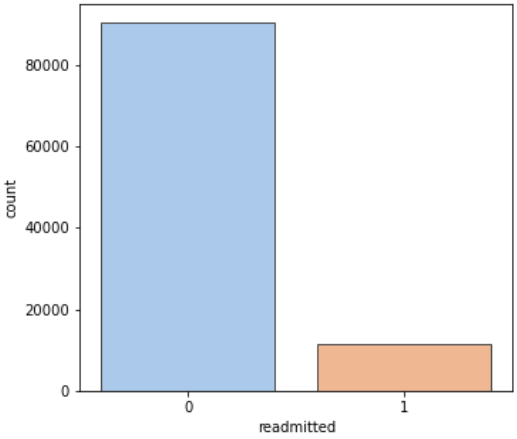
The primary focus of our work will be on predicting the readmission given the records of any patient i.e. the readmission feature will act as the target variable.
We would be focussing on analyzing the relationship between different categories of admission_type_id and admission_source_id with the target variable, which has not been explored by any previous works on the dataset.

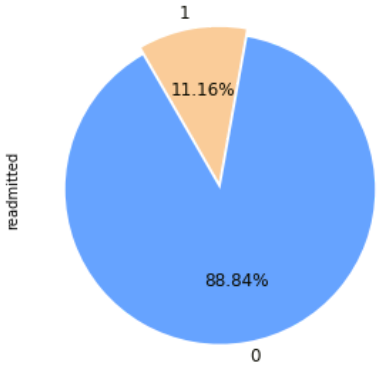
Besides this, we will also study the relations between various features such as type of admission (emergency, urgent) and age, as well as that between age and the admission source. Since we have medical data, we also aim to calculate the risk factor for readmission using domain-specific techniques such as the Cox Proportional Hazards model.

Appendix:-

1. https://www.kaggle.com/code/hkubra/predicting-the-readmission-of-diabetic-patient-s

Plots of whether a diabetic patient was readmitted or not (target variable)



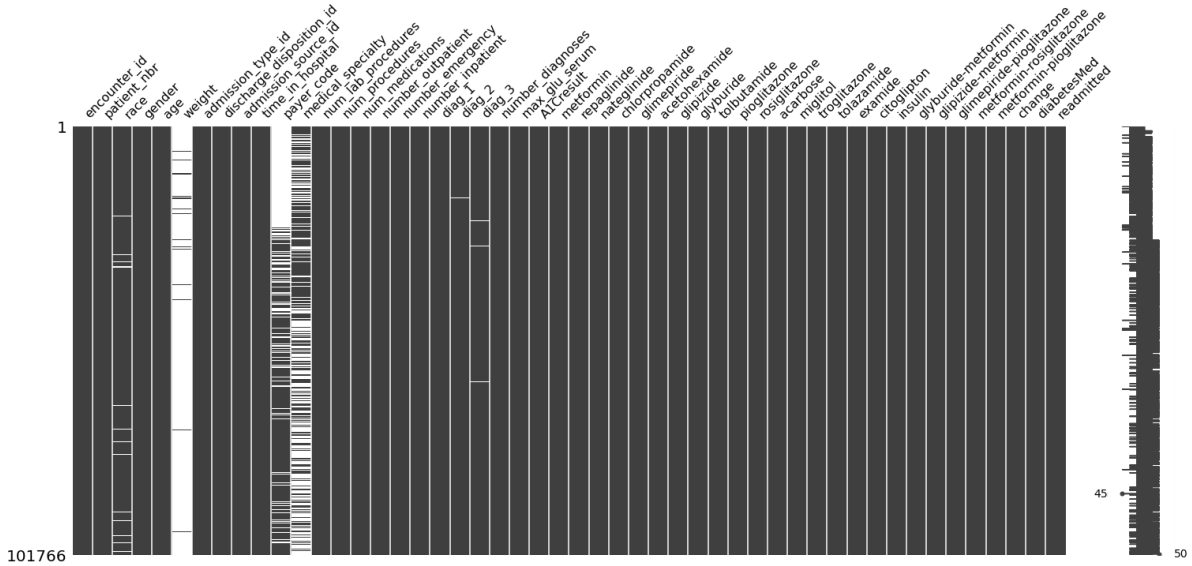


We can see that the number of records for class 1 is a lot less than for class 0 , hence class balancing techniques need to be used for good results

Resampling techniques _ Undersample majority class was used.

The number of not_readmitted (class 0) attributes used was undersampled so that we in the end have a len same as those that were readmitted

Missing value matrix (MSNO)

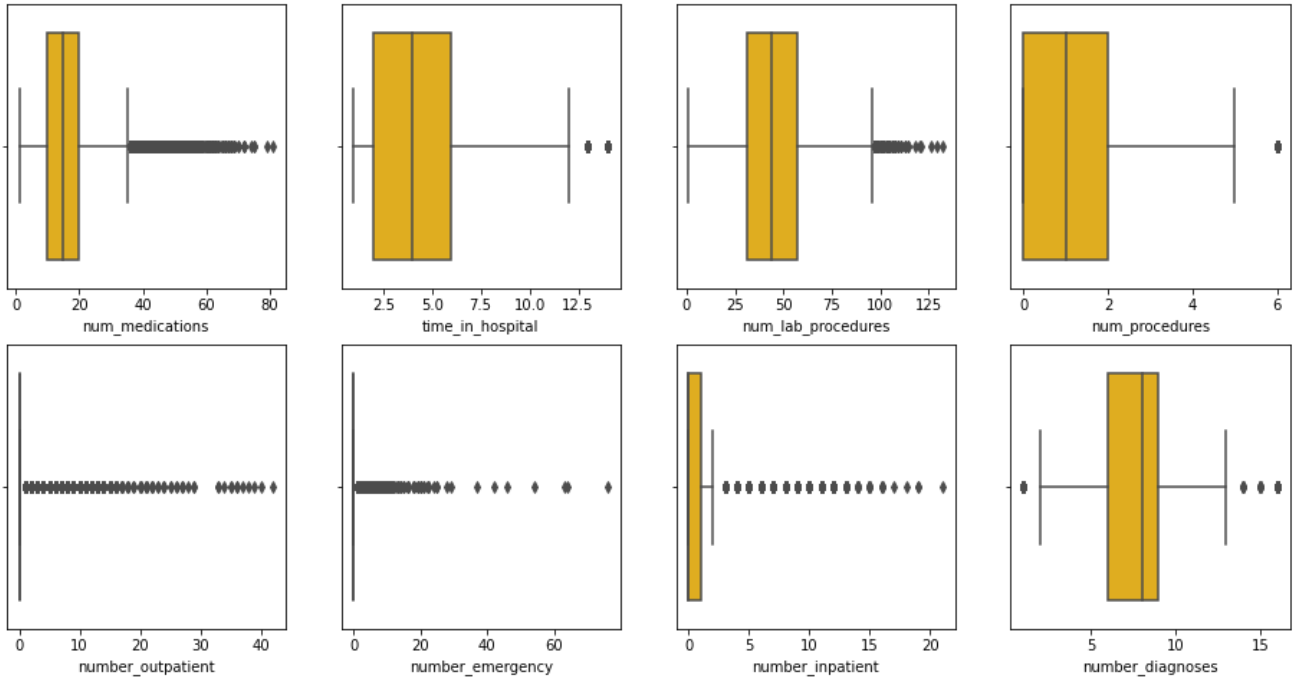


Order of columns from most to least missing values includes
['Weight' , medical_speciality , payer_code , race]

	#_Total_Missing_Value	%_Missing_Value_Rate	Data_Type	Unique_Value	Total_Unique_Value
weight	98569	0.9686	object	[nan, [75-100), [50-75), [0-25), [100-125), [2...	10
medical_specialty	49949	0.4908	object	[Pediatrics-Endocrinology, nan, InternalMedici...	73
payer_code	40256	0.3956	object	[nan, MC, MD, HM, UN, BC, SP, CP, SI, DM, CM, ...	18
race	2273	0.0223	object	[Caucasian, AfricanAmerican, nan, Other, Asian...	6
diag_3	1423	0.0140	object	[nan, 255, V27, 403, 250, V45, 38, 486, 996, 1...	790

Result : Due to high percentage of missing values, Weight, medical_speciality and payer_code were removed

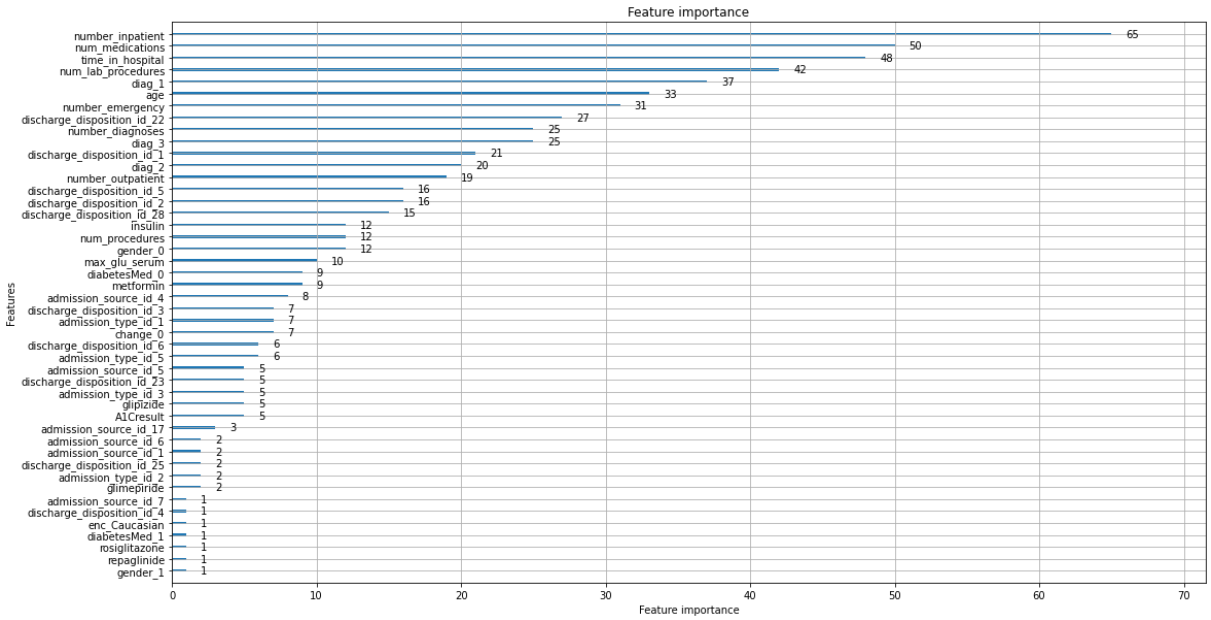
Outlier visualization with box plot



Use of Interquartile range to remove outliers was used

Method used:-
Define the normal data range with lower limit as Q1–1.5*IQR and upper limit as Q3+1.5*IQR. Any data point outside this range is considered as outlier and should be removed for further analysis.

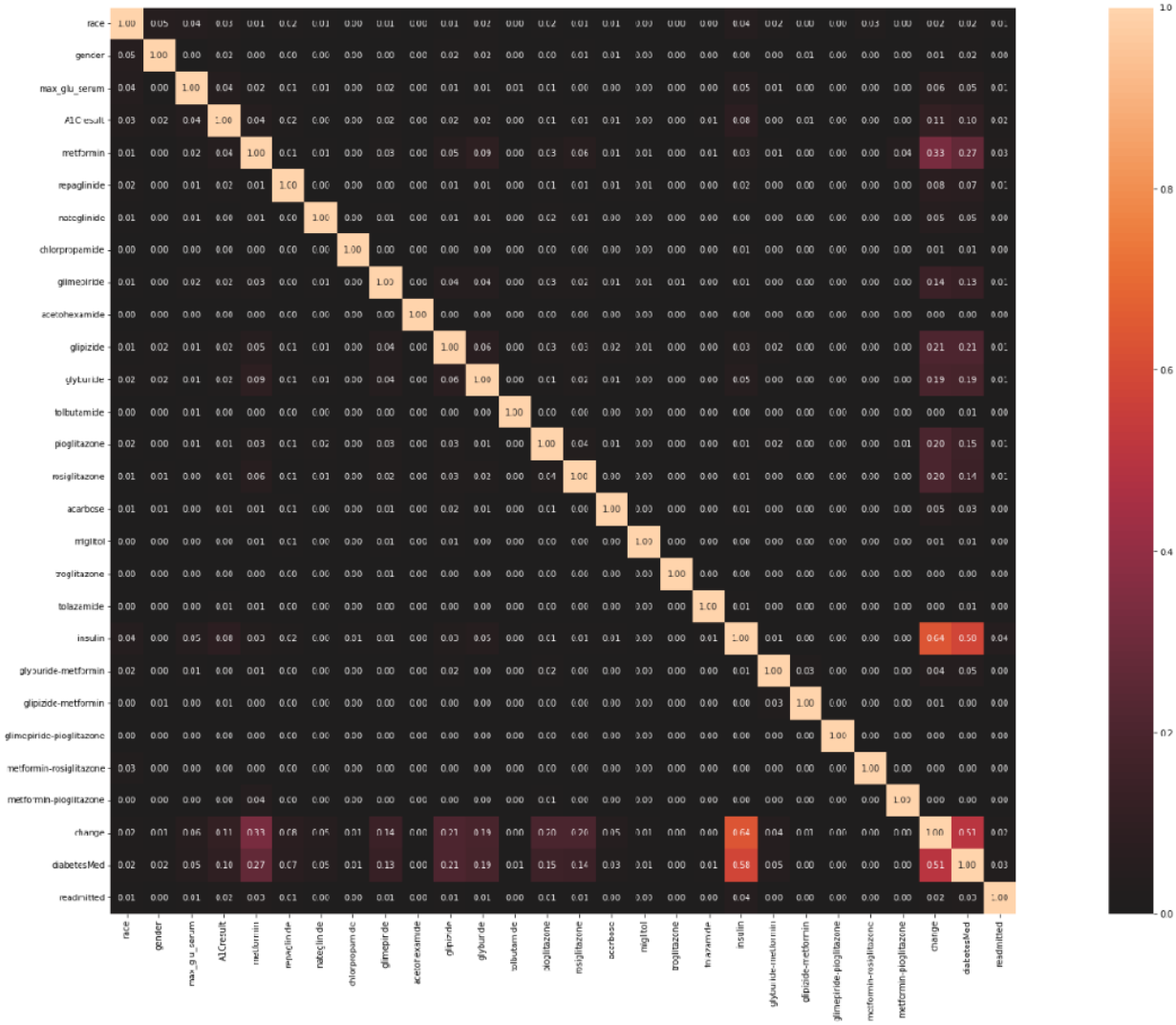
Feature Importance:-



Lightgbm library was used to get feature importance of various columns and the following graph was obtained

2.
https://github.com/Mohith-Kota/Diabetic-patients-Readmission-in-Hospitals-Prediction/blob/main/Report.pdf

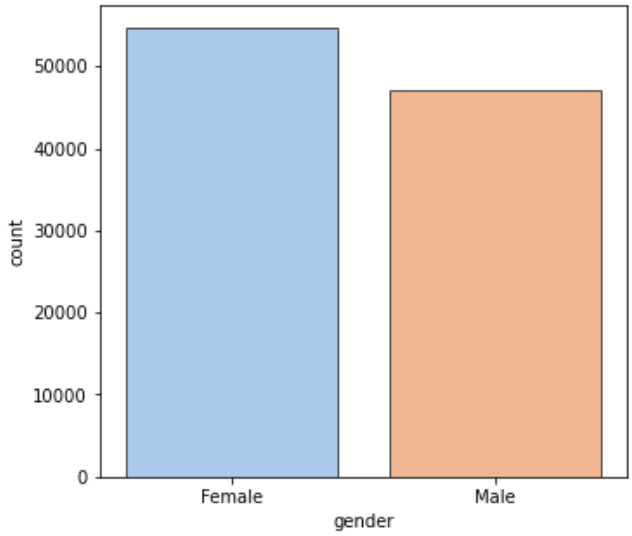
Correlation coefficient

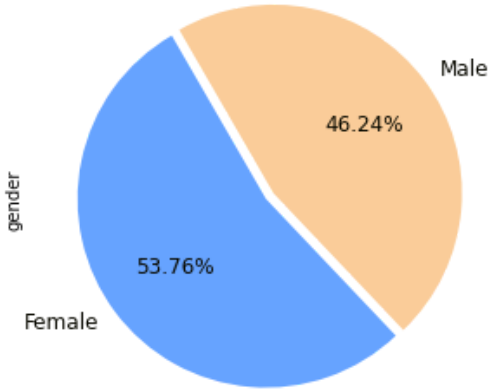


Results :

- The above heatmap shows that dependency among independent features is less which is good for model training.
- The highest correlation is among the Features ‘insulin’ and ‘Change’ which clearly shows that change in insulin dosage has considerable effect on the readmission.
- Also Features ‘diabetesMed’ and ‘insulin’ are closely correlated as Insulin is the most prescribed and prevalent medicine for diabetes.

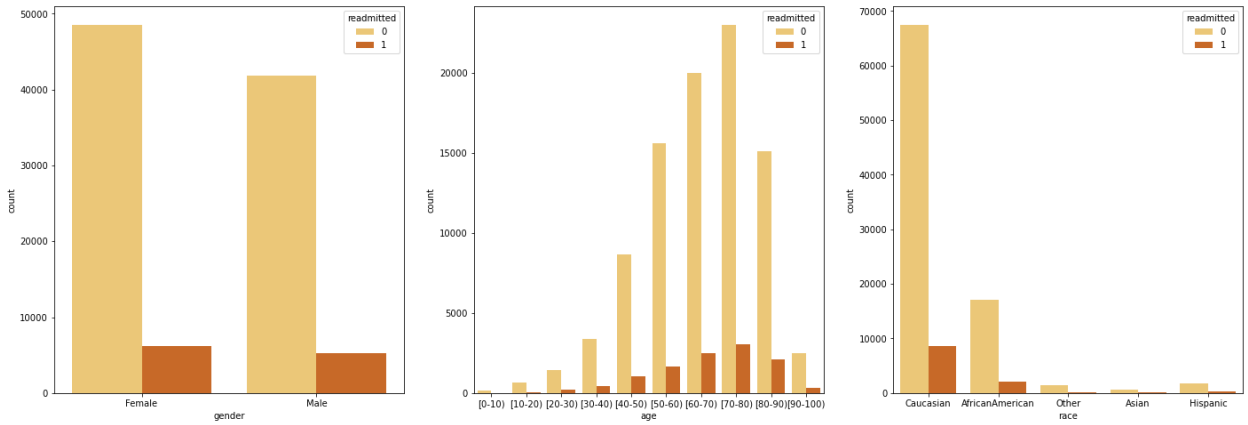
Gender Distribution





The model can be biased towards female re-admittion cases if it is trained on this dataset . Hence balancing on this female to male ratio needs to be done.

Gender, Age and Race v/s Readmitted patients



Since age is an ordinal attribute which has a meaningful sequence of values . Ordinal encoding was used on the ‘age’ column

3. https://www.kaggle.com/code/kavyarall/predicting-effective-treatments

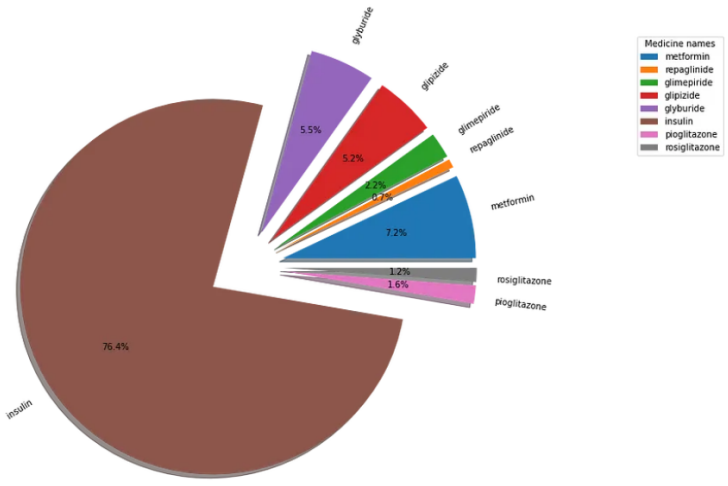
Problem of general Correlation test -

Feature / Response	Continuous	Categorical
Continuous	Pearson’s Correlation	LDA
Categorical	Anova	Chi-Square

With respect to the problem statement given, the output variable is observed to be the “readmitted” feature. The input variables are both Discrete Quantitative and Categorical and our output variable

is Categorical. Since we have a combination of Discrete Quantitative Variables and Categorical Variables, we cannot perform general Correlation tests.

4. https://saurabhraj5162.medium.com/diabetes-130-us-hospitals-for-years-1999-2008-hospital-readmission-823ff48272f9



This plot only considers medicine with at least 100 cases of upgradation. Clearly, of all the encounters where dosage were increased, Inuslin has the majority of the cases (76.1%), followed by Metformin. There are many medicines whose dosage were not increased, not even in a single case.

5. https://medium.com/analytics-vidhya/diabetes-130-us-hospitals-for-years-1999-2008-e18d69beea4d

Feature engineering:-

```
high_frequency = ['InternalMedicine', 'Family/GeneralPractice', 'Cardiology', 'Surgery-General',
                  'Emergency/Trauma', 'Urology','ObstetricsandGynecology','Psychiatry','Pulmonolo

low_frequency = ['Surgery-PlasticwithinHeadandNeck','Psychiatry-Addictive','Proctology','Dermato
                 'Neurophysiology','Resident','Pediatrics-Hematology-Oncology','Pediatrics-Emerge
                 'Pediatrics-Pulmonology','Surgery-Pediatric','AllergyandImmunology','Pediatrics-
                 'Endocrinology-Metabolism','PhysicianNotFound','Surgery-Colon&Rectal','OutreachS
                 'Surgery-Maxillofacial','Rheumatology','Anesthesiology-Pediatric','Obstetrics','

pediatrics = ['Pediatrics','Pediatrics-CriticalCare','Pediatrics-EmergencyMedicine','Pediatrics-
               'Pediatrics-Neurology','Pediatrics-Pulmonology', 'Anesthesiology-Pediatric', 'Car

psychic = ['Psychiatry-Addictive', 'Psychology', 'Psychiatry',  'Psychiatry-Child/Adolescent', '

neurology = ['Neurology', 'Surgery-Neuro',  'Pediatrics-Neurology', 'Neurophysiology']


surgery = ['Surgeon', 'Surgery-Cardiovascular', \
          'Surgery-Cardiovascular/Thoracic', 'Surgery-Colon&Rectal', 'Surgery-General', 'Surgery
             'Surgery-Plastic', 'Surgery-PlasticwithinHeadandNeck',  'Surgery-Thoracic',\
             'Surgery-Vascular', 'SurgicalSpecialty', 'Podiatry']

ungrouped = ['Endocrinology','Gastroenterology','Gynecology','Hematology','Hematology/Oncology',
            'Oncology','Ophthalmology','Otolaryngology','Pulmonology','Radiology']


missing = ['?']
```

i. Grouping similar categories

Endocrinology →glands, Gastroenterology →stomach , Gynecology →women reproduction system, Hematology →Blood
Hematology/Oncology →Blood, Hospitalist →one who takes care of admitted patients, Oncology →cancer, Ophthalmology →eye, otolaryngology →ears, nose, and throat
Pulmonology →respiratory, Radiology — diagnosing and treating injuries and diseases using medical imaging (radiology) procedures (exams/tests) such as X-rays

ii. Domain knowledge provided in the description file to create fewer categories.

```
data['diag_1'] = data['diag_1'].apply(lambda x : 'other' if (str(x).find('V') != -1 or str(x).fi
                                      else ('circulatory' if int(float(x)) in range(390, 460)
                                      else     ('respiratory' if int(float(x)) in range(460, 5
                                      else     ('digestive'   if int(float(x)) in range(520, 5
                                      else     ('diabetes'    if int(float(x)) == 250
                                      else     ('injury'      if int(float(x)) in range(800, 1
                                      else ('musculoskeletal' if int(float(x)) in range(710, 7
                                      else ('genitourinary'   if int(float(x)) in range(580, 6
                                      else ('neoplasms'       if int(float(x)) in range(140, 2
                                      else ('pregnecy'        if int(float(x)) in range(630, 6
                                      else 'other')))))))))
```

1.Circulatory → 390–459, 785 → Diseases of the circulatory system
2.Respiratory → 460–519, 786 → Diseases of the respiratory system

3.Digestive → 520–579, 787 → Diseases of the digestive system
4.Diabetes → 250.xx → Diabetes mellitus
5.Injury → 800–999 → Injury and poisoning
6.Musculoskeletal → 710–739 → Diseases of the musculoskeletal system and connective tissue etc.

iii. Creating new features

```
number_of_changes = []
for i in tqdm(range(len(data))) :
    changeCount = 0
    for col in drugList :
        if data.iloc[i][col] in ['Down', 'Up'] :
            changeCount += 1
    number_of_changes.append(changeCount)

data['number_of_changes'] = number_of_changes
```

a) Health_index = (1 / (number_emergency + number_inpatient + number_outpatient))

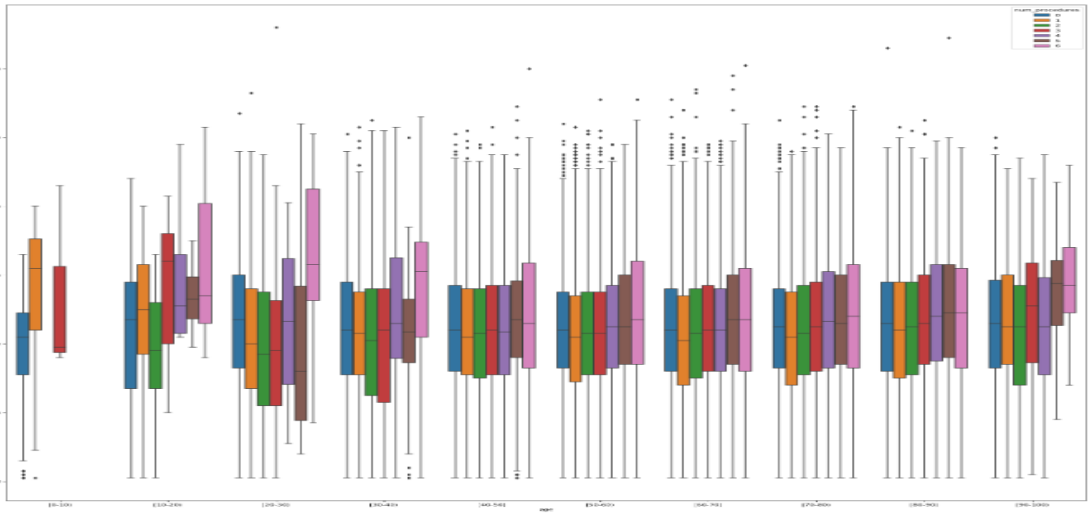
b) severity_of_disease = (time_in_hospital + num_procedures + num_medications + num_lab_procedures + number_of_diagnoses)

c) Number of changes :- (changes in proportion of medicines for patients)

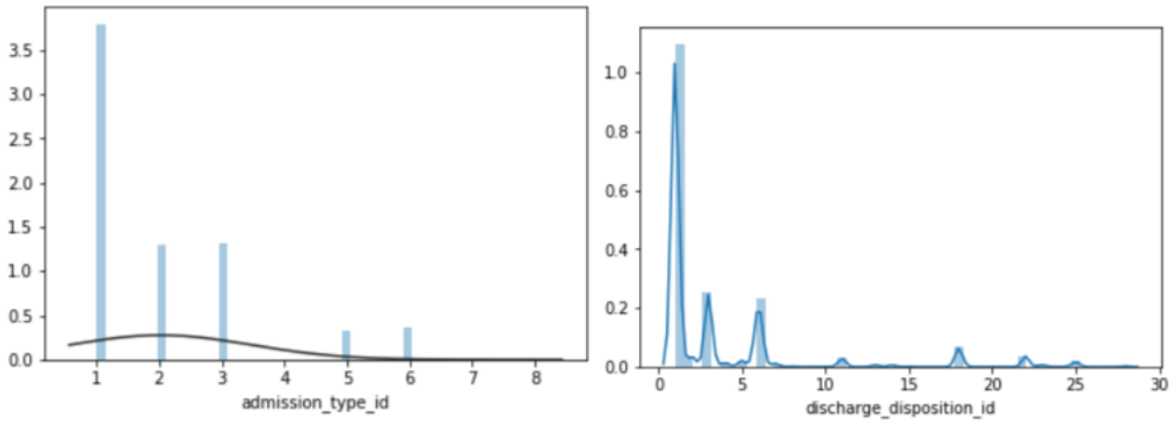
iv) Removing least significant features using Chi Squared test and Spearman's correlation coefficient

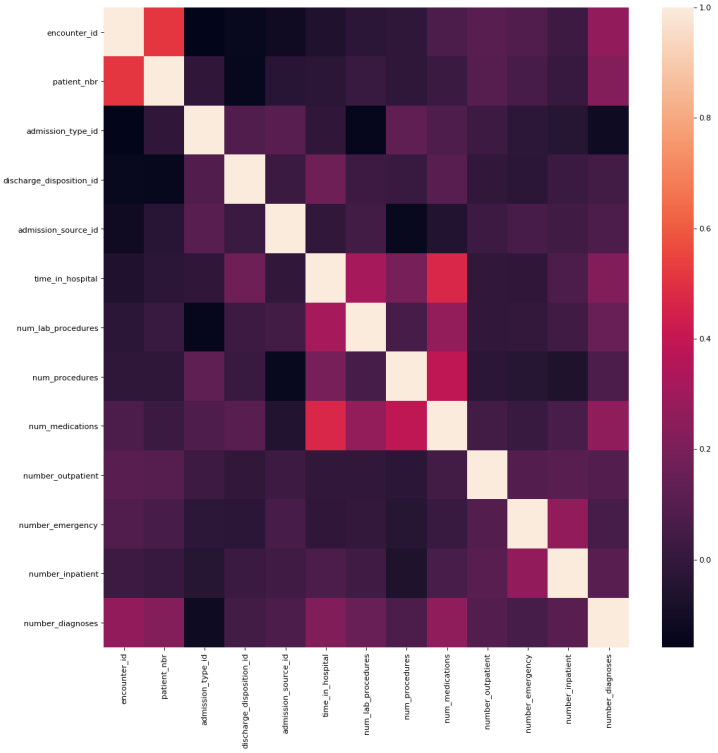
6. [My_RoadMap_Data_Science/data_analysis/Diabetes 130-US hospitals for years 1999-2008 Data.ipynb at master · VolodymyrGavrysh/My_RoadMap_Data_Science (github.com)]

Boxplots for num_procedure for various age groups shows significant number of outliers for 50-60 and 60-70 year age groups



Distribution plots for ‘admission_type_id’ and ‘discharge_disposition_id’.





It was observed by the author of this code from the correlation matrix that 3 factors are substantially to the time spent by patients in hospital- number of procedures, number of medications, number of lab procedures.