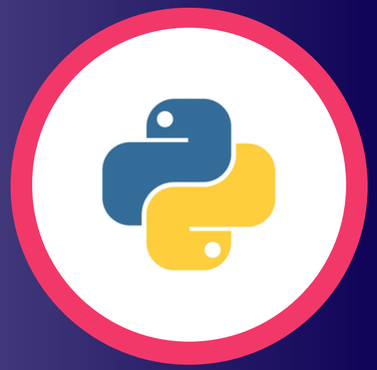
Alhag Ali, Mahmoud

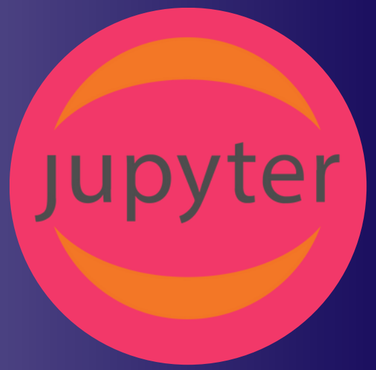
Topic: 2

Tools that I used



PYTHON

Programming Language



JUPYTER NOTEBOOK

a web-based interactive computing platform

STREAMLIT

an open-source Python framework for data scientists and AI/ML engineers to deliver interactive data apps

Main steps of the Presentation

1 Dimensionality Reduction

2 Data Imputation

3 Clustering

4 Classification

5 Over & Undersampling

Choose number of rows to display

3

1 31611

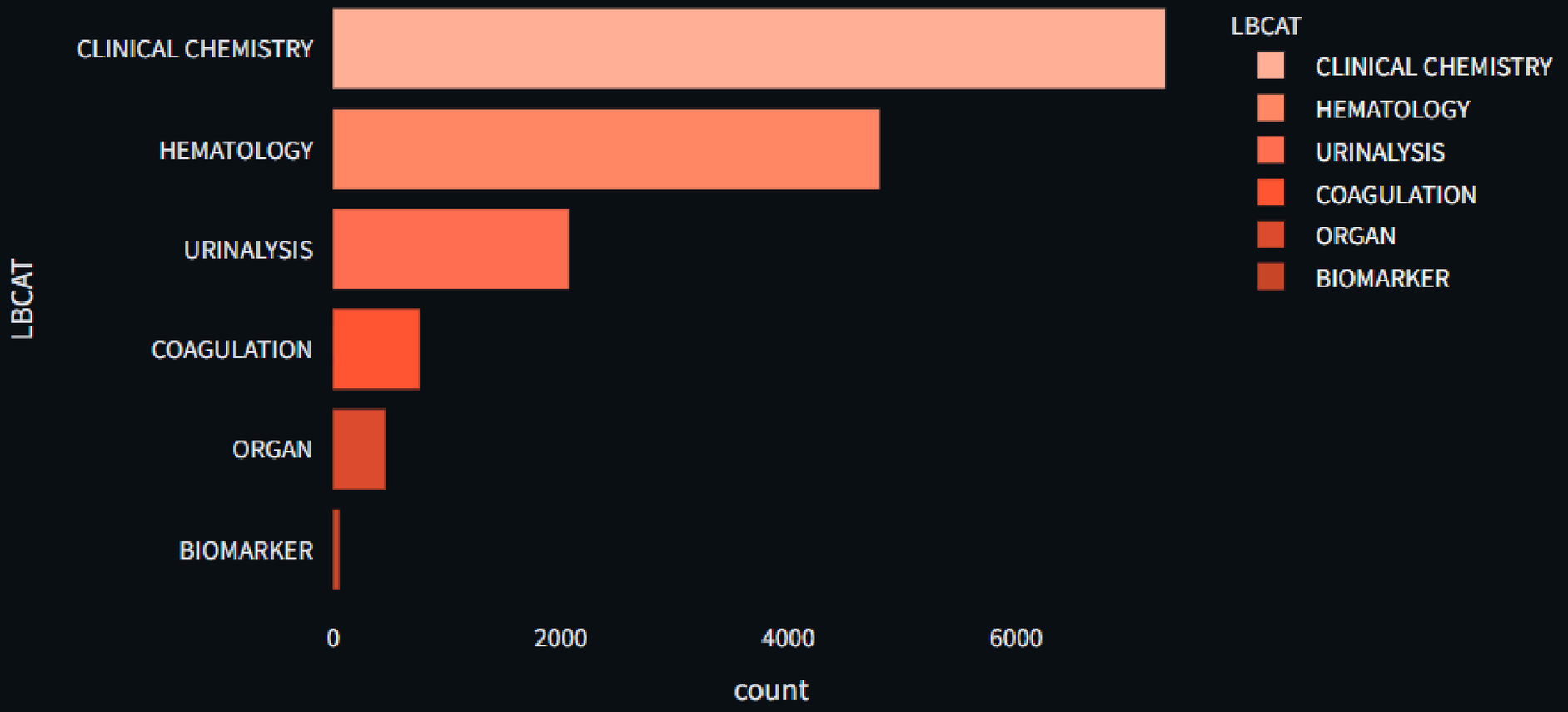
Select columns to show

SPREFID × USUBJID × CLLOC × STUDY_WEEK × CLORRES × CLRESCAT × CLSTAT × CLREASND × CLCAT × CLSEV × CLRFTDTC × CLTPTNUM × CLDY × SEX × trial_set_descrip… ×

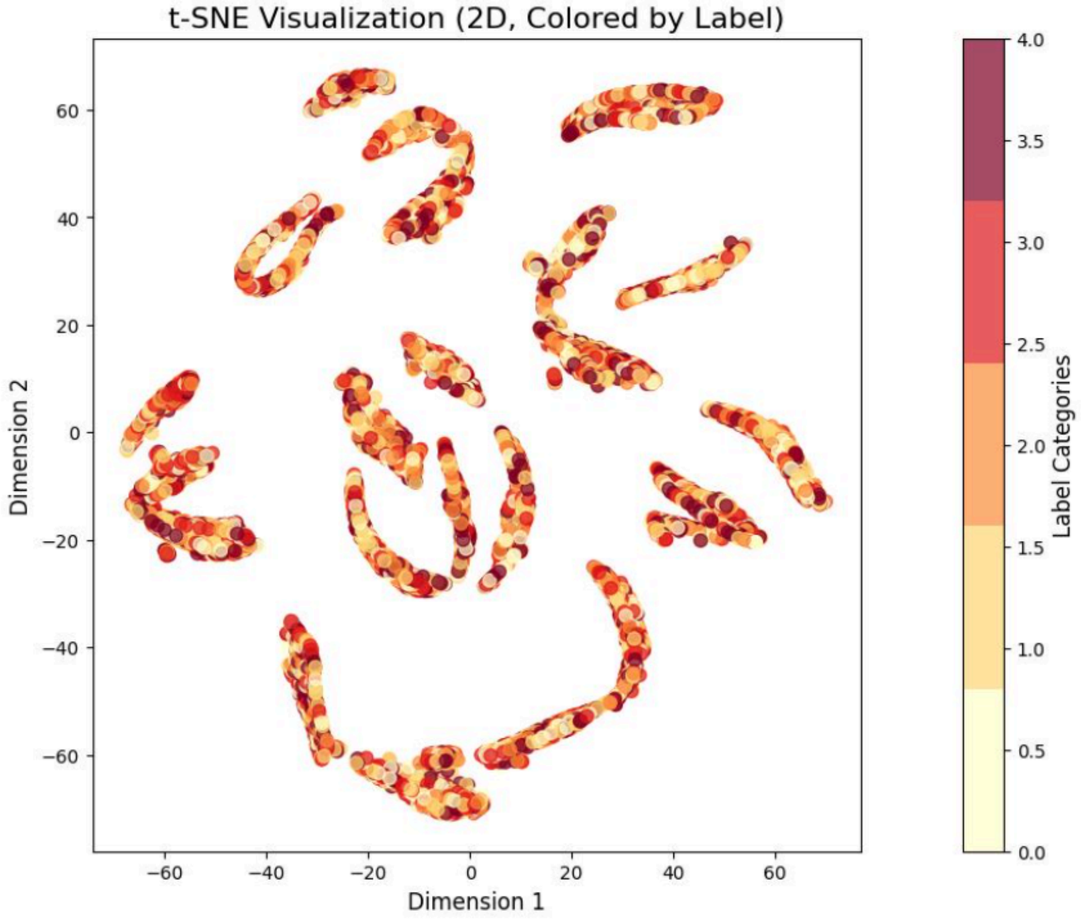
	SPREFID	USUBJID	CLLOC	STUDY_WEEK	CLORRES	CLRESCAT	CLSTAT	CLREASND	CL
0	HCD-12	HCD-12-0083	None	None	NORMAL	CLINICAL SIGNS	None	None	De
1	HCD-12	HCD-12-0083	None	None	NORMAL	CLINICAL SIGNS	None	None	De
2	HCD-12	HCD-12-0083	None	None	NORMAL	CLINICAL SIGNS	None	None	De

	SEX	count
0	M	15,838
1	F	15,773

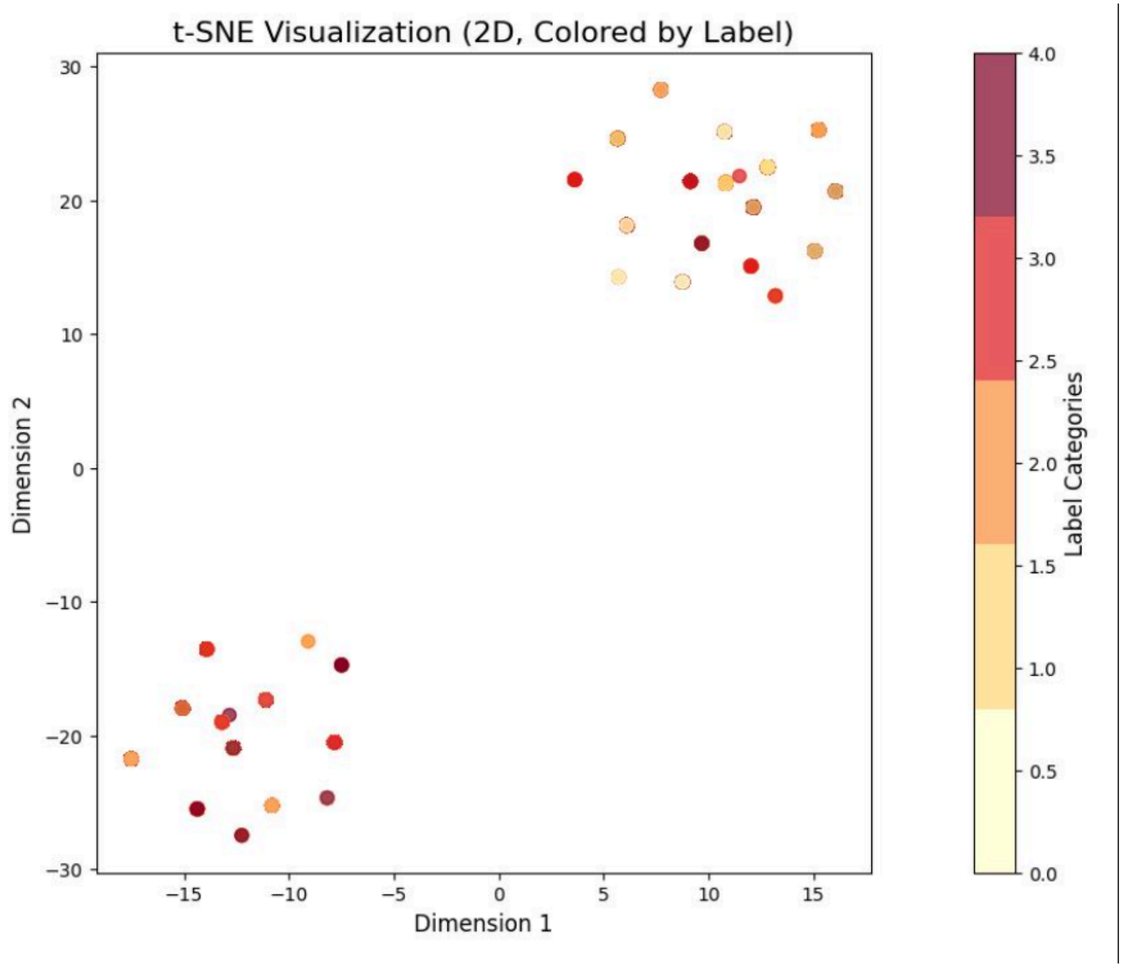
Anzahl der getesteten Parameter



Dimensionality Reduction

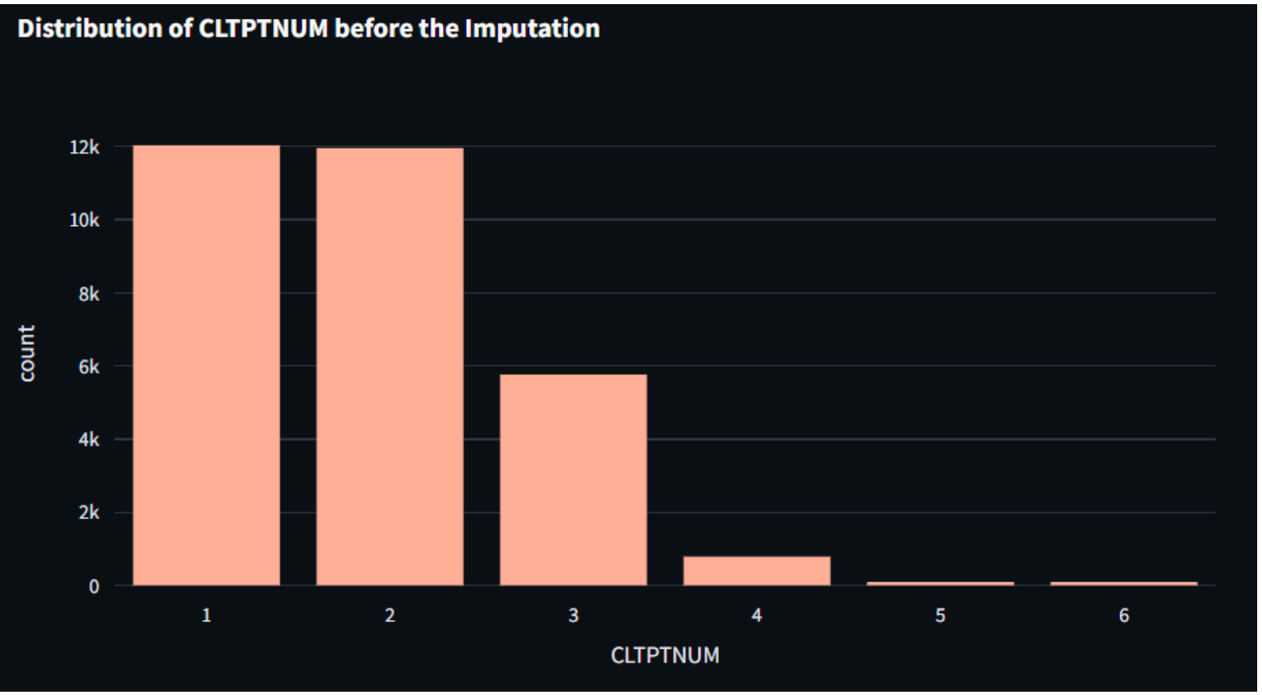


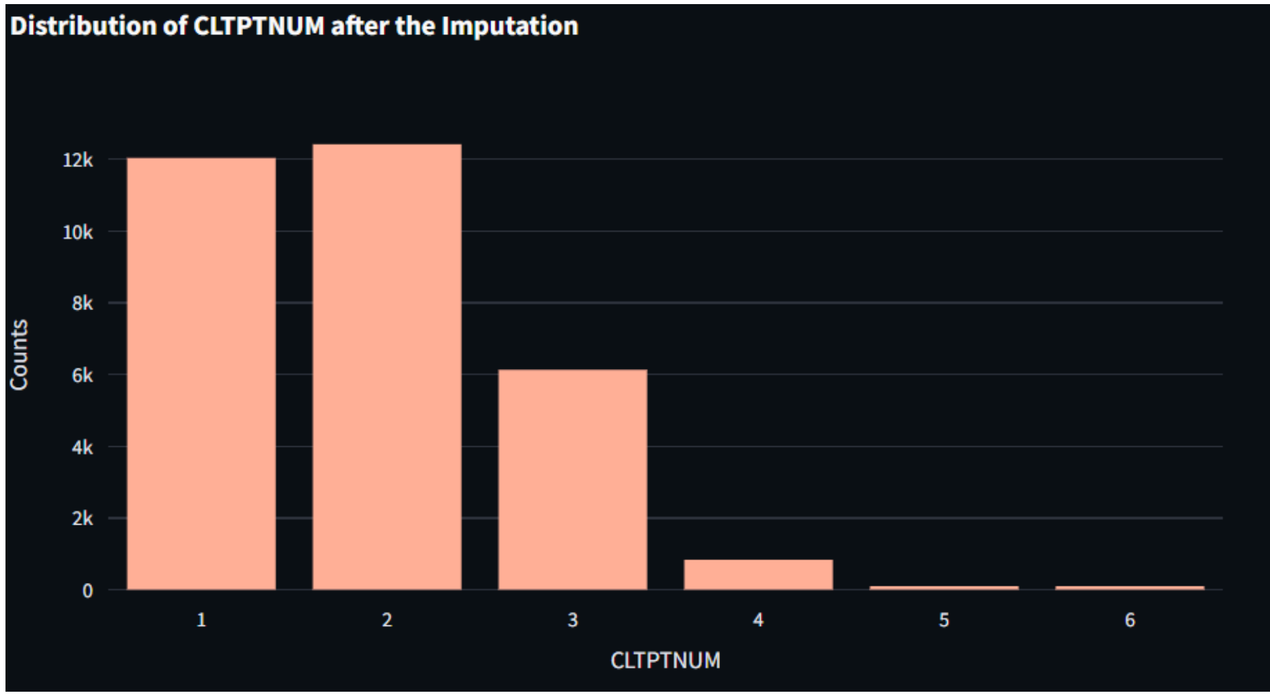
All features



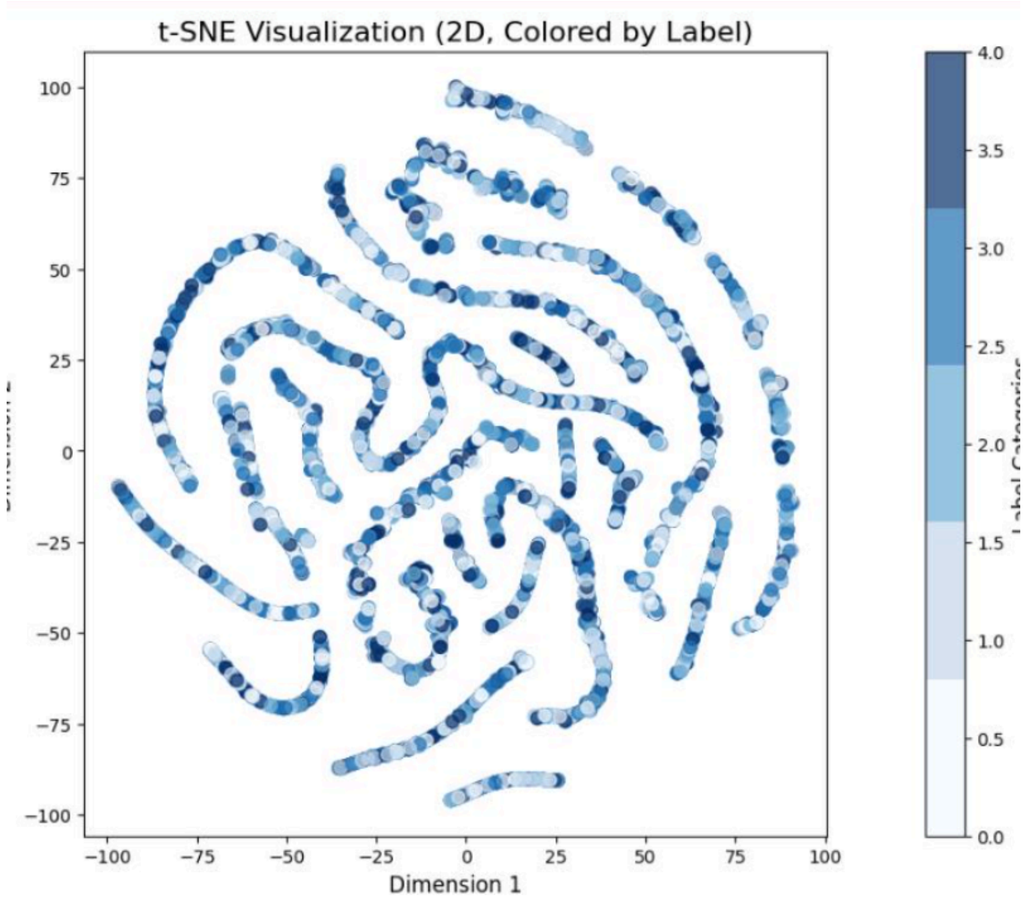
One feature

Distribution of CLTPNUM before and after Imputation

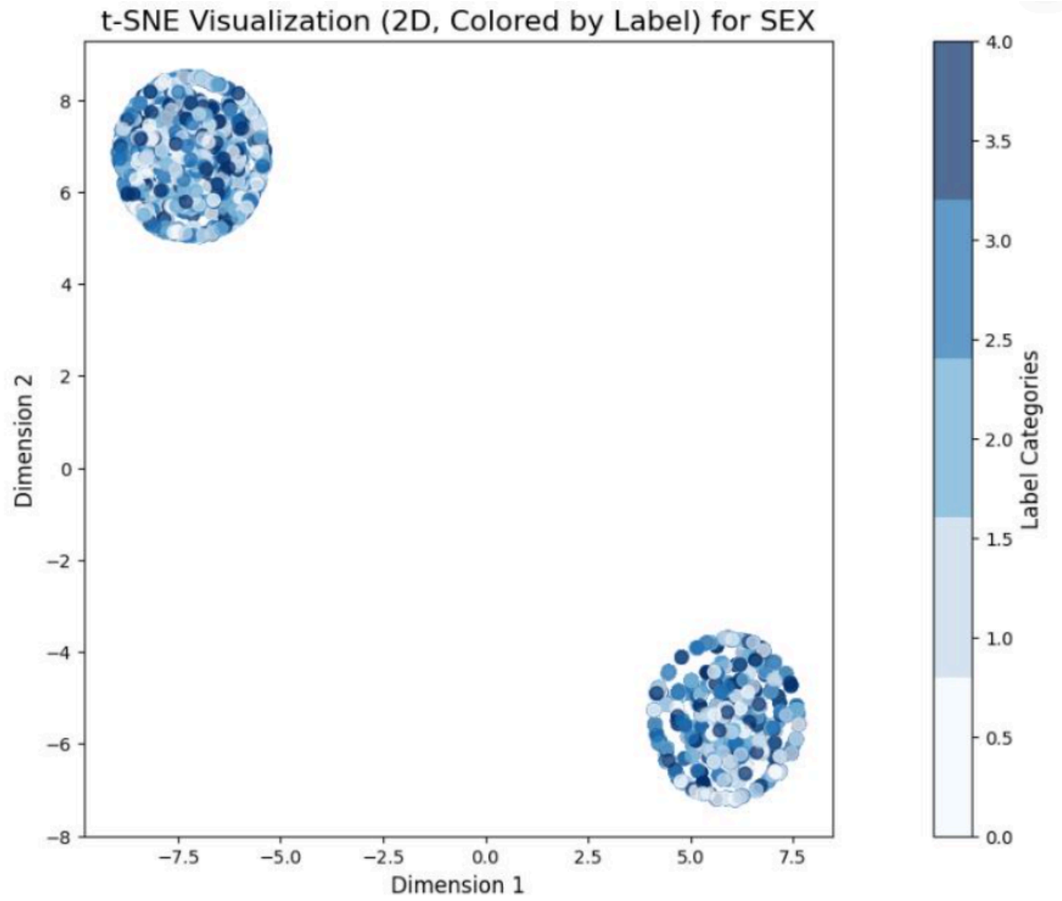




Dimensionality Reduction after imputation



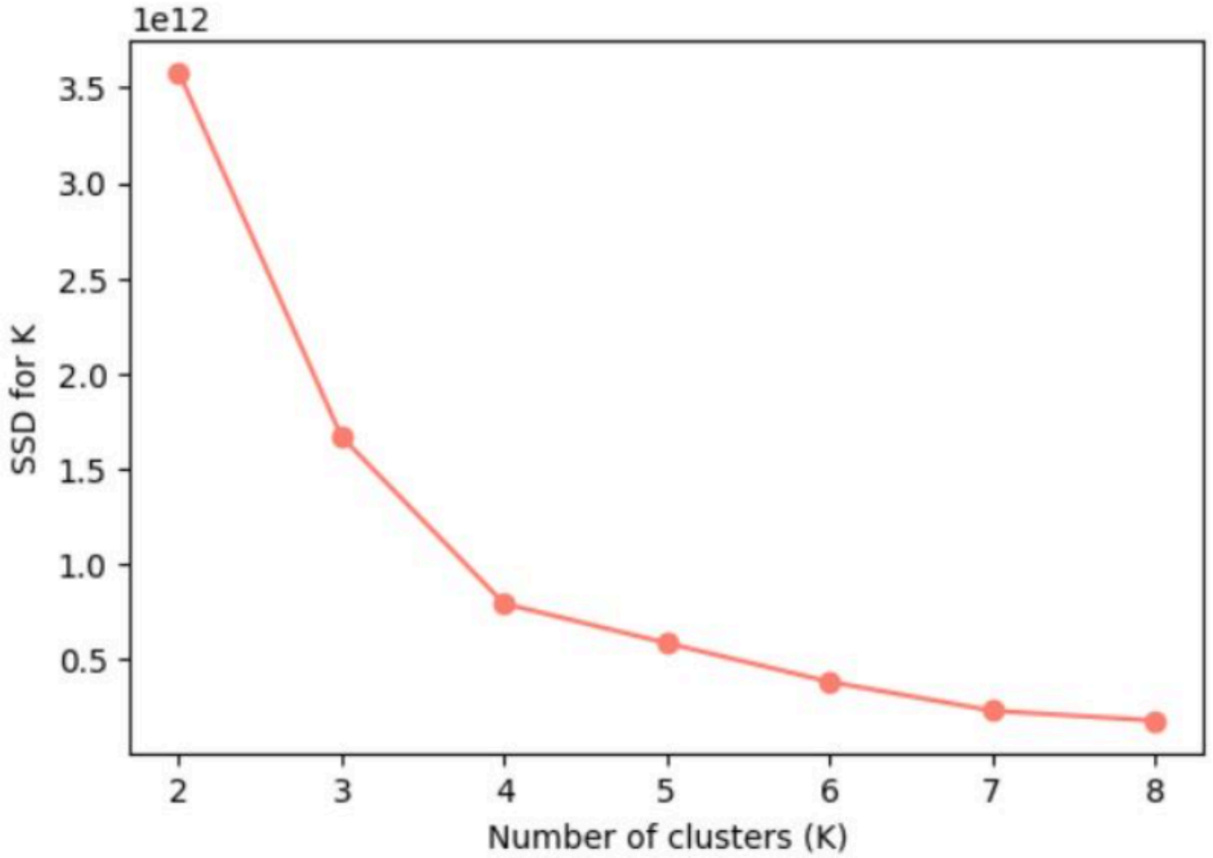
All features



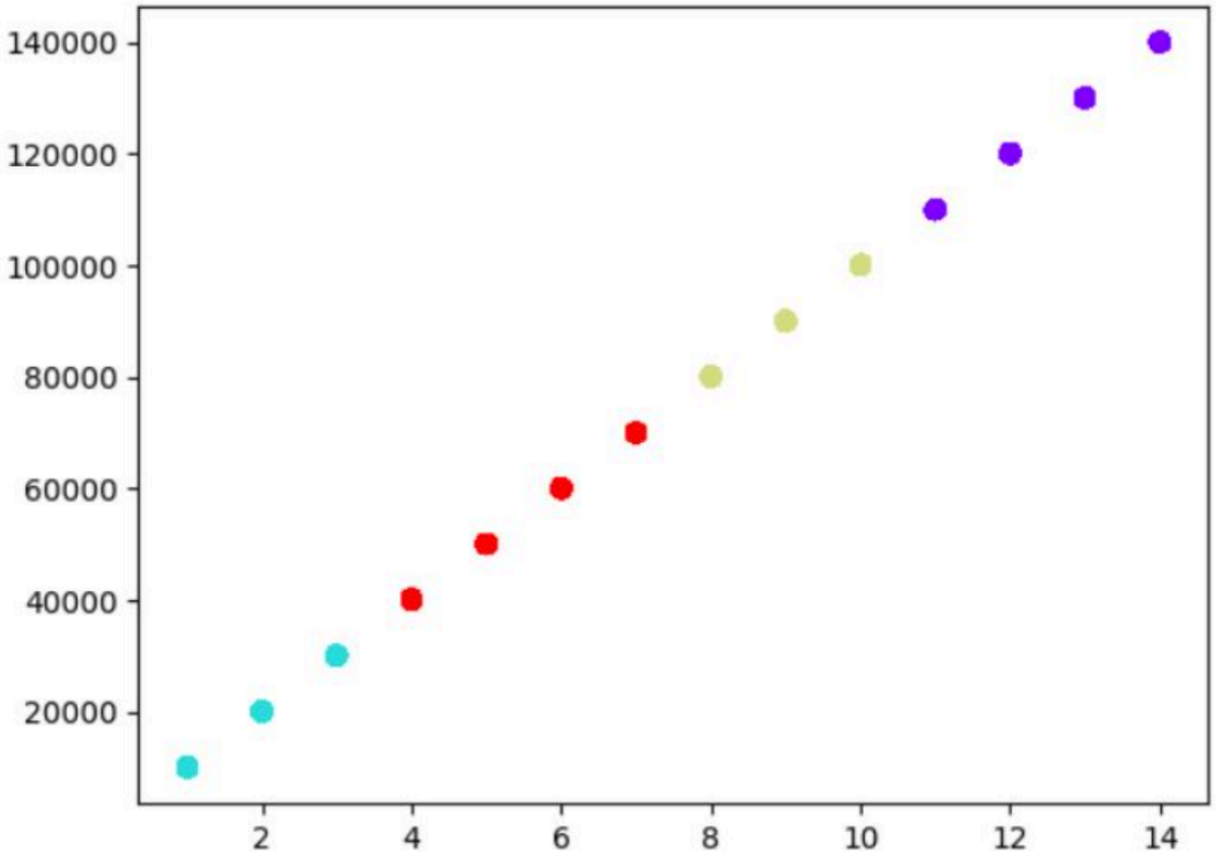
One feature

Elbow Method

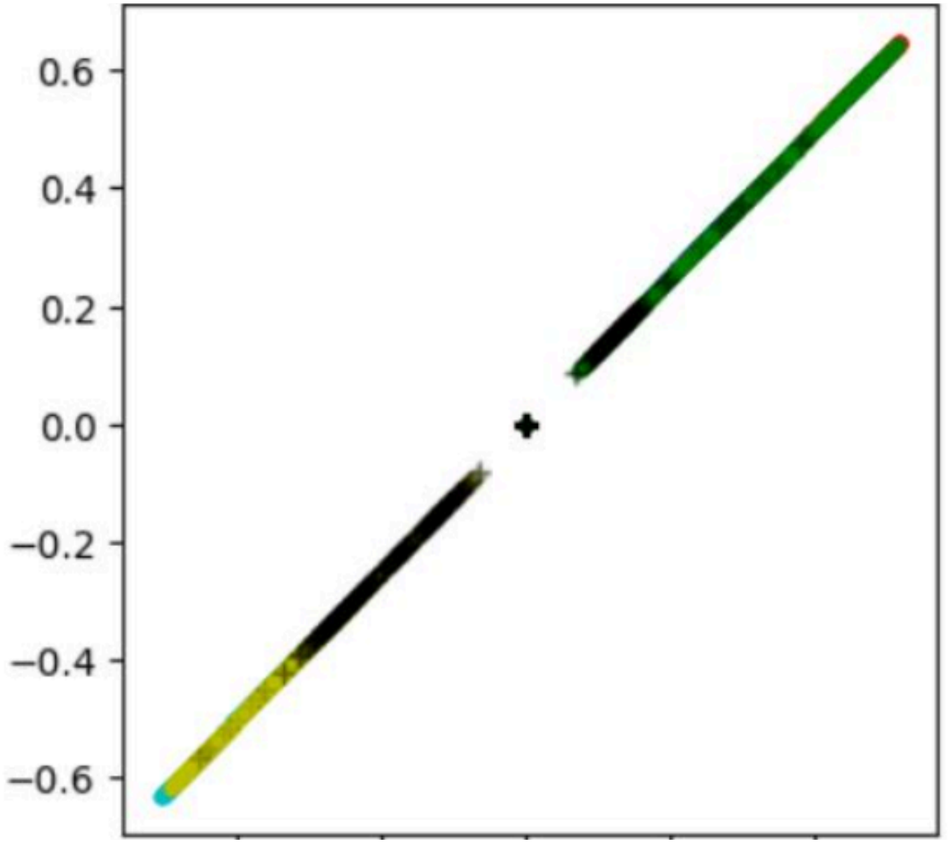
A heuristic used in determining the number of clusters in a data set



Clustering



KMeans



OPTICS

Silhouette Coefficient

A metric used to evaluate the quality of clustering in computer science

01.

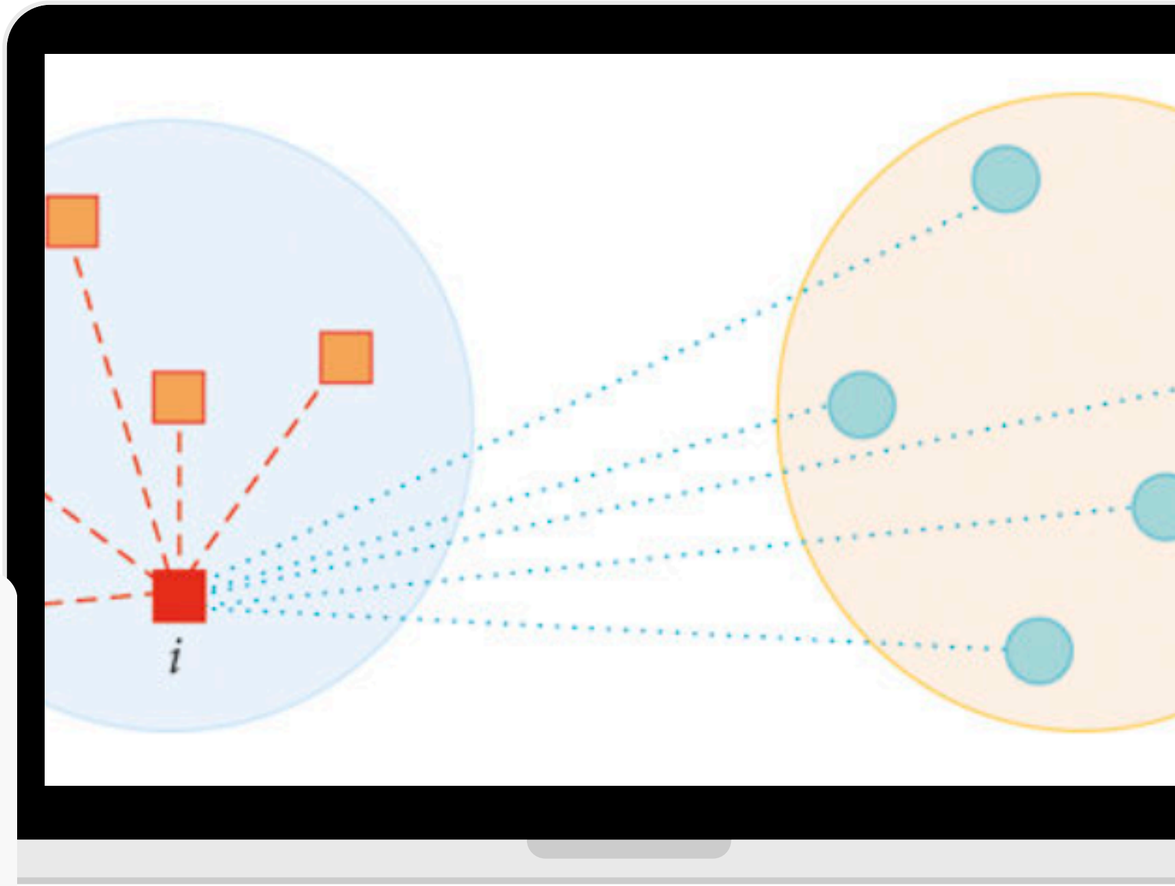
Cohension

Measures the distance of the one point to other points in the same Cluster

02.

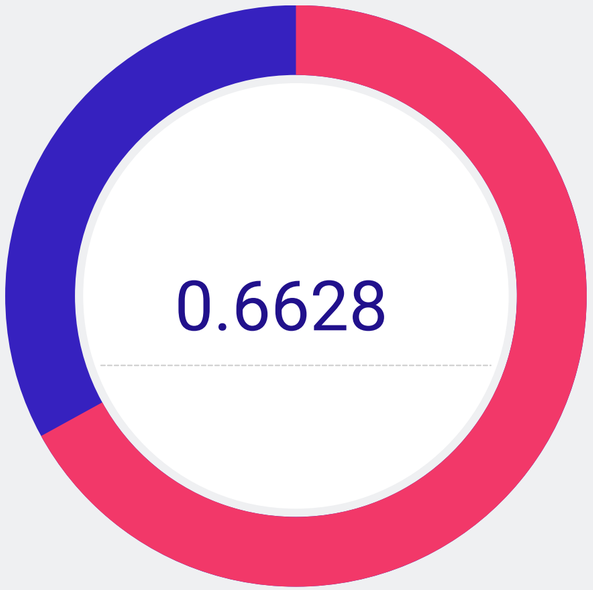
Separation

Measure the distance between one point in one cluster with points in other clusters



https://www.mdpi.com/1099-4300/23/6/759

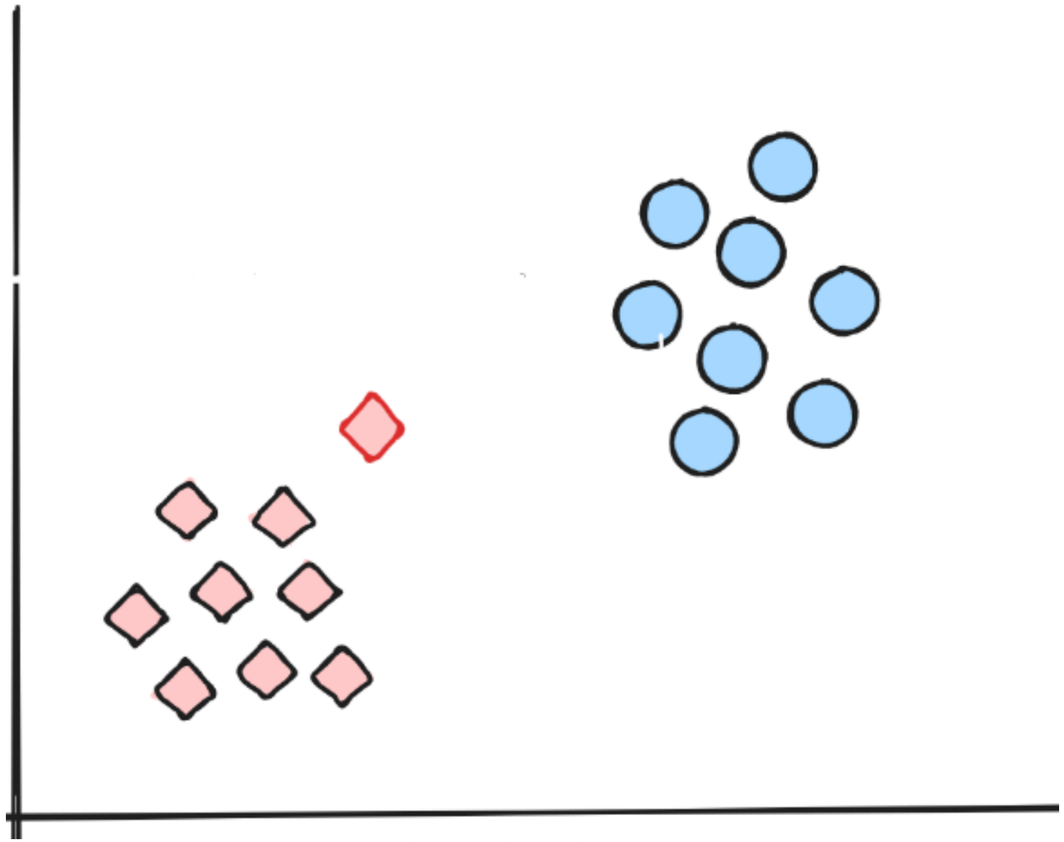
Silhoutte Coefficients



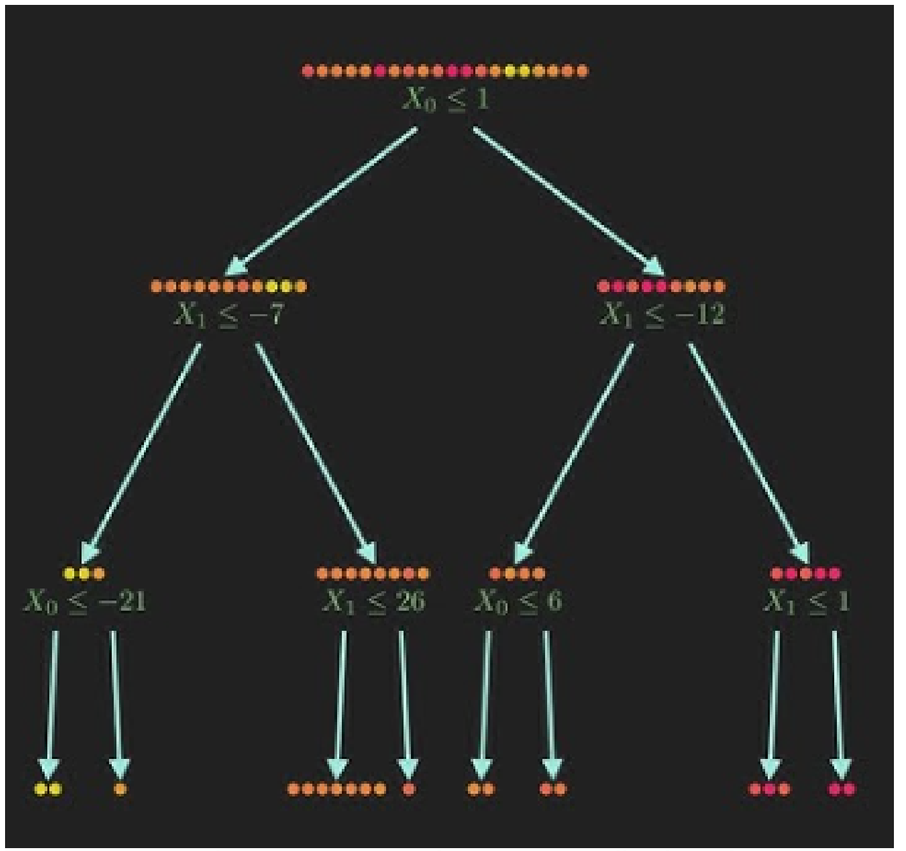
KMeans

OPTICS

Classification



Source: https://thatgirlcoder.com



Source: https://www.youtube.com/Normalized Nerd

Accuracy measure

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

F1 score

Accuracy simply measures how often the classifier correctly predicts

78%

Desicion Tree Classifier

79%

K Nearst Neighbor

Classification and Data Imbalance

Select Target Column

SPREFID

Select Feature Columns

BWORRES × BWORRESU × BWDY × SEX ×

Select Data Imbalance Handling

- (•) None
- ( ) Undersampling
- ( ) Oversampling

Train Model

Classification Report

	precision	recall	f1-score	support
HCD-01	0.1794	0.1942	0.1865	206.000000
HCD-02	0.1057	0.0985	0.1020	132.000000
HCD-03	0.1465	0.1368	0.1415	212.000000
HCD-04	0.2955	0.3186	0.3066	204.000000
HCD-05	0.1275	0.1313	0.1294	198.000000

Select Target Column

SPREFID

Select Feature Columns

BWORRES × BWORRESU × BWDY × SEX ×

Select Data Imbalance Handling

- ( ) None
- (•) Undersampling
- ( ) Oversampling

Train Model

Classification Report

	precision	recall	f1-score	support
HCD-01	0.1866	0.1214	0.1471	206.000000
HCD-02	0.1460	0.1515	0.1487	132.000000
HCD-03	0.1569	0.1132	0.1315	212.000000
HCD-04	0.2733	0.2304	0.2500	204.000000
HCD-05	0.1360	0.0859	0.1053	198.000000

Select Target Column

SPREFID

Select Feature Columns

BWORRES × BWORRESU × BWDY × SEX ×

Select Data Imbalance Handling

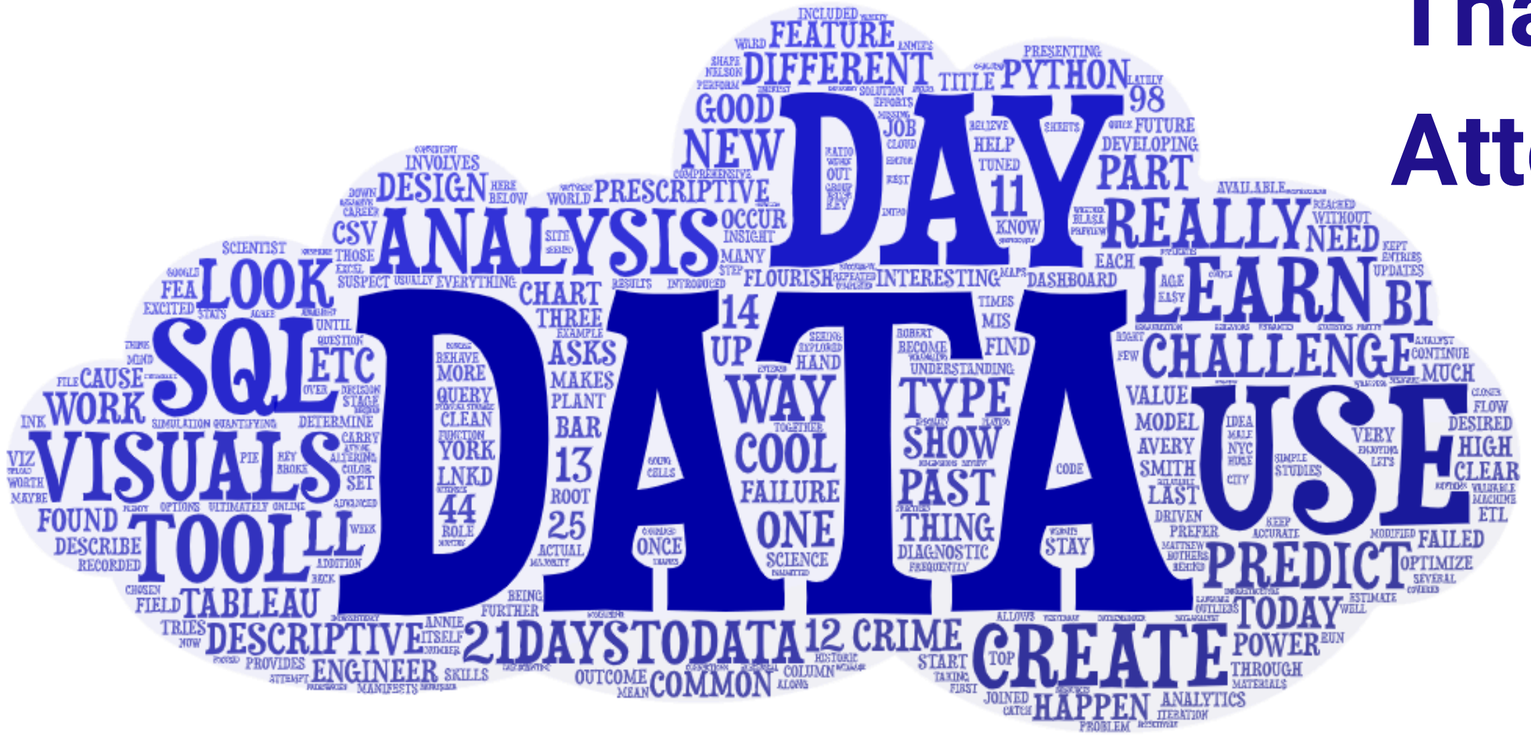
- ( ) None
- ( ) Undersampling
- (•) Oversampling

Train Model

Classification Report

	precision	recall	f1-score	support
HCD-01	0.1765	0.1602	0.1679	206.000000
HCD-02	0.1319	0.1439	0.1377	132.000000
HCD-03	0.1627	0.1274	0.1429	212.000000
HCD-04	0.2984	0.2794	0.2886	204.000000
HCD-05	0.1294	0.1111	0.1196	198.000000

Thank you for your Attention



Source: linkedin/Brandon Hopkins