|  |  |
| --- | --- |
| Algorithm: Fuzzy C Means Clustering | |
| USN: 1MS18CS024 | NAME: Aravind S |
| USN: 1MS18CS029 | NAME: Arun J Kennedy |

**Description of the Algorithm: <<Write 2-3 Paragraphs about the Algorithm>>**

In hard clustering, point belongs to any one cluster or group. In soft clustering, instead of putting each data point into separate clusters, a probability of that point to be in that cluster assigned. In soft clustering or fuzzy clustering, each data point can belong to multiple clusters along with its probability score or likelihood.

Fuzzy C-Means clustering is a soft clustering approach, where each data point is assigned a likelihood or probability score to belong to that cluster.

Fuzzy c-means clustering has can be considered a better algorithm compared to the k-Means algorithm. Unlike the k-Means algorithm where the data points exclusively belong to one cluster, in the case of the fuzzy c-means algorithm, the data point can belong to more than one cluster with a likelihood.

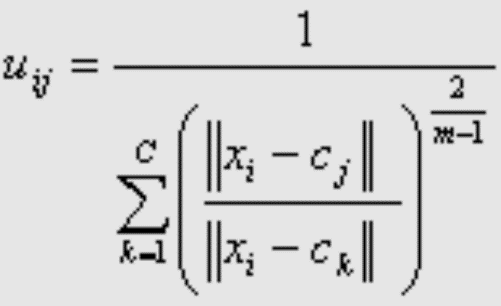
**Algorithm Pseudocode:**

The algorithm is composed of the following steps:

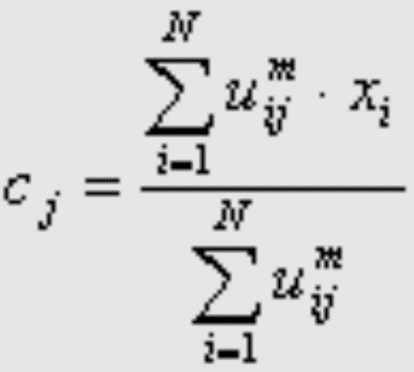
Step 1: Randomly select cluster centre

Step 2: Initialize U= [ uij ] matrix, U(0)

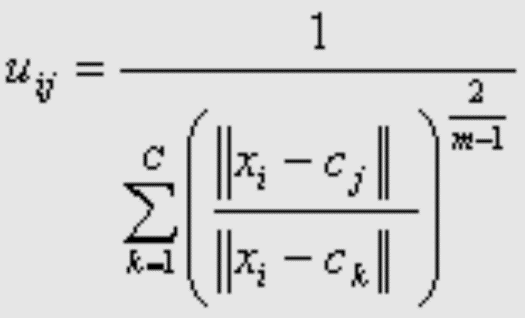
Calculate the uij using:



Step 3: At k-step: calculate the centers vectors C(k)=[cj] with U(k)

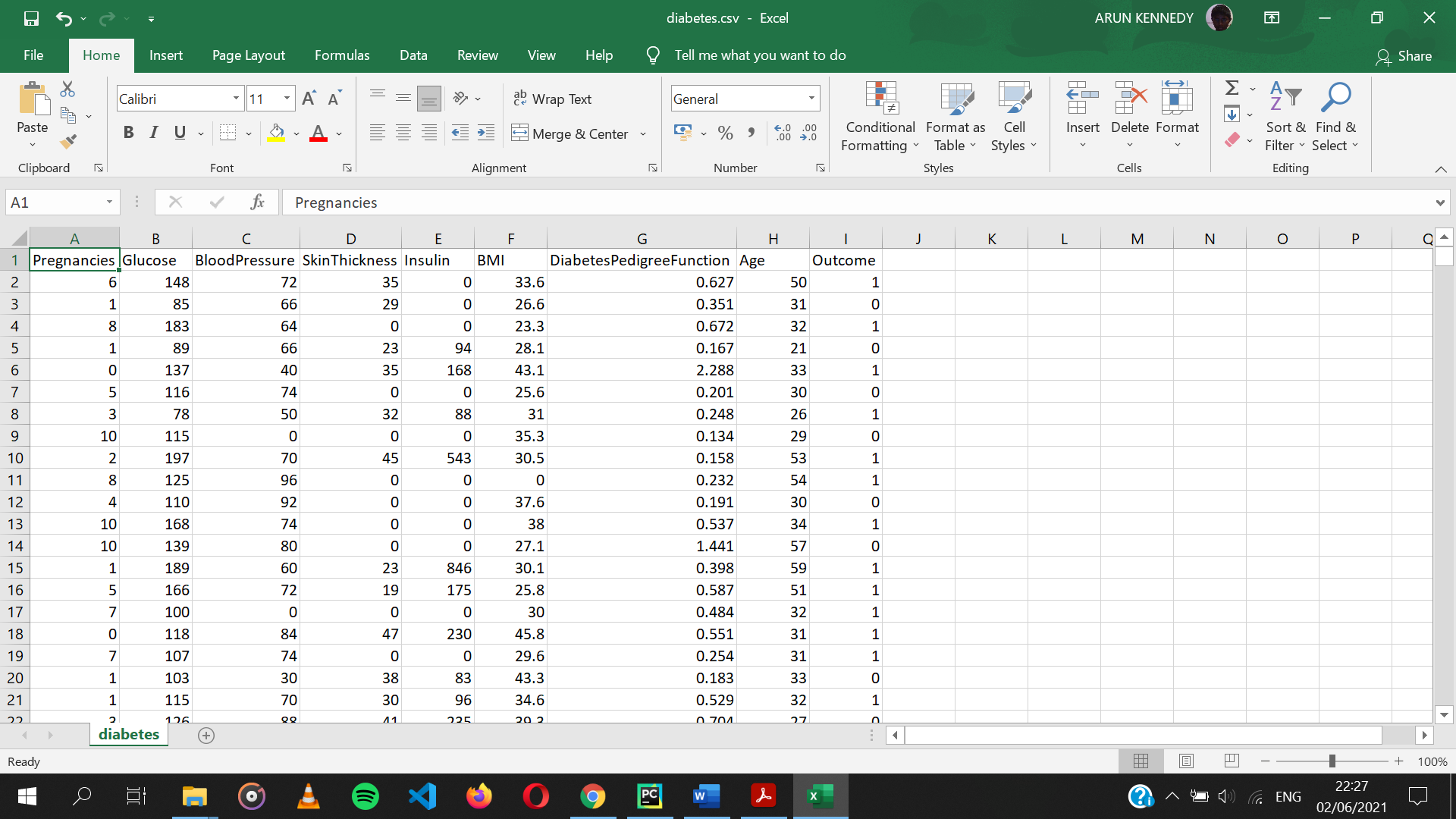


Step 4: Update U(k), U(k+1)



Step 5: If || U(k+1) - U(k)||< ε or the minimum J is achieved, then STOP; otherwise return to step 2.

**Data set Used: (Attach Screen shot of the few rows and also the Kaggle/Dataset link)**



**Dataset link:** <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

**Challenges faced during the implementation of the program:**

**Problem 1:** If we observe the dataset, we can see the Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI) value in rows equal to 0. We know that in real life these Values can never be equal to zero. Therefore, we can conclude that these values are wrong.

**Solution:** Due to this we have to clean the dataset before training the model. We do this by finding the mean of all the non-zero values of an attribute and replace 0s of the respective column by the mean. In real world Insulin and Skin thickness is dependent on Glucose and BMI respectively, so we group them accordingly and find respective mean of the median of each group to replace 0.

The accuracy of prediction also improves by cleaning the data.\

**Problem 2:** The dataset consists of 8 attributes and 1 outcome. It is a multidimensional dataset. It is effectively hard plot a multidimensional dataset in a 2d plot.

**Solution:** We transform the 8 attributes into 2 principal components using PCA to obtain x-axis and y-axis of the plot. We then plot outcome of 0 with red color point and 1 with green color point.

**Advantages & Disadvantages of the Algorithm:**

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| Gives best result for overlapped data set and comparatively better then k-means algorithm. | Apriori specification of the number of clusters. |
| Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center. | With lower value of β we get the better result but at the expense of more number of iteration. |
|  | Euclidean distance measures can unequally weight underlying factors. |

**Applications of the Algorithm:**

Clustering problems have applications in surface science, biology, medicine, psychology, economics, and many other disciplines.

* In the field of bioinformatics, clustering is used for a number of applications. One use is as a pattern recognition technique to analyze gene expression data from microarrays or other technology.
* Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise.
* In marketing, customers can be grouped into fuzzy clusters based on their needs, brand choices, psycho-graphic profiles, or other marketing related partitions.

**Output: (Screen shots)**

1. **Visualization of the dataset:**

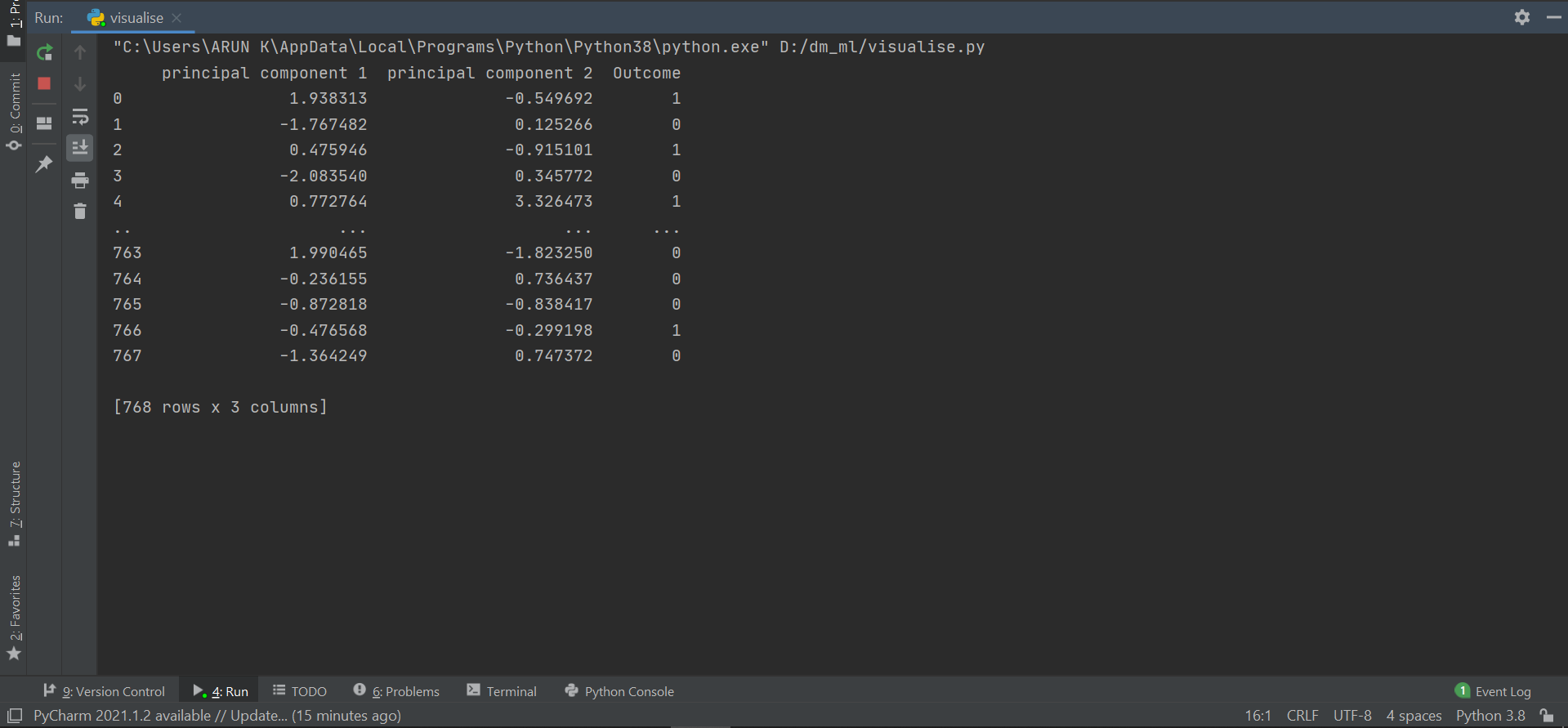


Fig 1: Console output of Visualization Program

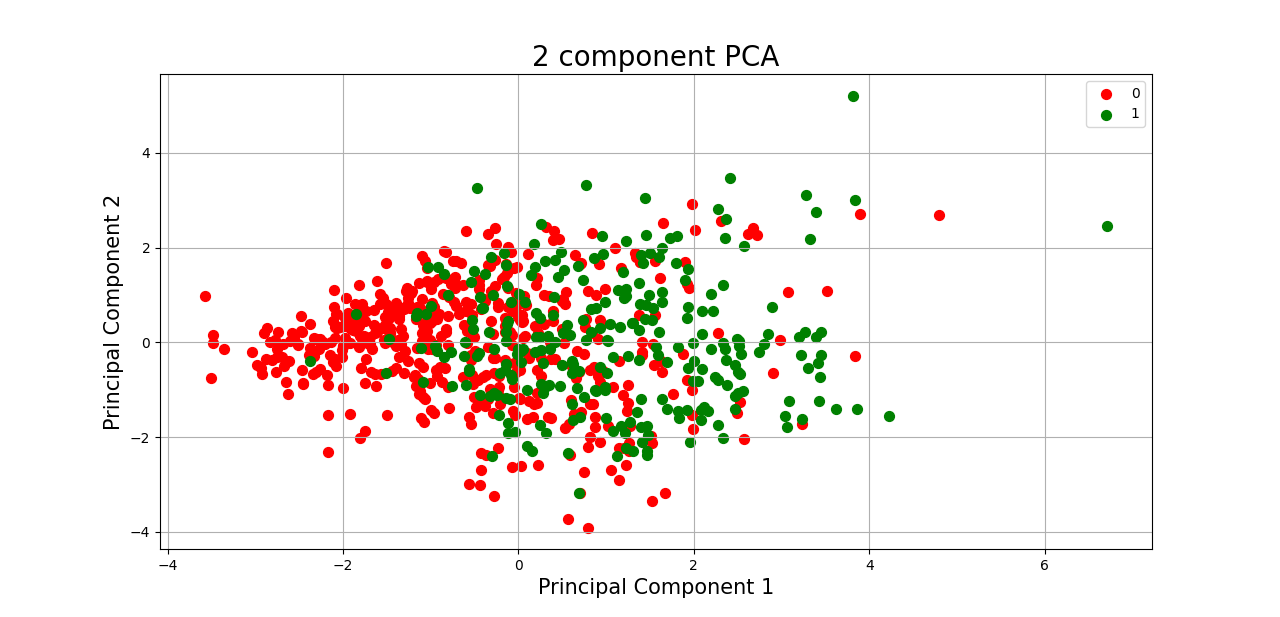


Fig 2: Scatter Plot output of Visualization program

In this plot we can observe that the 8 dimensions of the dataset is reduced to 2 components. We can also see that the red and green points are almost overlapped but the density of red points is more towards left and density of green points is more towards right.

1. **Output of Fuzzy c-means clustering program**

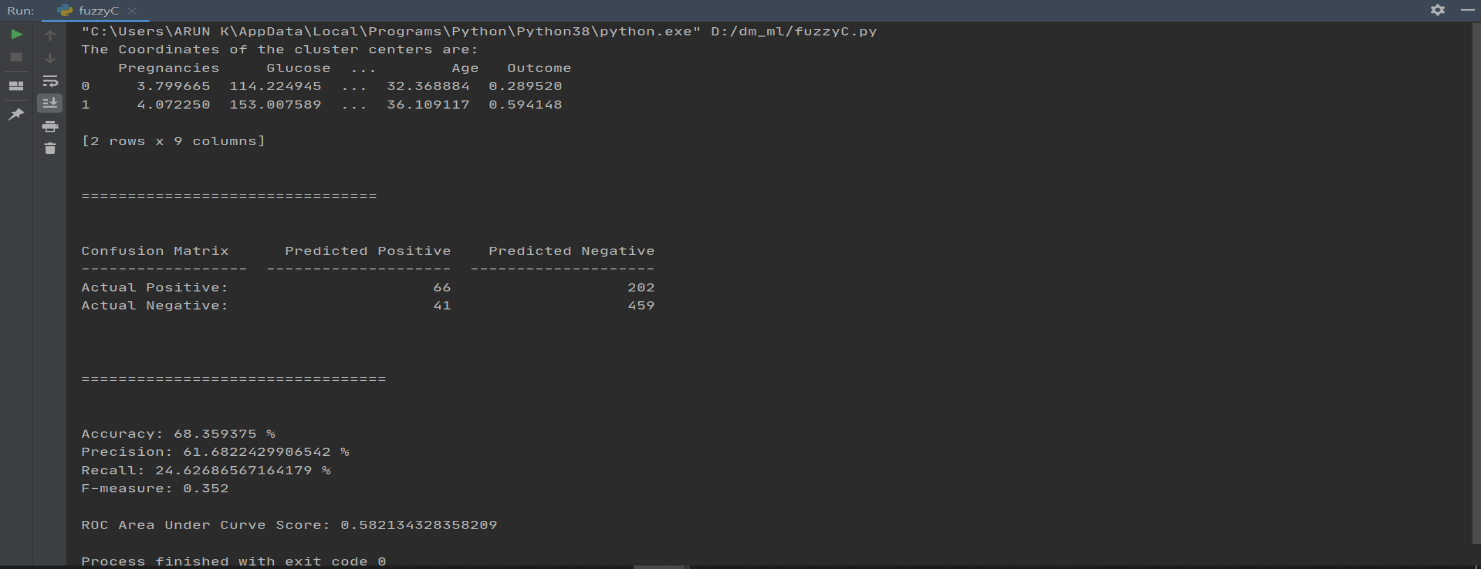


Fig 3: Console output of the program.

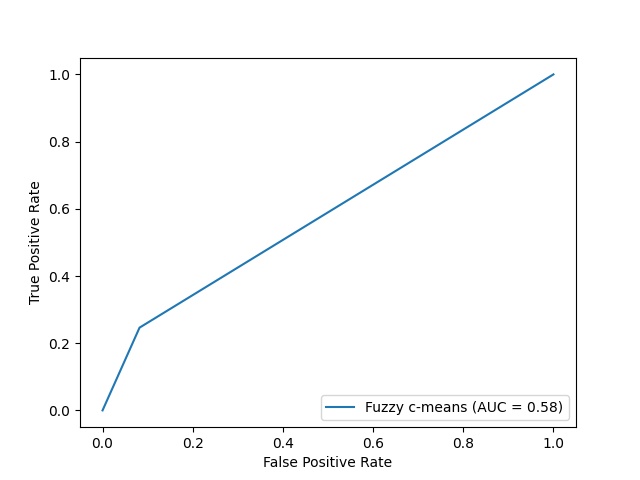


Fig 4: ROC Curve of fuzzy c-means clustering.

We can see that the AUC is 0.58 which is not very good prediction. From the confusion matrix we can also see that the prediction classifies most of the cases as negative. We also see that True Positive is 66 and False Positive is 41, Since they are very close the AUC of 0.58 makes sense.

**References:**

<https://pypi.org/project/fuzzy-c-means/>

<https://en.wikipedia.org/wiki/Fuzzy_clustering>

<https://scikit-earn.org/stable/modules/generated/sklearn.metrics.RocCurveDisplay.html#sklearn.metrics.RocCurveDisplay>