Report of the BSPC Project **Jan-2024**

Brain Stroke Patients Clustering **Github:@izyyaz**

Objectives : the main objective of this project is to perform clustering analysis on the brain stroke patients data, which is publicly available in Kaggle. The clustering results can help with providing custom health care services for the different categories clustered, and help pay attention to patients with risky medical conditions. The side objective is to analyze the data and bring insights related to the patients attributes.

The dataset consists of 10 features and 1 target column indicates the stroke condition. I've created a new dataframe after exploring the features, and only chose 4 features for clustering, which were age, heart disease, AGL, and bmi. For more information about the dataset and features processing details of the project, I highly recommend viewing the jupyter notebook.

Attribute Information

1) gender: "Male", "Female" or "Other". 2) age: age of the patient. 3) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension. 4) heart disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease. 5) Ever-married: "No" or "Yes".

6) work type: "children", "Govtjov", "Private" or "Self-employed". 7) Residencetype: "Rural" or "Urban".

8) avg glucose level: average glucose level in blood

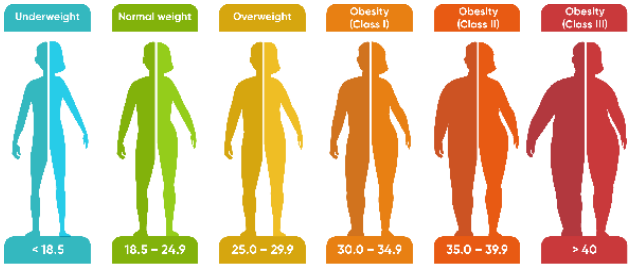
9) BMI: body mass index

10) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown". 11) stroke: 1 if the patient had a stroke or 0 if not

Here are some charts needed to understand some feature

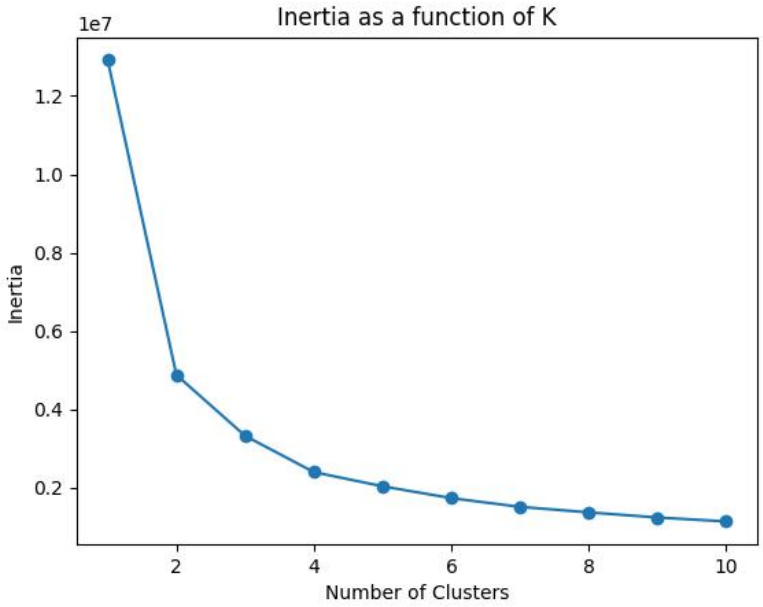
LEVEL	mg/dl	mmol/L	RISK	SUGGESTED ACTION
DANGER - HIGH	315+	17.4	VERY HIGH	MEDICAL ATTENTION
HIGH	280	15.6	HIGH	MEDICAL ATTENTION
HIGH	250	13.7	HIGH	MEDICAL ATTENTION
HIGH	215	11	HIGH	MEDICAL ATTENTION
BORDERLINE	180	10	MEDIUM	CONSULT DOCTOR
BORDERLINE	150	8.2	MEDIUM	CONSULT DOCTOR
BORDERLINE	120	7	MEDIUM	CONSULT DOCTOR
NORMAL	108	6	NO RISK	NO ACTION NEEDED
NORMAL	72	4	NO RISK	NO ACTION NEEDED
LOW	70	3.9	MEDIUM	CONSULT DOCTOR
DANGER - LOW	50	2.8	HIGH	MEDICAL ATTENTION

Source: CMI Health, Blood glucose levels



Source: USZ, BMI Calculator

For the clustering part, I chose KMeans Algorithm. I used the famous inertia as a function of k way (Elbow Method) to determine the appropriate number of clusters, and my choice was 4, obviously, as you can see in the following figure



After feeding data into the KMeans (k=4) model, I added the cluster labels from the model to the new dataframe that I had previously created. Subsequently, I created new subtables each one containing only one cluster in the cluster column. Each subtable was named based on the cluster class it contained (e.g. a dataframe that contains only cluster 1 would be named df1).

I've explored the different subtables created and compared the attributes of each, and here's a brief comparison between the different clusters (subtables). It's a result of statistical analysis.

Table 0:

Age: we can notice that the average age among the patients is 60 years old. With 39 and 82 min and max values, respectively.

Heart Disease: A tendency of 7% to heart disease.

Average Glucose Level: The estimated mean average glucose level is 82 mg/dl, which is considered normal (refer to the BGL chart). It also has 55 and 117 min and max AGL values, respectively.

BMI: The estimated mean bmi is 29, which is considered overweight (refer to the bmi chart). It also has 14 and 48 min and max bmi values, respectively.

Table 1:

Age: 60 years old is the average age (again) among the patients, but it has 1 and 82 min and max ages, respectively. Although cluster 0 and 1 share the same average age, cluster 1 has a wider range of ages.

Heart Disease: Cluster 1 patients exhibit the highest tendency for heart disease(17%) among the other clusters, making their medical situation serious.

Average Glucose Level: With an average glucose level of 211 mg/dl, Cluster 1 is in a precarious medical state since this value represents a high level-high risk (based on the BGL chart). Additionally, with min-max values of 161 and 271 respectively, it has the most dangerous range among all other clusters.

BMI: The mean BMI is 32, which falls under the Obesity Class 1 category on the BMI chart. The min and max values for BMI were found to be 14 and 48, respectively.

Table 2:

Age: Yes, youth! Patients in cluster 2 estimated an average age of 20. With 0.08 and 40 min-max ages.

Heart Disease : Youth's cluster patients have the lowest risk of heart disease (0.06%) compared to other clusters.

Average Glucose Level: The mean AGL is 82 mg/dl, which is considered normal. With 55 and 115 being the min-max values, which is an appropriate range too.

BMI: The mean bmi is 25, which is barely normal (refer to the chart). It has 14 and 48 min-max values.

Table 3:

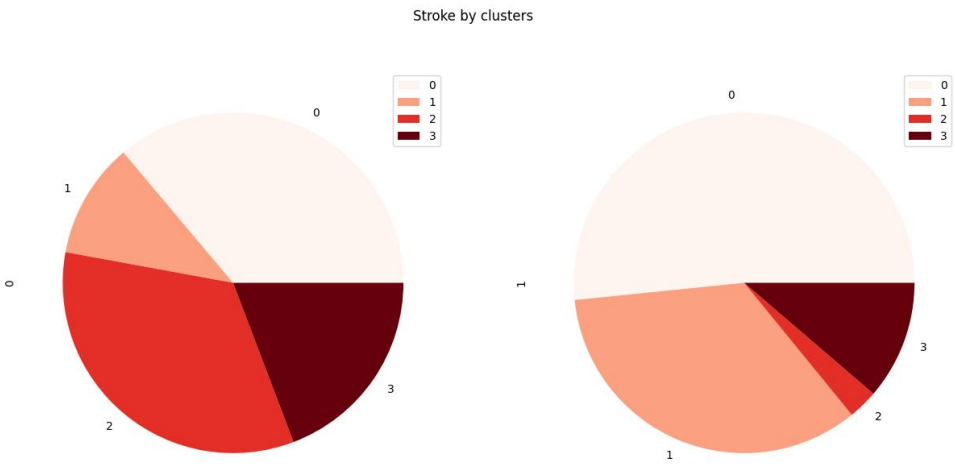
Age: The average age in cluster 3 is 39 years old. With 0.08 and 82 min-max ages.

Heart Disease: Patients in cluster 3 had a 3% tendency to heart disease.

Average Glucose Level: The estimated mean AGL is124 mg/dl. It belongs to the borderline category in the chart. It has 99 and 173 min-max values, which is a medium risk range.

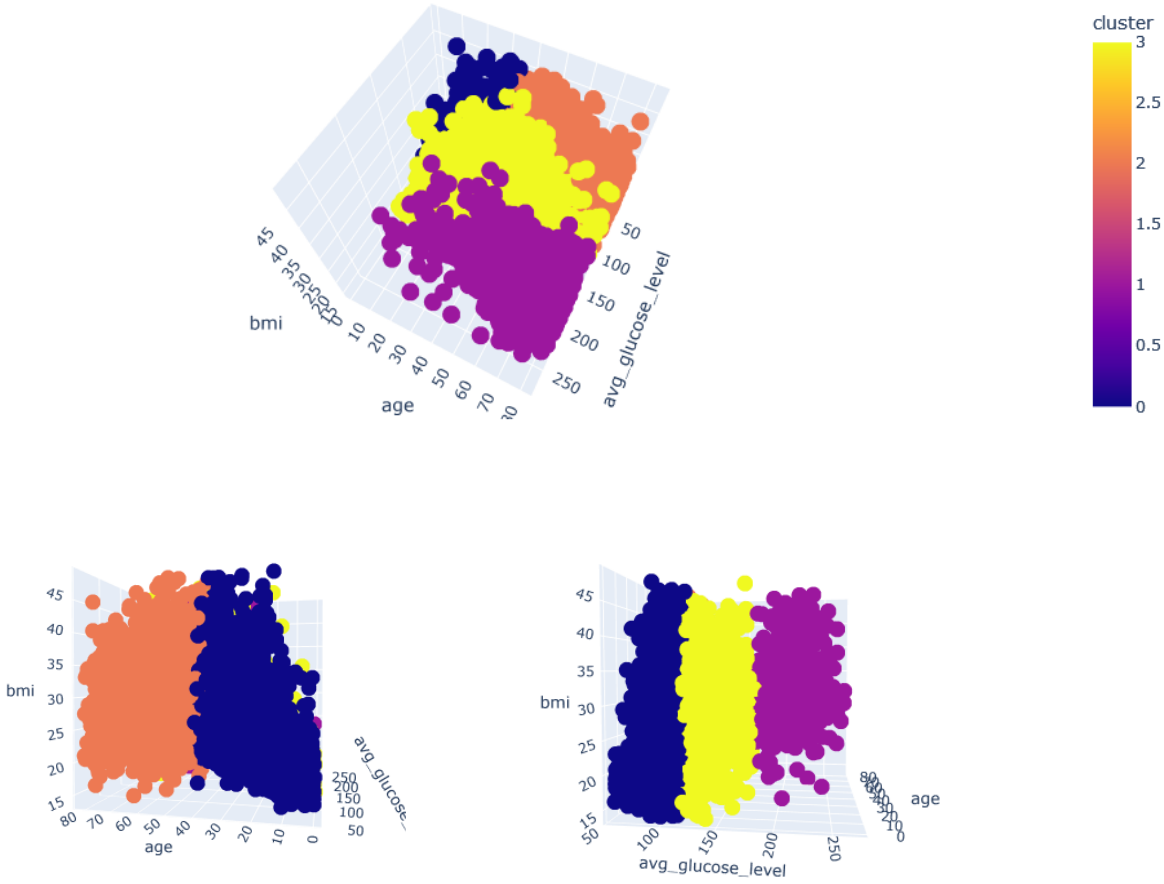
BMI: The mean bmi is 25, which is barely natural, again. It has 14 and 48 min-max values.

After successfully clustering data, I've added the cluster column to the original data, in order to compare the stroke conditions between the clusters. I've made a groupby dataframe that had the columns (cluster & stroke). The results are better visualized in the following figure.



As you can see, the figure consists of two pie plots, each is segmented by the clusters. The plot to the left indicates the non-stroke conditions, while the plot to the right indicates the stroke conditions.

I selected the features (AGL, age, and bmi) and visualized them on the 3d scatter plot, with highlighting colors indicate the clusters.



These snapshots are from the plotly figure in the notebook. They're also available as html file in the zipped project folder /Figures. You can play with it there.

Finally, how would the project influence the development of health care?

The results of the project can be instrumental in medical studies and treatment as they provide a detailed understanding of different clusters of brain stroke patients. This information can help in tailoring healthcare services to specific patient groups, identifying high-risk patients, and developing targeted treatment plans. For example, by understanding the characteristics of each cluster, healthcare providers can better allocate resources and focus on preventive measures for patients with higher risk factors such as heart disease or high glucose levels. Additionally, this data can contribute to research on the relationship between patient attributes and stroke outcomes, ultimately leading to improved medical interventions and care strategies.