Olaru Bogdan-Ioan

Project 2

The dataset used for this project can be found on *UC Irvine Machine Learning Repository* at this link. The original dataset consists of 5 different folders, each with 100 files, with each file representing a single subject/person. Each file is a recording of brain activity for 23.6 seconds. The corresponding time-series is sampled into 4097 data points. Each data point is the value of the EEG recording at a different point in time. To obtain the actual dataset they divided and shuffled every 4097 data points into 23 chunks, each chunk contains 178 data points for 1 second, and each data point is the value of the EEG recording at a different point in time. So now we have 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column), the last column represents the label y {1,2,3,4,5}. All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure. Only subjects in class 1 have epileptic seizure. Our motivation for creating this version of the data was to simplify access to the data via the creation of a .csv version of it. Although there are 5 classes most authors have done binary classification, namely class 1 (Epileptic seizure) against the rest.

In order to solve this problem as a clustering one, I chose to use Gaussian Mixture Model and DBSCAN. In the next part we will have a look at the how they work.

A Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset. In the simplest case, GMMs can be used for finding clusters in the same manner as k-means. So, GMM is also a type of clustering algorithm. As its name implies, each cluster is modelled according to a different Gaussian distribution. This flexible and probabilistic approach to modelling the data means that rather than having hard assignments into clusters like k-means, we have soft assignments. This means that each data point could have been generated by any of the distributions with a corresponding probability. In effect, each distribution has some ‘responsibility’ for generating a particular data point.

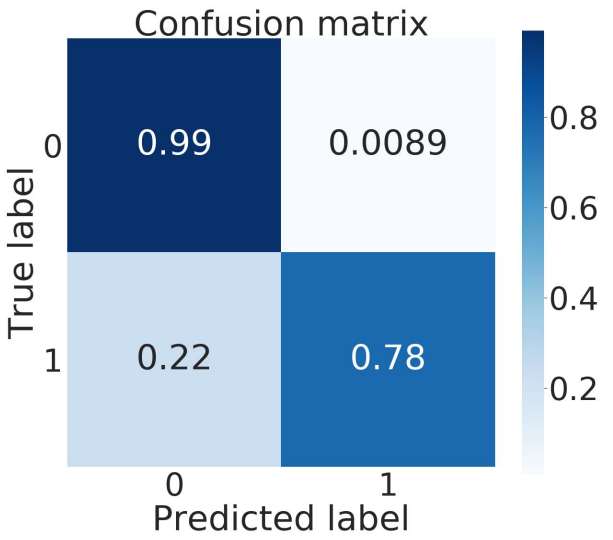
Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. Based on a set of points DBSCAN groups together points that are close to each other

based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

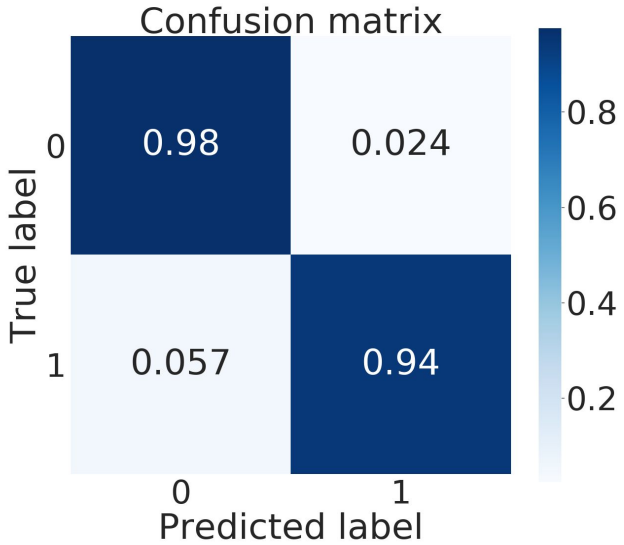
First I removed the 'Unnamed' column and I mapped the classes 5, 4, 3, and 2 to 0. Then I shuffled and splitted the dataset into 80% training+validation and 20% test. And then I splitted the training+validation dataset in 80% training and 20% validation. Now, let's see some results and comment about them:

1. DBSCAN

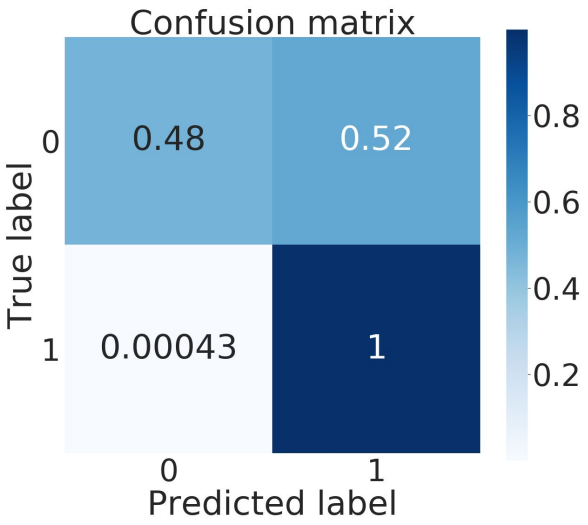
- eps=10, min_samples=30 -> acc: 94.8%



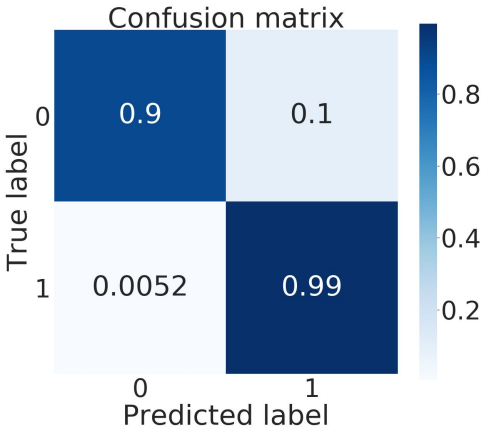
- eps=7, min_samples=20 -> acc: 96.91%



- eps=3, min_samples=20 -> acc: 58.09%



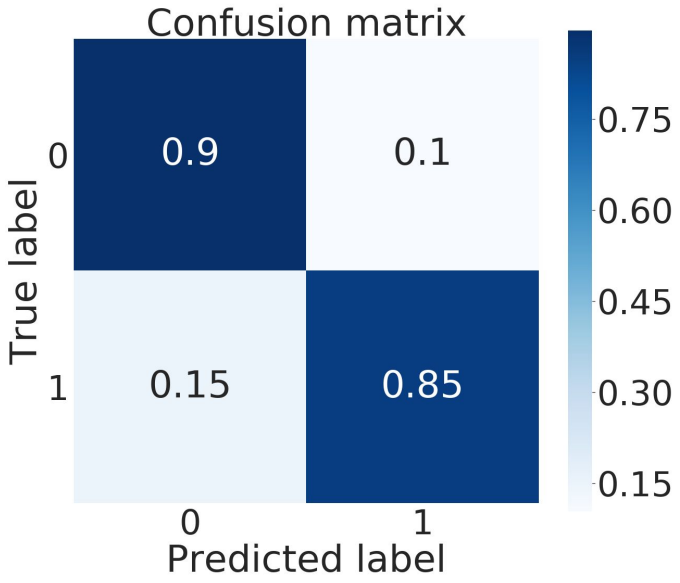
- eps=5, min_samples=20 -> acc: 91.9%



So, if we are aiming to have a better overall accuracy, the best results are obtained using an eps of 7 and around 20 min_samples. But if we are aiming to obtain a better accuracy for epileptic seizures, then we have the best results with an eps of 5-6, and around 15-20 min_samples. So far the results are very good, taking in consideration that the dataset is unbalanced.

2. Gaussian Mixture Model

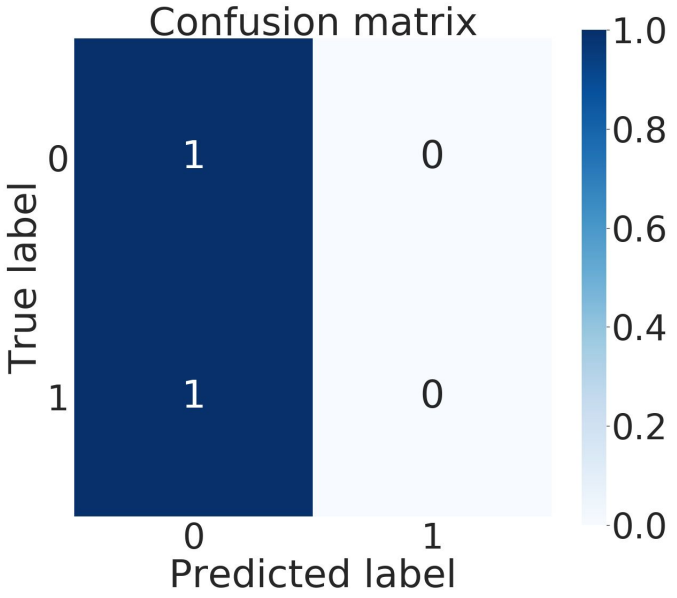
- n_components=2, covariance_type='full', max_iter=200, tol=1e-4
  -> val_acc: 0.89%
  -> acc: 0.89%



- n_components=2, covariance_type='full', max_iter=100, tol=1e-5

  -> val_acc: 0.79%

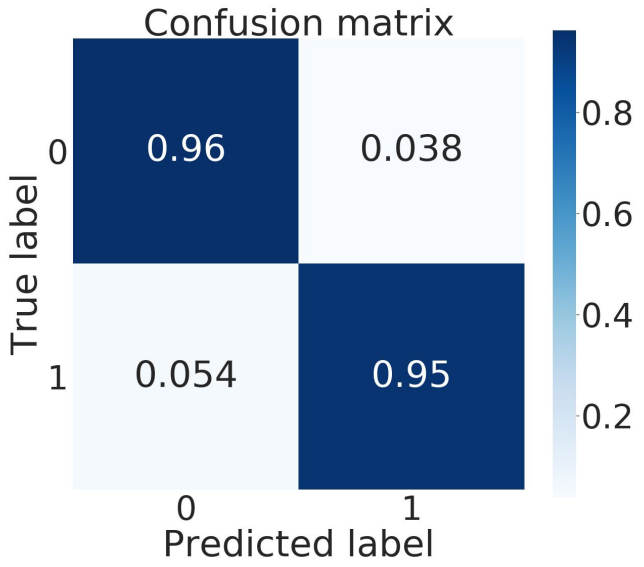
  -> acc: 0.79%



- n_components=2, covariance_type='spherical', max_iter=200, tol=1e-4

  -> val_acc: 95.54%

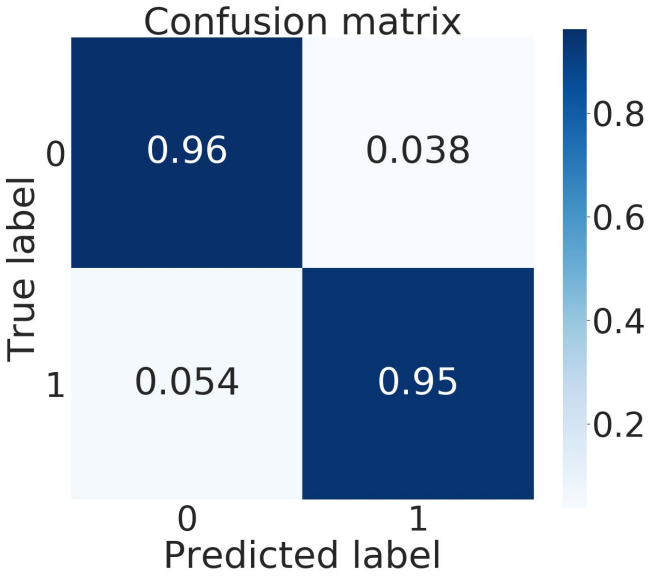
  -> acc: 95.87%



- n_components=2, covariance_type='spherical', max_iter=2000, tol=1e-10

  -> val_acc: 95.54%

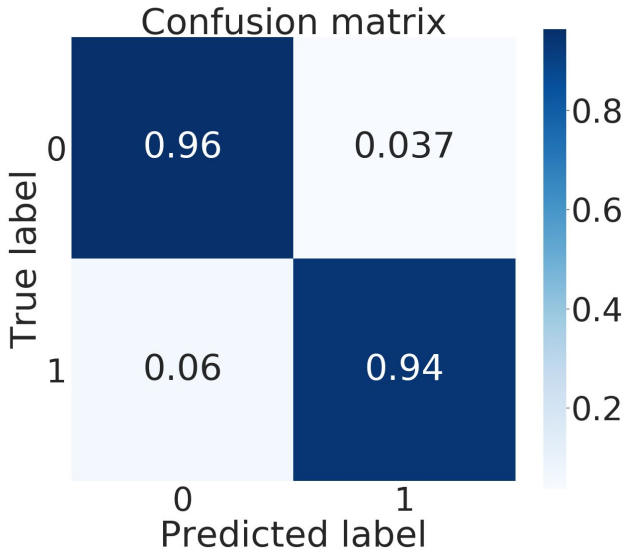
  -> acc: 95.87%



- n_components=2, covariance_type='spherical', max_iter=50, tol=1e-1

  -> val_acc: 95.6%

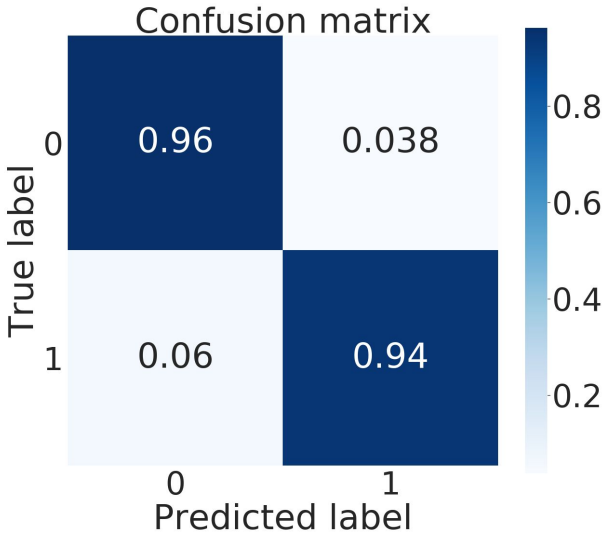
  -> acc: 95.83%



- n_components=2, covariance_type='diag', max_iter=200, tol=1e-3

  -> val_acc: 95.6%

  -> acc: 95.74%



So, GMM obtains a really good overall accuracy, and in some cases a really well balanced accuracy between the 2 classes. Almost all results are good, and it also performs very well on unscaled data.

As a conclusion, I think that DBSCAN still has a better accuracy in finding the epileptic seizures, and that's what matters the most. And the overall accuracy is almost the same.