stroke prediction

Ayala Bouhnik-Gelbord 206654873
Maayan Sulimani 313563009

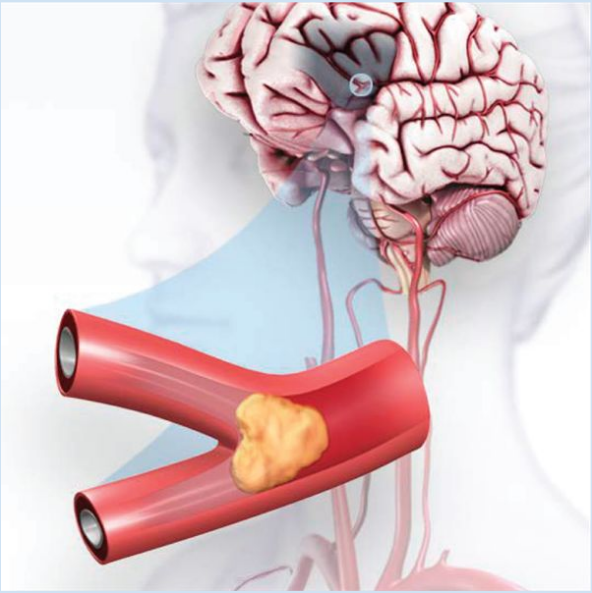
- This project has been done as part of `Machine Learning' course, led by Prof. Lee-Ad Gottleib

What is stroke?

Stroke is a **disease that affects the arteries leading to and within the brain**. It is the number 5 cause of death and a leading cause of disability in the United States. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or bursts (or ruptures).

What are the effects of stroke?

The brain is an extremely complex organ that controls various body functions. If a stroke occurs and blood flow can't reach the region that controls a particular body function, that part of the body won't work as it should.



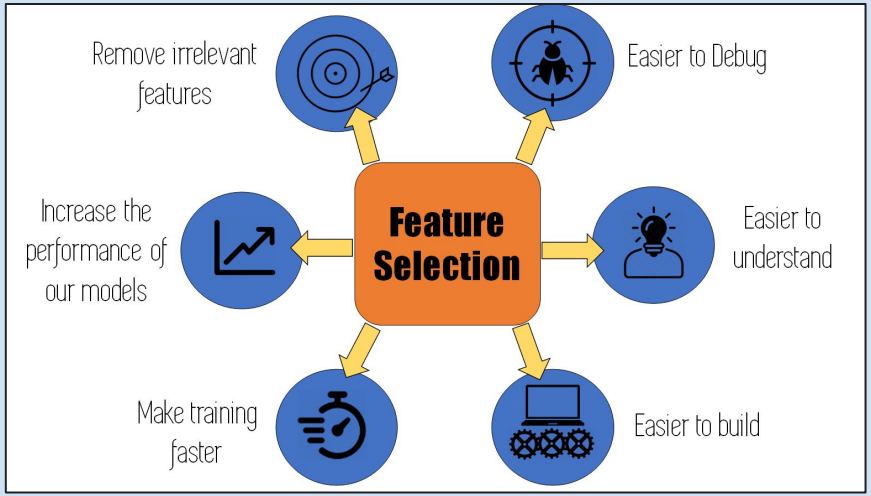
project target-

In order to try to reduce the death from stroke, we will try to predict which people are more likely to have stroke.

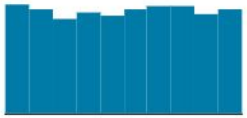
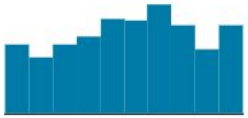
For this project we used Kaggle dataset that includes 11 clinical features for predicting stroke events.

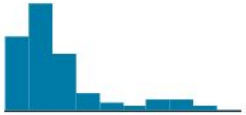
The features are-

1. Gender
2. Age
3. Hypertension binary feature
4. Heart disease binary feature
5. Has the patient ever been married?
6. Work type of the patient
7. Residence type of the patient
8. Average glucose level in blood
9. Body Mass Index
10. Smoking status of the patient
11. Stroke event



A little bit about our data..

id	gender	age	hypertension	heart_disease
Unique id	Gender	Age	Hypertension binary feature	Heart disease binary feature
 67 … 72.9k	Female 59% Male 41% Other (1) 0%	 0.08 … 82	 0 … 1	 0 … 1

ever_married	work_type	Residence_type	avg_glucose_level	bmi
Has the patient ever been married?	Work type of the patient	Residence type of the patient	Average glucose level in blood	Body Mass Index
 true 3353 66% false 1757 34%	Private 57% Self-employed 16% Other (1366) 27%	Urban 51% Rural 49%	 55.1 … 272	N/A 4% 28.7 1% Other (4868) 95%

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv("/content/healthcare-dataset-stroke-data.csv")
data.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Lets Explore Our Data-

```
<bound method DataFrame.info of          id  gender   age  ...   bmi   smoking_status stroke
0      9046    Male  67.0  ...  36.6  formerly smoked      1
1     51676  Female  61.0  ...   NaN     never smoked      1
2     31112    Male  80.0  ...  32.5     never smoked      1
3     60182  Female  49.0  ...  34.4           smokes      1
4      1665  Female  79.0  ...  24.0     never smoked      1
...     ...     ...   ...  ...   ...              ...    ...
5105  18234  Female  80.0  ...   NaN     never smoked      0
5106  44873  Female  81.0  ...  40.0     never smoked      0
5107  19723  Female  35.0  ...  30.6     never smoked      0
5108  37544    Male  51.0  ...  25.6  formerly smoked      0
5109  44679  Female  44.0  ...  26.2          Unknown      0

[5110 rows x 12 columns]>
```

```
[34] data.isnull().sum()
```

```
id                     0
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  201
smoking_status         0
stroke                 0
dtype: int64
```

As we can see 'bmi' column holds some missing values.

So we need to fill the null values:

```
data['bmi'].fillna(data['bmi'].mean(), inplace = True)
```

```
data.isnull().sum()
```

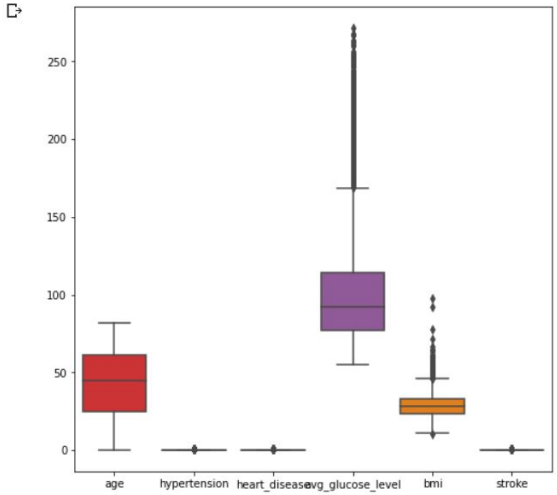
```
gender                 0
age                    0
hypertension           0
heart_disease          0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                    0
smoking_status         0
stroke                 0
dtype: int64
```

After we explorer the data, we decided to remove two columns- ‘id’ and ‘ever married’.

So now we have 10 attributes.

We will use subplot function to find the outliers in our data.

```
fig, ax = plt.subplots(figsize = (8, 8))
sns.boxplot(data = data, palette = 'Set1')
plt.show();
```



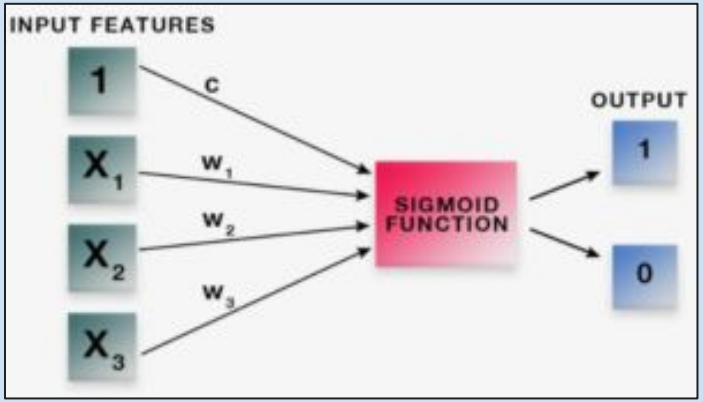
Algorithms-

We would want to know, based on the 11 features, what are the chances to have a stroke.

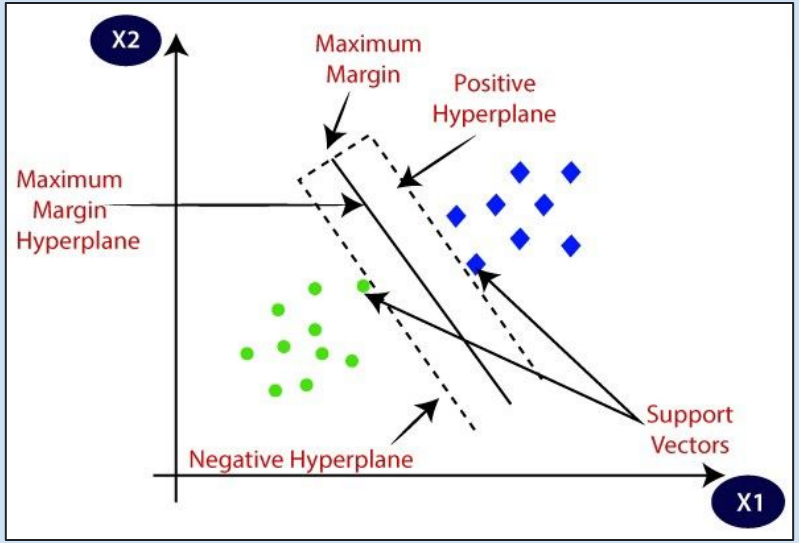
<u>We will use the following technics:</u>

- Logistic Regression
- SVM
- Decision Trees
- K- nearest neighbour

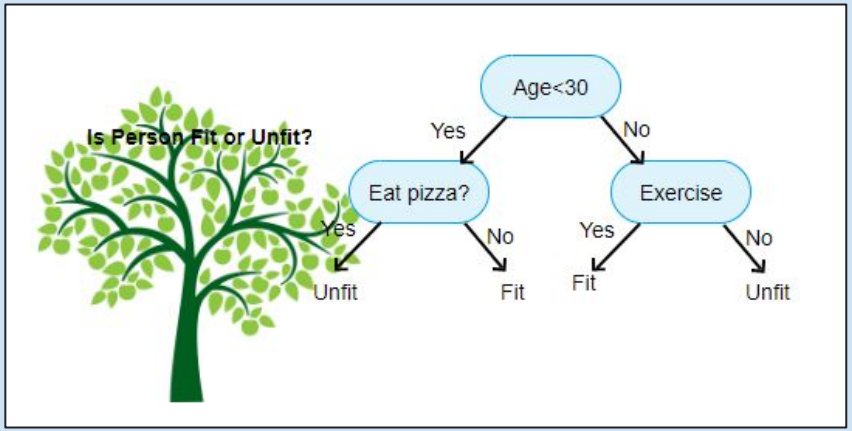
Logistic Regression-



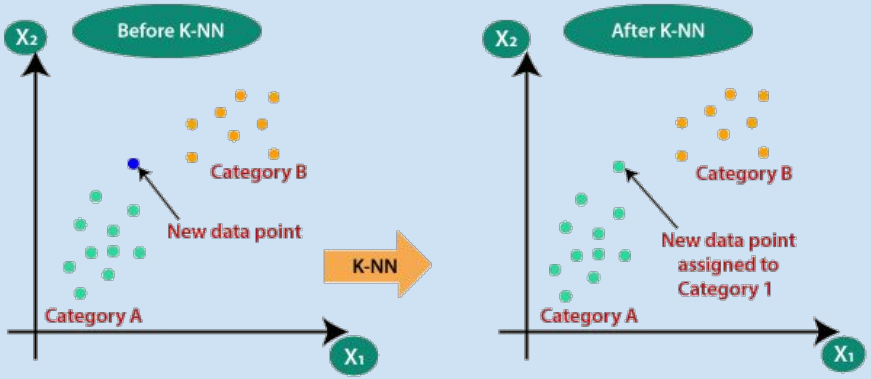
SVM-


X2
Maximum
Margin
Positive
Hyperplane
Maximum
Margin
Hyperplane
Support
Vectors
Negative Hyperplane
X1

Decision tree-



K- Nearest Neighbour-



results-

Algorithms-	Logistic regression	SVM	Decision Tree	K- Nearest Neighbour
Accuracy-	**94.8914431673052 4 %**	**94.8914431673052 4 %**	92.33716475095785 %	94.6360153256705 %

מקורות וחומרי עזר-

- https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
- https://github.com/riddhi-jain