**Hospital Readmission (1999-2008)**

Diabetes is a chronic disease that affects the way the body processes blood sugar (glucose). There are two main types of diabetes: type 1 and type 2. Type 1 diabetes is usually diagnosed in children and young adults, and occurs when the body's immune system mistakenly attacks and destroys the cells in the pancreas that produce insulin. Type 2 diabetes is the most common form of diabetes and usually occurs in adults, although it is increasingly being diagnosed in children and adolescents. In type 2 diabetes, the body does not produce enough insulin or does not use it properly.

According to the International Diabetes Federation, in 2021, there were approximately 537 million people living with diabetes worldwide. This number is expected to rise to 643 million by 2045. Diabetes is also a leading cause of death, with an estimated 4.2 million deaths globally attributed to the disease in 2019. In addition, diabetes can cause a wide range of complications, including heart disease, stroke, kidney disease, blindness, and amputations. Therefore, it is crucial to manage and prevent diabetes through lifestyle changes, medication, and regular monitoring of blood sugar levels.

The given data

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 encounter\_id 101766 non-null int64

1 patient\_nbr 101766 non-null int64

2 race 101766 non-null object

3 gender 101766 non-null object

4 age 101766 non-null object

5 weight 101766 non-null object

6 admission\_type\_id 101766 non-null int64

7 discharge\_disposition\_id 101766 non-null int64

8 admission\_source\_id 101766 non-null int64

9 time\_in\_hospital 101766 non-null int64

10 payer\_code 101766 non-null object

11 medical\_specialty 101766 non-null object

12 num\_lab\_procedures 101766 non-null int64

13 num\_procedures 101766 non-null int64

14 num\_medications 101766 non-null int64

15 number\_outpatient 101766 non-null int64

16 number\_emergency 101766 non-null int64

17 number\_inpatient 101766 non-null int64

18 diag\_1 101766 non-null object

19 diag\_2 101766 non-null object

20 diag\_3 101766 non-null object

21 number\_diagnoses 101766 non-null int64

22 max\_glu\_serum 101766 non-null object

23 A1Cresult 101766 non-null object

24 metformin 101766 non-null object

25 repaglinide 101766 non-null object

26 nateglinide 101766 non-null object

27 chlorpropamide 101766 non-null object

28 glimepiride 101766 non-null object

29 acetohexamide 101766 non-null object

30 glipizide 101766 non-null object

31 glyburide 101766 non-null object

32 tolbutamide 101766 non-null object

33 pioglitazone 101766 non-null object

34 rosiglitazone 101766 non-null object

35 acarbose 101766 non-null object

36 miglitol 101766 non-null object

37 troglitazone 101766 non-null object

38 tolazamide 101766 non-null object

39 examide 101766 non-null object

40 citoglipton 101766 non-null object

41 insulin 101766 non-null object

42 glyburide-metformin 101766 non-null object

43 glipizide-metformin 101766 non-null object

44 glimepiride-pioglitazone 101766 non-null object

45 metformin-rosiglitazone 101766 non-null object

46 metformin-pioglitazone 101766 non-null object

47 change 101766 non-null object

48 diabetesMed 101766 non-null object

49 readmitted 101766 non-null object

The project will be divided into the following stages:

1. Data Extraction: The first stage of the project will involve extracting data from the Edmunds API using Python. The data will include information on used car prices, specifications, and reviews.
2. Data Cleaning and Pre-processing: Once the data is extracted, the next stage will involve cleaning and pre-processing the dataset. This will involve handling missing values, removing duplicates, correcting data types, and handling outliers.
3. Exploratory Data Analysis: Once the data is cleaned and pre-processed, the next stage will involve exploring the data to gain insights and identify patterns. This will involve visualizations and statistical analysis.
4. Feature Engineering: Based on the insights gained from the exploratory data analysis, the next stage will involve feature engineering to create new features that can help improve the performance of the machine learning model. 5. Model Building and Evaluation: The final stage of the project will involve building and evaluating a machine learning model that can predict the price of a used car. The model will be trained on a subset of the data and evaluated using various performance metrics such as accuracy, precision, recall, and F1 score

The given dataset has many terms which can be divided into different subsets:

**Patient data:** Race, Gender, Age, Weight

**Hospital Entry data:**Admission type, Discharge disposition, Admission source, Time in hospital

**Medicine and other health data:** Number of diagnoses, Primary and Secondary diagnoses , Medications prescribed

**Test :** Glucose serum test, Hemoglobin test

**Readmission :** <30 days, >30 days or No readmission.

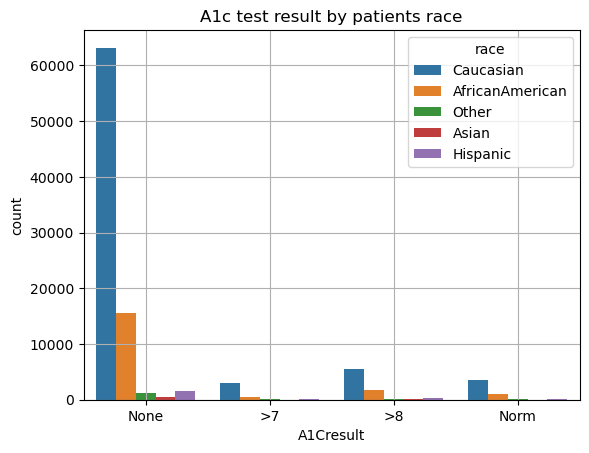
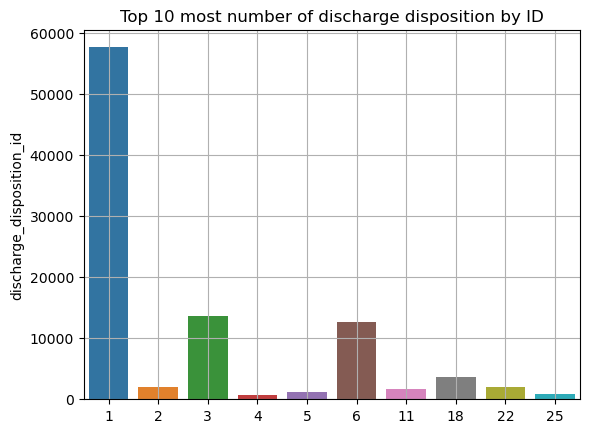
To clean the data we have multiple levels

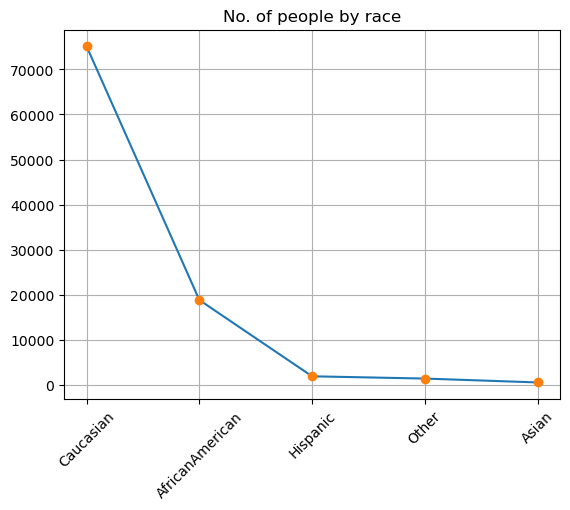
1. First we have to remove the null values

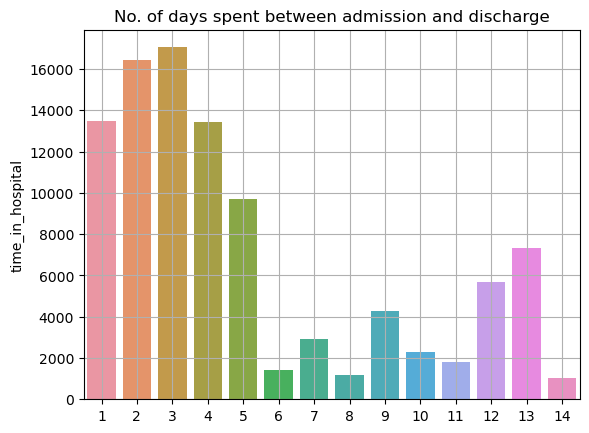
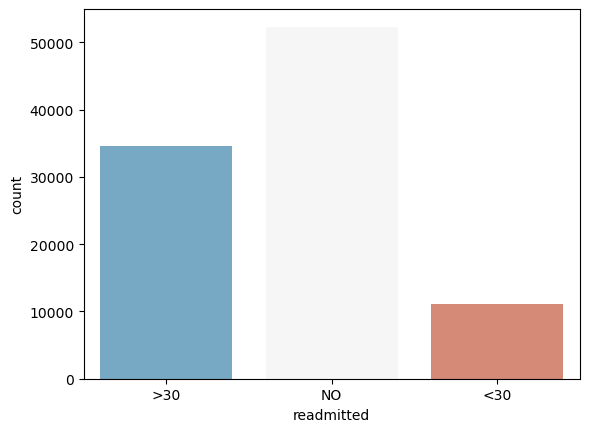
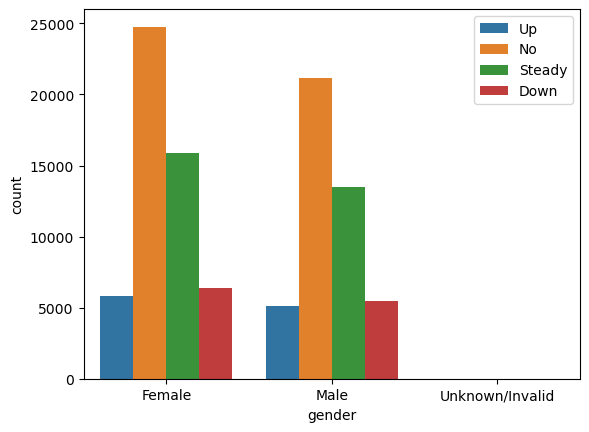
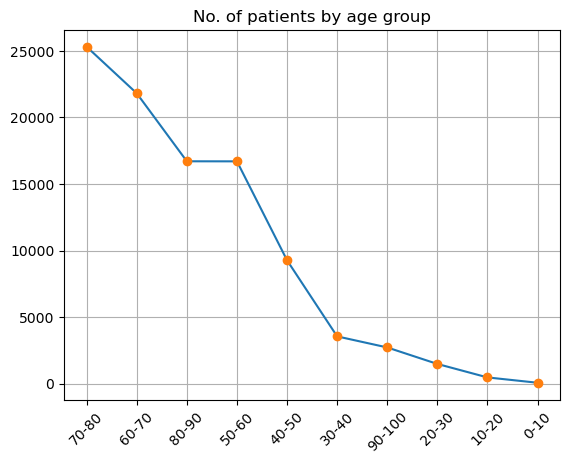
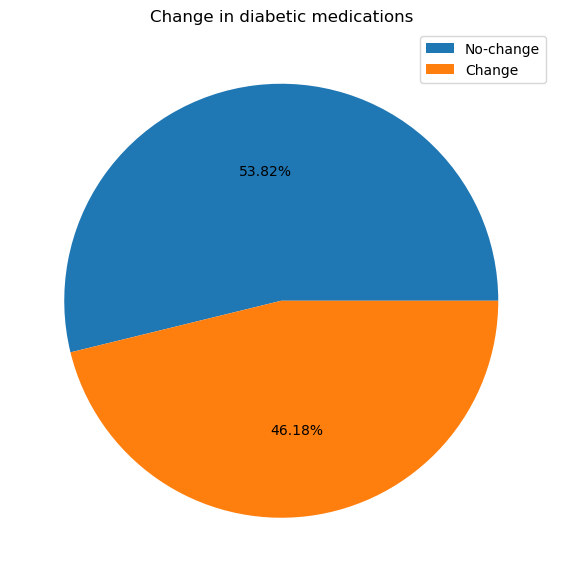
2. Drop the columns which have lot of null values

3. Drop the duplicate values

4. Clean the string impurities using the replace

Data visualization





Machine learning

**For machine learning process we have to first draw the correlation heatmap to check the correlation.**

**Form X and Y parts form ml model.**

**Preform a train test split.**

**Standardise the data using standard scaler.**

**Use degree of dimensionality to use only required columns.**

**Use randomunsampler to balance the data.**

**Execute different machine learning model.**

**Using different Naive Bias and classifiers to see weather it improves the ml model output.**

|  |  |  |
| --- | --- | --- |
| ML Model | Training Score | Testing Score |
| Logistic regression | 0.558 | 0.555 |
| Kneighebours Classifier | 0.662 | 0.468 |
| Decision Tree | 0.861 | 0.434 |
| Random Forest | 0.863 | 0.513 |
| Bernoullie NB | 0.539 | 0.533 |
| Gaussian NB | 0.533 | 0.532 |
| ADA Boost | 0.566 | 0.558 |
| Gradient Boost | 0.576 | 0.571 |
| XG boost | 0.67 | 0.549 |
| XG boost (depth=3) | 0.585 | 0.56 |
| XG boost (depth=2) | 0.574 | 0.561 |

Conclusion

After preforming different ml processes we realized the most ideal model for this data is logistic regression model which was bit expected but after using different ensemble learning we realized using Gaussian NB we get the best model