Data analysis:

General :

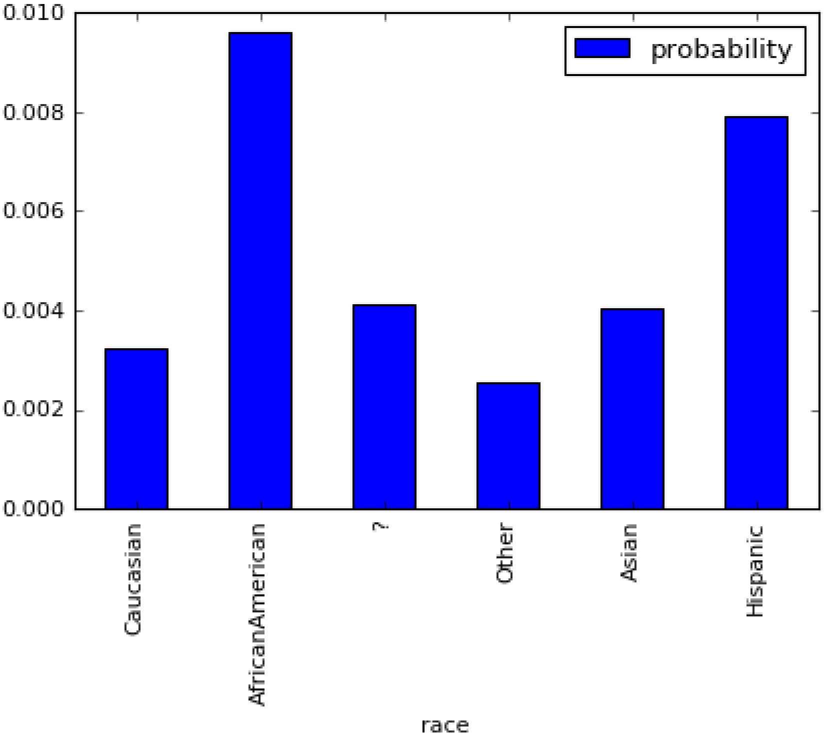
- The data is composed of clinical records collected across U.S hospitals , regardin patient connected to diabetes.
- Data contains 101766 samples , and 50 features.
- There are 71518 unique patients.
- There are 40 nominal columns , and 8 numerical columns.(+ **encounter_id** and **patient_nbr** which are serial numbers).
- 23 of the nominal features are related to medications dosage change.
- Some feature characterize the patient demographic and medically , (e.g : race , age , admission source) and others characterize the medical treatment the patient received at the hospital (e.g : A1c test results , diagnoses , etc.)
- Since there are many patient with more than one encounter , i only used each patient's first encounter.

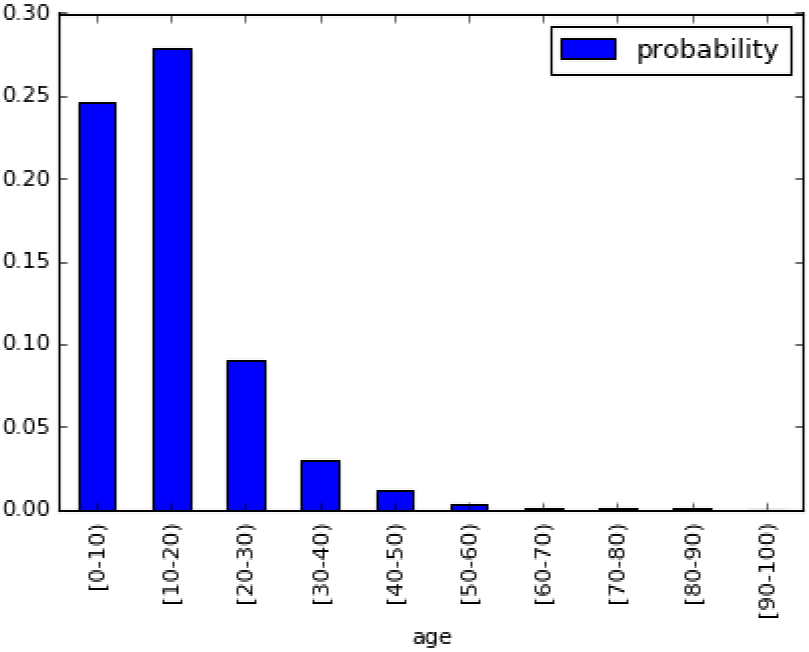
Features connection:

- In order to find connection between the numerical features i checked correlation (results attached in csv file) , it showed high correlation between 'time in hospital' , 'num diagnoses' , 'num medication' and other similar features (which is quite reasonable , and also makes them candidates for dimension reduction).
- In order to find connections between nominal features i checked how much feature A existence effects feature B probability(minimum 100 samples required).

  Some interesting examples :

  1. Being african american increases the probability of Hypertensive renal disease and Heart disease
  2. Age 40-50 increases the probability of Symptoms involving respiratory system
  3. Age 10-20 increases the probability of Diabetes with ketoacidosis, type I. (Results attached in csv file).


0.010
0.008
0.006
0.004
0.002
0.000
probability
Caucasian
AfricanAmerican
?
Other
Asian
Hispanic
race


0.30
0.25
0.20
0.15
0.10
0.05
0.00
probability
[0-10)
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
age

Unsupervised methods :

- I will attempt to cluster the patients based on the patient demographic and medical features, and see if clusters have some distinction.
- I will attempt to cluster patient medical treatment features and see if clusters have some distinction.
- Finally , i will check if somehow these 2 types of clusters are related.
- Detailed Information regarding the methods will be in a separate doc.