Name: Rahul Vijayvargiya
Student: 245784

Task:
Heart failure clinical records Link for dataset:

[Dataset Link]

The collection contains data related to cases of heart failure. Models learned from this dataset were used to assess patient survival and aid treatment selection. The collection is adapted for classification and clustering tasks – for this assignment it will be used for clustering.

Sol.

Hello, We have a dataset of clinical Heart Disease and people who got affected with respect their age, gender and cause of death and survive from the disease,

Here we are going to use clustering, unsupervised machine learning problem, with respect K-Means and DBSCAN

K-Means:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.

DBSCAN:

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers). The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

Is DBSCAN better than KMeans?

K-Means

K-means has difficulty with non-globular clusters and clusters of multiple sizes.

DBSCAN

DBSCAN is used to handle clusters of multiple sizes and structures and is not powerfully influenced by noise or outliers.

Here we go with our report and data preprocessing part:

```
# DataFrame First 5 Rows
df.head()
```

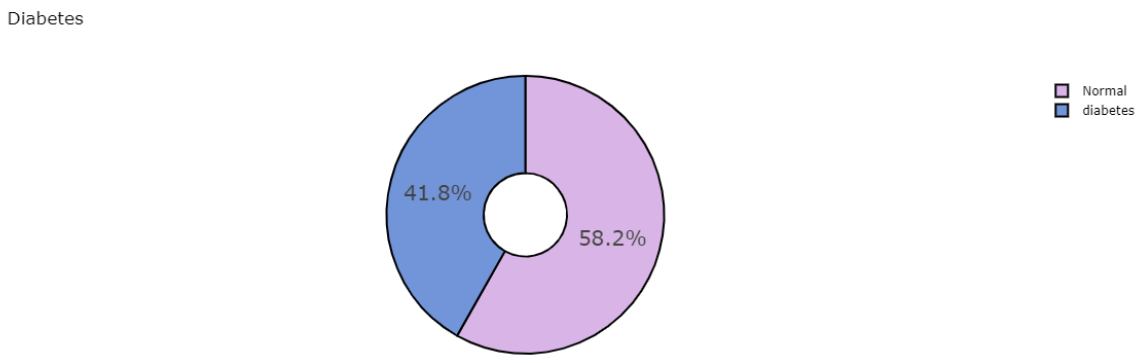
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8

First fives rows of of our data set,

Index(['age', 'anaemia', 'creatinine_phosphokinase', 'diabetes',
'ejection_fraction', 'high_blood_pressure', 'platelets',
'serum_creatinine', 'serum_sodium', 'sex', 'smoking', 'time',
'DEATH_EVENT'], dtype='object')

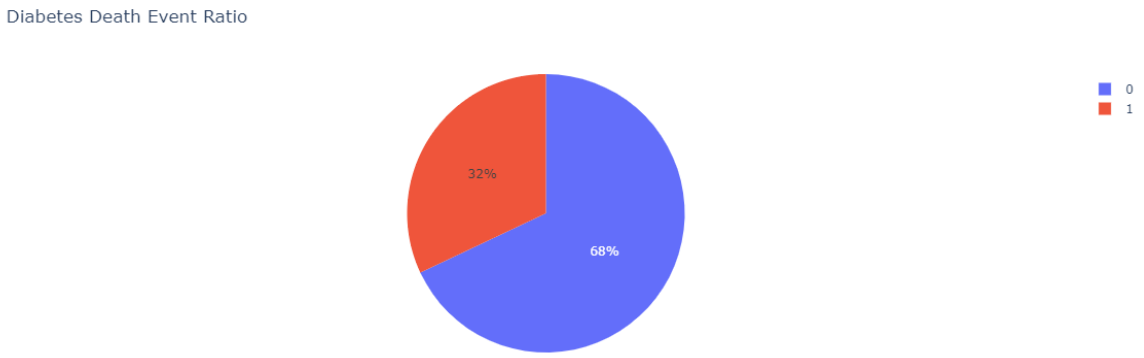
As you can see above the columns name of our dataset tells us about age, the disease and sex and death, he or she survived or died from diease

1.



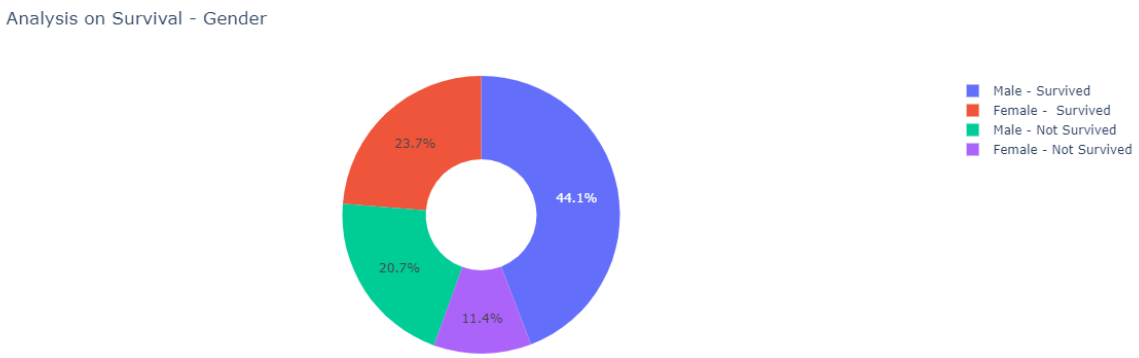
Above you see total no. of people are affected from diabetes or normal person in the dataset

2. plot



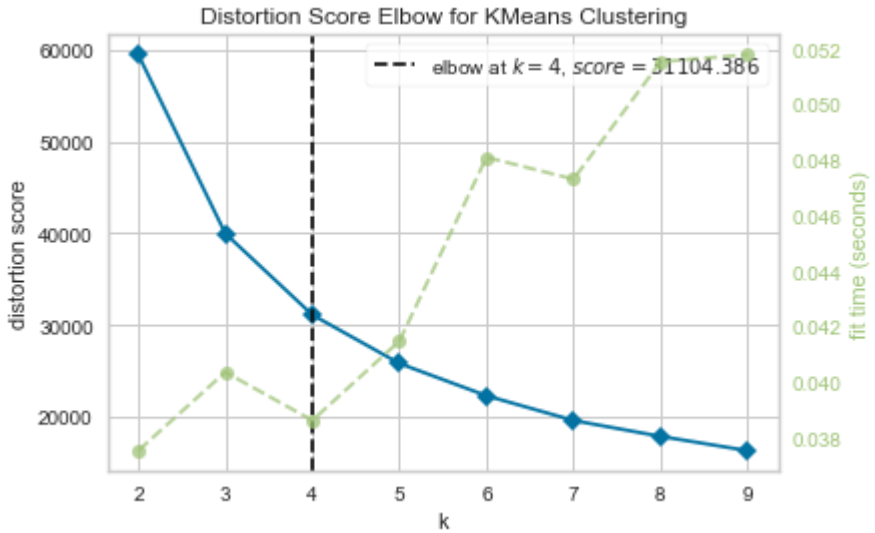
People died from diabetes event ratio

3. plot



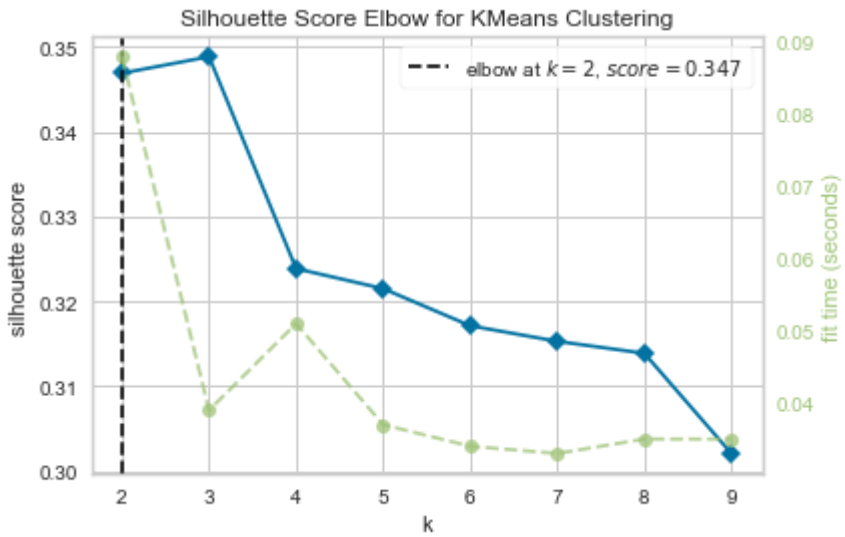
Male v/s Female Survival ratio

4. Metrics Distortion score



It suggest us 4 cluster

5. Silhouette score



It suggest us two cluster

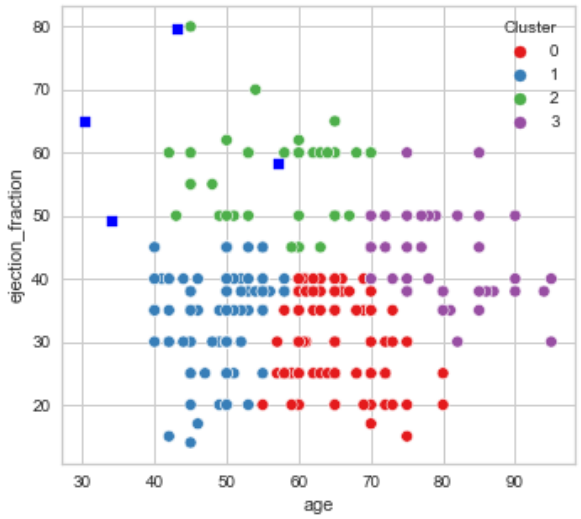
Since the choice of method is unsupervised learning, so we do not have any label data to form cluster, for k-means we provide k value, as per k value it will gonna form a cluster

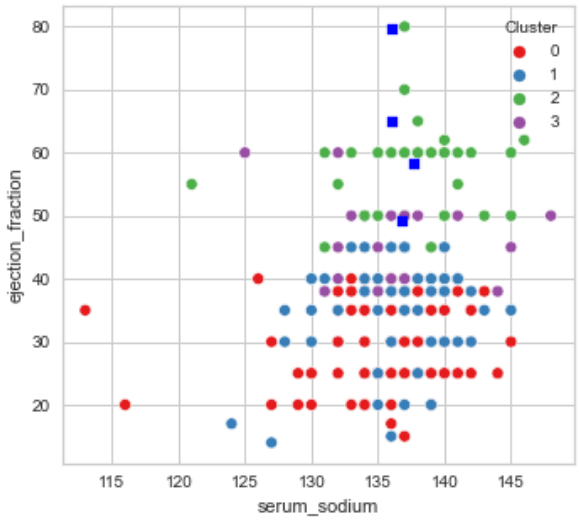
4 Cluster:

```
# initialise and fit K-Means model, we want 4 clusters to form

KM_4_clusters = KMeans(n_clusters=4, init='k-means++').fit(x)

KM4_clustered = x.copy()
KM4_clustered.loc[:,'Cluster'] = KM_4_clusters.labels_
```

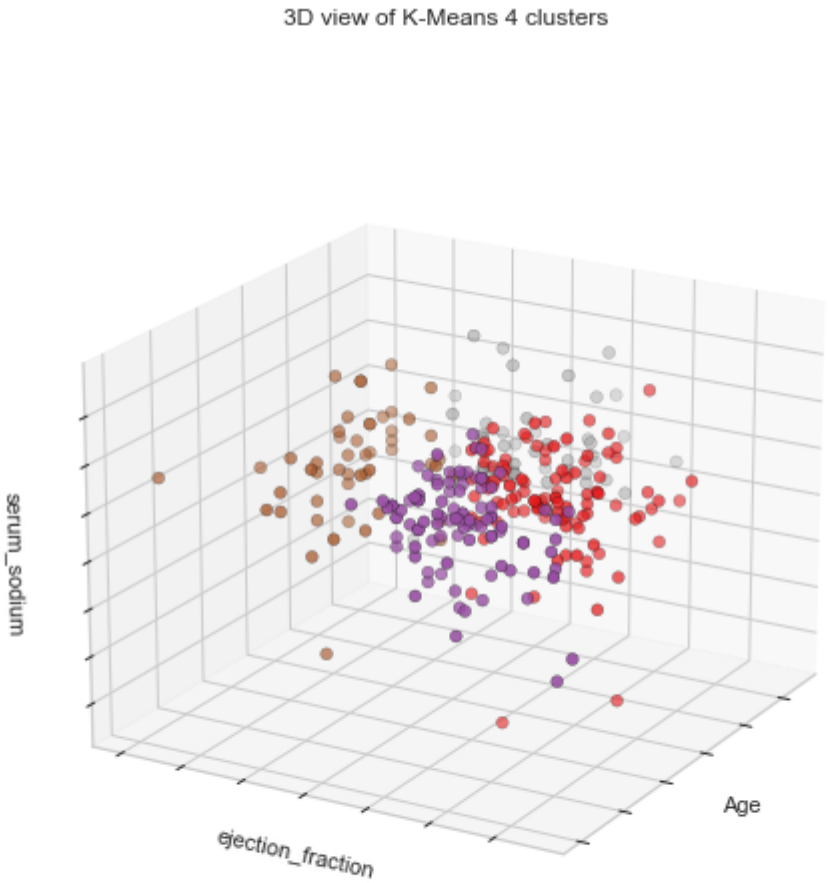
24] ✓ 0.9s

2D View:





3D view:



DBSCAN, it's a density based clustering works on epsilon and minimum sample in cluster to form a cluster

```python
# eps= 5 and min_samples= 5, Density base Clustring and with 4 min samples needed to make a cluster

DBS_clustering = DBSCAN(eps=5, min_samples=5).fit(x)

DBSCAN_clustered = x.copy()
DBSCAN_clustered.loc[:,'Cluster'] = DBS_clustering.labels_ # append labels to points
```

[28] ✓ 0.4s

```python
DBSCAN_clust_sizes = DBSCAN_clustered.groupby('Cluster').size().to_frame()
DBSCAN_clust_sizes.columns = ["DBSCAN_size"]
DBSCAN_clust_sizes
```

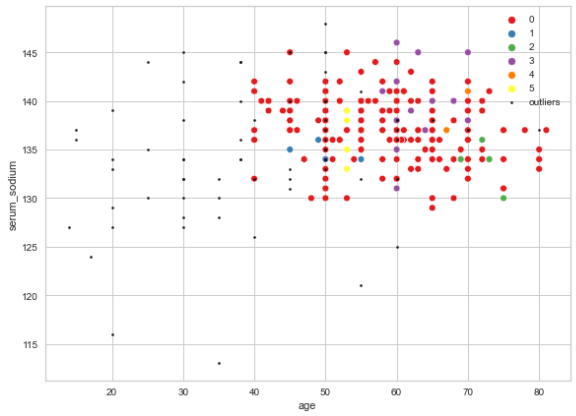
[29] ✓ 0.1s

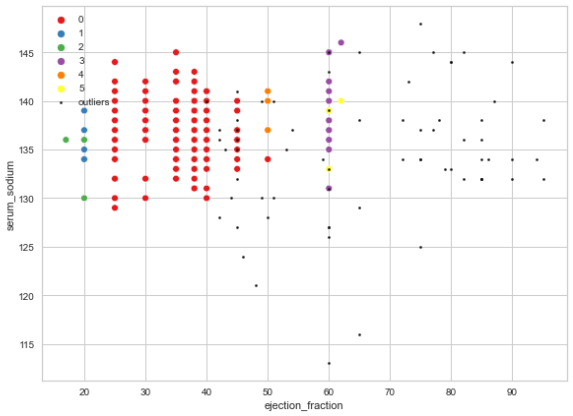
Cluster	DBSCAN_size
-1	68
0	186
1	7
2	6
3	22
4	5
5	5

DBSCAN created 5 clusters plus outliers cluster (-1). There are 68 outliers.

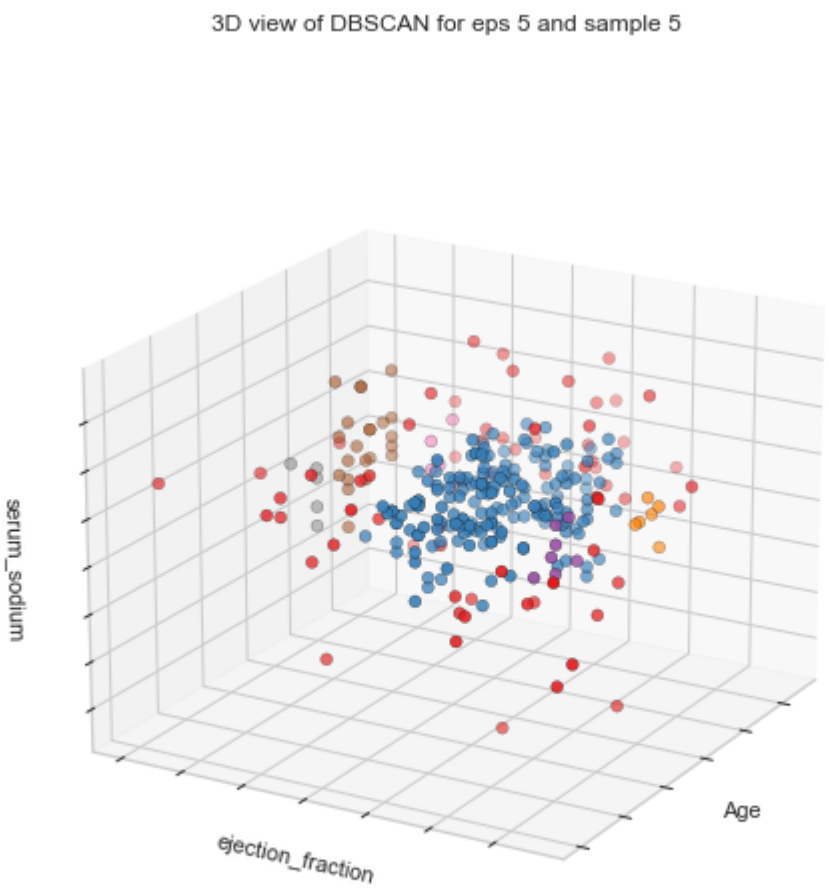
As you can see a epsilon value is 5 which means the diameter or cluster is 5 and minimum samples means that it should have at least 5 samples, minimum 5 samples to form a cluster

2D View:



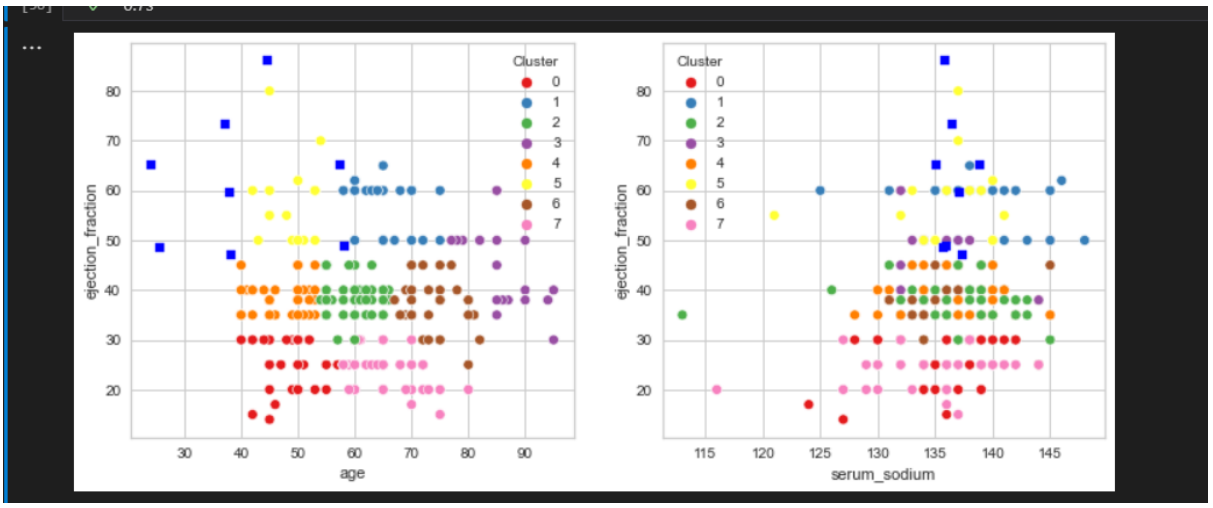


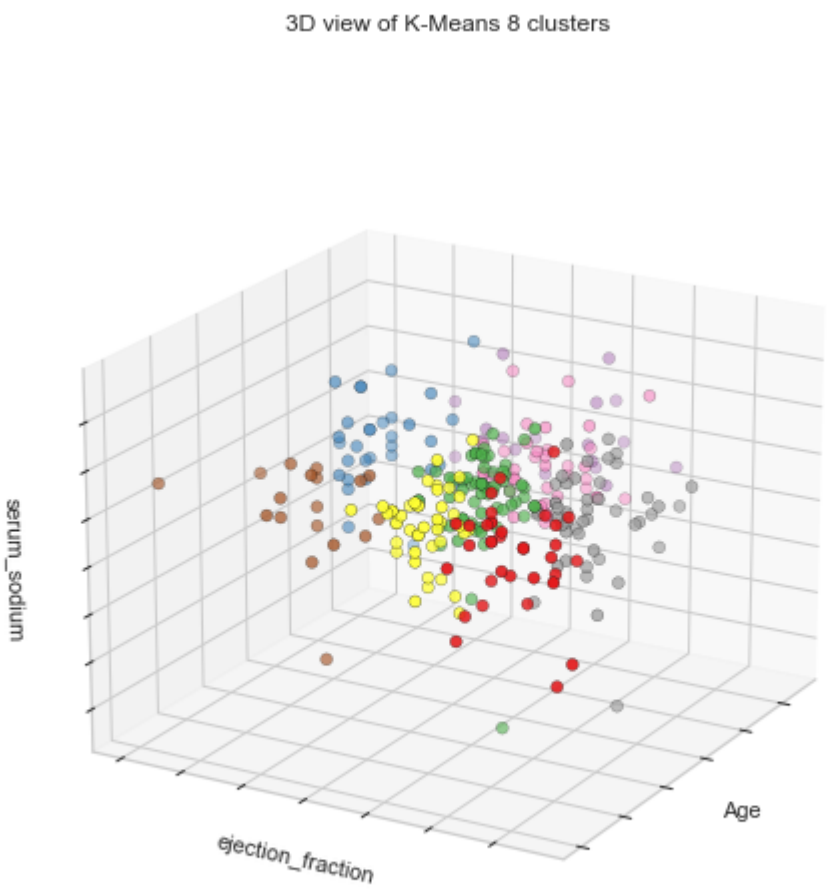
3D View:



Experiments:

With K-means form 8 clusters





8 clusters in 3D Pane

DBSCAN:
EPS: 10 and Sample Size 10, it formed a 1 big cluster and -1 stands for outliers

Cluster	DBSCAN_size
-1	8
0	291

3D view of DBSCAN for eps 10 and sample 10