Stroke prediction (2)

1. Review
2. Decision Tree
3. ML process
4. Results

Stroke prediction (2)

1. Review

2. Decision Tree

3. ML process

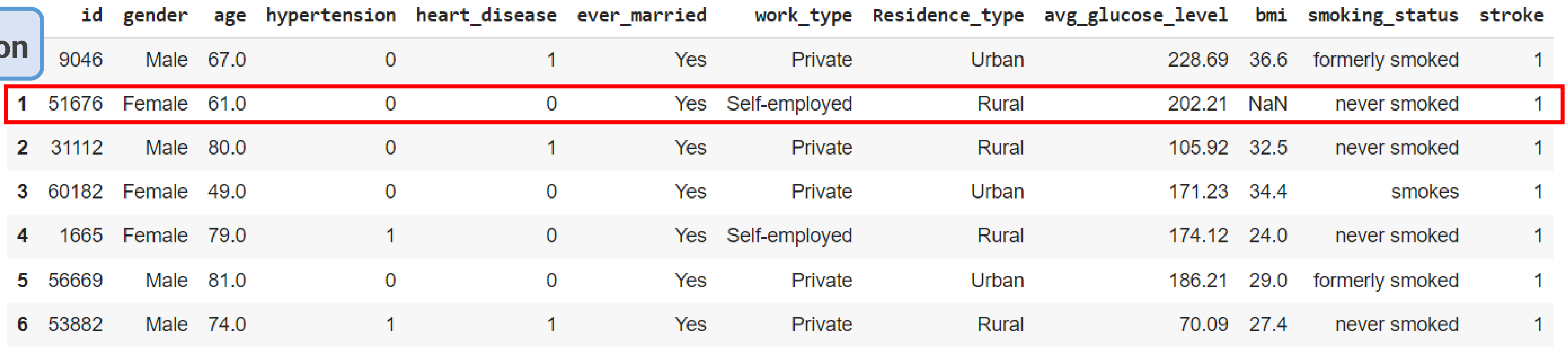
4. Results

Stroke prediction (2)

1. Review

Data

Person

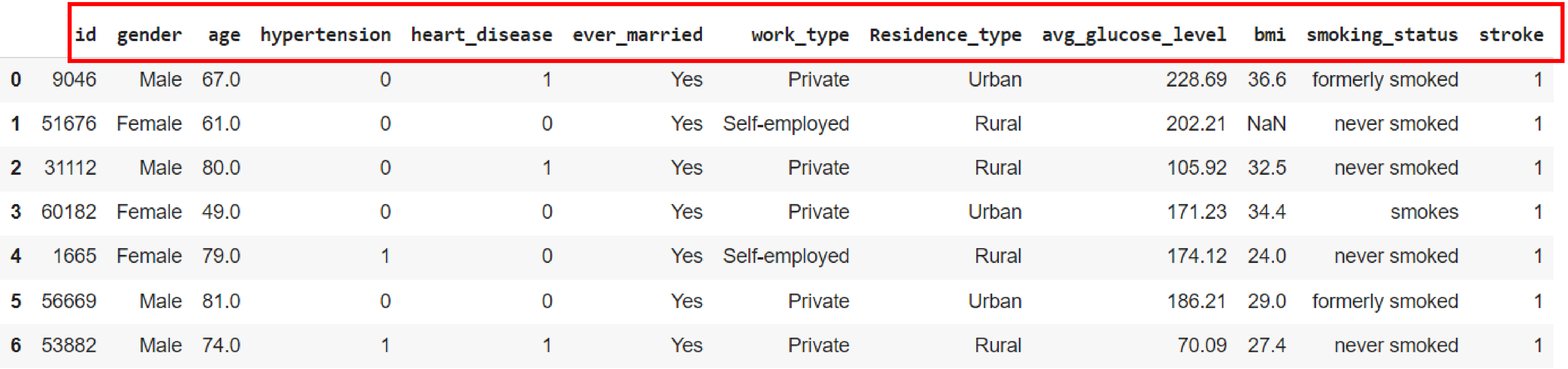


	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	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
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1

1. Review

Data

Features Or Attributes

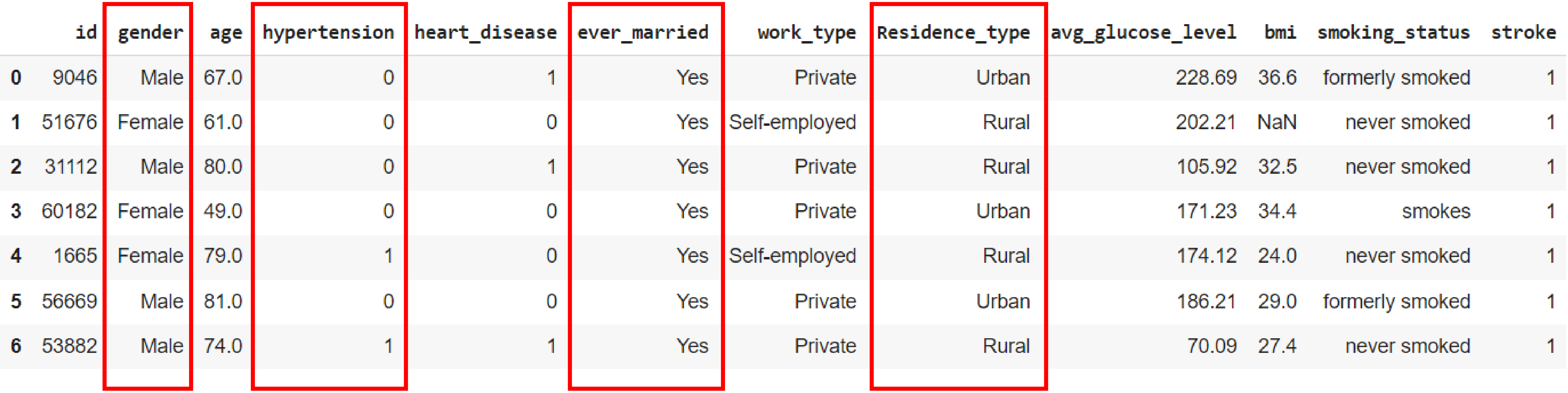


	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	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
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1

1. Review

Data

Features Or Attributes

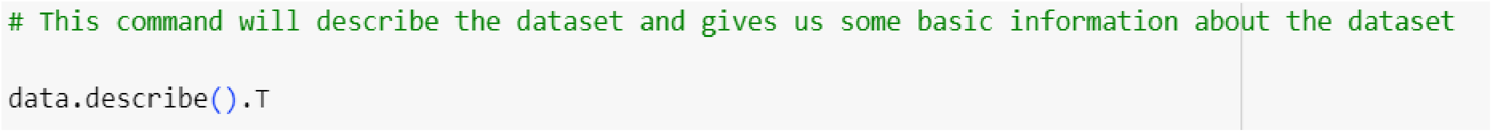


	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	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
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1

1. Review

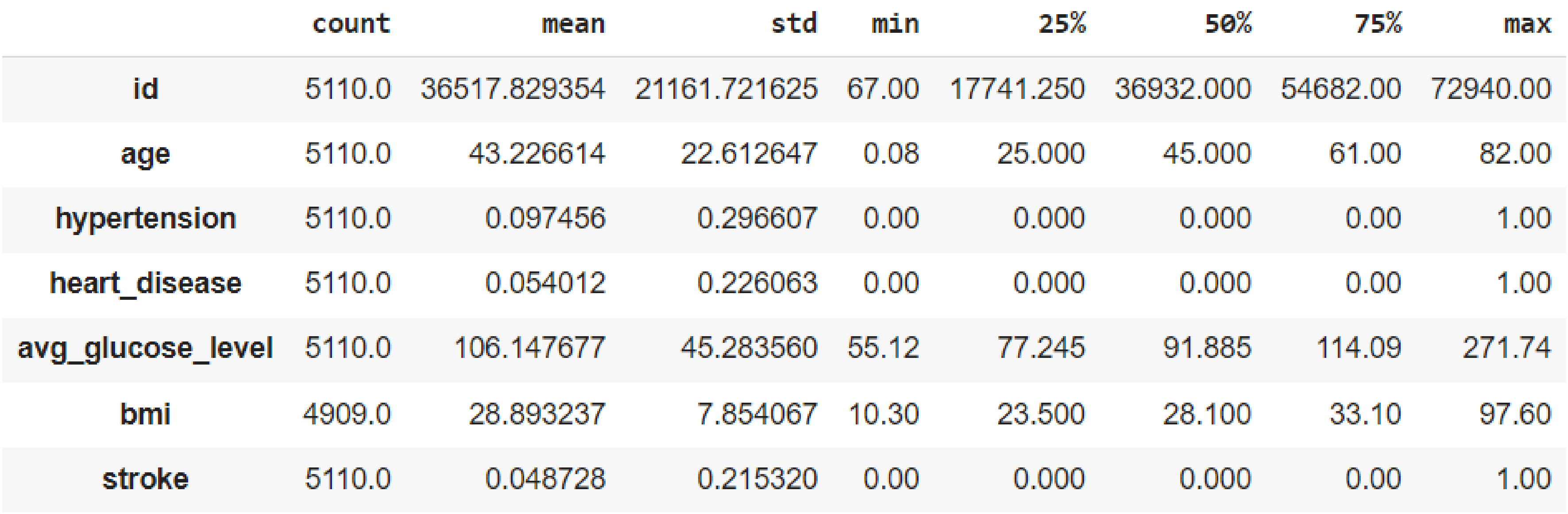
Data

Analyzing



```
# This command will describe the dataset and gives us some basic information about the dataset

data.describe().T
```

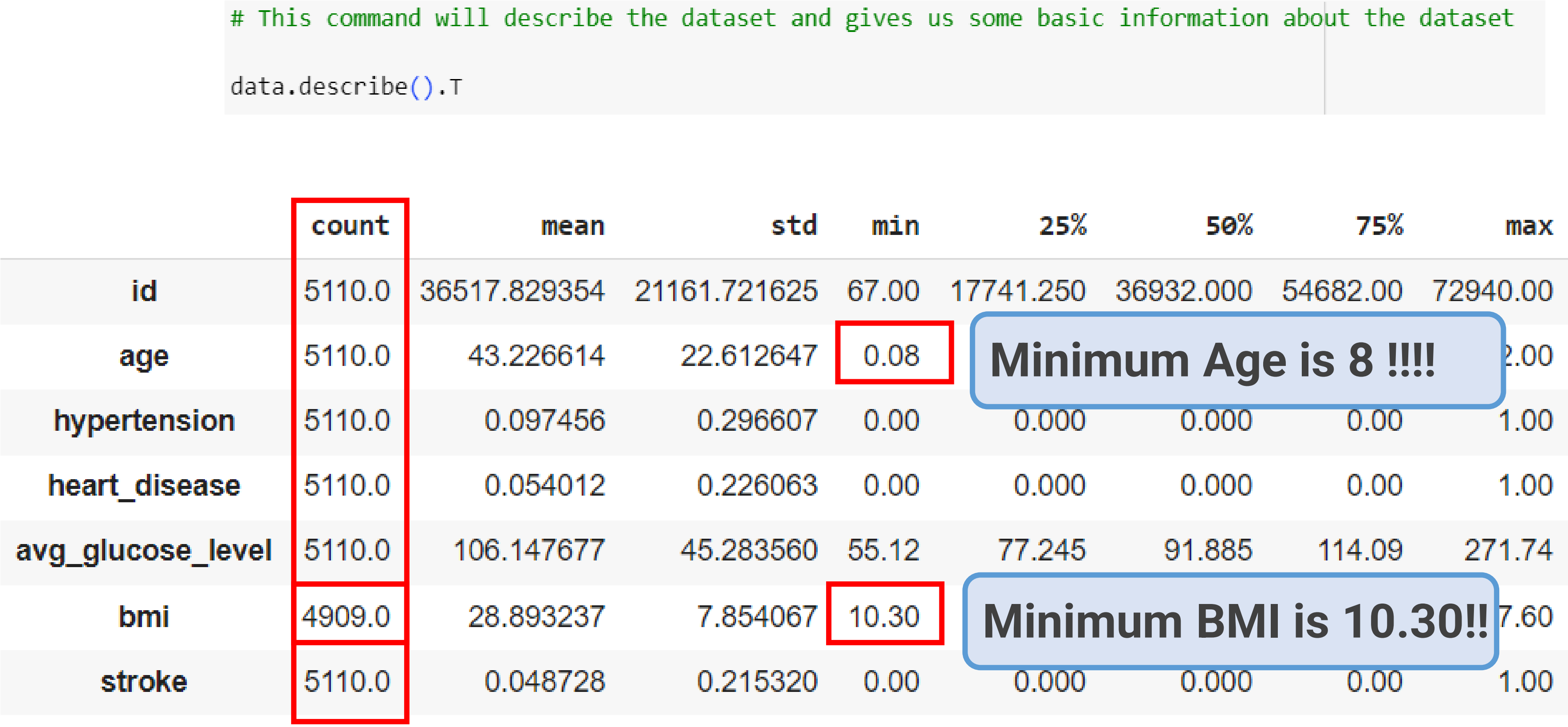


	count	mean	std	min	25%	50%	75%	max
id	5110.0	36517.829354	21161.721625	67.00	17741.250	36932.000	54682.00	72940.00
age	5110.0	43.226614	22.612647	0.08	25.000	45.000	61.00	82.00
hypertension	5110.0	0.097456	0.296607	0.00	0.000	0.000	0.00	1.00
heart_disease	5110.0	0.054012	0.226063	0.00	0.000	0.000	0.00	1.00
avg_glucose_level	5110.0	106.147677	45.283560	55.12	77.245	91.885	114.09	271.74
bmi	4909.0	28.893237	7.854067	10.30	23.500	28.100	33.10	97.60
stroke	5110.0	0.048728	0.215320	0.00	0.000	0.000	0.00	1.00

1. Review

Data

Analyzing



```
# This command will describe the dataset and gives us some basic information about the dataset

data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
id	5110.0	36517.829354	21161.721625	67.00	17741.250	36932.000	54682.00	72940.00
age	5110.0	43.226614	22.612647	0.08	[illegible]	[illegible]	[illegible]	2.00
hypertension	5110.0	0.097456	0.296607	0.00	0.000	0.000	0.00	1.00
heart_disease	5110.0	0.054012	0.226063	0.00	0.000	0.000	0.00	1.00
avg_glucose_level	5110.0	106.147677	45.283560	55.12	77.245	91.885	114.09	271.74
bmi	4909.0	28.893237	7.854067	10.30	[illegible]	[illegible]	[illegible]	7.60
stroke	5110.0	0.048728	0.215320	0.00	0.000	0.000	0.00	1.00

Minimum Age is 8 !!!!

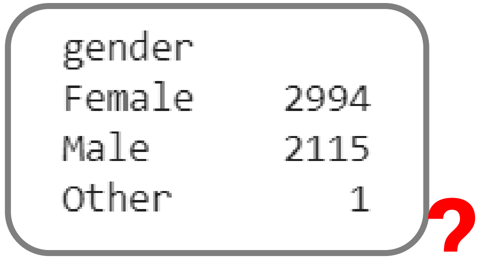
Minimum BMI is 10.30!!

We have missing data!

1. Review

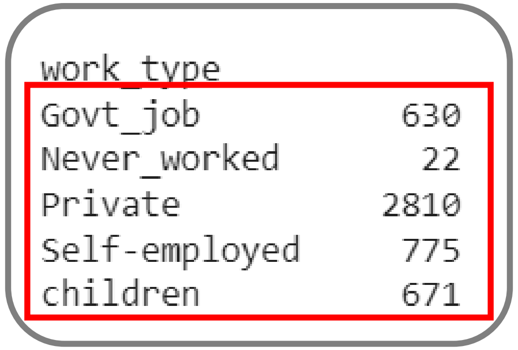
Data

Analyzing

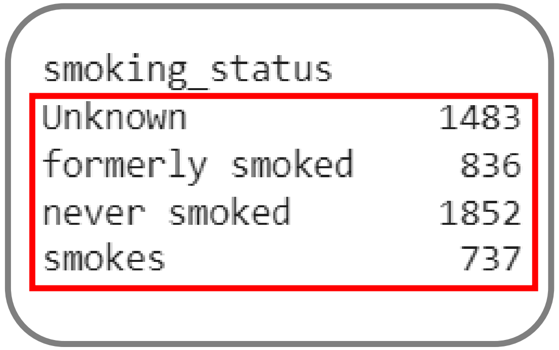


```
gender
Female    2994
Male      2115
Other        1
```

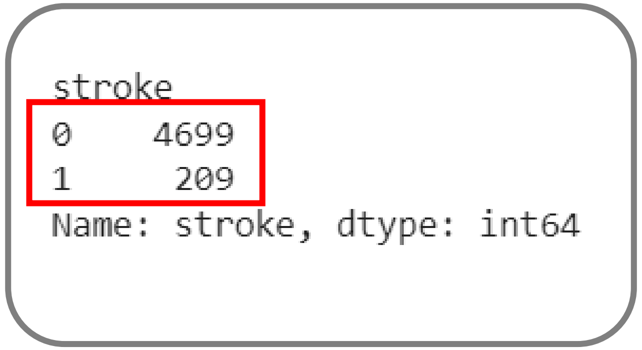
???



```
work_type
Govt_job          630
Never_worked       22
Private          2810
Self-employed     775
children          671
```



```
smoking_status
Unknown            1483
formerly smoked     836
never smoked       1852
smokes              737
```



```
stroke
0    4699
1     209
Name: stroke, dtype: int64
```

Are they truly useful?

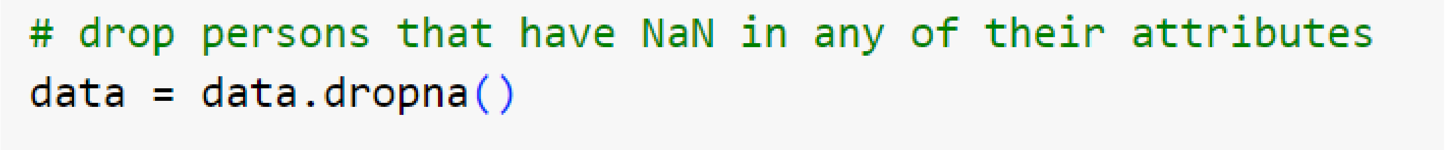
1. Review

Data

Analyzing

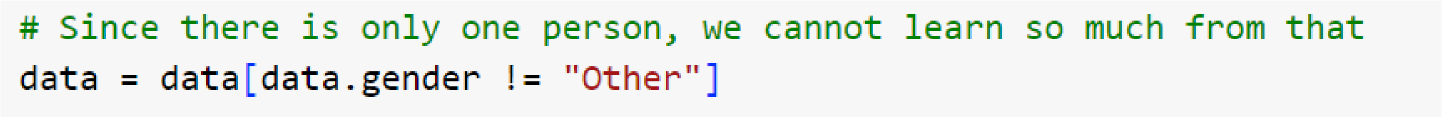
Preprocessing

- Ignore people that we don’t have all the information about them



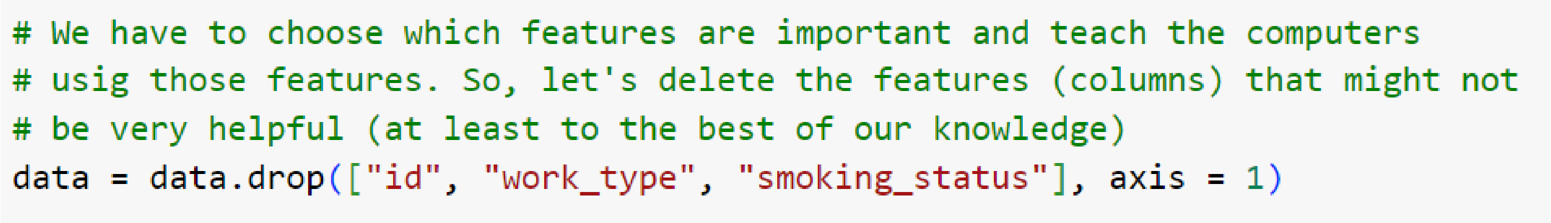
```
# drop persons that have NaN in any of their attributes
data = data.dropna()
```

- Delete the one person with “Other” gender



```
# Since there is only one person, we cannot learn so much from that
data = data[data.gender != "Other"]
```

- Deleting the columns (features) that are not helpful



```
# We have to choose which features are important and teach the computers
# usig those features. So, let's delete the features (columns) that might not
# be very helpful (at least to the best of our knowledge)
data = data.drop(["id", "work_type", "smoking_status"], axis = 1)
```

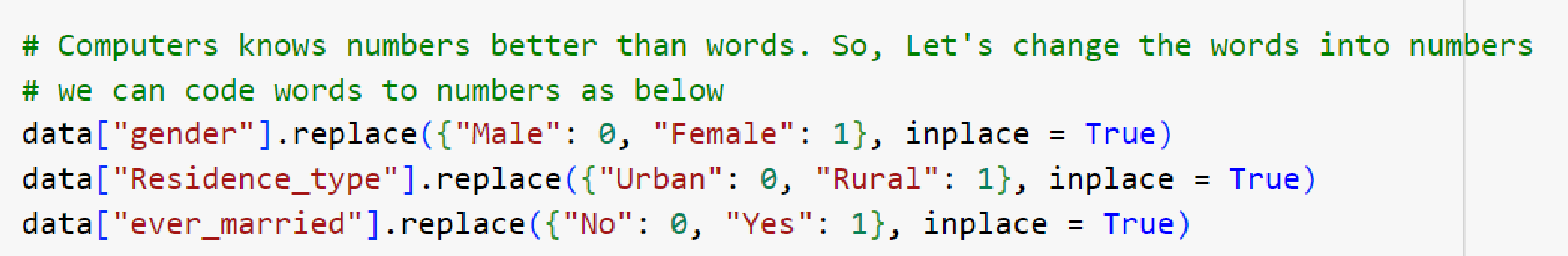
1. Review

Data

Analyzing

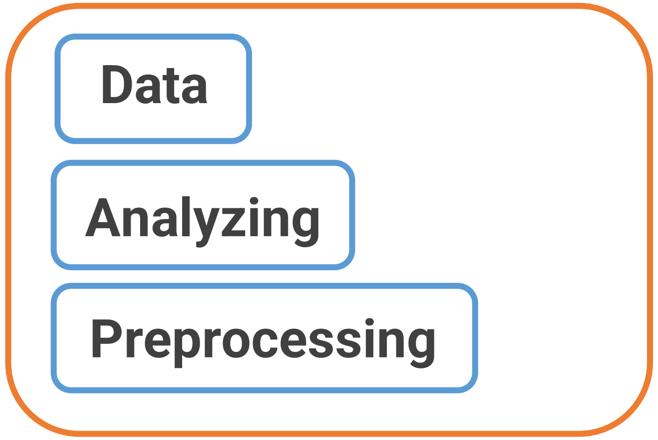
Preprocessing

- Computers understand numbers better than words, So let’s use numbers instead of words!



```
# Computers knows numbers better than words. So, Let's change the words into numbers
# we can code words to numbers as below
data["gender"].replace({"Male": 0, "Female": 1}, inplace = True)
data["Residence_type"].replace({"Urban": 0, "Rural": 1}, inplace = True)
data["ever_married"].replace({"No": 0, "Yes": 1}, inplace = True)
```

1. Review



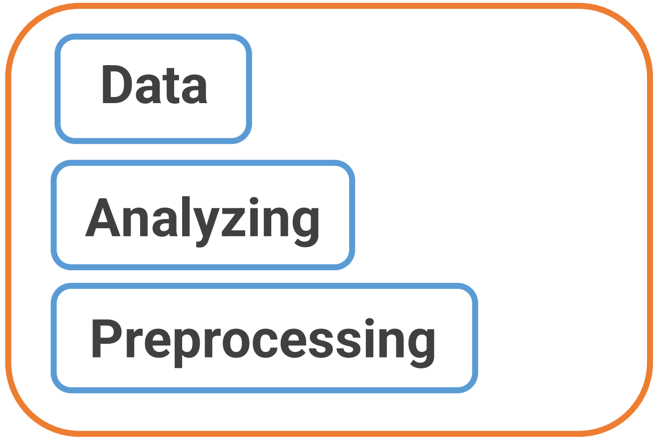
Data

Analyzing

Preprocessing

Data Science

1. Review

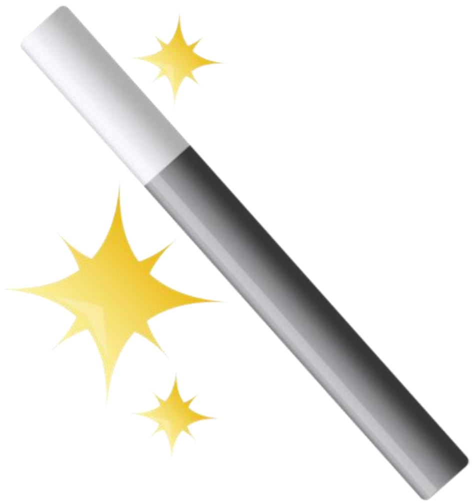


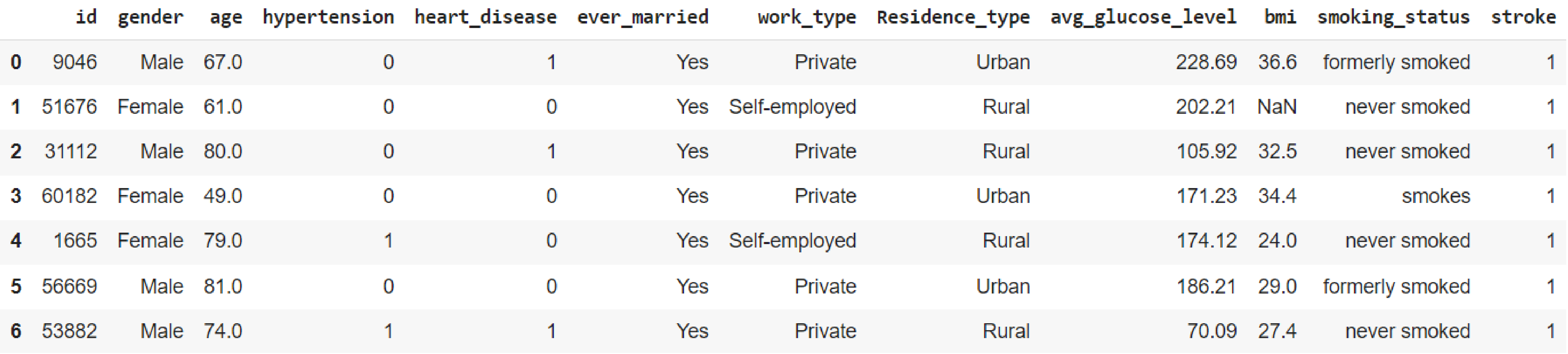
Data

Analyzing

Preprocessing

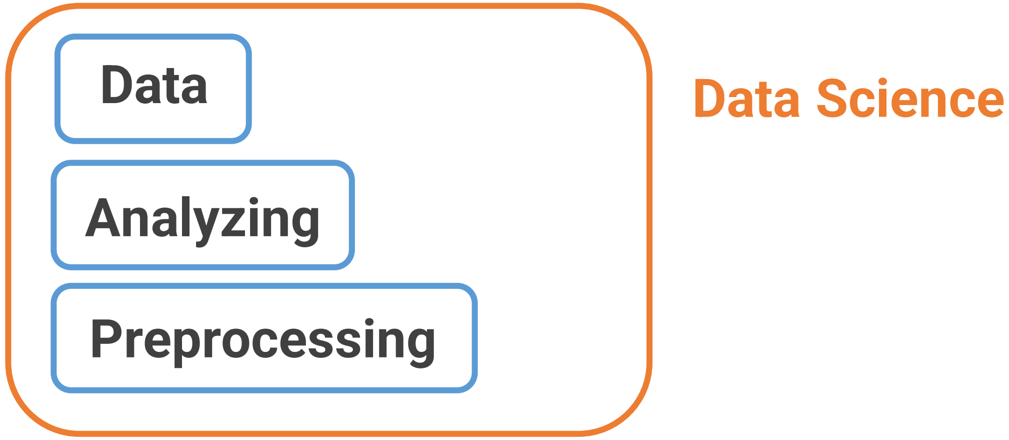
Data Science





	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
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1

1. Review

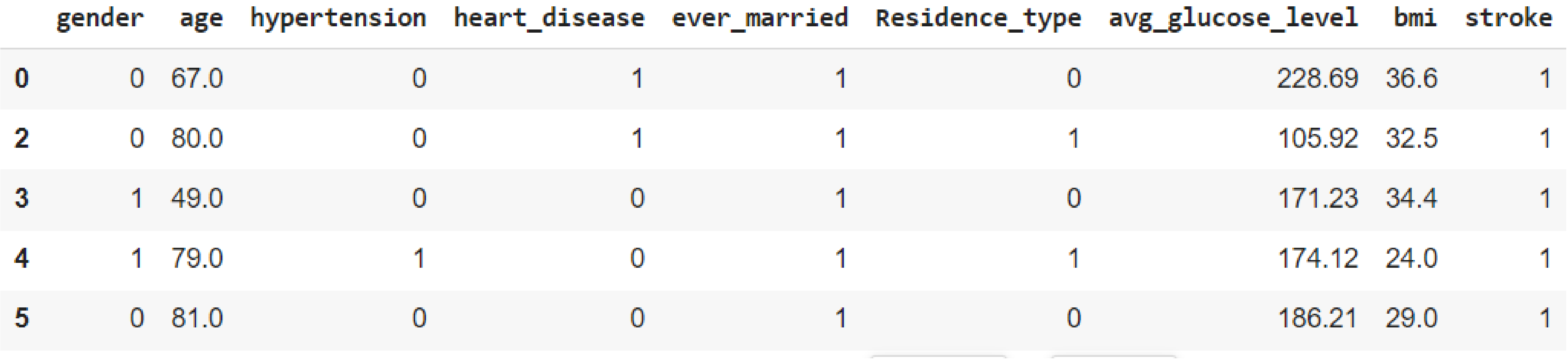


Data

Analyzing

Preprocessing

Data Science



	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke
0	0	67.0	0	1	1	0	228.69	36.6	1
2	0	80.0	0	1	1	1	105.92	32.5	1
3	1	49.0	0	0	1	0	171.23	34.4	1
4	1	79.0	1	0	1	1	174.12	24.0	1
5	0	81.0	0	0	1	0	186.21	29.0	1

1. Review

Data

Analyzing

Preprocessing

Select ML algorithm

1. Review

2. Decision Tree

3. ML process

4. Results

Stroke prediction (2)

Stroke prediction (2)

1. Review
2. Decision Tree
3. ML process
4. Results

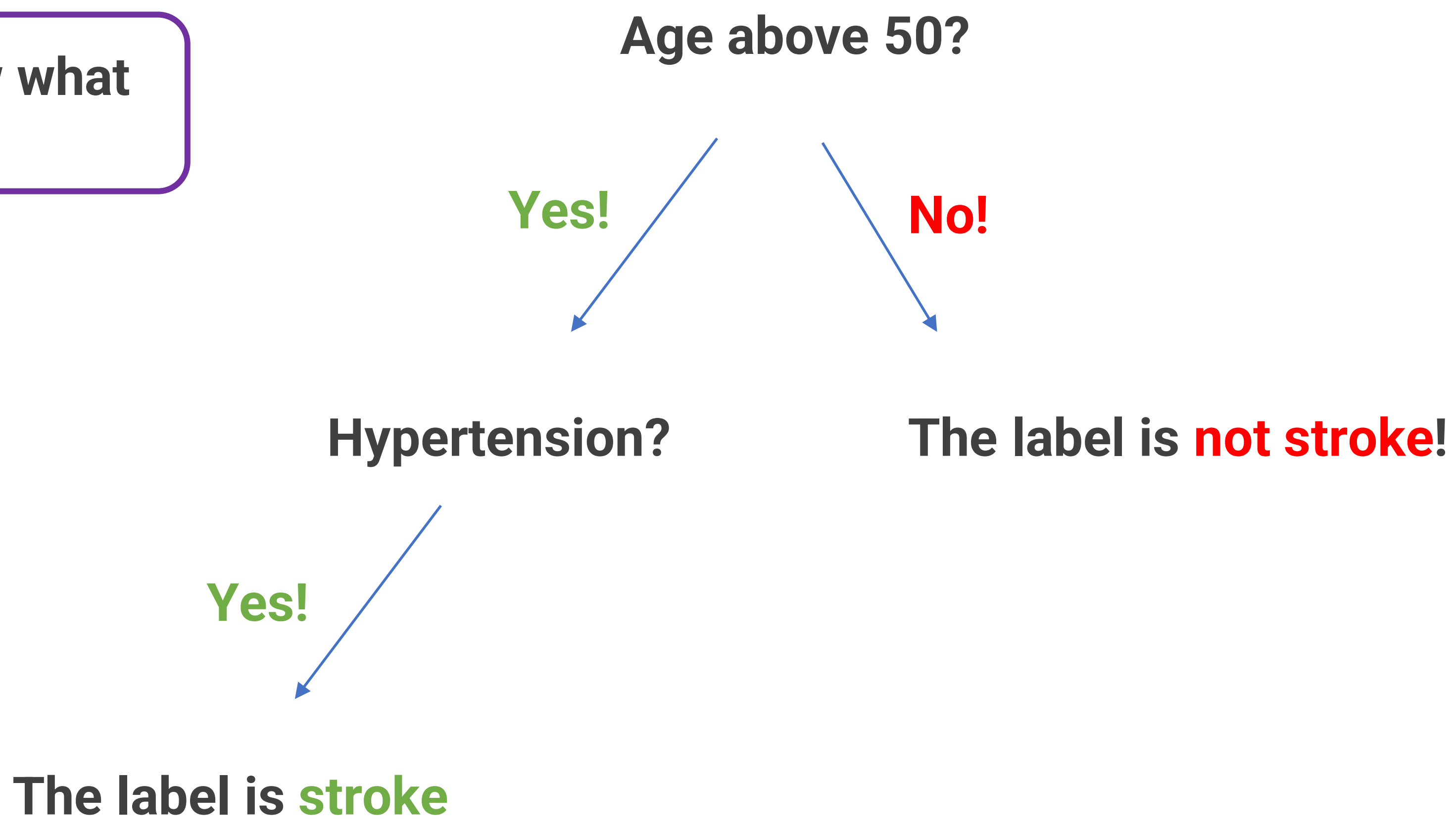
1. Review
2. **Decision Tree**
3. ML process
4. Results

Stroke prediction (2)

2. Decision Tree

2. Decision Tree

How can it know what to ask First?



Age above 50?

Yes!

No!

Hypertension?

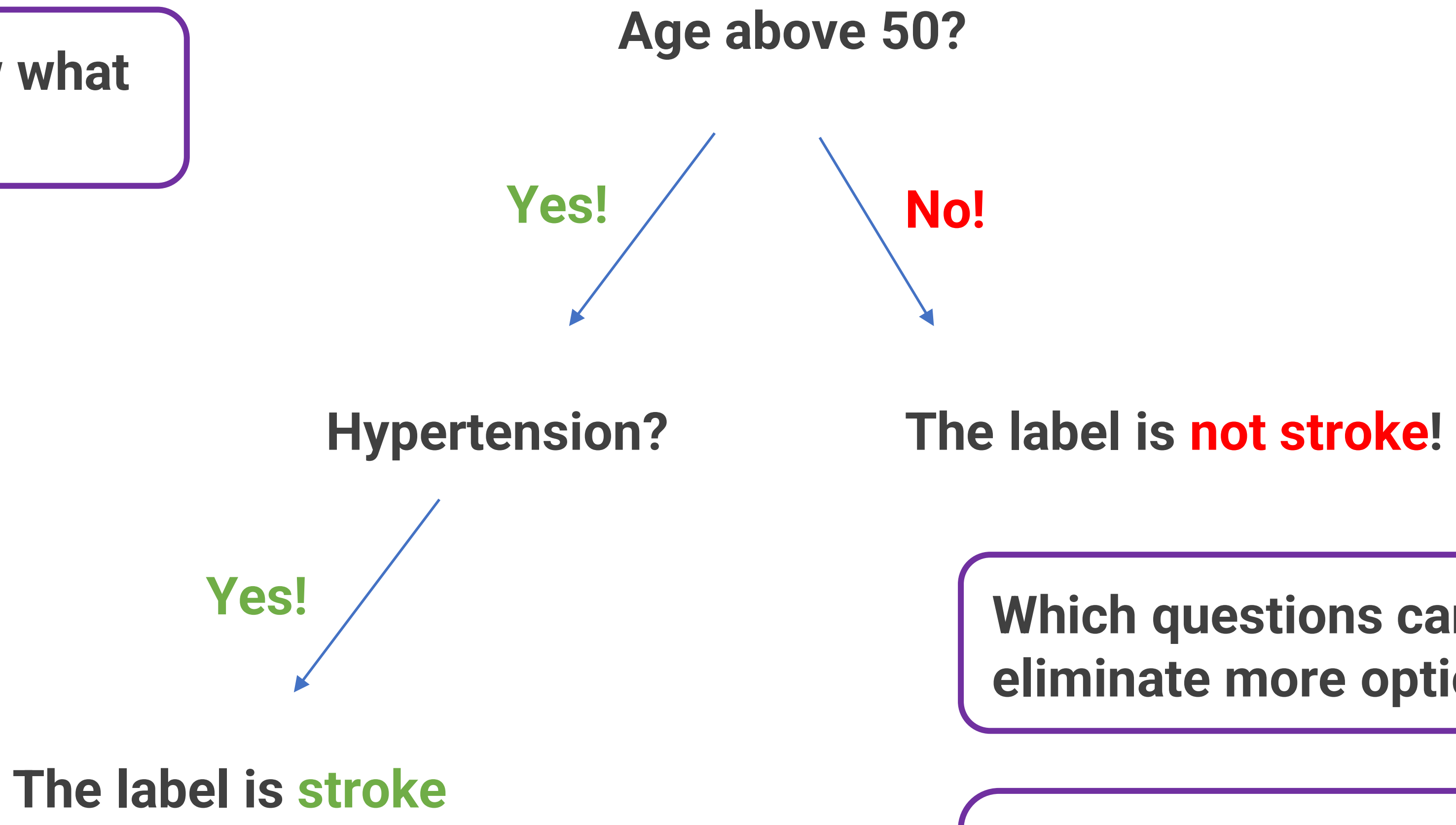
The label is **not stroke**!

Yes!

The label is **stroke**

2. Decision Tree

How can it know what to ask First?



Age above 50?

Yes!

No!

Hypertension?

The label is not stroke!

Yes!

The label is stroke

Which questions can eliminate more options?

Which questions can divide our options in two equal groups?

2. ML process

Data

Analyzing

Preprocessing

Select ML algorithm

1. Review

2. Decision Tree

3. ML process

4. Results

Stroke prediction (2)

Stroke prediction (2)

1. Review
2. Decision Tree
3. ML process
4. Results

1. Review

2. Decision Tree

3. ML process

4. Results

Stroke prediction (2)

3. ML process

Data

Analyzing

Preprocessing

Select ML algorithm

3. ML process

Data

Analyzing

Preprocessing

Select ML algorithm

Training the AI (model)

3. ML process

Data

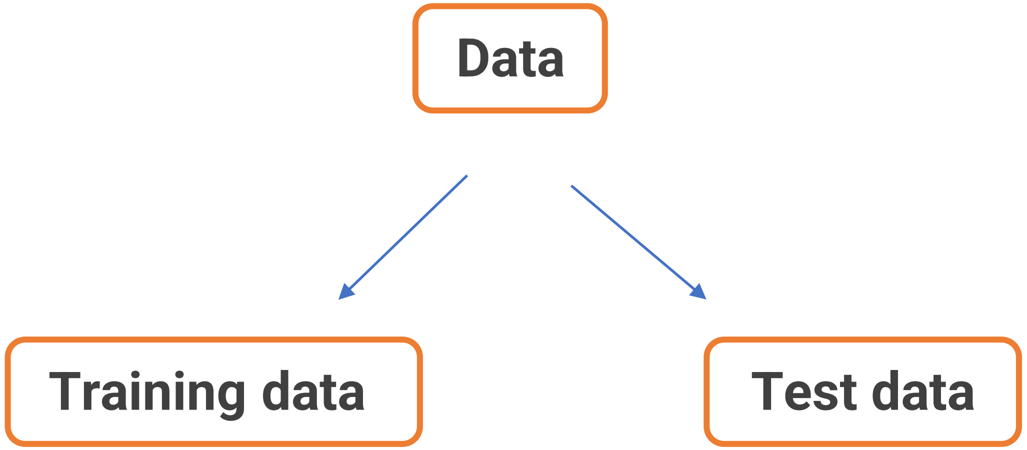
Analyzing

Preprocessing

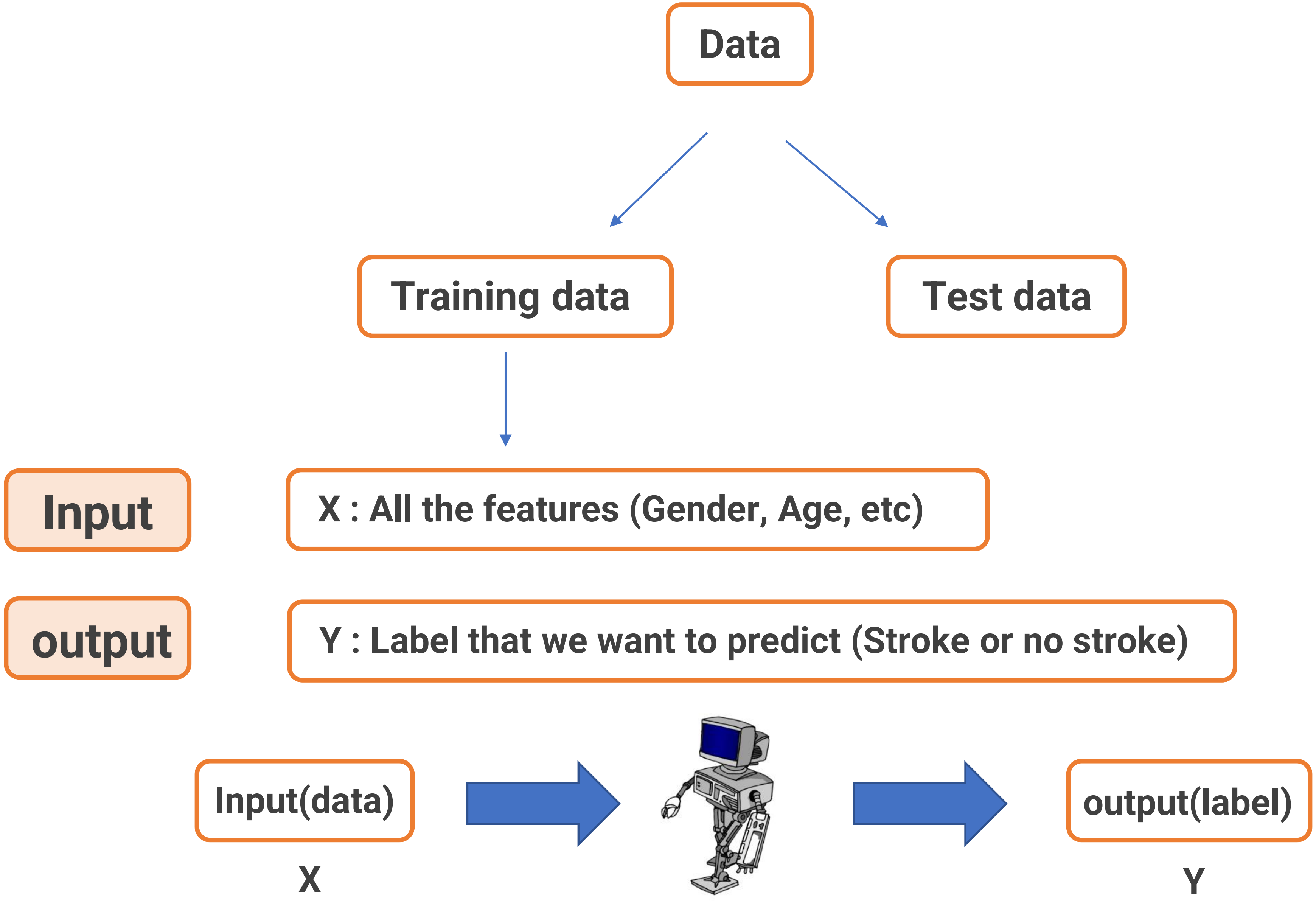
Select ML algorithm

Training the AI (model)

We want to train the AI, and after it learns, we want to take an exam to make sure it has learned!



3. ML process



Data

Training data

Test data

Input

X : All the features (Gender, Age, etc)

output

Y : Label that we want to predict (Stroke or no stroke)

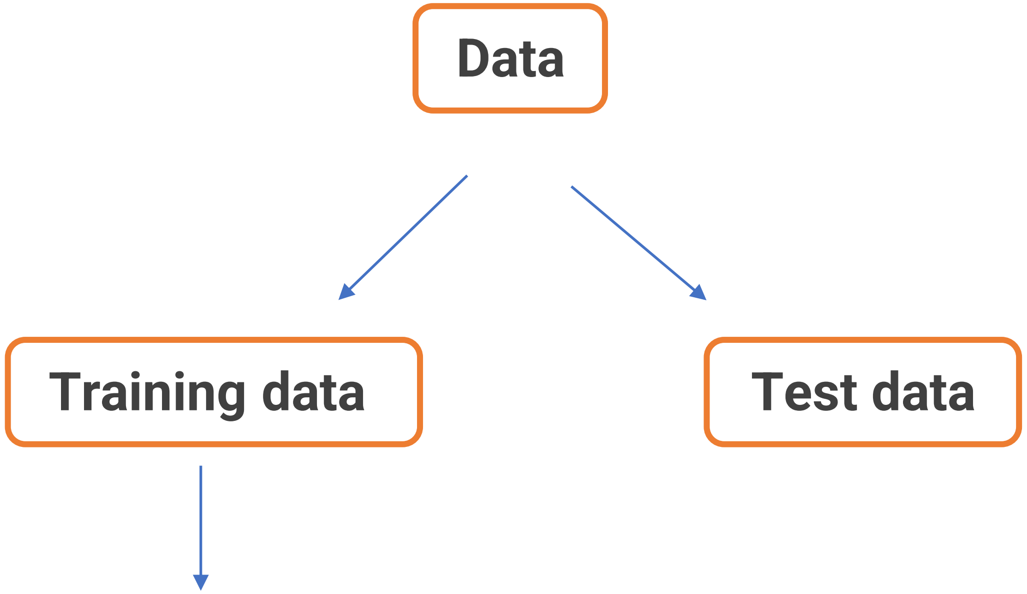
Input(data)

X

output(label)

Y

3. ML process



Data

Training data

Test data

Input

X : All the features (Gender, Age, etc)

output

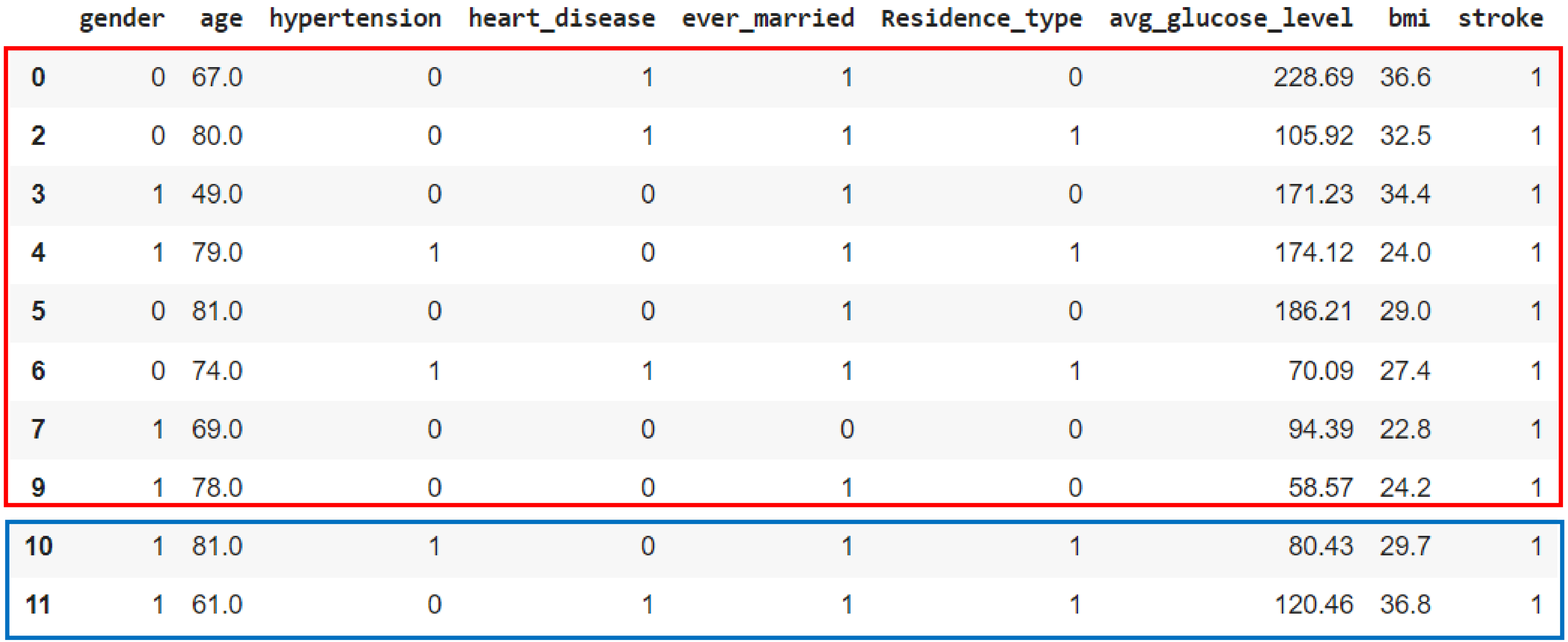
Y : Label that we want to predict (Stroke or no stroke)

By having the information about the features of each person, what would be the label (stroke or no stroke)

3. ML process

4908 persons

80% for training

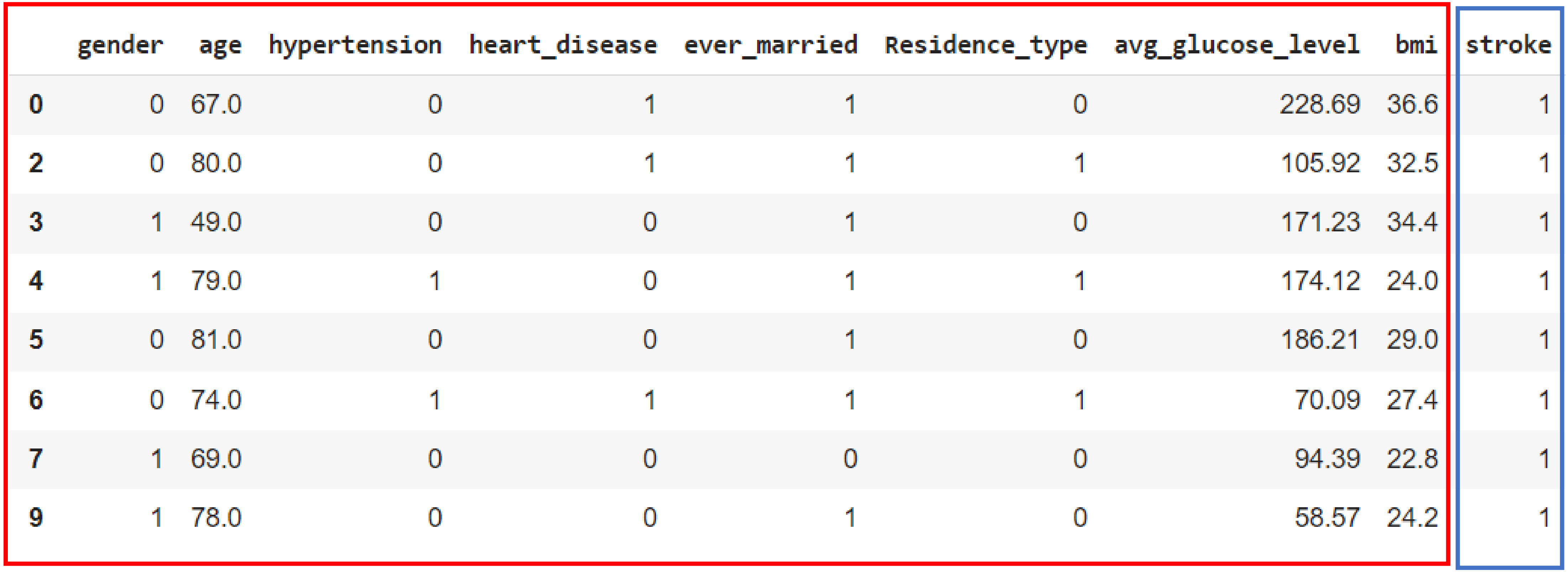


	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke
0	0	67.0	0	1	1	0	228.69	36.6	1
2	0	80.0	0	1	1	1	105.92	32.5	1
3	1	49.0	0	0	1	0	171.23	34.4	1
4	1	79.0	1	0	1	1	174.12	24.0	1
5	0	81.0	0	0	1	0	186.21	29.0	1
6	0	74.0	1	1	1	1	70.09	27.4	1
7	1	69.0	0	0	0	0	94.39	22.8	1
9	1	78.0	0	0	1	0	58.57	24.2	1
10	1	81.0	1	0	1	1	80.43	29.7	1
11	1	61.0	0	1	1	1	120.46	36.8	1

20% for Testing

3. ML process

Training Data



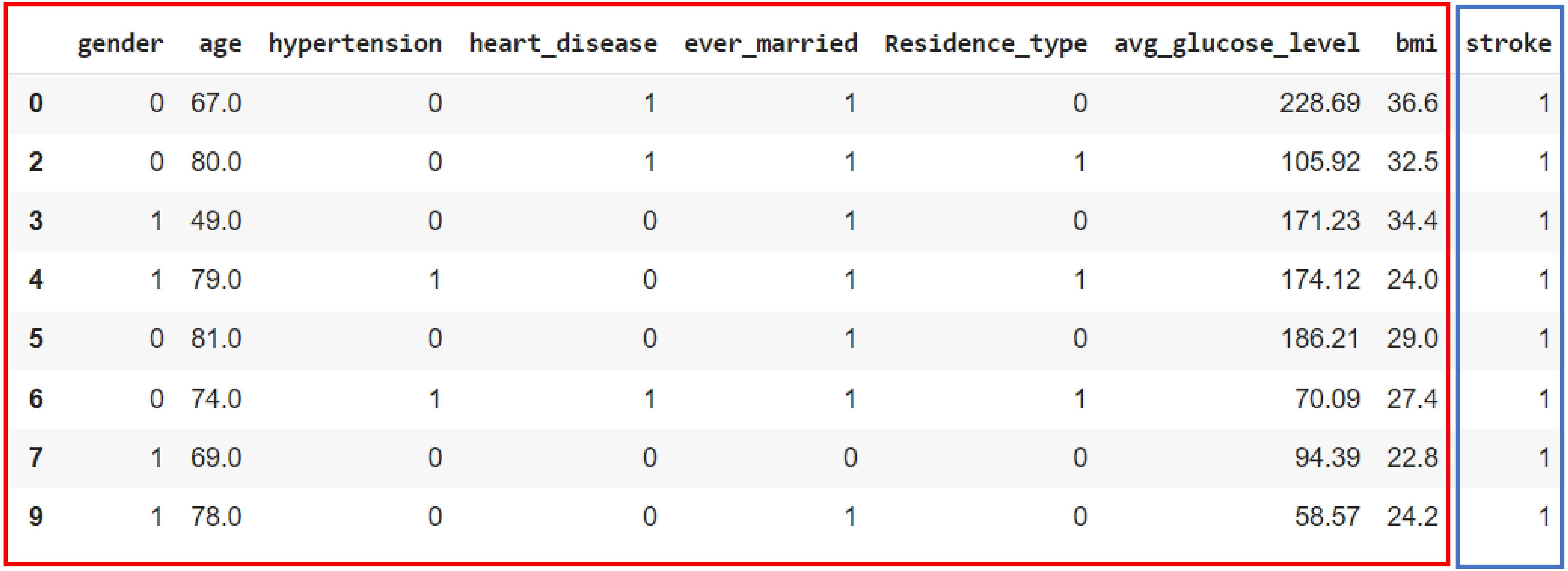
	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke
0	0	67.0	0	1	1	0	228.69	36.6	1
2	0	80.0	0	1	1	1	105.92	32.5	1
3	1	49.0	0	0	1	0	171.23	34.4	1
4	1	79.0	1	0	1	1	174.12	24.0	1
5	0	81.0	0	0	1	0	186.21	29.0	1
6	0	74.0	1	1	1	1	70.09	27.4	1
7	1	69.0	0	0	0	0	94.39	22.8	1
9	1	78.0	0	0	1	0	58.57	24.2	1

X _train

Y _train

3. ML process

Training Data



	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke
0	0	67.0	0	1	1	0	228.69	36.6	1
2	0	80.0	0	1	1	1	105.92	32.5	1
3	1	49.0	0	0	1	0	171.23	34.4	1
4	1	79.0	1	0	1	1	174.12	24.0	1
5	0	81.0	0	0	1	0	186.21	29.0	1
6	0	74.0	1	1	1	1	70.09	27.4	1
7	1	69.0	0	0	0	0	94.39	22.8	1
9	1	78.0	0	0	1	0	58.57	24.2	1

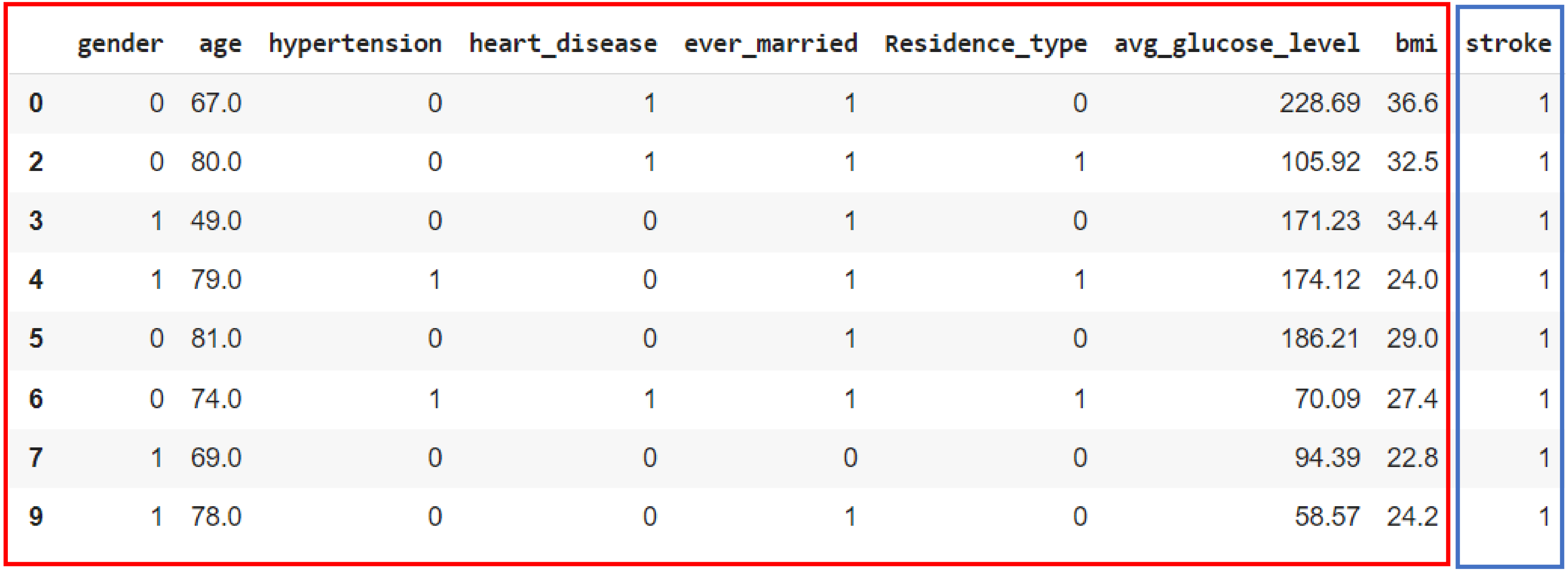
X _train

Y _train

By having X, try to predict Y

3. ML process

Training Data



	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke
0	0	67.0	0	1	1	0	228.69	36.6	1
2	0	80.0	0	1	1	1	105.92	32.5	1
3	1	49.0	0	0	1	0	171.23	34.4	1
4	1	79.0	1	0	1	1	174.12	24.0	1
5	0	81.0	0	0	1	0	186.21	29.0	1
6	0	74.0	1	1	1	1	70.09	27.4	1
7	1	69.0	0	0	0	0	94.39	22.8	1
9	1	78.0	0	0	1	0	58.57	24.2	1

X _train

Y _train

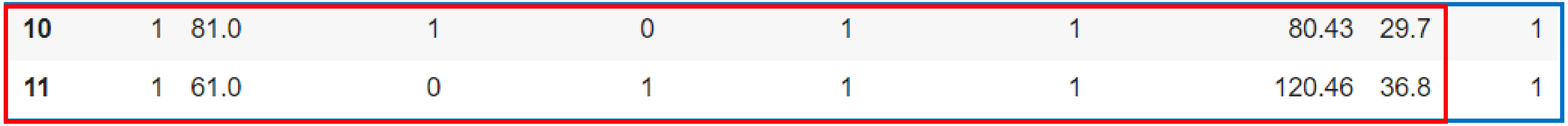
By having X, try to predict Y

3. ML process

Same thing for test data

X _test

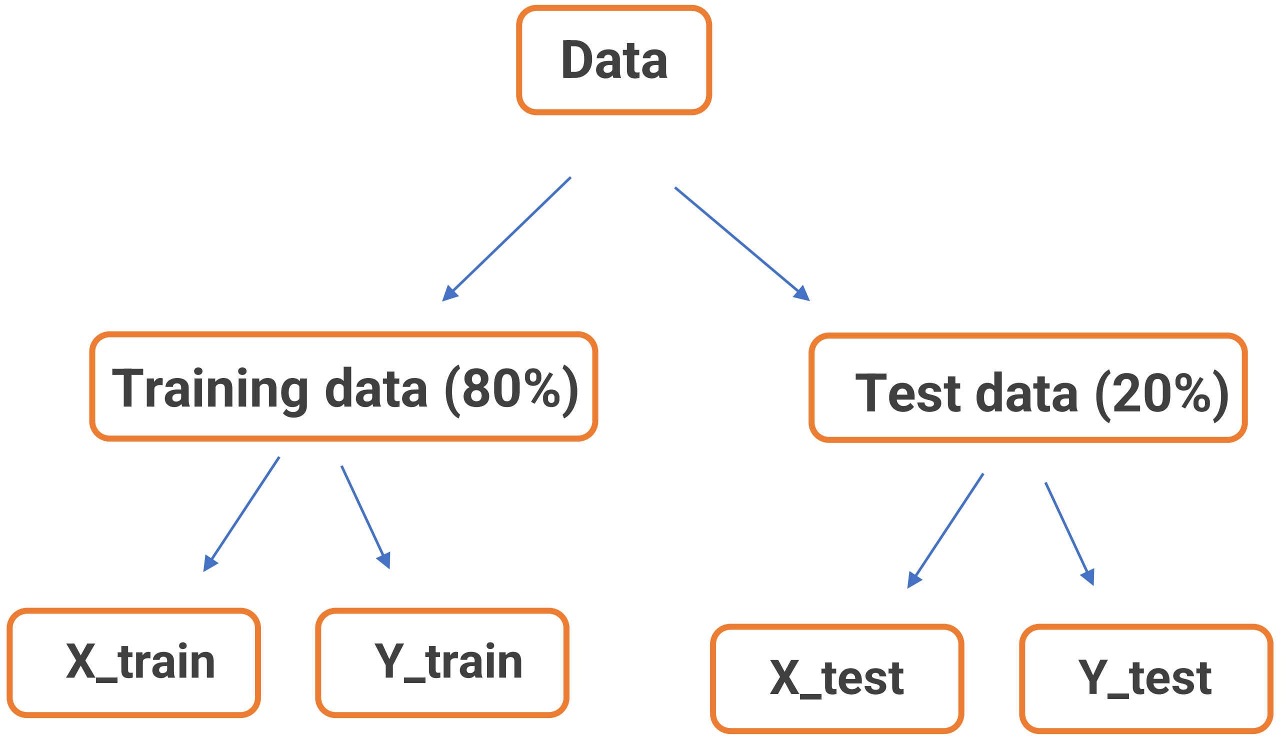
Y _test



10	1	81.0	1	0	1	1	80.43	29.7	1
11	1	61.0	0	1	1	1	120.46	36.8	1

20% for Testing

3. ML process



Data

Training data (80%)

Test data (20%)

X_train

Y_train

X_test

Y_test

X : All the features (Gender, Age, etc)

Y : Label that we want to predict (Stroke or no stroke)

3. ML process

- Data
- Analyzing
- Preprocessing
- Select ML algorithm
- Training the AI (model)
- Testing the model (taking the exam!)

3. ML process

Data

Analyzing

Preprocessing

Select ML algorithm

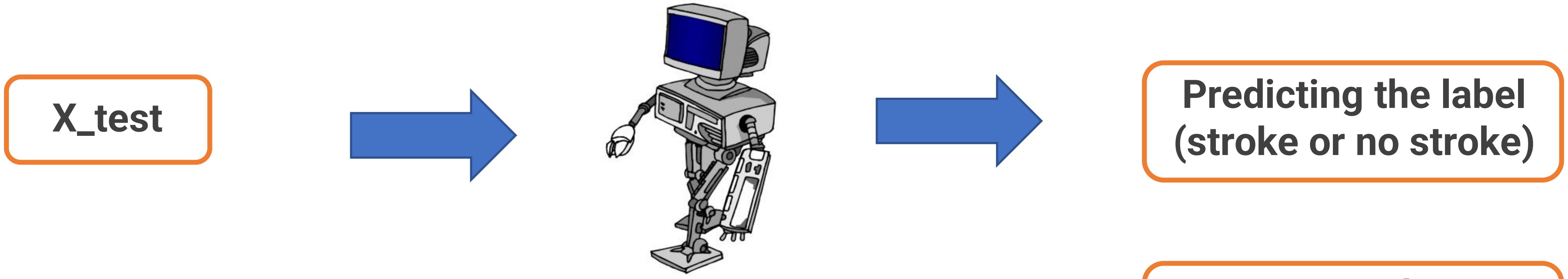
Training the AI (model)

Testing the model (taking the exam!)

We will use Test Data X_test and Y_test

3. ML process

Testing the model (taking the exam!)



X_test

Predicting the label (stroke or no stroke)

How can we know it predict good or bad

X : All the features (Gender, Age, etc)

We can compare it to the ACTUAL Label

Y_test

3. ML process

Data

Analyzing

Preprocessing

Select ML algorithm

Training the AI (model)

Testing the model (taking the exam!)

3. ML process

- Data
- Analyzing
- Preprocessing
- Select ML algorithm
- Training the AI (model)
- Testing the model (taking the exam!)
- Analyzing the results

1. Review
2. Decision Tree
3. **ML process**
4. Results

Stroke prediction (2)

Stroke prediction (2)

1. Review
2. Decision Tree
3. ML process
4. Results

1. Review

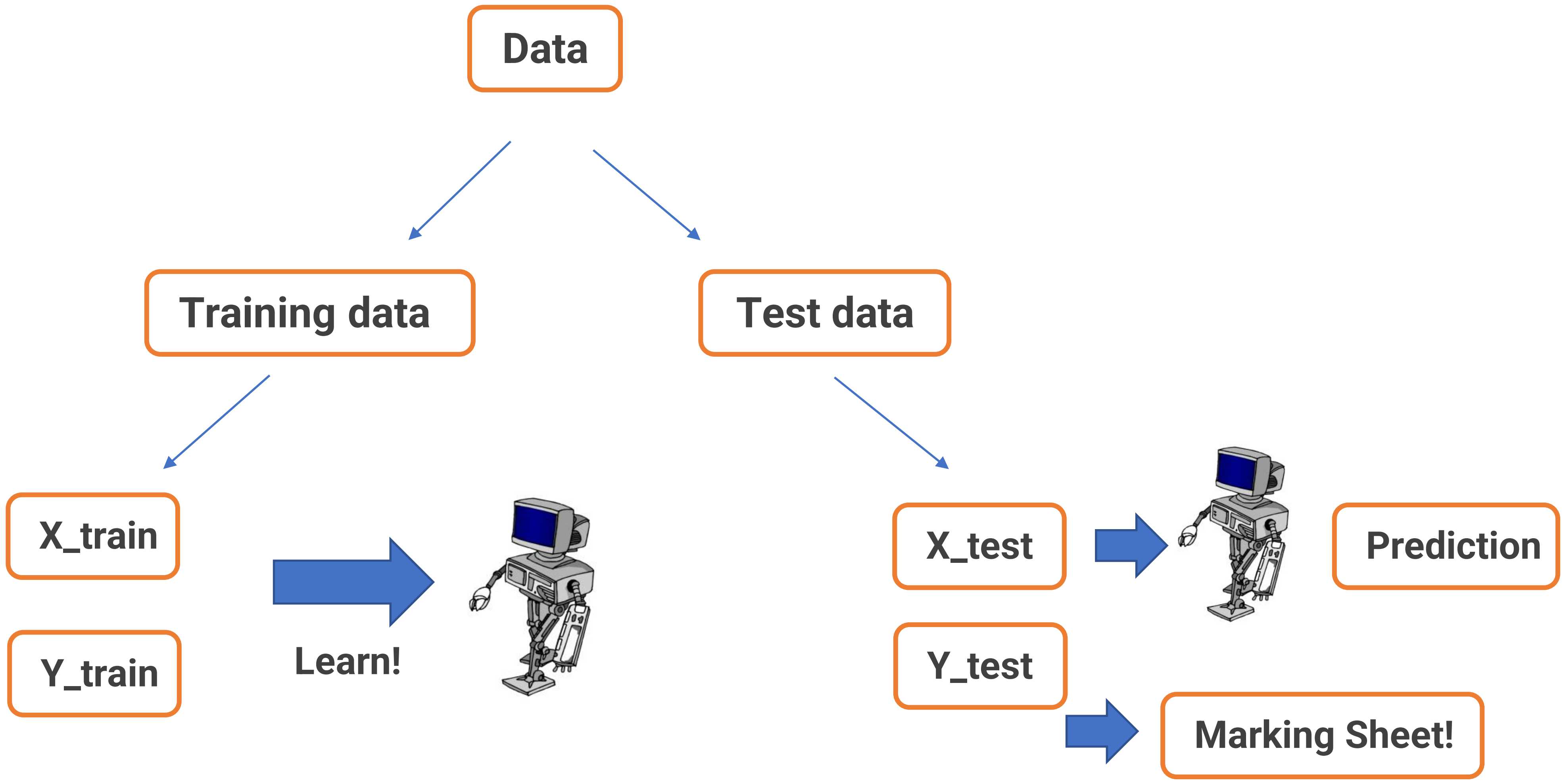
2. Decision Tree

3. ML process

4. Results

Stroke prediction (2)

4. Results



Data

Training data

Test data

X_train

Y_train

Learn!

X_test

Y_test

Prediction

Marking Sheet!

4. Results

We were unable to specify the rules, it was too complicated, and we didn’t know!!

Do we have the rules now?

Let’s visualize it!

1. Review

2. Decision Tree

3. ML process

4. Results

Stroke prediction (2)

1. Review
2. Decision Tree
3. ML process
4. Results

Stroke prediction (2)

Summary

- Data: for training an AI we need data
- Analyzing: give us a good insight about the data and help us to prepare the data for training purposes in preprocess stage.
- Preprocessing: preparing the data for training the model:
  - Ignoring unnecessary data
  - Ignoring unnecessary features
  - Convert to numbers
  - Other changes based on what we found in the previous stage
- Select ML algorithm: Decision tree (20 Questions)
- Training AI: spiliting the data into test data and train data. We use train data to train our AL
- Testing the model: using test data to take the exam and see if out AI works well
- Analyzing the result and extracting valuable information

Questions?

Homework

- Complete the project.
- Why can we predict no stroke with high accuracy, but we cannot predict a stroke with the same accuracy?
- What can we do to improve our AI?