DATA MINING FINAL REPORT

STROKE RISK PREDICTION

Professor: Chih-Chou Chiu

Benoit Kham
112015007

余珮綺
112598028

楊玉英
112578408

詹端安
111749407

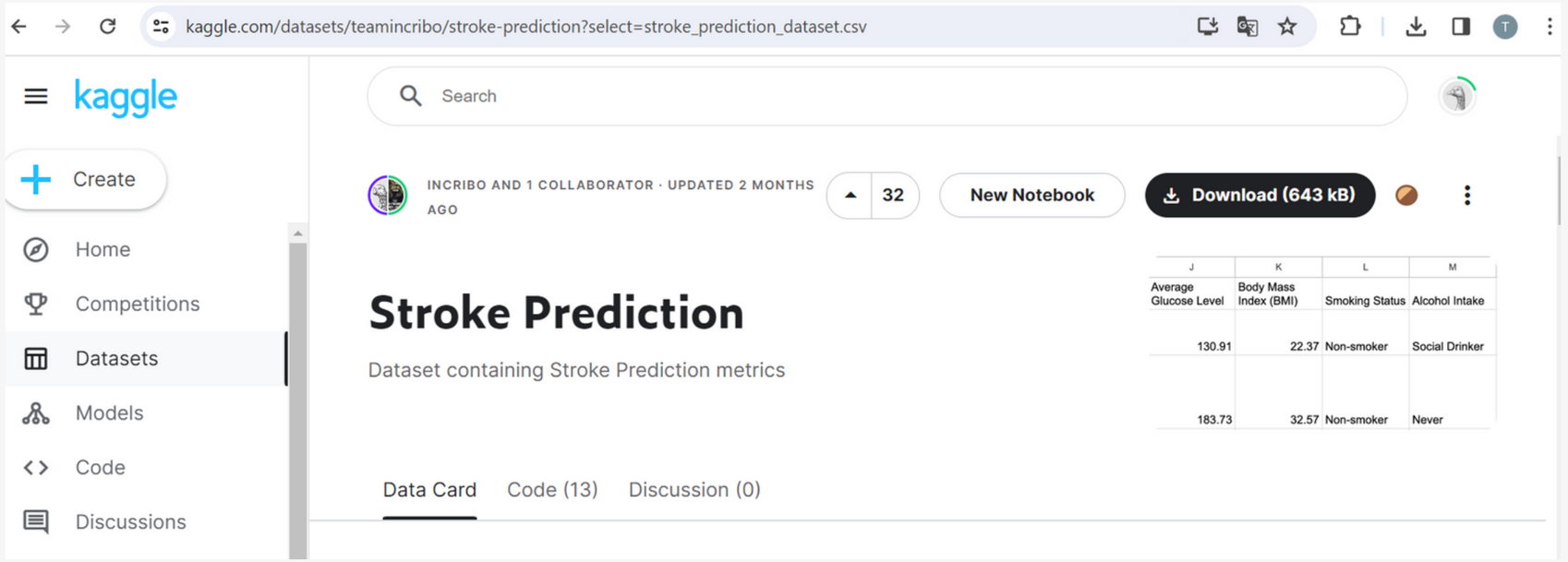
Agenda

1. Introduction of Data Source
2. Explanation of Data Variables
3. Explanation of Research Topic
4. Data Cleaning Process
5. Variable Selection Process
6. Cluster
7. Data Splitting
8. Building a Predict Model
9. Summary

1

Introduction

1 | Introduction of Data Source



The dataset is available on **Kaggle** and consists of **15,000 records** with **22 features**

To **predict the likelihood of a stroke occurrence** in individuals.

1 | Introduction of Data Source

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Patient ID	Patient Name	Age	Gender	Hyperten	Heart Disease	Marital Status	Work Type	Residenc	Average (	Body Mass Inde	Smoking	Alcohol I	Physical /	Stroke Hi	Family Hi	Dietary H	Stress Le	Blood Pr	Chole
2	18153	Mamooty Khurana	56	Male	0	1	Married	Self-employed	Rural	130,91	22,37	Non-smok	Social Drin	Moderate	0	Yes	Vegan	3,48	140/108	HDL: 6
3	62749	Kaira Subramaniam	80	Male	0	0	Single	Self-employed	Urban	183,73	32,57	Non-smok	Never	Low	0	No	Paleo	1,73	146/91	HDL: 6
4	32145	Dhanush Balan	26	Male	1	1	Married	Never Worked	Rural	189	20,32	Formerly S	Rarely	High	0	Yes	Paleo	7,31	154/97	HDL: 5
5	6154	Ivana Baral	73	Male	0	0	Married	Never Worked	Urban	185,29	27,5	Non-smok	Frequent D	Moderate	0	No	Paleo	5,35	174/81	HDL: 7
6	48973	Darshit Jayaraman	51	Male	1	1	Divorced	Self-employed	Urban	177,34	29,06	Currently S	Rarely	Low	0	Yes	Pescataria	6,84	121/95	HDL: 6
7	29307	Advika Kota	62	Female	0	0	Single	Private	Urban	91,6	37,47	Currently S	Social Drin	High	0	No	Gluten-Fre	4,85	132/64	HDL: 8
8	25525	Elakshi Karan	40	Female	1	0	Married	Private	Urban	77,83	28,2	Currently S	Never	Low	1	No	Vegetariar	6,38	178/105	HDL: 3
9	4809	Shalv Dugar	61	Female	0	1	Divorced	Government Jo	Rural	194,73	26,44	Formerly S	Rarely	Moderate	1	No	Vegan	5,85	179/72	HDL: 6
10	7372	Raghav Handa	72	Female	1	1	Married	Self-employed	Rural	72,99	30,1	Formerly S	Rarely	High	1	No	Vegetariar	0,73	141/106	HDL: 7
11	37504	Krish Kulkarni	82	Male	0	0	Divorced	Self-employed	Urban	111,23	28,83	Currently S	Rarely	High	1	Yes	Non-Vege	8,39	119/88	HDL: 3
12	15298	Neelofar Devan	41	Male	0	1	Divorced	Government Jo	Urban	94,9	36,74	Formerly S	Frequent D	Low	1	Yes	Pescataria	1,56	91/80	HDL: 3
13	36017	Anaya Koshy	72	Female	0	0	Divorced	Private	Urban	155,32	30,87	Currently S	Frequent D	Moderate	0	Yes	Paleo	8,71	127/110	HDL: 5
14	66924	Ahana Lalla	30	Female	0	1	Divorced	Government Jo	Urban	163,15	19,36	Formerly S	Frequent D	Moderate	0	Yes	Non-Vege	9,19	114/67	HDL: 8
15	46821	Zaina Chaudhary	80	Female	0	0	Single	Private	Urban	136,06	25,19	Formerly S	Never	High	1	No	Gluten-Fre	4,14	97/81	HDL: 4
16	54426	Tara Swaminathan	42	Male	0	1	Married	Self-employed	Rural	181,02	19,35	Formerly S	Frequent D	Low	1	No	Paleo	2,58	170/102	HDL: 7
17	86093	Azad Krishnan	86	Male	0	1	Married	Government Jo	Rural	130,71	31,83	Non-smok	Social Drin	High	0	No	Non-Vege	3,77	151/97	HDL: 5
18	9062	Mehul Ranganathan	31	Female	1	1	Single	Private	Rural	64,91	16,9	Formerly S	Frequent D	High	0	No	Vegetariar	3,26	175/84	HDL: 3
19	29940	Nishith Bhattacharyya	63	Female	1	1	Divorced	Self-employed	Rural	88,43	32,45	Currently S	Rarely	High	1	No	Non-Vege	8,29	103/86	HDL: 5
20	53292	Vritika Lala	40	Female	0	1	Single	Government Jo	Urban	199,01	31,22	Non-smok	Rarely	Moderate	1	Yes	Vegan	0,82	120/60	HDL: 5
21	23954	Taran Khatri	25	Male	0	0	Married	Private	Urban	71,38	39	Non-smok	Rarely	Moderate	0	Yes	Gluten-Fre	0,46	170/64	HDL: 7
22	73140	Nitara Kapadia	40	Female	1	1	Divorced	Government Jo	Rural	171,67	18,12	Formerly S	Social Drin	Low	1	No	Vegetariar	9,7	98/106	HDL: 4
23	62785	Manjari Dhaliwal	33	Female	0	0	Divorced	Government Jo	Urban	72,85	37,58	Formerly S	Never	Low	0	No	Non-Vege	5,65	119/105	HDL: 6
24	44810	Saira Loyal	64	Female	1	1	Divorced	Never Worked	Rural	72,66	24,82	Currently S	Frequent D	Moderate	1	No	Non-Vege	2,27	134/70	HDL: 4
25	35165	Emir Rajan	69	Male	0	1	Married	Private	Rural	149,46	27,2	Non-smok	Frequent D	Low	0	Yes	Vegan	8,4	163/94	HDL: 5
26	79771	Nayantara Issac	80	Female	0	1	Divorced	Never Worked	Urban	154,25	15,42	Non-smok	Never	High	0	Yes	Vegan	9,56	178/74	HDL: 5
27	36975	Jhanvi Brar	24	Female	0	0	Married	Self-employed	Urban	79,89	17,58	Currently S	Social Drin	High	1	No	Vegetariar	6,48	151/65	HDL: 7
28	81347	Gokul Bhakta	47	Female	0	1	Divorced	Government Jo	Rural	83,76	20,14	Non-smok	Never	Low	1	No	Gluten-Fre	9,15	102/85	HDL: 3

Wide range of variables that are suspected to influence the likelihood of a stroke, such as Gender, Marital Status, Work Type, Residence Type, Smoking Status, Alcohol Intake, Physical Activity, Family History, Dietary Habits, Symptoms

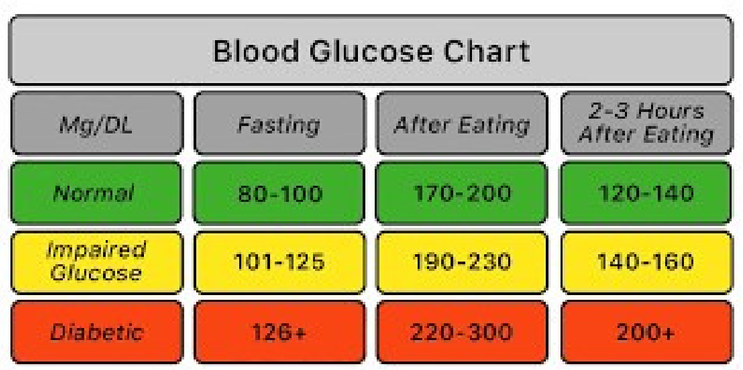
2

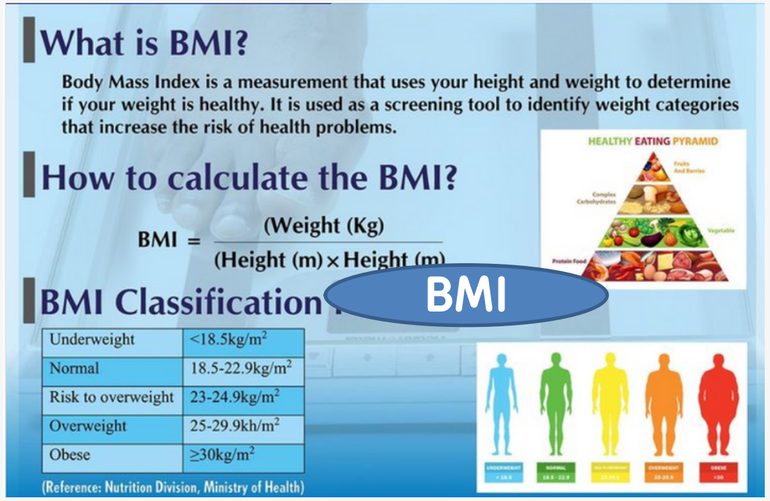
Explanation

of Data Variables

2 | Explanation of Data Variables

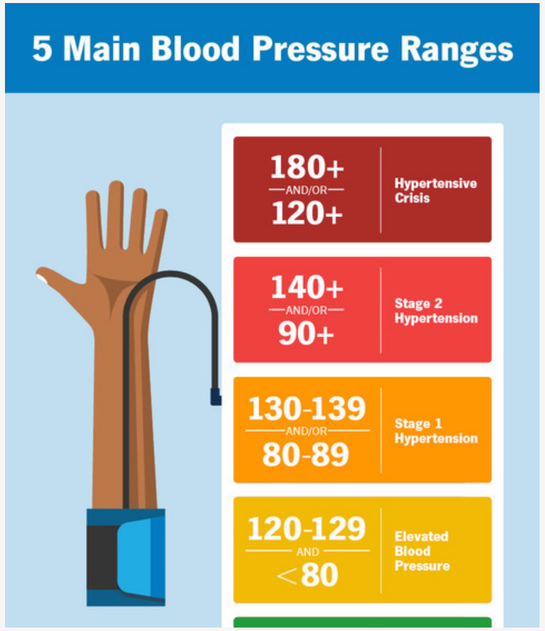
Variables	Name	Definition	Types	Uses
X1	Patient ID	*Unique identifier for each patient.*	1-15,000	Enables the tracking and differentiation of individual patients within the dataset
X2	Patient Name	*Name of the patient*	Names	Provides personal identification for reference purposes
X3	Age	*Age of the patient*	Numbers	Indicates the patient's age for demographic and age-related health assessments
X4	Gender	*Gender of the patient*	Male Female	Incorporates gender information for gender-specific health analyses and considerations
X5	Hypertension	*Presence of hypertension*	0: No 1: Yes	Flags whether the patient has a history of hypertension, a critical factor in cardiovascular health
X6	Heart Disease	*Presence of heart disease*	0: No 1: Yes	Indicates whether the patient has a history of heart disease, an essential cardiovascular health indicator
X7	Marital Status	*Marital status of the patient*	Single Married Divorced	Incorporates socio-demographic information for potential correlations with health outcomes
X8	Work Type	*Type of work the patient is engaged in*	Private Government Never worked Self-employed	Considers occupational factors that may impact health and lifestyle.
X9	Residence Type	*Type of residence*	Urban Rural	Reflects the patient's living environment, which can influence health behaviors and outcomes
X10	Average Glucose Level	*Average glucose level in the patient's blood.*	Numbers	A key indicator of blood sugar control and diabetes risk
X11	Body Mass	*Body Mass Index of*	Numbers	Evaluates the patient's weight status, a

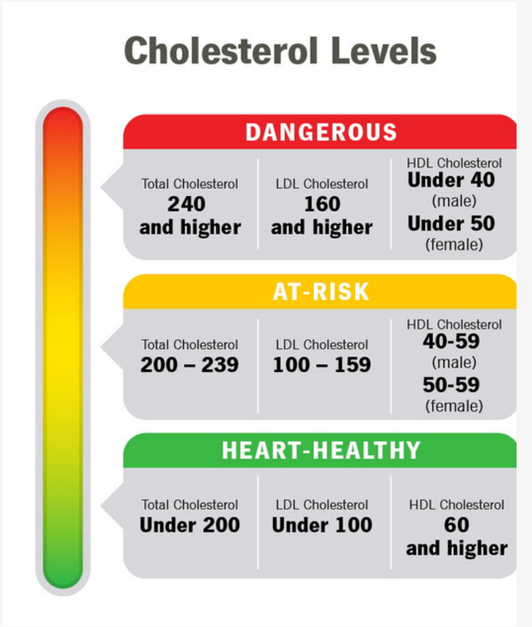




2 | Explanation of Data Variables

X12	Smoking Status	*Smoking status of the patient*	Non-smoker Formerly-smoked Currently Smokes	Captures information on tobacco use, a significant factor in respiratory and health
X13	Alcohol Intake	*Alcohol consumption status*	Social Drinker Never Rarely Frequent Drinker	Considers lifestyle factors related to alcohol consumption and potential impacts on health.
X14	Physical Activity	*Level of physical activity*	High Moderate Low	Assesses the patient's activity level, which is crucial for overall health and well-being
X15	Stroke History	*History of Stroke*	0: No 1: Yes	Indicates whether the patient has a history of stroke, a critical neurological event
X16	Family History	*Family History of Stroke*	Yes No	Considers genetic factors and family history related to stroke risk.
X17	Dietary Habits	*Dietary habits of the patient*	Vegan Paleo Pescatarian Gluten-Free Vegatarian Non-vegatarian	Captures information about the patient's dietary preferences, which can impact health outcomes
X18	Stress Levels (***)	*Stress levels of the patient.*	Numbers	Considers the patient's perceived stress levels, which can affect overall health.
X19	Blood Pressure Levels (****)	*Blood pressure levels of the patient*	Numbers	Essential for assessing health and risk factors.
X20	Cholesterol Levels (*****)	*Cholesterol levels of the patient*	Numbers	A key indicator of health
X21	Symptoms	*Symptoms reported by the patient*	Difficulty Speaking, Headache, and others	Capture additional health-related information.
Y	Diagnosis	*The target variable indicating the*	Stroke/No Stroke	





3

Explanation

of Research Topic

3 | Explanation of Research Topic

Main Objective

identifying and understanding the **factors** that contribute to the **likelihood of a patient experiencing a stroke**

Risk assessment
Evaluate the risk of stroke for an individual based on their demographic information (age, gender), lifestyle factors (smoking, alcohol intake, physical activity), or medical history (hypertension, heart disease).

Early detection
Identify patterns or combinations of factors that may indicate an increased risk of stroke, allowing for early intervention and preventive measures.

Treatment planning
Provide insights into the potential impact of specific health indicators (average glucose level, BMI, blood pressure, cholesterol levels) on stroke risk, assisting in the development of personalized treatment plans.

Public health strategies
Understand the prevalence of stroke risk factors in different populations, helping to develop targeted public health campaigns and interventions.

Patient education
Educate individuals about modifiable risk factors (such as lifestyle choices) to empower them to make informed decisions and adopt healthier behaviors

Resource allocation
Assist healthcare providers in allocating resources more efficiently by focusing on high-risk populations and tailoring interventions based on individual characteristics

Genetic and environmental factors
Explore the impact of family history, dietary habits, stress levels, and residence type on stroke risk, considering both genetic and environmental influences

Monitoring and follow-up
Establish a system for regular monitoring and follow-up of individuals identified as high-risk, ensuring timely interventions and adjustments to their healthcare plans

RESEARCH METHODS

Step 1

Data Preprocessing
Handle any missing or incomplete data.

Step 2

Variable Selection Process
Identify important variables that contribute to the prediction of the diagnosis.

Step 3

Choose a Cluster Model
Select a cluster model suitable by comparing the calculated center positions of various clusters under different clustering methods

Step 4

Split the Data
Split the dataset into training and testing sets.

Step 5

Building a predict model

4

Data Cleaning

Process

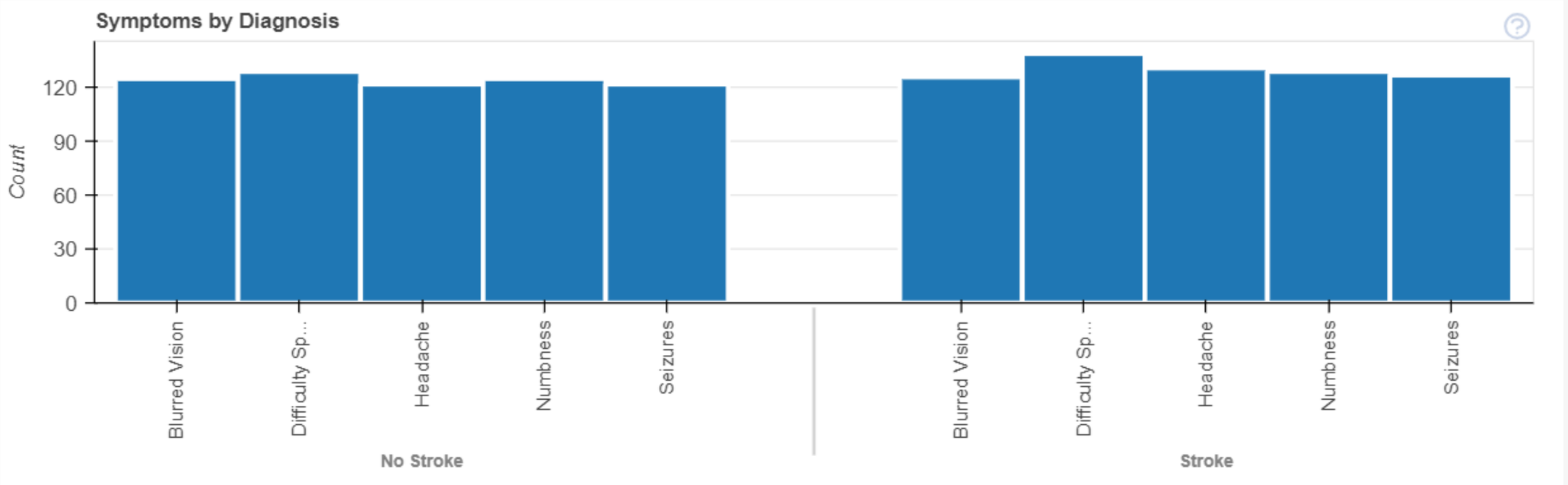
4 | Data Cleaning Process

	Patient ID	Patient Name	Age	Gender	Hypertension	Heart Disease	Marital Status	Work Type	Residence Type	Average Glucose Level	...	Alcohol Intake	Physical Activity
0	18153	Mamooty Khurana	56	Male	0	1	Married	Self-employed	Rural	130.91	...	Social Drinker	Moderate
1	62749	Kaira Subramaniam	80	Male	0	0	Single	Self-employed	Urban	183.73	...	Never	Low
2	32145	Dhanush Balan	26	Male	1	1	Married	Never Worked	Rural	189.00	...	Rarely	High
3	6154	Ivana Baral	73	Male	0	0	Married	Never Worked	Urban	185.29	...	Frequent Drinker	Moderate
4	48973	Darshit Jayaraman	51	Male	1	1	Divorced	Self-employed	Urban	177.34	...	Rarely	Low

5 rows × 22 columns

Blood Pressure Levels	Cholesterol Levels	Symptoms	Diagnosis
140/108	HDL: 68, LDL: 133	Difficulty Speaking, Headache	Stroke
146/91	HDL: 63, LDL: 70	Loss of Balance, Headache, Dizziness, Confusion	Stroke
154/97	HDL: 59, LDL: 95	Seizures, Dizziness	Stroke
174/81	HDL: 70, LDL: 137	Seizures, Blurred Vision, Severe Fatigue, Head...	No Stroke

4 | Data Cleaning Process



4 | Data Cleaning Process

	Patient ID	Patient Name	Age	Gender	Hypertension	Heart Disease	Marital Status	Work Type	Residence Type	Average Glucose Level	...	Alcohol Intake	Physical Activity	Stroke History	Family History of Stroke	Dietary Habits	Stress Levels	Blood Pressure Levels	Cholesterol Levels	Symptoms	Diagnosis
0	18153	Mamooty Khurana	56	Male	0	1	Married	Self-employed	Rural	130.91	...	Social Drinker	Moderate	0	Yes	Vegan	3.48	140/108	HDL: 68, LDL: 133	Difficulty Speaking, Headache	Stroke
1	62749	Kaira Subramaniam	80	Male	0	0	Single	Self-employed	Urban	183.73	...	Never	Low	0	No	Paleo	1.73	146/91	HDL: 63, LDL: 70	Loss of Balance, Headache, Dizziness, Confusion	Stroke
2	32145	Dhanush Balan	26	Male	1	1	Married	Never Worked	Rural	189.00	...	Rarely	High	0	Yes	Paleo	7.31	154/97	HDL: 59, LDL: 95	Seizures, Dizziness	Stroke
3	6154	Ivana Baral	73	Male	0	0	Married	Never Worked	Urban	185.29	...	Frequent Drinker	Moderate	0	No	Paleo	5.35	174/81	HDL: 70, LDL: 137	Seizures, Blurred Vision, Severe Fatigue, Head...	No Stroke
4	48973	Darshit Jayaraman	51	Male	1	1	Divorced	Self-employed	Urban	177.34	...	Rarely	Low	0	Yes	Pescatarian	6.84	121/95	HDL: 65, LDL: 68	Difficulty Speaking	Stroke

5 rows × 22 columns

```
Index(['Female', 'Male'], dtype='object')
Index(['Divorced', 'Married', 'Single'], dtype='object')
Index(['Government Job', 'Never Worked', 'Private', 'Self-employed'], dtype='object')
Index(['Rural', 'Urban'], dtype='object')
Index(['Currently Smokes', 'Formerly Smoked', 'Non-smoker'], dtype='object')
Index(['Frequent Drinker', 'Never', 'Rarely', 'Social Drinker'], dtype='object')
Index(['High', 'Low', 'Moderate'], dtype='object')
Index(['No', 'Yes'], dtype='object')
Index(['Gluten-Free', 'Keto', 'Non-Vegetarian', 'Paleo', 'Pescatarian',
       'Vegan', 'Vegetarian'],
      dtype='object')
Index(['No Stroke', 'Stroke'], dtype='object')
```

4 | Data Cleaning Process

	Patient ID	Patient Name	Age	Gender	Hypertension	Heart Disease	Marital Status	Work Type	Residence Type	Average Glucose Level	...	Alcohol Intake	Physical Activity	Stroke History	Family History of Stroke	Dietary Habits	Stress Levels	Blood Pressure Levels	Cholesterol Levels	Symptoms	Diagnosis
0	18153	Mamooty Khurana	56	Male	0	1	Married	Self-employed	Rural	130.91	...	Social Drinker	Moderate	0	Yes	Vegan	3.48	140/108	HDL: 68, LDL: 133	Difficulty Speaking, Headache	Stroke
1	62749	Kaira Subramaniam	80	Male	0	0	Single	Self-employed	Urban	183.73	...	Never	Low	0	No	Paleo	1.73	146/91	HDL: 63, LDL: 70	Loss of Balance, Headache, Dizziness, Confusion	Stroke
2	32145	Dhanush Balan	26	Male	1	1	Married	Never Worked	Rural	189.00	...	Rarely	High	0	Yes	Paleo	7.31	154/97	HDL: 59, LDL: 95	Seizures, Dizziness	Stroke
3	6154	Ivana Baral	73	Male	0	0	Married	Never Worked	Urban	185.29	...	Frequent Drinker	Moderate	0	No	Paleo	5.35	174/81	HDL: 70, LDL: 137	Seizures, Blurred Vision, Severe Fatigue, Head...	No Stroke
4	48973	Darshit Jayaraman	51	Male	1	1	Divorced	Self-employed	Urban	177.34	...	Rarely	Low	0	Yes	Pescatarian	6.84	121/95	HDL: 65, LDL: 68	Difficulty Speaking	Stroke

5 rows × 22 columns

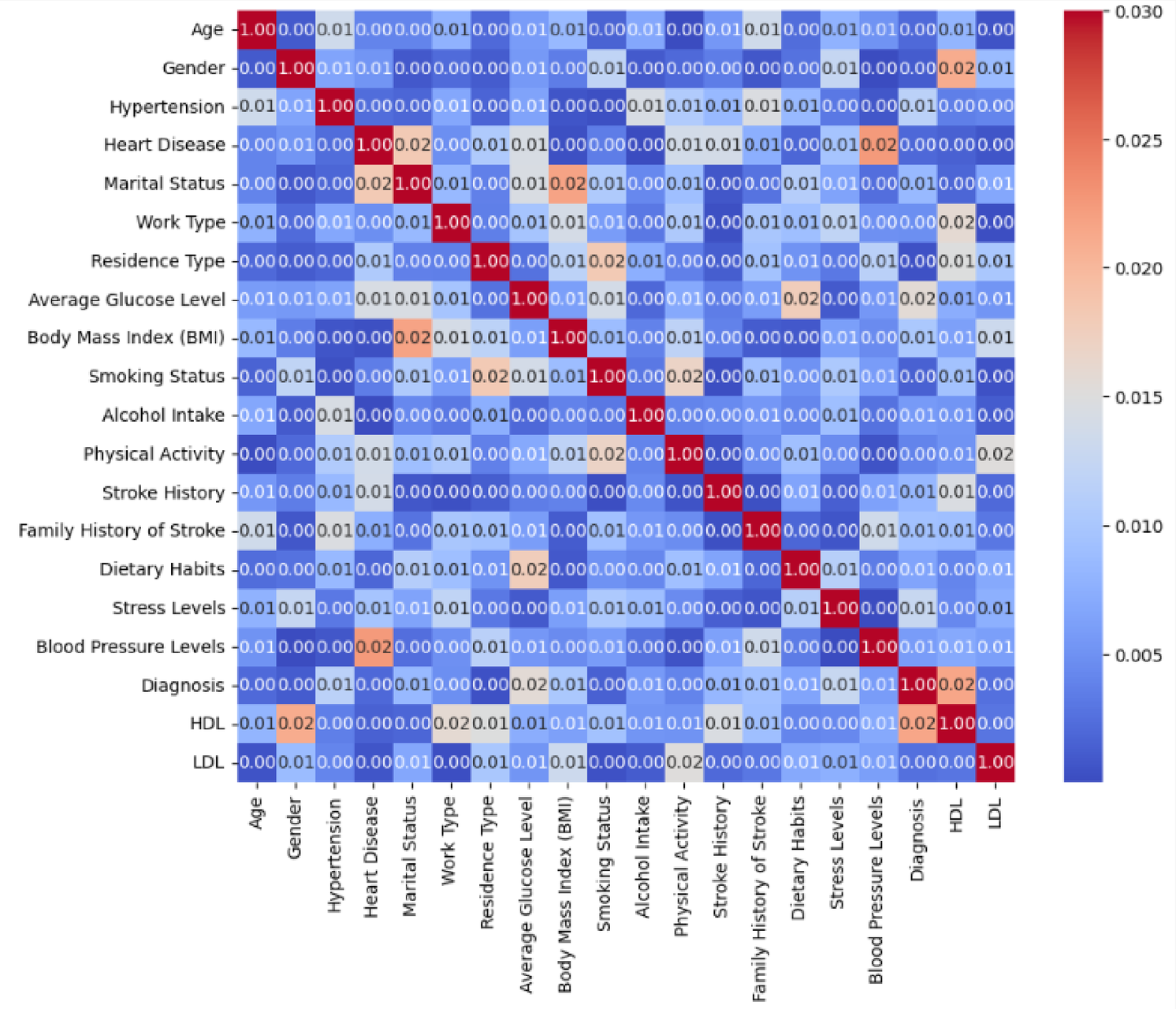
	Age	Gender	Hypertension	Heart Disease	Marital Status	Work Type	Residence Type	Average Glucose Level	Body Mass Index (BMI)	Smoking Status	Alcohol Intake	Physical Activity	Stroke History	Family History of Stroke	Dietary Habits	Stress Levels	Blood Pressure Levels	Diagnosis	HDL	LDL
0	56	1	0	1	1	3	0	130.91	22.37	2	3	2	0	1	5	3.48	1.30	1	68	133
1	80	1	0	0	2	3	1	183.73	32.57	2	1	1	0	0	3	1.73	1.60	1	63	70
2	26	1	1	1	1	1	0	189.00	20.32	1	2	0	0	1	3	7.31	1.59	1	59	95
3	73	1	0	0	1	1	1	185.29	27.50	2	0	2	0	0	3	5.35	2.15	0	70	137
4	51	1	1	1	0	3	1	177.34	29.06	0	2	1	0	1	4	6.84	1.27	1	65	68

5

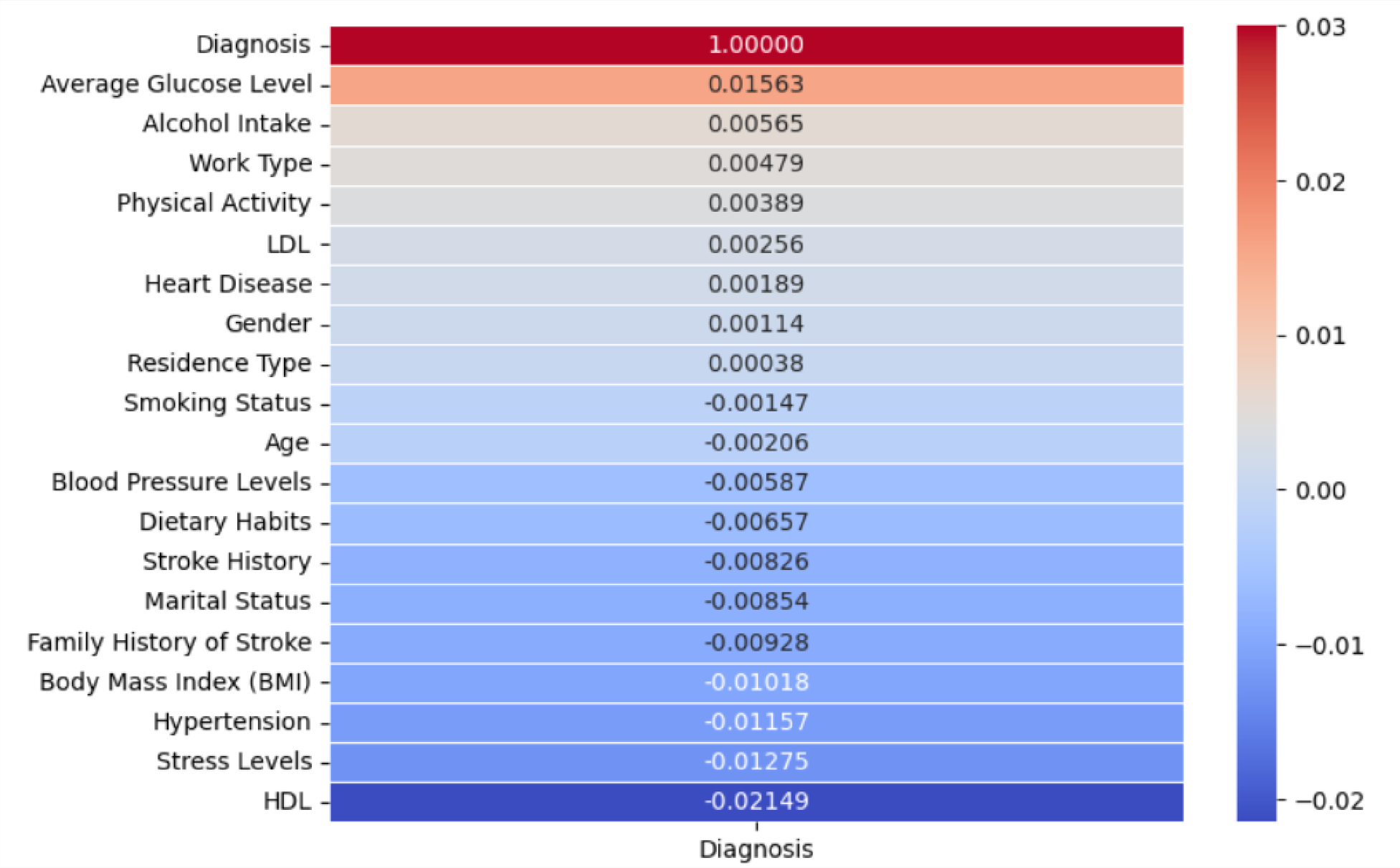
Variable Selection

Process

5 | Variable Selection Process



5 | Variable Selection Process

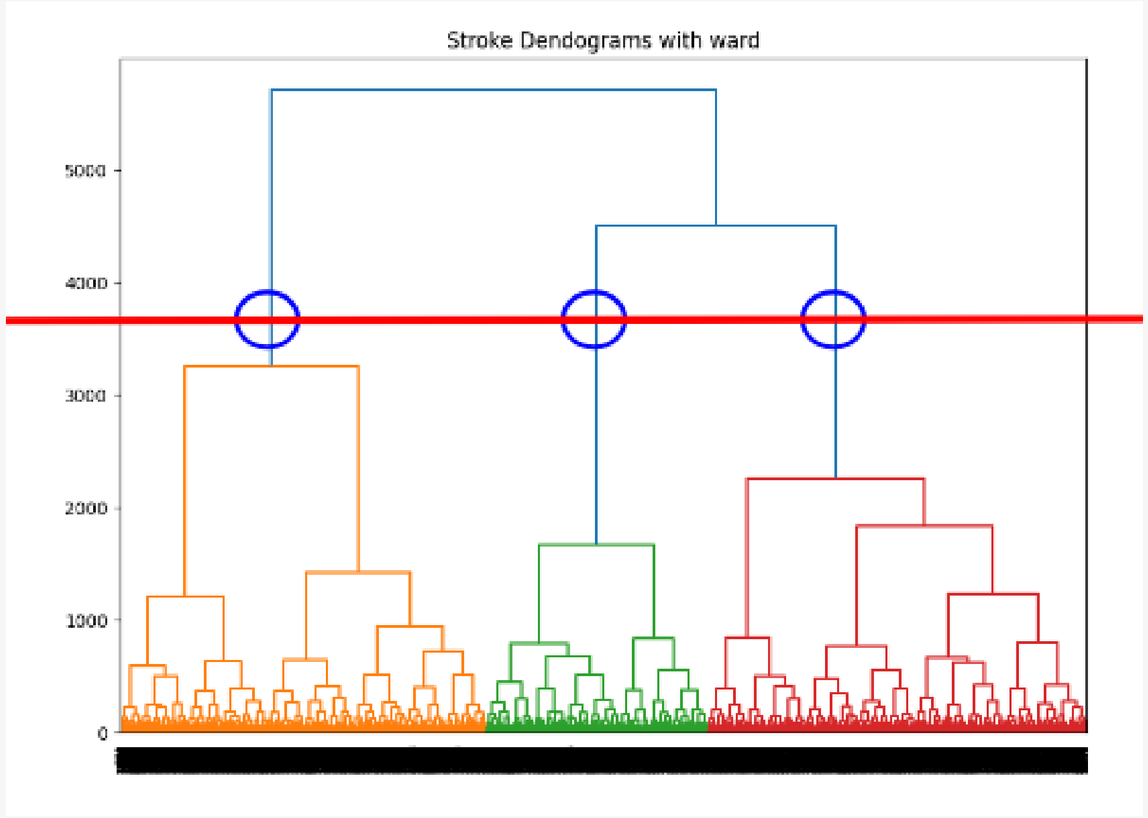


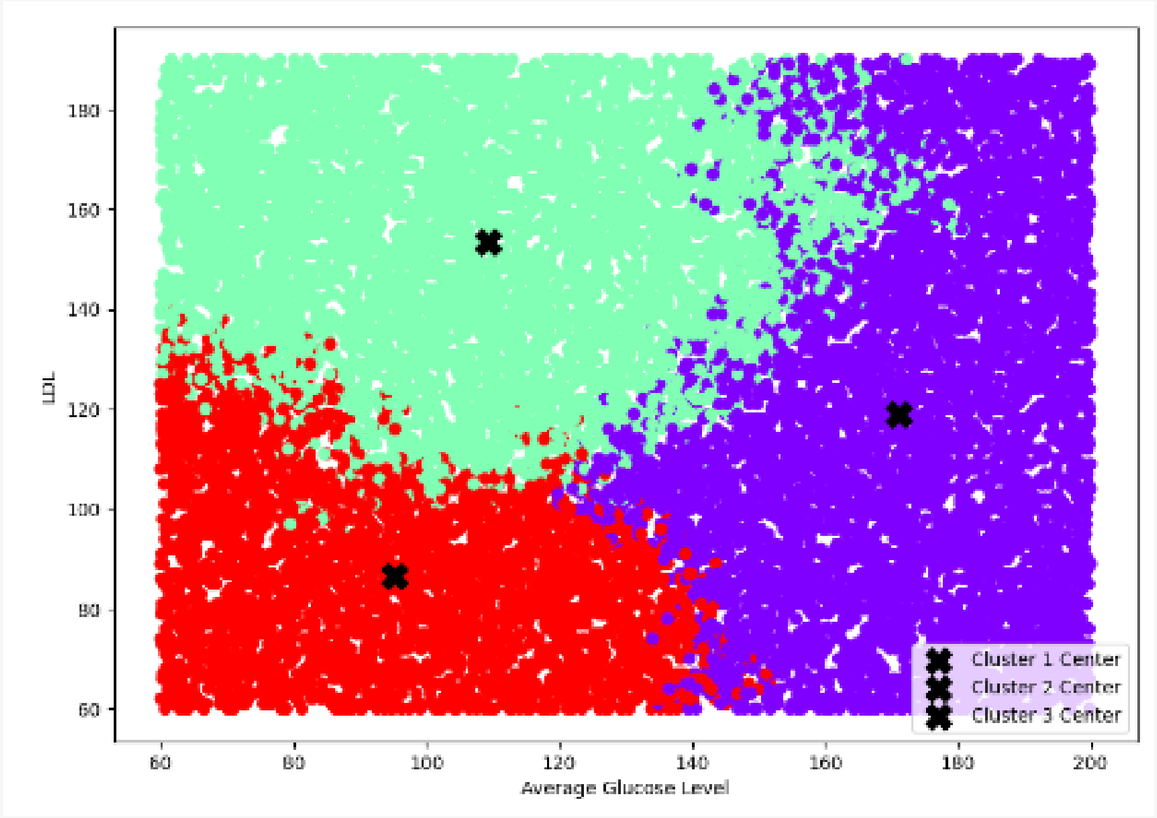
6

Cluster

6 | Cluster – Ward's Method

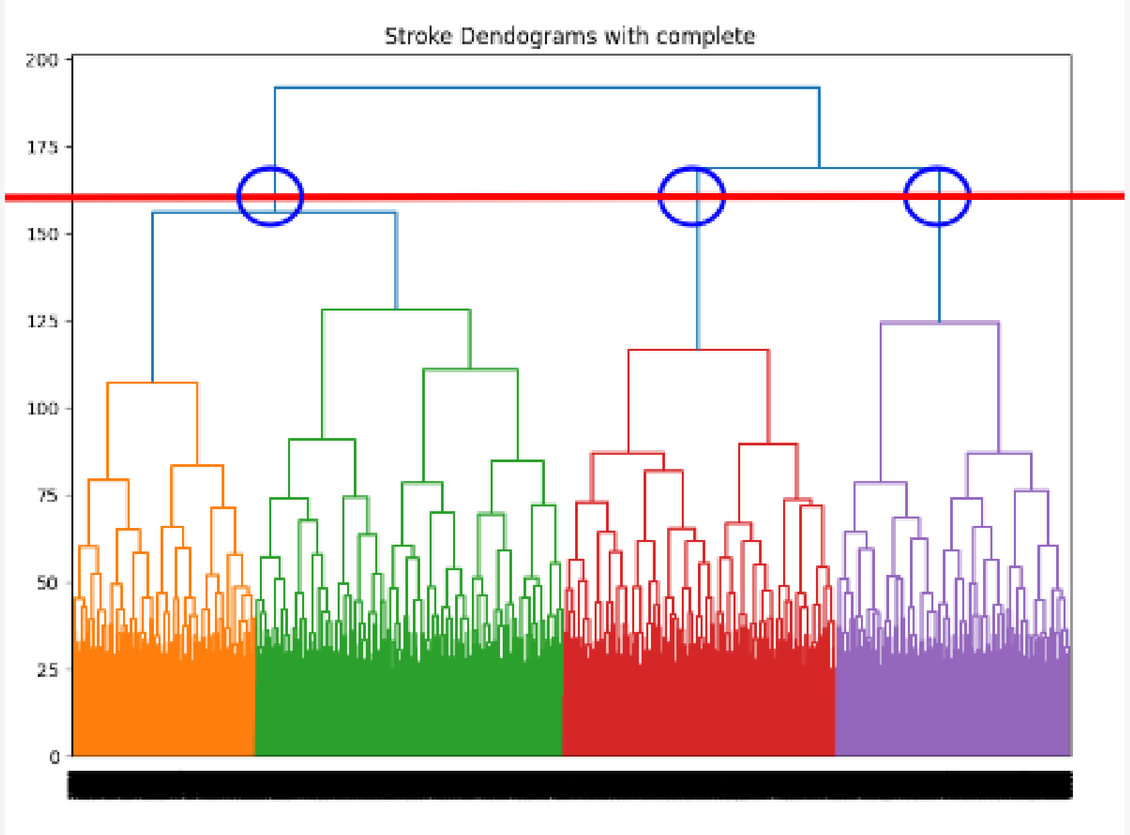
Use **Ward's** Method to categorize the selected data into 3 groups.

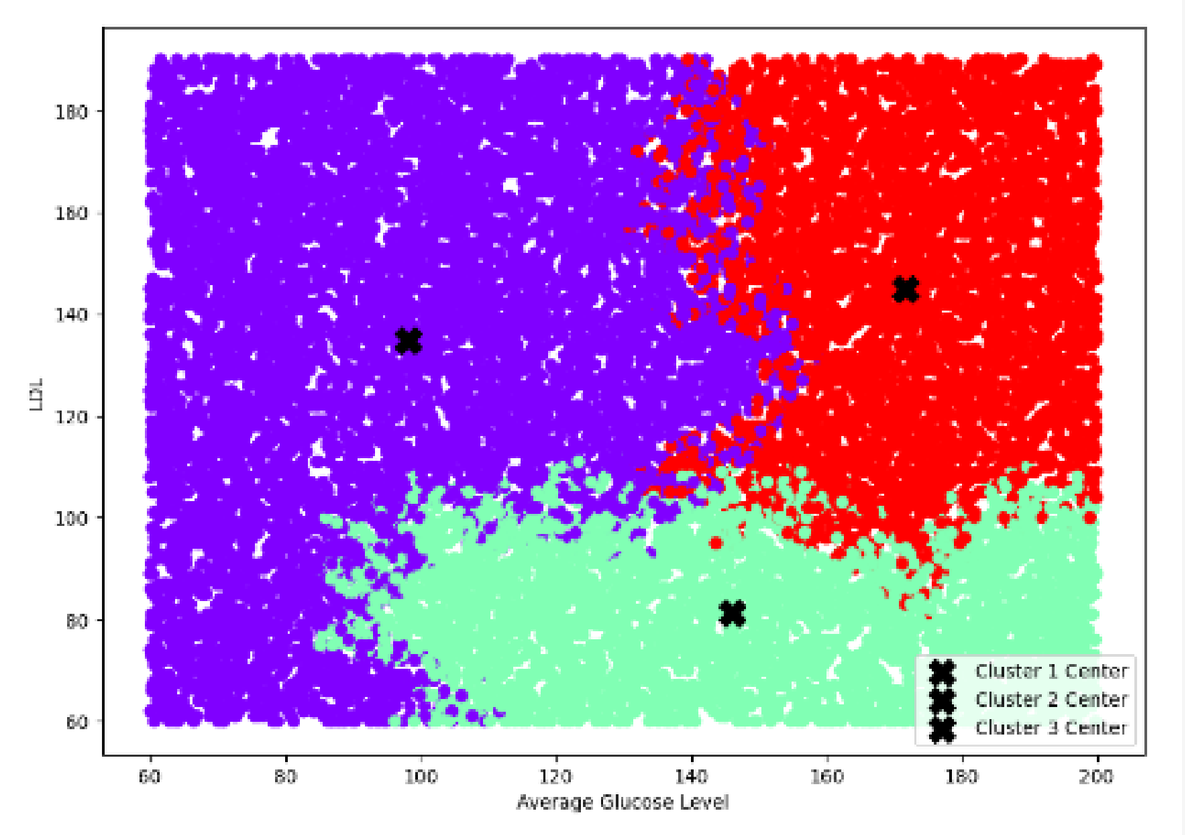




6 | Cluster – Complete Linkage Method

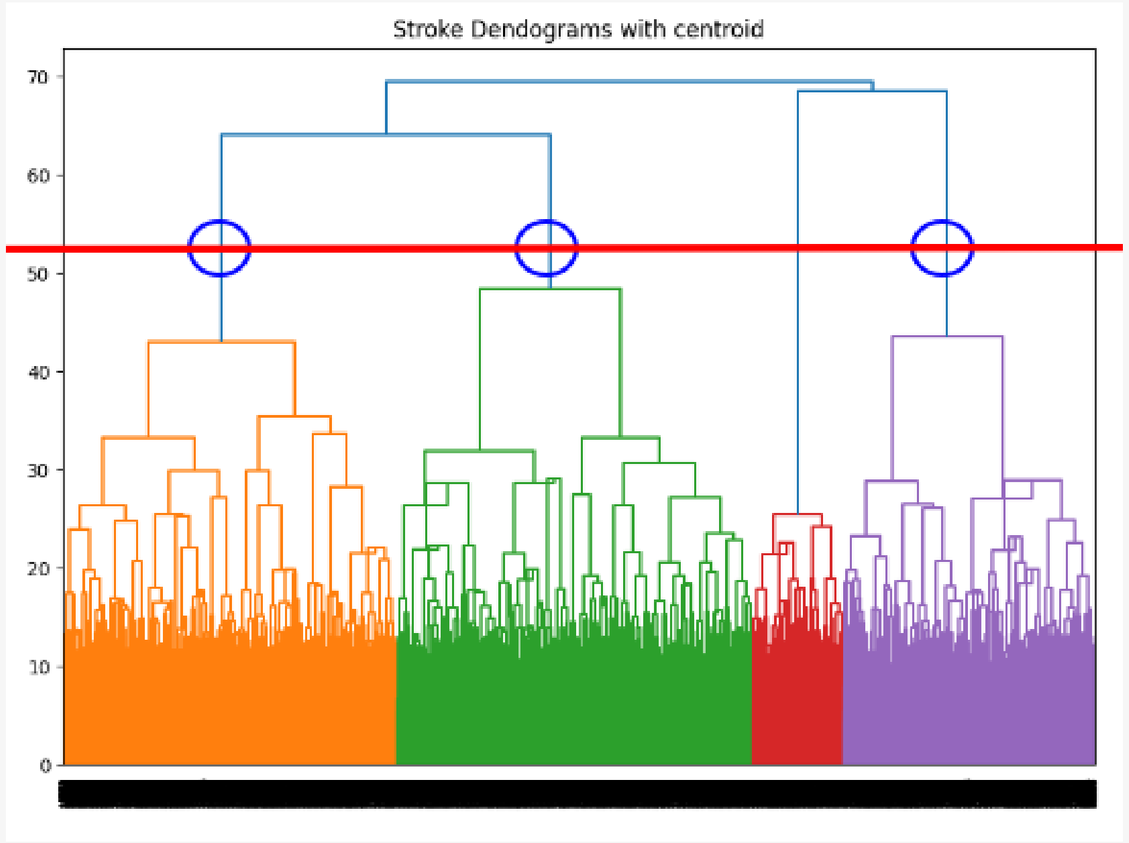
Use **Complete Linkage** Method to categorize the selected data into 3 groups.

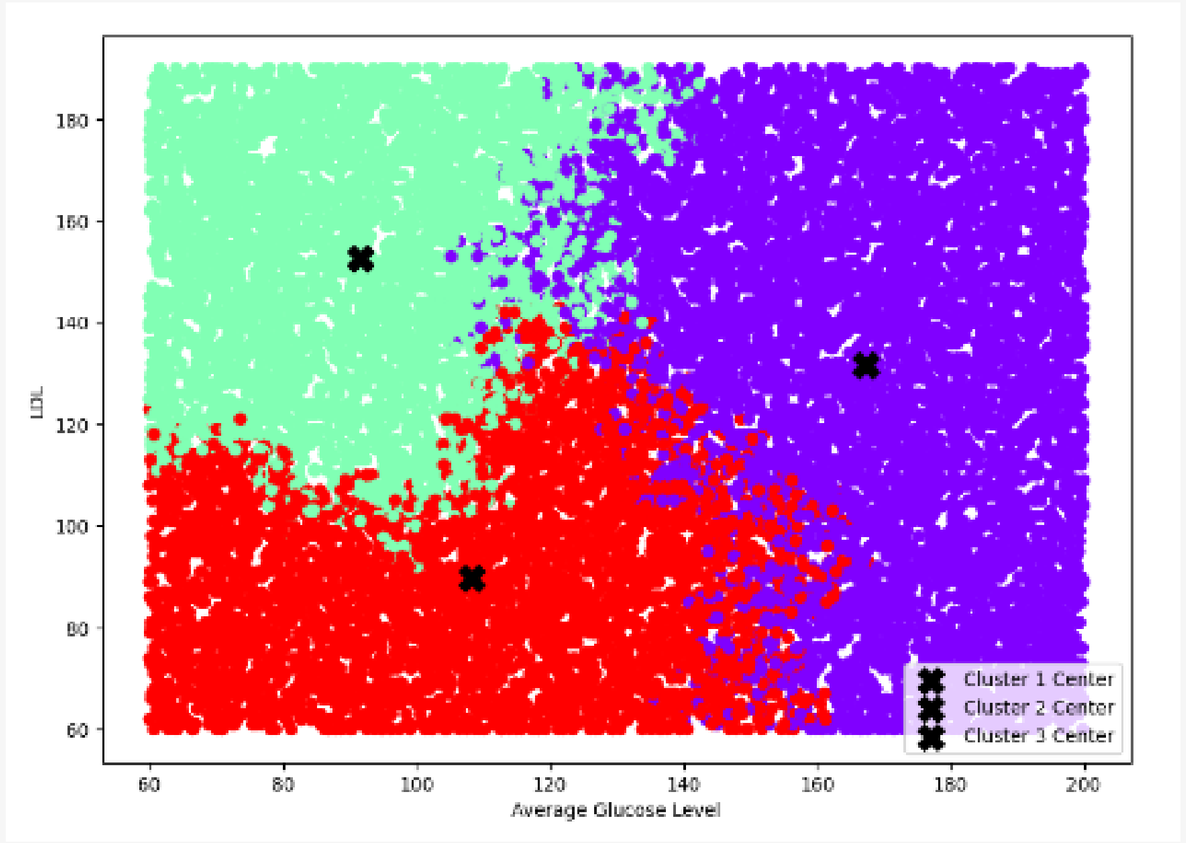




6 | Cluster – Centroid Method

Use **Centroid** Method to categorize the selected data into 3 groups.

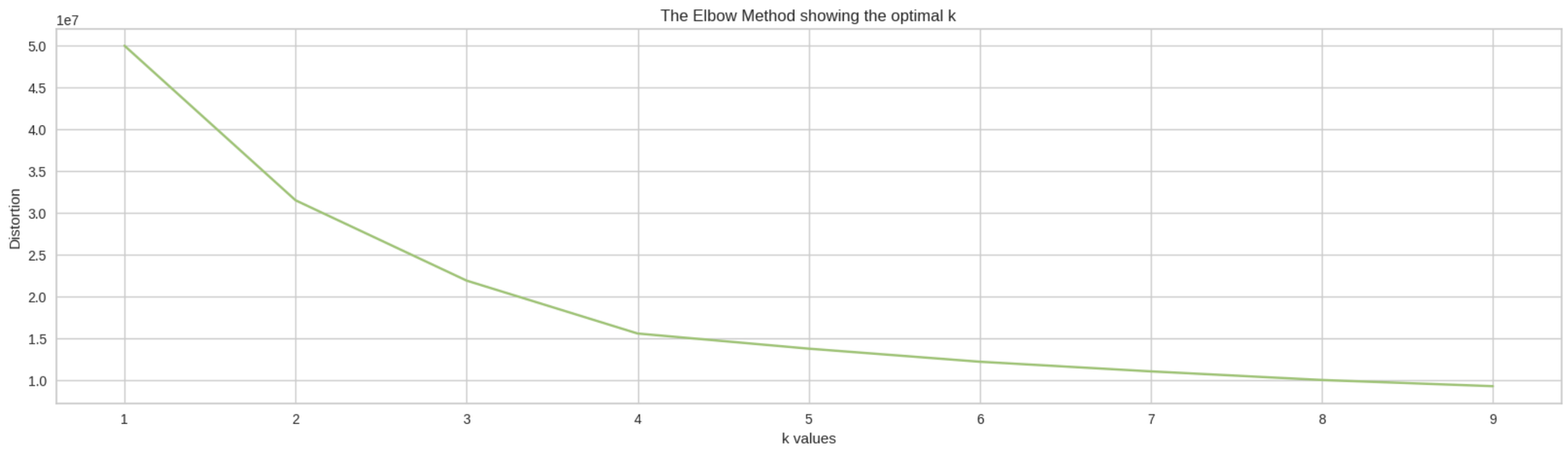




6 | Cluster – K-means

Use **K-means** to categorize the selected data into 4 groups.

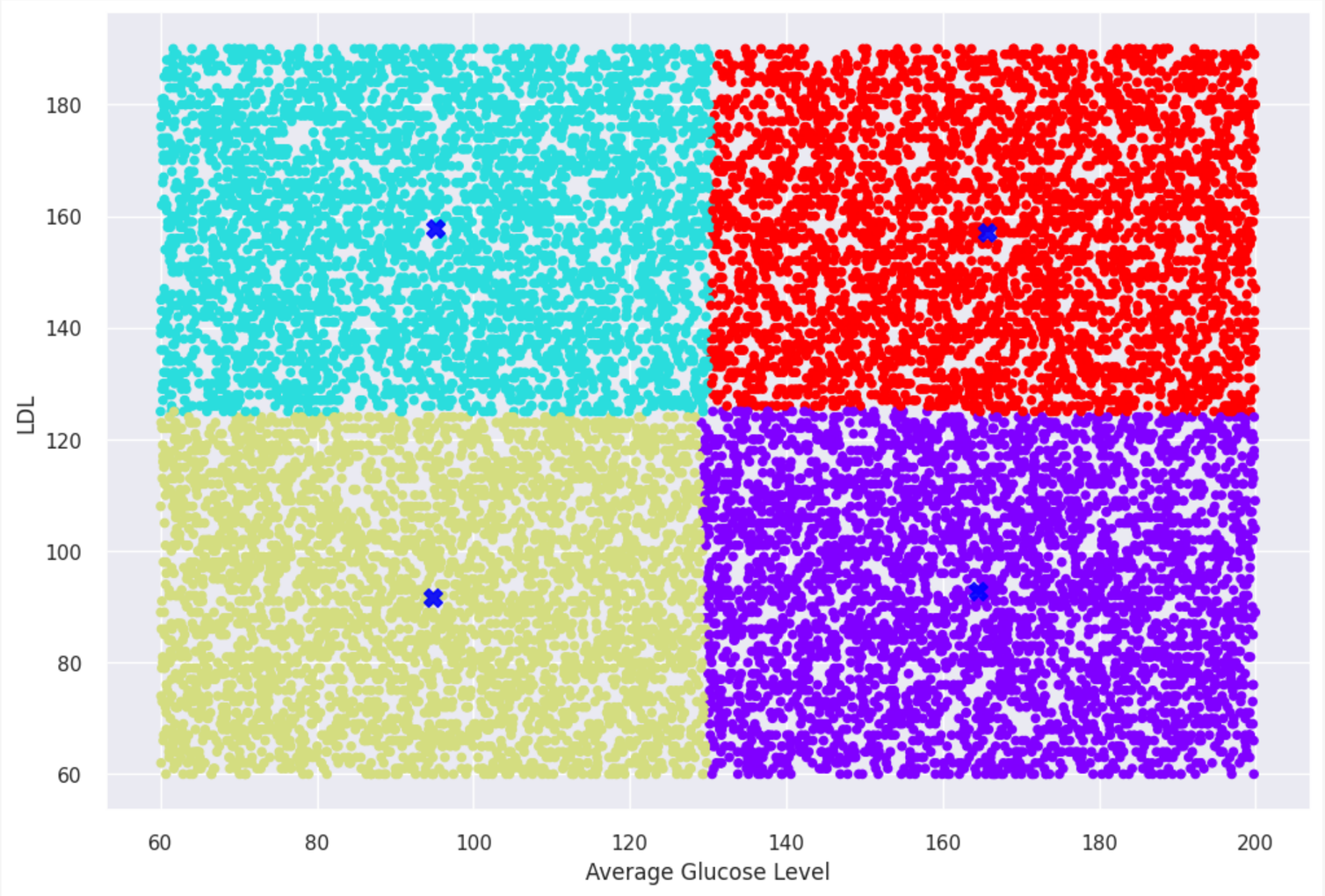
Elbow method



Silhouette score

6 | Cluster – K-means

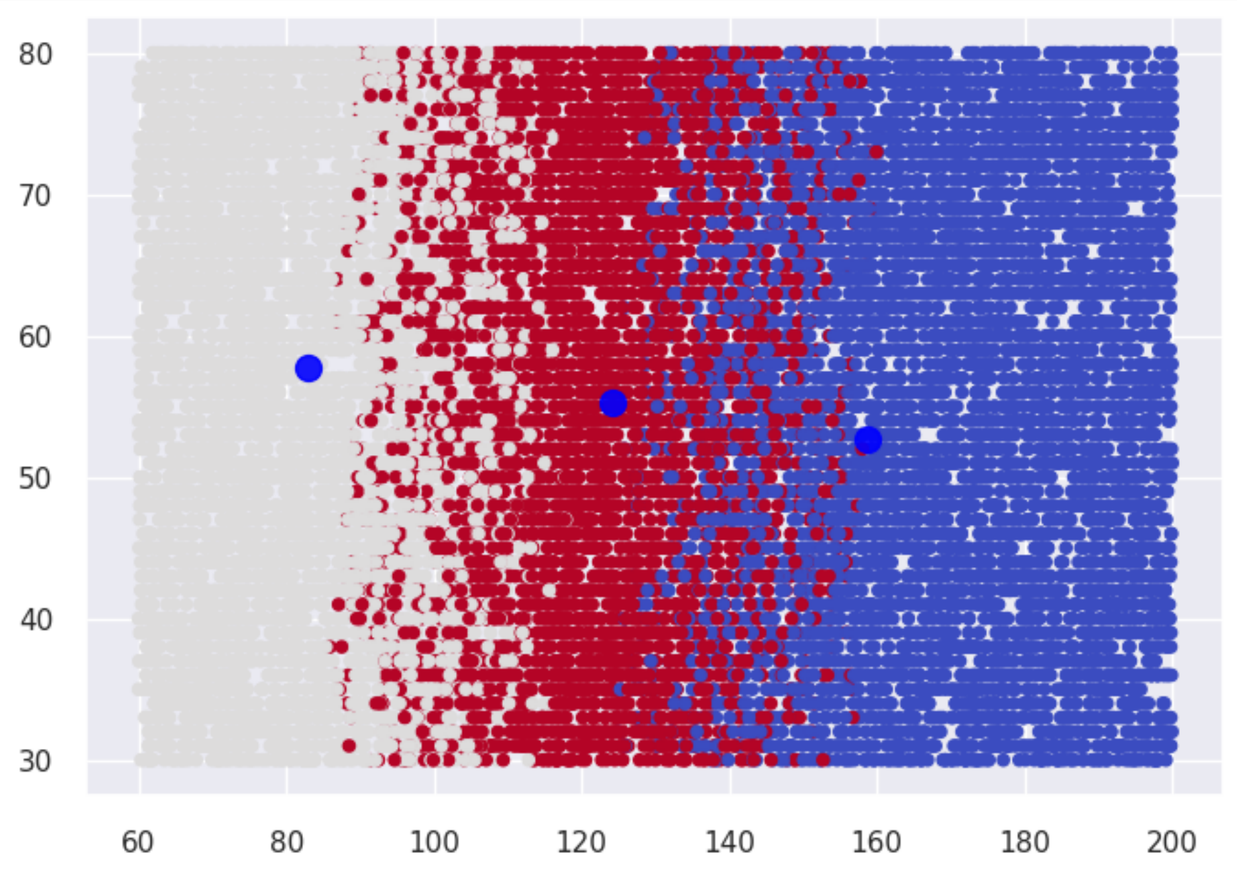
Use **K-means** to categorize the selected data into 4 groups.



6 | Cluster – Self-Organizing Map (SOM)

Use **SOM** to categorize the selected data.

```
s = som.SOM(neurons=(1,3), dimentions=6, n_iter=500, learning_rate=0.2)
s.train(samples)
print("SOM Cluster centres:", s.weights_)
print("SOM labels:", s.labels_)
result_SOM = s.predict(samples)
```



Comparison of all cluster methods

	Ward	Complete Linkage	Centroid	K-means	SOM
Cluster's samples	Cluster 1: 5691 Cluster 2: 5865 Cluster 3: 3444	Cluster 1: 7397 Cluster 2: 3505 Cluster 3: 4098	Cluster 1: 5417 Cluster 2: 4471 Cluster 3: 5112	Cluster 1: 3703 Cluster 2: 3877 Cluster 3: 3744 Cluster 4: 3676	Cluster 1: 4005 Cluster 2: 5506 Cluster 3: 5489
Intra-Cluster Distance	Avg dis C1: 40.759 Avg dis C2: 37.375 Avg dis C3: 32.006 **total avg: 36.714**	Avg dis C1: 42.982 Avg dis C2: 35.198 Avg dis C3: 34.007 **total avg: 37.396**	Avg dis C1: 43.348 Avg dis C2: 33.893 Avg dis C3: 35.780 **total avg: 37.540**	Avg dis C1: 30.967 Avg dis C2: 31.153 Avg dis C3: 30.967 Avg dis C4: 31.302 **total avg:31.097**	Avg dis C1: 55.173 Avg dis C2: 63.013 Avg dis C3: 64.966 **total avg: 61.051**
Inter-Cluster Distance	Average Inter-Cluster Distance: 73.887	Average Inter-Cluster Distance: 71.501	Average Inter-Cluster Distance: 71.816	Average Inter-Cluster Distance: 77.038	Average Inter-Cluster Distance: 51.349

Explaining the Meaning of Each K-means Cluster

Cluster	Average Glucose Level	HDL	LDL	Stress Levels	Hypertension	Body Mass Index (BMI)
0	164.467707	55.115312	92.850122	5.008844	0.243316	27.340945
1	95.143260	55.046428	157.739489	5.037908	0.247098	27.579123
2	94.678745	54.808494	91.654380	4.989228	0.255342	27.384479
3	165.752603	55.294886	157.040261	5.054684	0.250272	27.589570

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Average Glucose Level	High	Low	Low	High
LDL	Low	High	Low	High

7

Data Splitting

Combine Diagnosis columns (Model's target) and separate the data into different data frames based on cluster results.

	Average Glucose Level	HDL	LDL	Stress Levels	Hypertension	Body Mass Index (BMI)	Cluster	Diagnosis
0	130.91	68	133	3.48	0	22.37	3	1
1	183.73	63	70	1.73	0	32.57	0	1
2	189.00	59	95	7.31	1	20.32	0	1
3	185.29	70	137	5.35	0	27.50	3	0
4	177.34	65	68	6.84	1	29.06	0	1

```
df1 = split_used_df[split_used_df['Cluster'] == 0]
df2 = split_used_df[split_used_df['Cluster'] == 1]
df3 = split_used_df[split_used_df['Cluster'] == 2]
df4 = split_used_df[split_used_df['Cluster'] == 3]
```

Split each data frame into training data and testing data.

```
from sklearn.model_selection import train_test_split
#1
X1 = df1.drop("Diagnosis", axis=1)
y1 = df1["Diagnosis"]
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1, test_size=0.2, random_state=42)

# 2
X2 = df2.drop("Diagnosis", axis=1)
y2 = df2["Diagnosis"]
X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2, test_size=0.2, random_state=42)


# 3
X3 = df3.drop("Diagnosis", axis=1)
y3 = df3["Diagnosis"]
X3_train, X3_test, y3_train, y3_test = train_test_split(X3, y3, test_size=0.2, random_state=42)

# 4
X4 = df4.drop("Diagnosis", axis=1)
y4 = df4["Diagnosis"]
X4_train, X4_test, y4_train, y4_test = train_test_split(X4, y4, test_size=0.2, random_state=42)
```

Confirm that the data has been successfully split

```
print(X1_train.head())
print("===============")
print(y1_train.head())
```

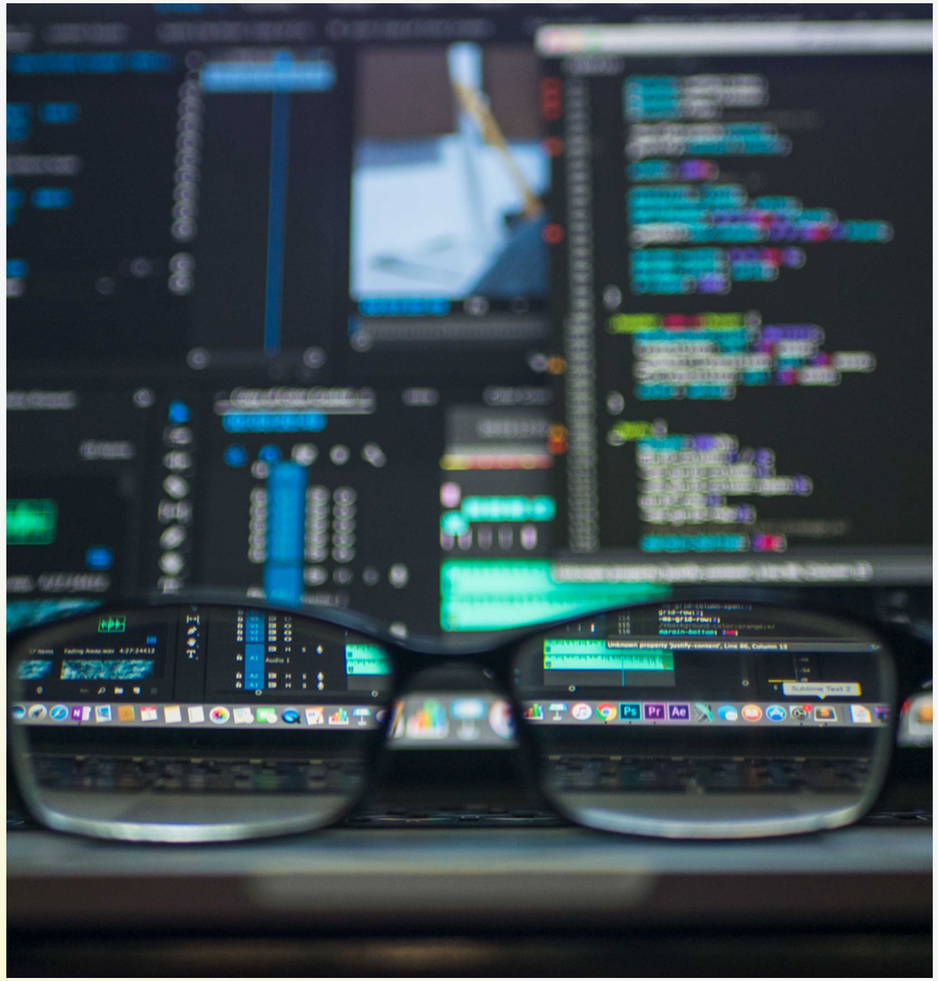
```
       Average Glucose Level  HDL  LDL  Stress Levels  Hypertension  \
12223                 131.56   38   73           1.23             0
1899                  177.77   79   63           8.88             0
20                    171.67   41  114           9.70             1
8395                  141.36   37   74           3.96             0
11487                 155.52   65  109           6.04             0

       Body Mass Index (BMI)  Cluster
12223                  19.54        0
1899                   25.32        0
20                     18.12        0
8395                   27.08        0
11487                  28.62        0
===============
12223    1
1899     0
20       0
8395     1
11487    0
Name: Diagnosis, dtype: int8
```

8

Building

a predict model



Discriminant:

```
[174] temp_discriminant_1 = y1_train.values.flatten()
      print(Counter(temp_discriminant_1))
      ### smote : balance training data
      from imblearn.over_sampling import SMOTE
      random_state = np.random.randint(0, 4294967295)
      sm = SMOTE(random_state = random_state)
      print(f"random_state used for this run: {random_state}")

      X1_train_SM, y1_train_SM = sm.fit_resample(X1_train, y1_train)
      print(Counter(y1_train_SM))
      clf1 = LinearDiscriminantAnalysis()
      clf1.fit(X1_train_SM, y1_train_SM)

      #Result of y prediction
      y1_predicted = clf1.predict(X1_test)

      ##Confusion matrix
      from sklearn.metrics import confusion_matrix
      confusion_matrix = confusion_matrix(y1_test, y1_predicted)
      print(confusion_matrix)

      accuracy_discriminant_1 = accuracy_score(y1_test, y1_predicted)
      print(f"Accuracy: {accuracy_discriminant_1}")
```

Cluster 1:

```
[[170 190]
 [192 184]]
Accuracy: 0.480978260869565 24
```

Cluster 2:

```
[[217 176]
 [180 176]]
Accuracy: 0.5246995994659546
```

Cluster 3:

```
[[185 187]
 [174 195]]
Accuracy: 0.5128205128205128
```

Cluster 4:

```
[[188 198]
 [203 187]]
Accuracy: 0.4832474226804124
```

Decision Tree:

```
temp_DecisionTree_1=y1_train.values.flatten()
print(Counter(temp_DecisionTree_1))
from imblearn.over_sampling import SMOTE
random_state_smote = np.random.randint(0, 4294967295)
sm = SMOTE(random_state=random_state_smote)
print(f"random_state_smote used for this run: {random_state_smote}")

X1_train_SM, y1_train_SM = sm.fit_resample(X1_train, y1_train)
print(Counter(y1_train_SM))

random_state = np.random.randint(0, 4294967295)
# Create a CART template
clf1= DecisionTreeClassifier(max_depth =5, random_state = random_state)
# Train the model on the training data
clf1.fit(X1_train_SM, y1_train_SM)
y1_predicted = clf1.predict(X1_test)
##Confusion matrix
from sklearn.metrics import accuracy_score, confusion_matrix
confusion_matrix = confusion_matrix(y1_test, y1_predicted)
print(confusion_matrix)

# Calculate the accuracy of the model
accuracy_DecisionTree_1 = accuracy_score(y1_test, y1_predicted)
print("Accuracy:", accuracy_DecisionTree_1)
print(f"random_state used for this run: {random_state}")
```

Cluster 1:

```
[[ 94 266]
 [ 81 295]]
Accuracy: 0.5285326086956522
```

Cluster 2:

```
[[175 218]
 [153 203]]
Accuracy: 0.5046728971962616
```

Cluster 3:

```
[[215 157]
 [202 167]]
Accuracy: 0.5155195681511471
```

Cluster 4:

```
[[129 257]
 [100 290]]
Accuracy: 0.5399484536082474
```

Neural Network:

```
temp_neural_network_1=y1_train.values.flatten()
print(Counter(temp_neural_network_1))
from imblearn.over_sampling import SMOTE
random_state_smote = np.random.randint(0, 4294967295)
sm = SMOTE(random_state=random_state_smote)
print(f"random_state_smote used for this run: {random_state_smote}")

X1_train_SM, y1_train_SM = sm.fit_resample(X1_train, y1_train)
print(Counter(y1_train_SM))

random_state = np.random.randint(0, 4294967295)
# Create a MLP template
clf1= MLPClassifier(alpha=0.5, max_iter=1000, random_state=random_state)
# Train the model on the training data
clf1.fit(X1_train_SM, y1_train_SM)
y1_predicted = clf1.predict(X1_test)
##Confusion matrix
from sklearn.metrics import accuracy_score, confusion_matrix
confusion_matrix = confusion_matrix(y1_test, y1_predicted)
print(confusion_matrix)

# Calculate the accuracy of the model
accuracy_neural_network_1 = accuracy_score(y1_test, y1_predicted)
print("Accuracy:", accuracy_neural_network_1)
print(f"random_state used for this run: {random_state}")
```

Cluster 1:

```
[[129 231]
 [127 249]]
Accuracy: 0.5135869565217391
```

Cluster 2:

```
[[389   4]
 [343  13]]
Accuracy: 0.5367156208277704
```

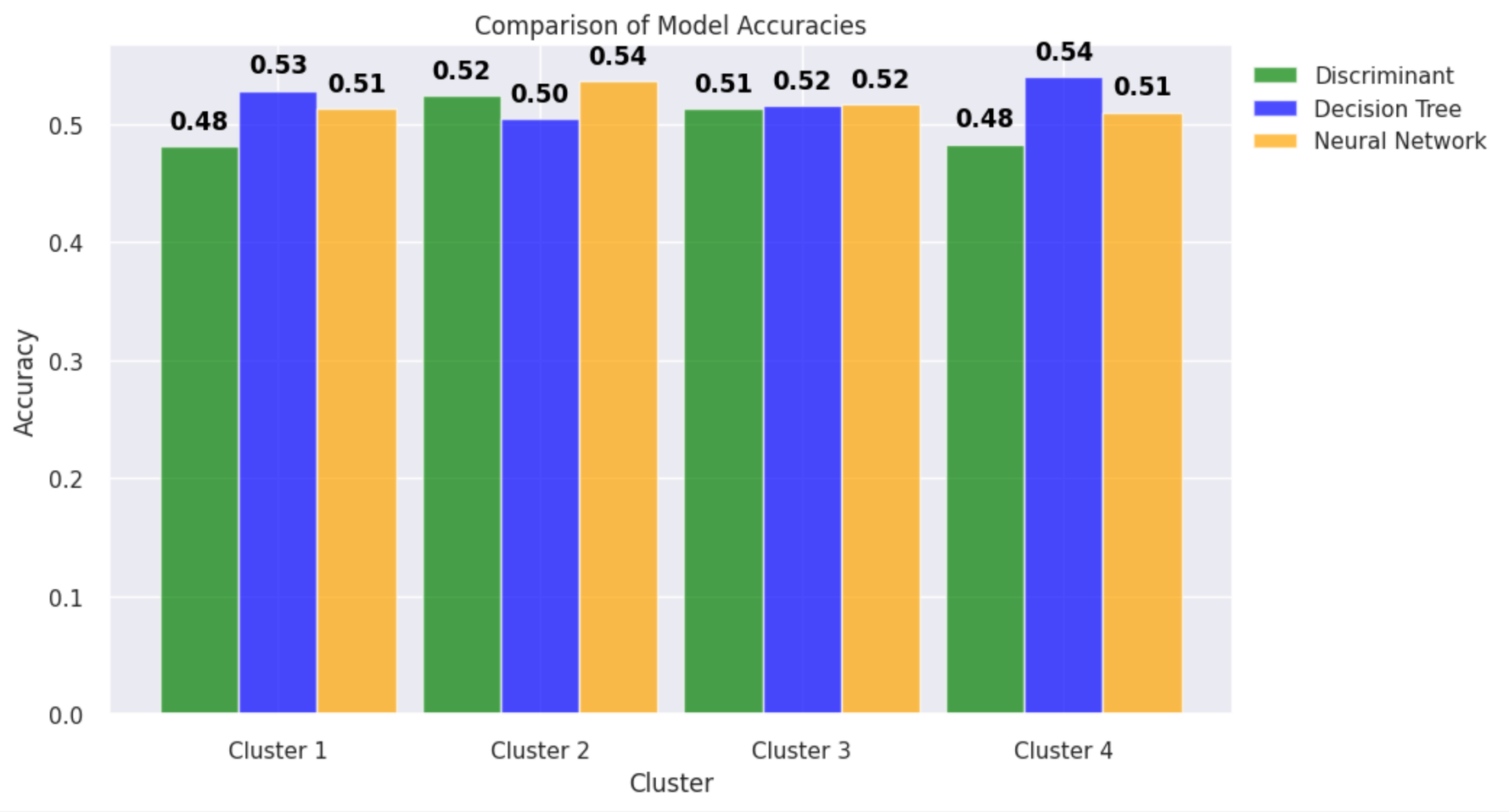
Cluster 3:

```
[[333  39]
 [319  50]]
Accuracy: 0.5168690958164642
```

Cluster 4:

```
[[225 161]
 [219 171]]
Accuracy: 0.5103092783505154
```

8 | Building a predict model



3 methods accuracies around 48% ~ 54%.

Decision Tree method have a slightly better of 53%, 50%, 52% and 54%

--> **Decided to adopt the Decision Tree method for further analysis and predictions.**

```
mean accuracy discriminant: 0.5004364489591112
mean accuracy decision_tree: 0.5221683819128271
mean accuracy neural_network: 0.5193702378791223
```

9

Summary

- Some features influencing stroke prediction, such as Average Glucose Level, HDL, LDL, Stress Levels, Hypertension, and BMI, were identified through thorough analysis.

(But in our analysis, these features have differences between people with stroke and without stroke, but this is a very very small difference, **only Average Glucose Level, HDL, LDL are slightly higher**).

- The model's accuracy is just a bit above 50%. It needs some improvements. We can make it better by getting more detailed data and looking at other things.

- Future Improvements:
  - Get more and better data.
  - Look into genes, the environment, and more health details.
  - Experiment with different ways of doing things, change settings
  - Make the model work better for certain groups or types of people.

Thanks for listening