Tomer Himi 204234454

Unsupervised Learning Project

Abstract

In this work I will investigate given data using unsupervised learning techniques. The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounter contains demographic data, medication information, hospitalizations and conclusions that were diagnosed during his stay at the hospital. The dataset is heterogeneous and inconsistent, with different missing values and dimensions. Thus, the analysis of the data becomes more complex, because the collected data isn't controlled.

My purpose is to show what I can learn from the dataset, using clustering methods.
I specified 4 known models: K-Means, DBSCAN, Agglomerative-Clustering and Gaussian Mixture.
My work flow is also separated to sections: data cleaning, statistics of the data set, pre-processing the data, clustering models testing and the last section is for conclusion and further notes.

Data Cleaning

The first step, after reading the csv file, was cleaning the data set.
The data set was spared with a lot of deficiencies and object types. This would not work on any clustering model. Thus, I decided on the following actions:

1) Replacing Nan's ids according to mappings description with new category called 'Unknown', mapped as zero for the features.
2) Dropping every feature that contains as less than 80% data.
The removed features (for the reason above) are:
weight- 96% are Nan's.
payer code- 40% are Nan's.
medical specialty- 49% are Nan's.
max glu serum- 94% are Nan's.
A1Cresult- 83% are Nan's.
3) Removing outliers, such as 'Unknown/Invalid' classification in the gender section.
4) Removing Nan's cases in some categories that 1.8% of the data contains Nan's.
5) Dealing with multiple encounters inside the data. In fact, those multiple encounters create a dependency and hit the statistical tests. Thus, I wanted to avoid those cases. I decided to take only the first encounter under consideration. I took only the first encounter, because the readmitted feature is imbalanced. If I will pick to take only the last encounter, it will increase the imbalanced of the data, due to the fact that last encounter of a patient means the value of his readmitted feature is 'NO'.

As a result of cleaning step, the data was decreased from 101,766 experiment participants to 70,413. The database has been cleaned from repeated information, Nan cases and data which contained a lot of deficiencies.

Statistics Of The Data Set

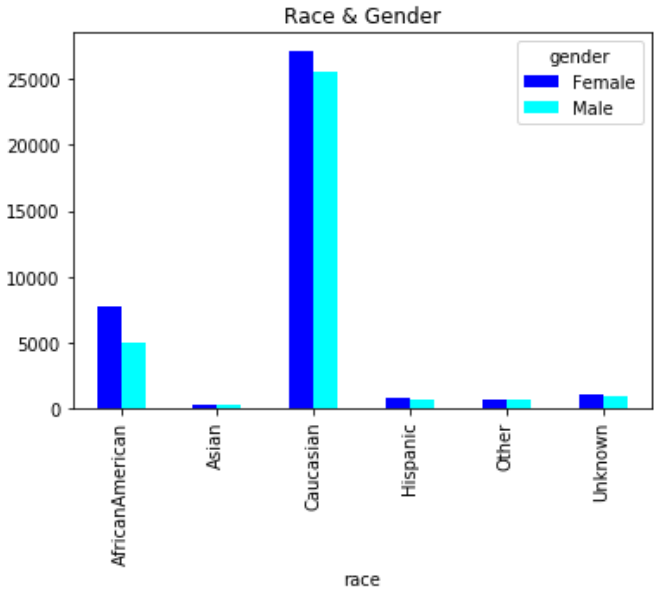
Performing preliminary analysis of the information I deal with is necessary to get to know the content of the data set and to understand what features I want to put an emphasis on. Initial conclusions about the data is done in order to focus more on the important features I can deduce most of the conclusions from them.

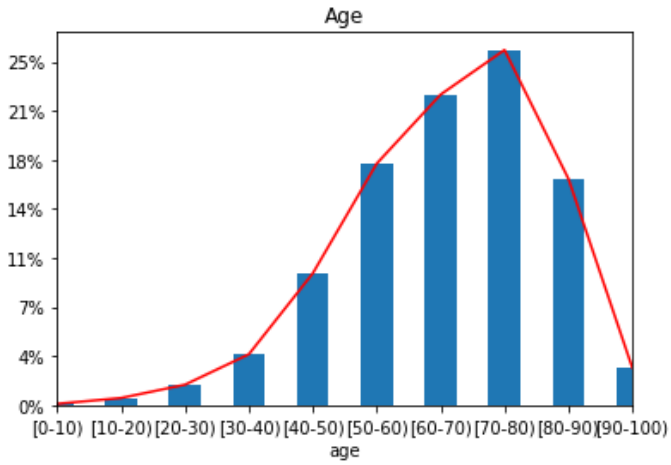
The data set contains two types of categories: person description of the patient, the hospital treatment of diabetes patient. I decided to investigate the statistics of those three types of categories and the connection between them.

Personal description

1) Most of the patients are on a certain age group [60-80]. In fact, the age section has negative asymmetric distribution, because the median is larger than the average of ages.
2) The gender section has almost uniform distribution, even though the face that there is a slight tendency towards females.
3) The race section is imbalance in favor of Caucasian and African American races.

It seems that most of the patients are Caucasians and African Americans adults at the mention age group. Nevertheless, this information isn't enough to conclude about younger ages or other races.

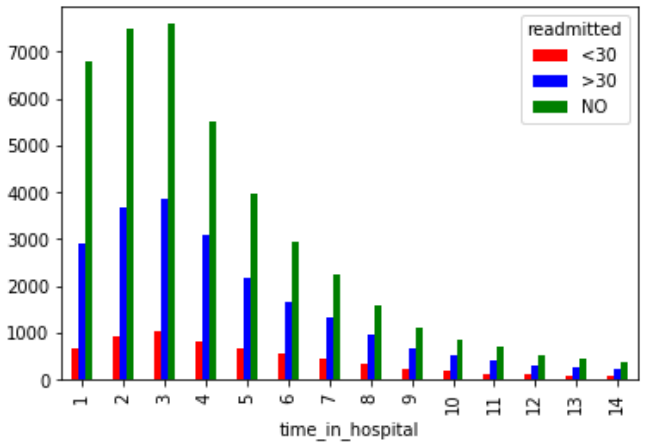




Treatment of diabetes patient

1) DiabetesMed category indicates where at least one diabetic medication prescribed. In late ages, it may imply that giving or changing dosage of drugs may be dangerous, whereas in young ages it may imply that the patients may not have diabetes at all.
2) 60% of the patients aren't readmitted. It may be for several reasons: the patients were cured form diabetes, the patients died; the medical tests found out the patients had no diabetes.
Moreover, the readmitted feature is imbalanced. I can deduce that fact, owing to the inconsistency in this feature: 31% patients are readmitted after 30 days, 8% are readmitted in less than 30 days and the remaining 60% aren't readmitted.
3) Sum of readmitted patients within less or more than 30 days (after first hospitalization) is almost equal to number of people who weren't hospitalized again.
4) Change feature, which indicated if there was change of medications, shows that patients, who change their drugs, readmitted again. Drugs change distribution, belong patients who readmitted again in less or more than 30 days, is a uniform distribution.
5) In all age scale groups the readmitted feature is remaining quite similar. The same thing happens when I check the connection between age and change of medications. Those facts may imply that the age criterion has no real impact.
6) Half of population stays in the hospital 2-6 days, when the average time in hospital is 4 days with a standard deviation of 3 days.
In fact, the more patient stays in the hospital, the higher patient's chance of returning to hospitalization. That does make sense logically, since the fact that the patient is treated for a long time, indicates a low level of patient's health which requires medical attention and follow-up.

More correlations between the different features will be in the next section of the project.



Pre-Processing The Data

In data cleaning section, I have wiped out some of the information that doesn't constitute useful information for the purpose of learning about the data set. In this section, I need to decrease number of categories with minimum data loss. I would like to take the processed data and encode the information stored in it to numerical information, so that I can run and

perform properly clustering models. The coding process is divided into two parts: handle type 'object' attributes and handle numeric features.

Handle 'object' type attributes

1) First step was coding some of 'object' type attributes to binary. The features, which were converted manually to binary, were: gender, change, diabetes Med, readmitted. For example, the gender category included only 'female' and 'male' values after data cleaning step. Thus, gender feature can be easily converted to 1 and 0 respectively.
2) I handle age category. Since I can't run scale groups in clustering models, I have decided to convert manually each age group to the average of the group.
3) There are 23 drugs columns after data cleaning. Those features indicate if a change in medicine was made during the patient visit in the hospital. I wanted to reduce the number of dimensions of the data set, while at the same time I wanted to minimize data loss. Therefore, I have decided to create a new attribute which sums the number of changes in medication dosage for each patient, and afterwards drop those features. I believe that change medication dosage during the patient visit in the hospital is related to diabetes disease.
4) I handle race category. Since I haven't any preference for one race or another, I can't order them. Thus, race attribute was converted into indicator columns ("dummy columns"). This is a way of encoding categorical features as a one numeric array.

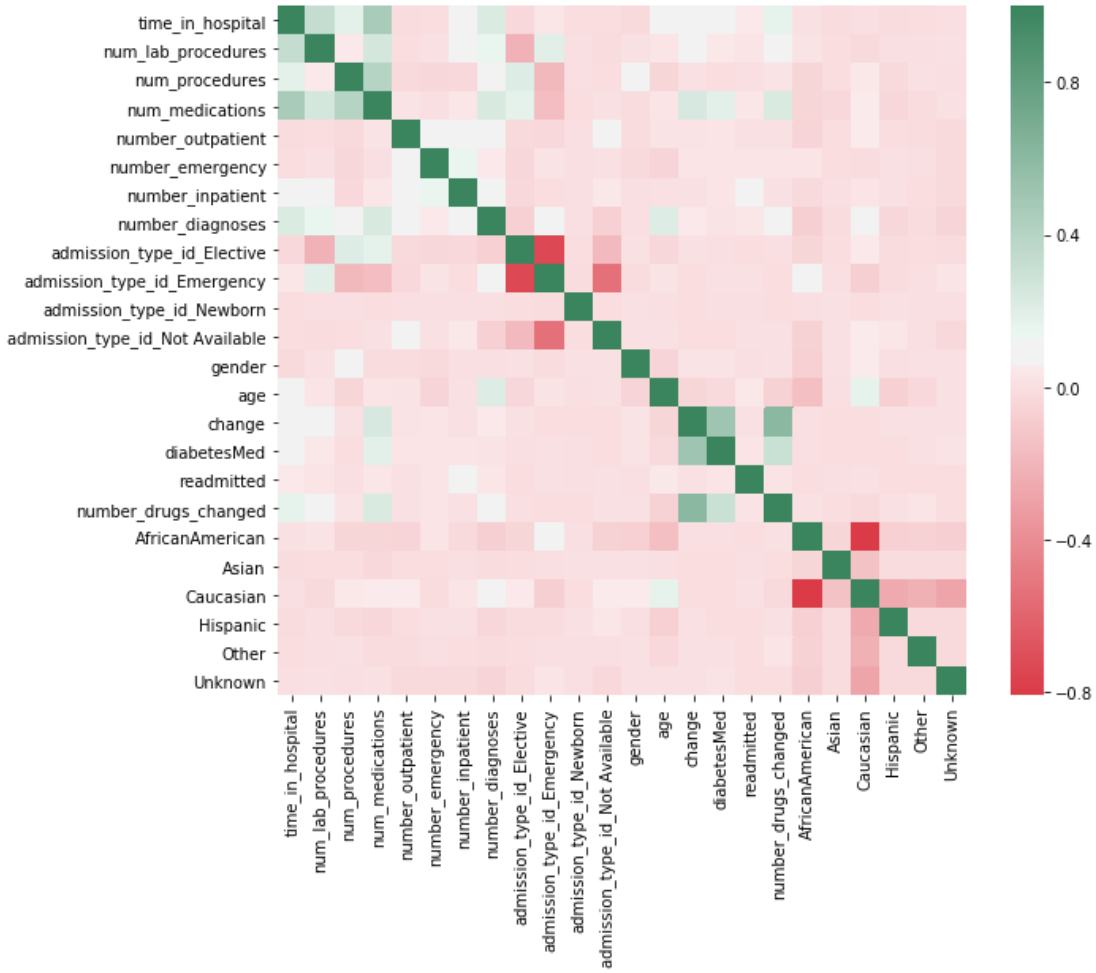
Handle numeric features

1) I decided to combine all similar categories in admission type id attribute in the following way: * 'Not Available', 'NULL' and 'Not Mapped' were combined to one category 'Not Available', which was marked as 0.
* 'Emergency', 'Urgent' and 'Trauma Centre' were combined to one category 'Emergency', which was marked as 1.
Therefore, number of categories in this feature was reduced to 4(from 8).
In fact, I took same approach here as race feature, but in two steps. First, give each numeric data a label, and then (second step) as we did with race and dummy columns.
2) There have been too many categories to deal with in some features. It could lead to bias results; hence I decided to drop those features that contain more than 5 categories. The removed features are: discharge disposition id, admission source id, diag 1, diag 2, diag 3. If I hadn't done so, the quality of our models would have been compromised by multiple clusters.
3) Standardization of numeric categories. This is a common requirement for many machine learning estimators. ML models might behave badly, if the features don't more or less look like standard normally distributed data.
The features I chose to be normalized: age, time in hospital, num lab procedures, num procedures, num medications, number outpatient, number emergency, number inpatient, number diagnoses, number drugs changed.

This is the process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

Correlations heatmap between the different features

Finding the correlations between each pair of features was intended to give an indication of important and interesting features that I want to explore and analyze.
The following heatmap shows the degree of correlation between any two categories by colors, i.e. green color represents perfect correlation, whereas the red color represents negative correlation.



Insights from the heatmap

1) There is a strong correlation between the time in hospital to the number of lab procedures, the number of procedures and the number of distinct medication administered during the encounter. That does make sense logically. Since the patient has to undergo a variety of lab tests and takes doses from different medications, the patient's stay at the hospital is more prolonged and vice versa.
2) There isn't direct correlation between number of diagnoses and the change in diabetic medications. There might be series of diagnoses about the patient's medical condition and it still won't change his diabetes meds.
3) There isn't almost correlation between readmitting to other features. It's quite surprising, due to the fact that people may assume there will be correlation between this attribute to change of medications or number of lab tests the patient have to go through.

4) Another strong correlation is performed between the change in medication to the diabetes medication and the drug dosage.

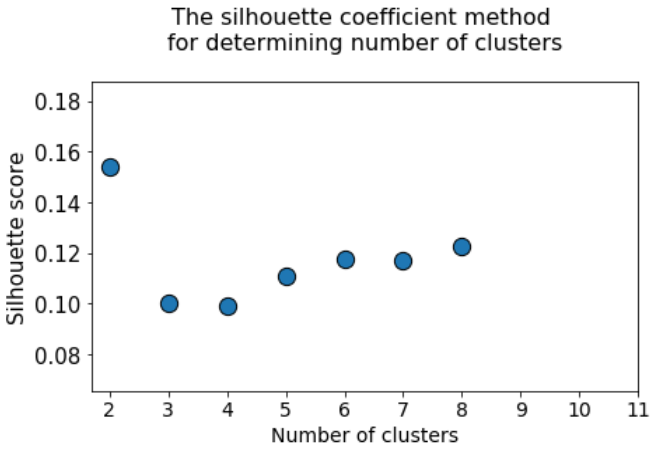
The data set after pre- processing includes the following columns:
time in hospital, num lab procedures, num procedures, num medications, number outpatient, number emergency, number inpatient, number diagnoses, admission type id elective, admission type id Emergency, admission type id Newborn, admission type id Not Available, gender, age, change, diabetes Med, readmitted, number drugs changed, African American, Asian, Caucasian, Hispanic, Other, Unknown.

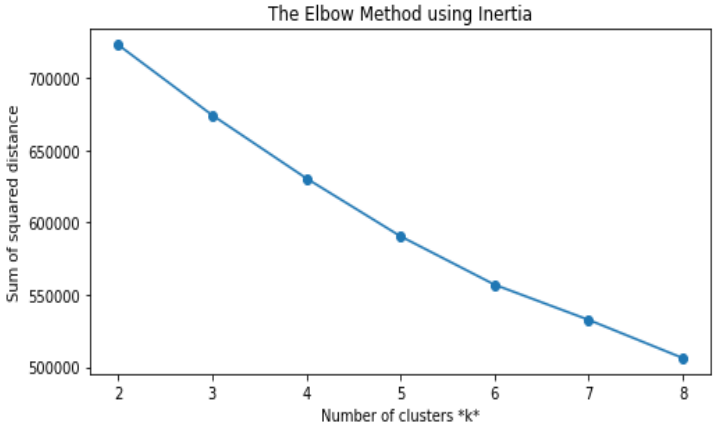
Finally, I got 24 normalized numeric columns after pre- processing of the data set.

Clustering Models Tests

In this section, I performed several clustering techniques I have learned during the course: K-Means, DBSCAN, Agglomerative Clustering and Gaussian Mixture. I analyzed each one of the models separately.
Furthermore, it has to be mentioned that I have used PCA algorithm. I've used this algorithm for several reasons: reduce high dimensions, interpret and visualize my data set. However, PCA algorithm won't be appropriate for every clustering model, such as DBSCAN. In this model, I have used MDS algorithm instead.

K-Means Clustering





K-Means clustering is a method of vector quantization, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

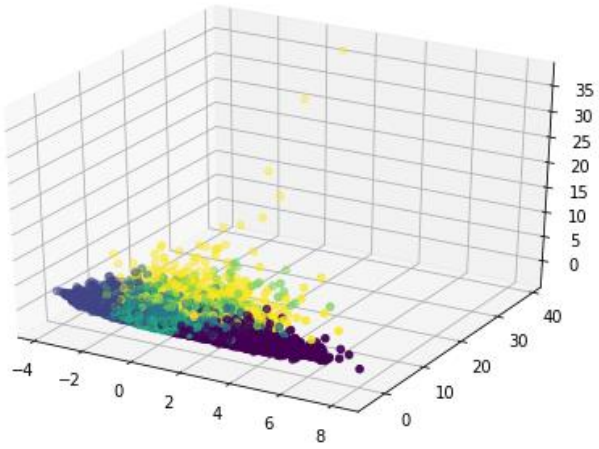
Statistical tests

In order to evaluate K-Means clustering I chose two statistical tests: elbow score and silhouette score. First of all, elbow score is used in order to find the optimal number of clusters. As depicted on the right diagram, curve looks like a hand and the number of clusters to be chosen over there should be equal to 6 as after that curve reaches a plateau.

Moreover, silhouette score is even a better measure to decide the number of clusters to be formulated from the data. It is calculated for each instance and the formula goes like this: (x-y)/ max(x,y). Silhouette score on the left plot shows that after k = 6 is the optimal number of clusters.

Results

1) Cluster 0 contains patients with high number of medicines (27), high number of diagnoses and relatively high drug changes. Above all, this cluster contains patients with much greater time in hospital than all other clusters- at least 50 percent.
2) As I noticed in the above heatmap, the strong correlation between time in hospital, number of lab tests, the number of procedures and the number of distinct medications administered during the encounter is presented here as well.
3) Not very surprising, the majority in each cluster is Caucasian since I already know that the data is imbalanced by race category.
4) Cluster 3 holds the oldest age group (71 years old) and has the highest number of diagnoses, while at the same time their medications are let unchanged.

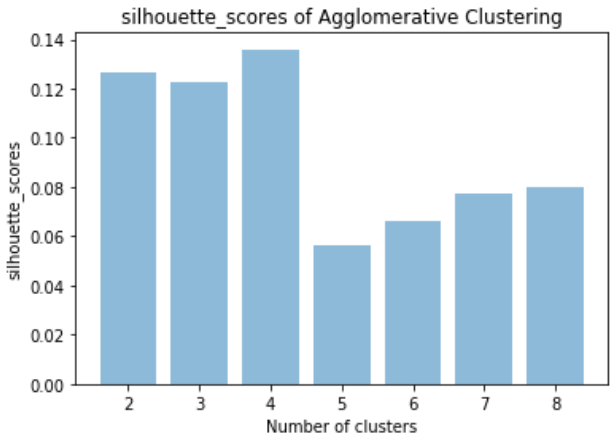


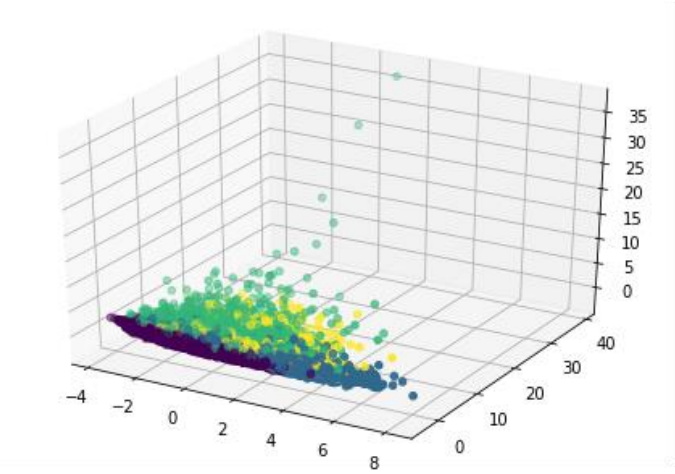
Agglomerative Clustering

Agglomerative clustering is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Statistical test

I have used silhouette score as a measure to decide the optimal number of clusters for Agglomerative model. My conclusion is that k = 4 is the best number of clusters, as you can see on the silhouette score diagram (left plot) k = 4 is the highest point.

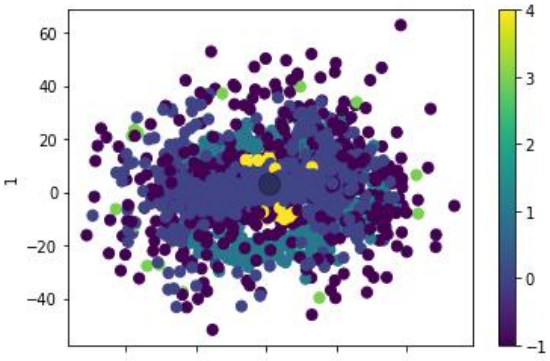


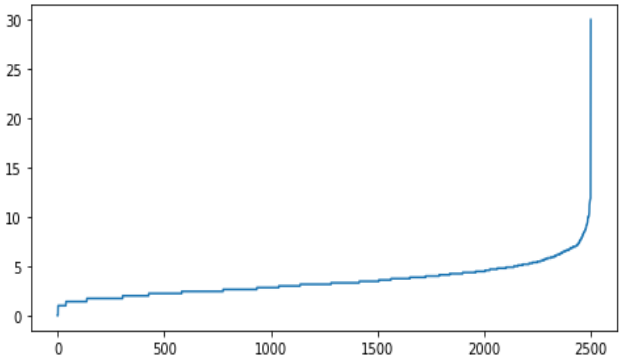


Results

1) Cluster 1 contains the highest number of drug changes and diagnoses, whereas cluster 0 stays stable.
2) The amount of patients in each cluster is quite interesting. Cluster 0 holds the highest amount of people.
3) Cluster 0 contains many people of the third age (60-70 years old) who don't change their drugs at all. Those patients have the lowest time in hospital, lowest inpatient rate, lowest outpatient rate and even the lowest emergency cases.

DBSCAN Clustering





DBSCAN clustering is a density-based clustering non-parametric algorithm. For a given set of points in some space, it groups together points with many nearby neighbors, marking as outliers points that lie alone in low-density regions.
It is important to mention that I haven't used PCA for this model, but have decided to use MDS due to the fact that it maintains the distribution of distances. Moreover, DBSCAN was run on a non-normalized data set, unlike other models that were run on this project.

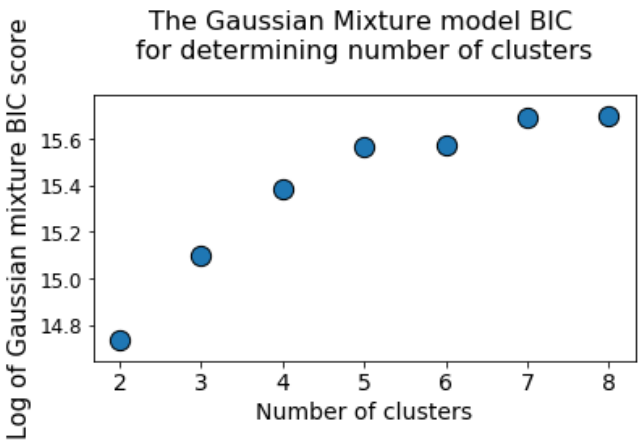
Statistical test

I found a suitable value for epsilon by calculating the distance to the nearest n points for each point (KNN), sorting and plotting the results. Then, I looked to see where the change was most pronounced and selected that as epsilon. In addition, the value of MinPts is determined by the other heuristics like MinPts = ln(n). According to this test, I have decided that epsilon = 6.2 (according to the right plot) and MinPts = 7.8 are the appropriate values. These values determine that k = 5 is the appropriate number of clusters.

Results

1) The inpatient and outpatient rates are quite stable between the different clusters.
2) Cluster 3 holds the highest number of diagnoses, although the number of drugs which was given to these people almost has not changed.

Gaussian Mixture Clustering



A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

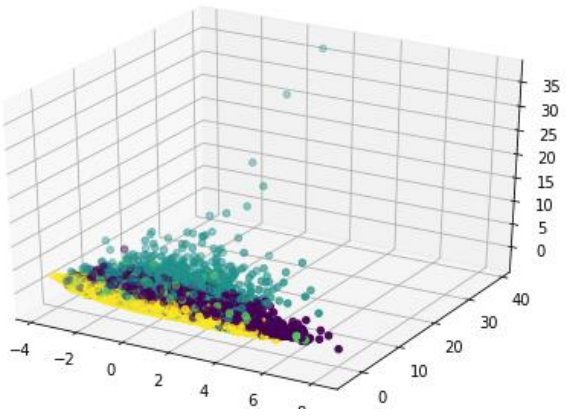
Statistical test

In order to evaluate GMM clustering I chose three statistical tests: V-Measure, Silhouette score and BIC criterion. First of all, V-Measure is defined as the harmonic mean of homogeneity and completeness of the clustering. Homogeneity is maximized when each cluster contains elements of as few different classes as possible.
It is important to note that V-Measure cluster labeling given a ground truth. V-Measure shows that k = 3 is the optimal number of clusters.
Furthermore, BIC (Bayesian information criterion) criterion is used to select the number of components in a Gaussian Mixture model in an efficient way. My conclusion from the BIC plot is that k = 3 is the optimal number of clusters as well.

Results

1) Cluster 2 shows the highest change in medication and hospital procedures. In addition, there isn't radical rate in this cluster compared to the others.
2) Cluster 1 shows that there is correlation between the time in hospital to the number of lab tests, the number of procedures and the number of distinct generic medications administered during the encounter. However, this cluster has zero change in drugs for average age group. It's quite surprising.

Gaussian mixture model gives me the best quality of clustering between the different models.



Conclusions

1) Caucasians react to diabetic diseases in the best way from all test races.
2) The more the age of a person increases, the more likely they are to have diabetes disease.
3) Some older people have a low probability of being cured by diabetes. Those patients have very high diagnoses, but the number of medicines has not changed.
4) Most older people have healed or remained in stable condition, i.e. they're able to cope with diabetes in a consistent and permanent treatment without drug change. Those patients have low risk of getting to emergency medical care.
5) Patients, who need to undergo many medical procedures replace many drugs and have many diagnoses, have a type of diabetes that is inconclusive. Moreover, it may be rare forms of diabetes that is less familiar in the world of medicine.
6) There's no diabetes treatment priority between hospitals and outpatient clinics. The inpatients and outpatients rates show numeric results that reflect the same condition.
7) There are default tests which constitute initial duty tests when a patient reaches hospitalization. When I speak specifically of diabetic diseases it seems that the number of average laboratory tests ranges from 40 to 60.
8) The more time a person stays in the hospital, the longer his chances of returning grow.
9) The higher number of days the patient has to stay in the hospital, the greater his dose of medicine increases. Most people consume up to 20 medications with remain in the hospital for up to 6 days.
10) If patient consumes a smaller number of drugs, his number of medical procedures will be reduced.