CLOTHES SIZE RECOMMENDATION SYSTEM

PRESENTER:

QUYNH THACH

TABLES OF CONTENT

01. INTRODUCTION

Size	XXS	XS	S	M	L	XL	2XL
Bust (cm)	72 - 76	76 - 80	80 - 84	84 - 88.5	88.5 - 93.5	93.5 - 98.5	98.5 - 103.5
Waist (cm)	60 - 64	64 - 68	68 - 72	72 - 76.5	76.5 - 81.5	81.5 - 86.5	86.5 - 91.5
Hip (cm)	82 - 86	86 - 90	90 - 94	94 - 98.5	98.5 - 103.5	103.5 - 108.5	108.5 - 113.5

02. DATA EXPLORATION & CLEANING

weight	age	height	size
62	28.0	172.72	XL
59	36.0	167.64	L
61	34.0	165.10	M

Kaggle

Number of records: 119734

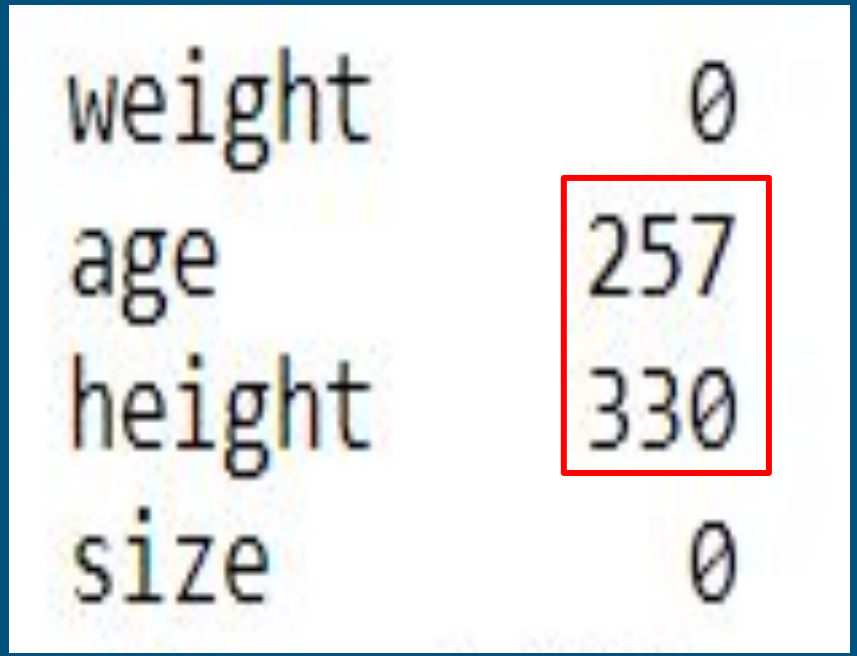
Number of columns: 4

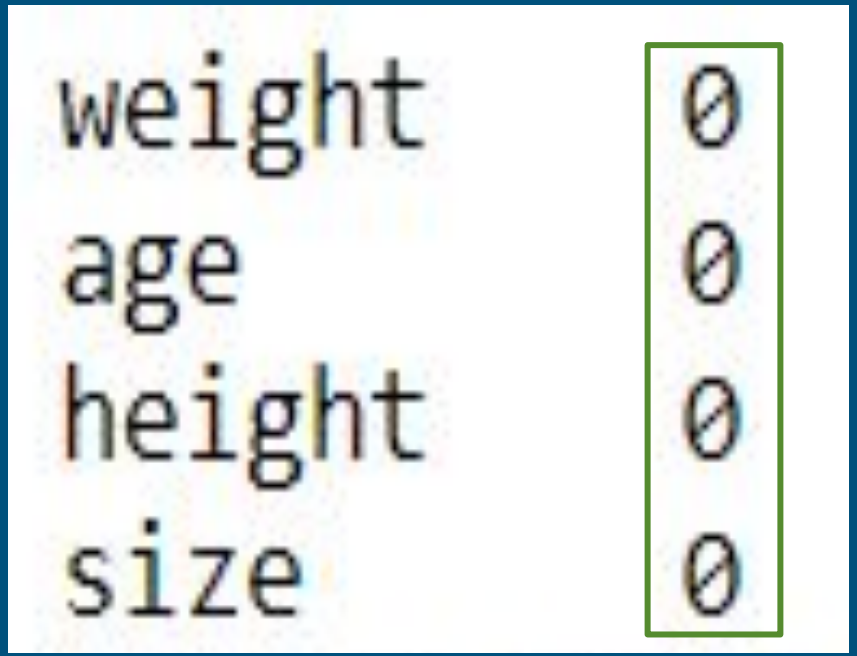
1) Weight - I/O
2) Age - I/O
3) Height - I/O
4) Size - I/O

XXS, S, M, L, XL, XXL, XXXL

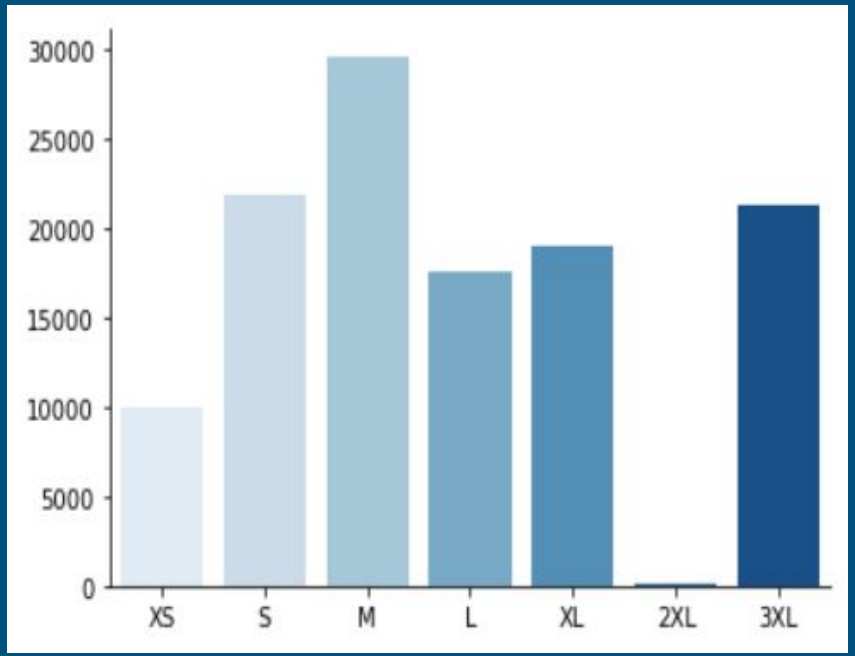
XS, S, M, L, XL, 2XL, 3XL

02. DATA EXPLORATION & CLEANING



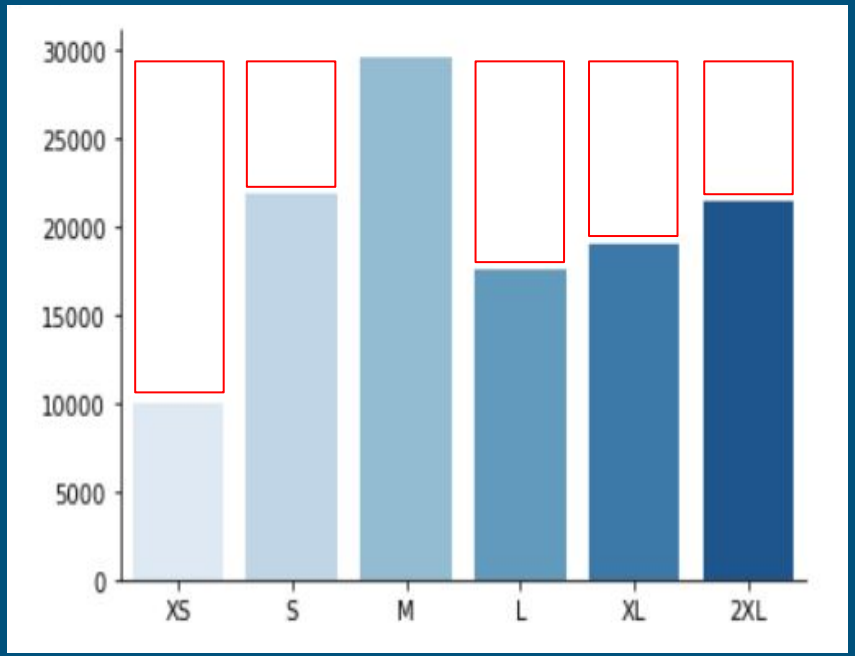


02. DATA EXPLORATION & CLEANING



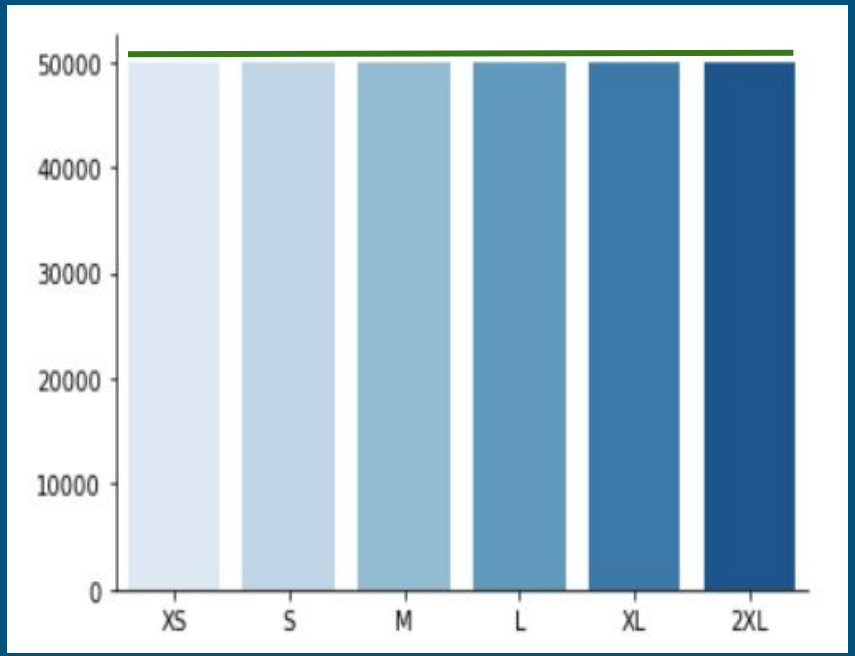
M	29575
S	21829
3XL	21259
XL	19033
L	17481
XS	9907
2XL	69

02. DATA EXPLORATION & CLEANING



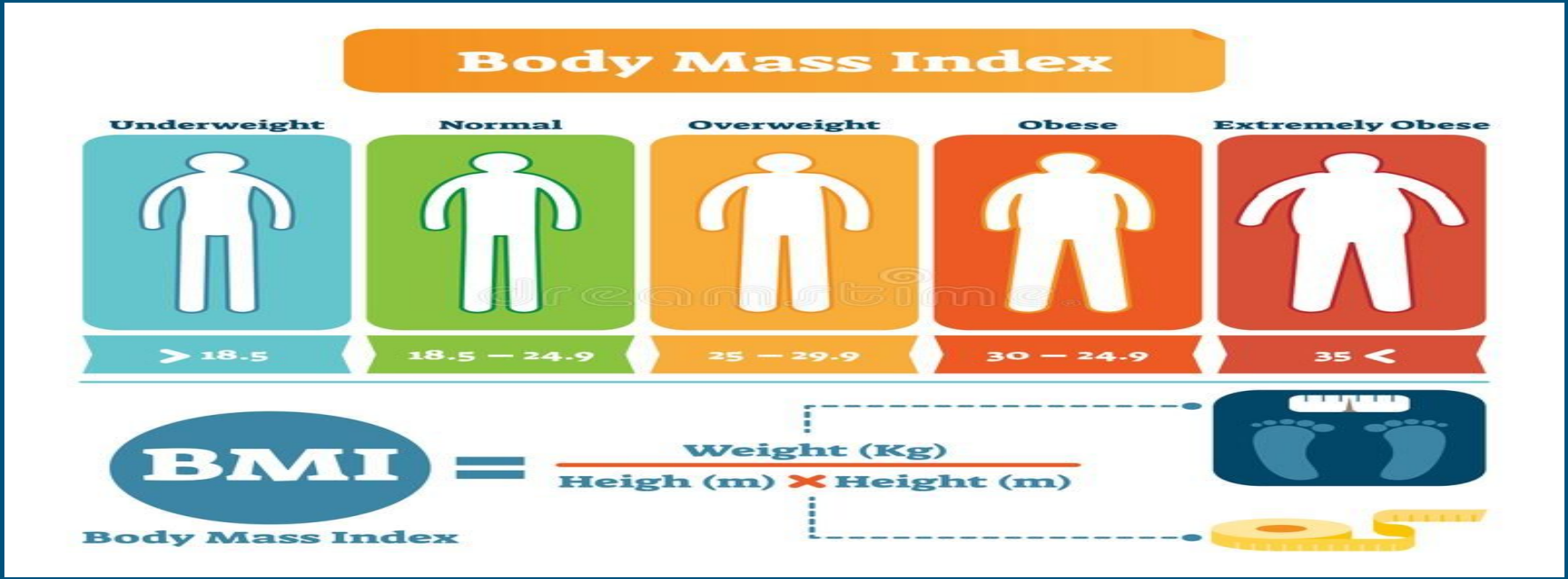
M	29575
S	21829
2XL	21328
XL	19033
L	17481
XS	9907

02. DATA EXPLORATION & CLEANING

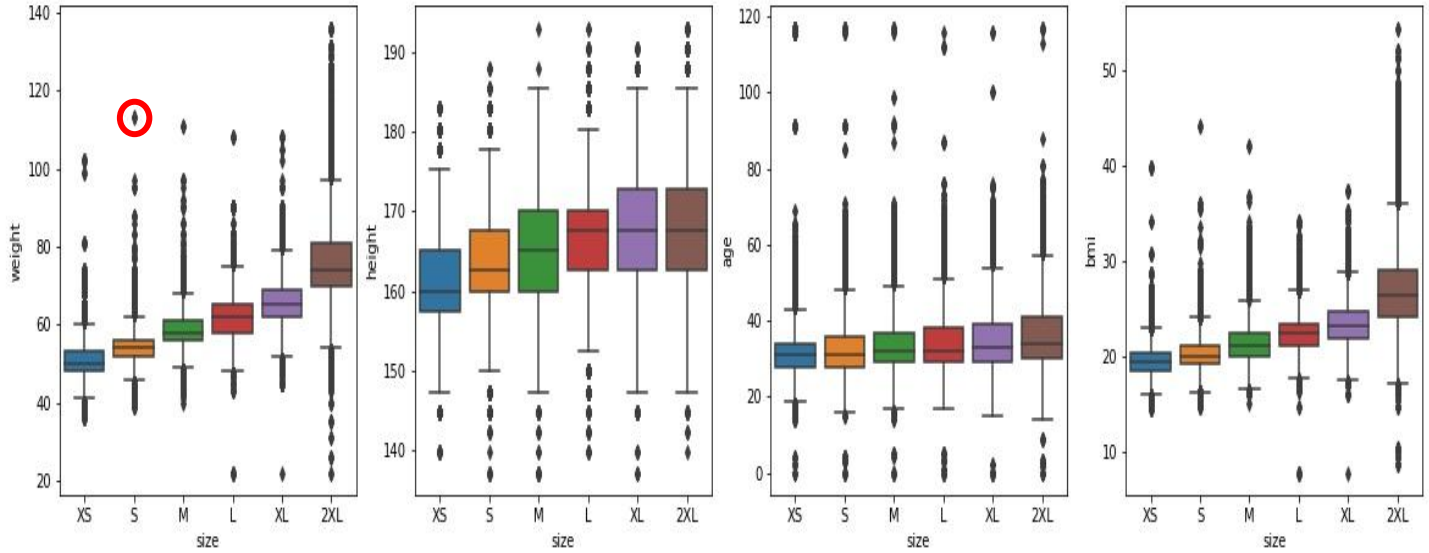


2XL	50000
XL	50000
S	50000
XS	50000
L	50000
M	50000

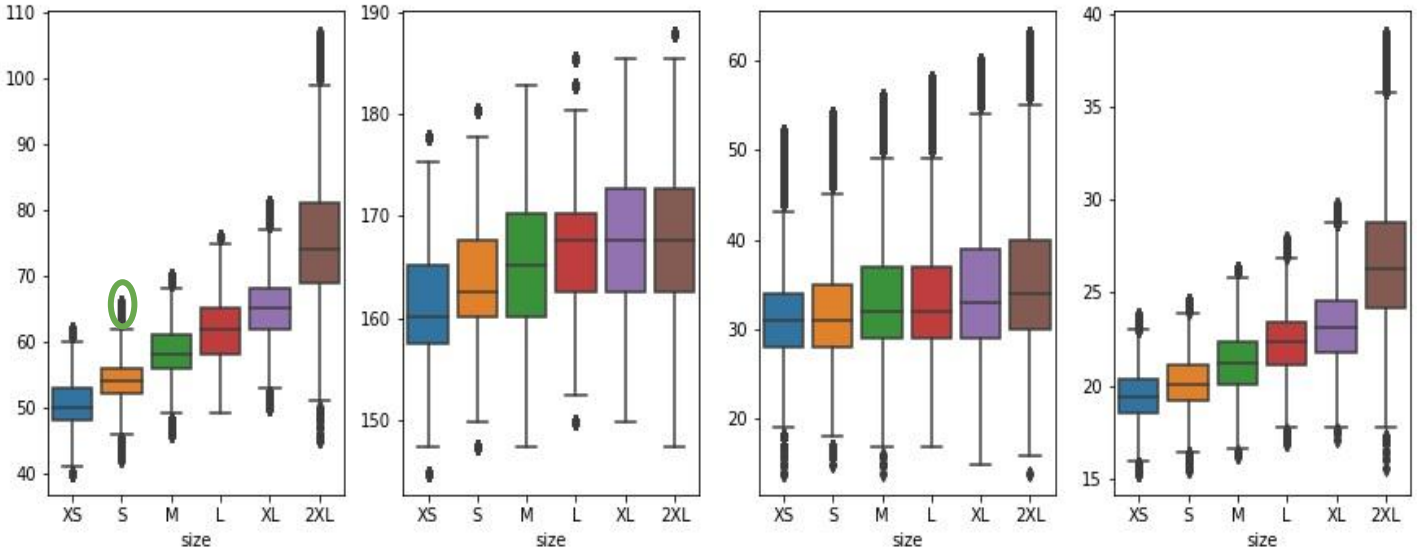
02. DATA EXPLORATORY & CLEANING



02. DATA EXPLORATION & CLEANING



02. DATA EXPLORATION & CLEANING

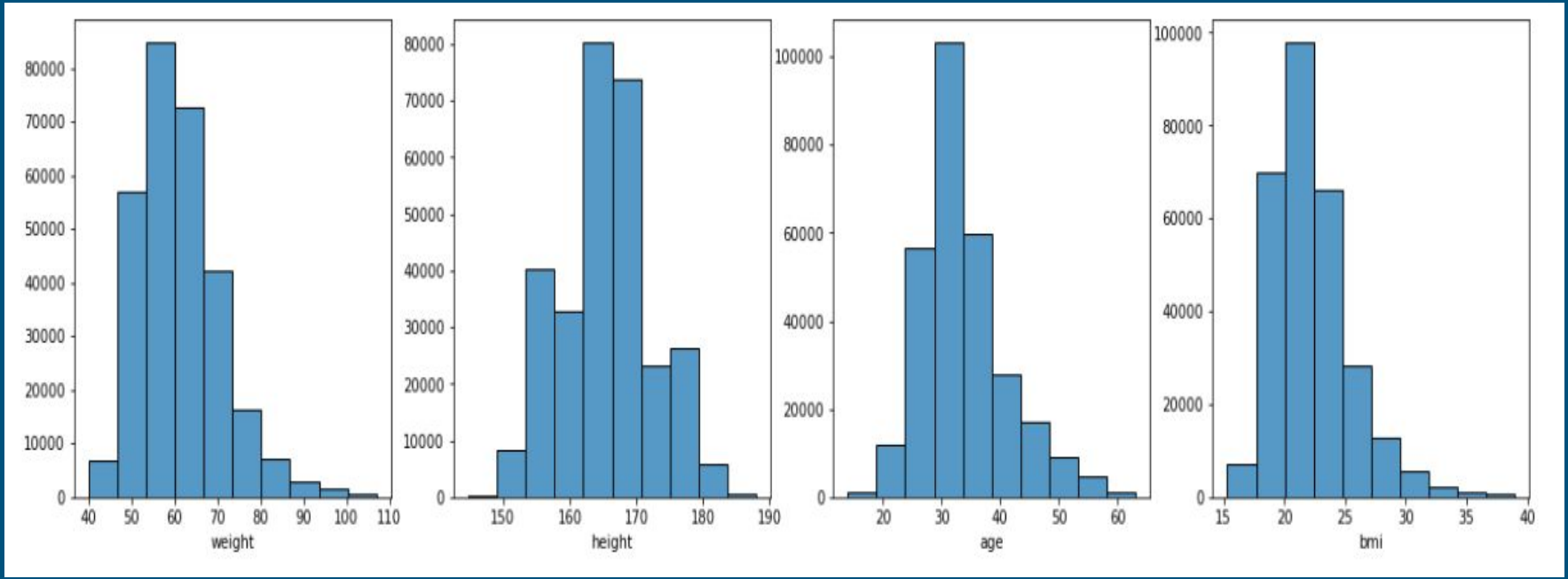


02. DATA EXPLORATION & CLEANING

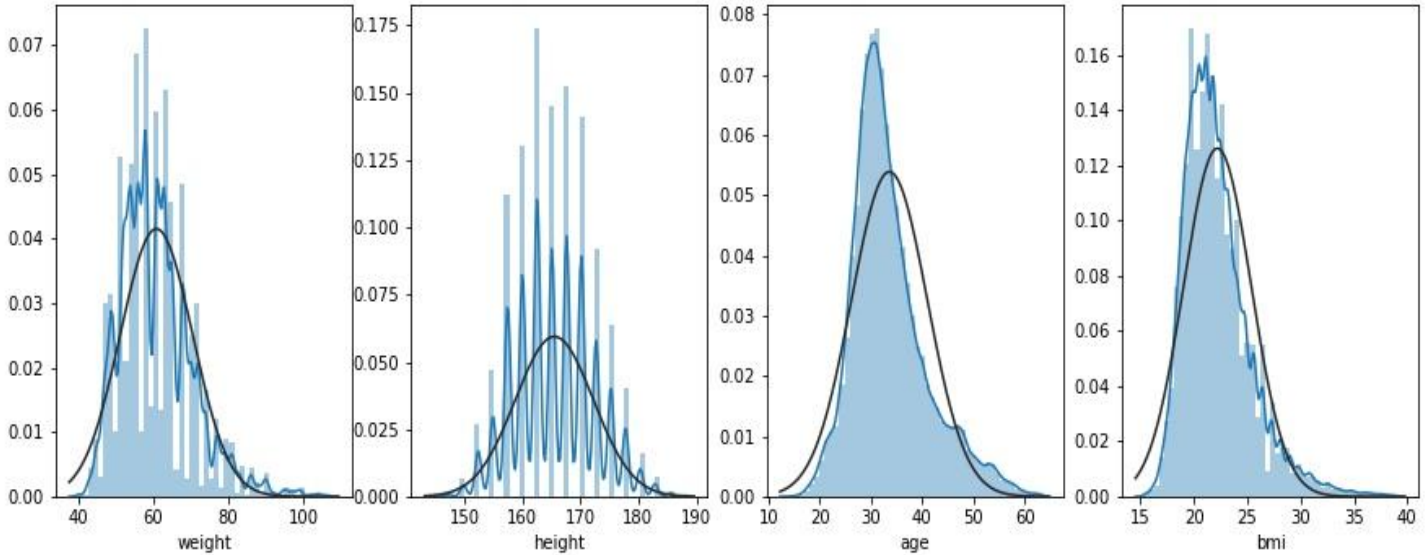
	weight	age	height
count	119153.000000	119153.000000	119153.000000
mean	61.756095	34.032714	165.807068
std	9.942877	8.148302	6.737797
min	22.000000	0.000000	137.160000
25%	55.000000	29.000000	160.020000
50%	61.000000	32.000000	165.100000
75%	67.000000	37.000000	170.180000
max	136.000000	117.000000	193.040000

	weight	age	height	bmi
count	291955.000000	291955.000000	291955.00000	291955.000000
mean	60.881030	33.559994	165.52207	22.204485
std	9.611322	7.404393	6.71491	3.161600
min	40.000000	14.000000	144.78000	15.300765
25%	54.000000	29.000000	160.02000	19.994665
50%	60.000000	32.000000	165.10000	21.705737
75%	65.000000	37.000000	170.18000	23.822169
max	107.000000	63.000000	187.96000	38.977129

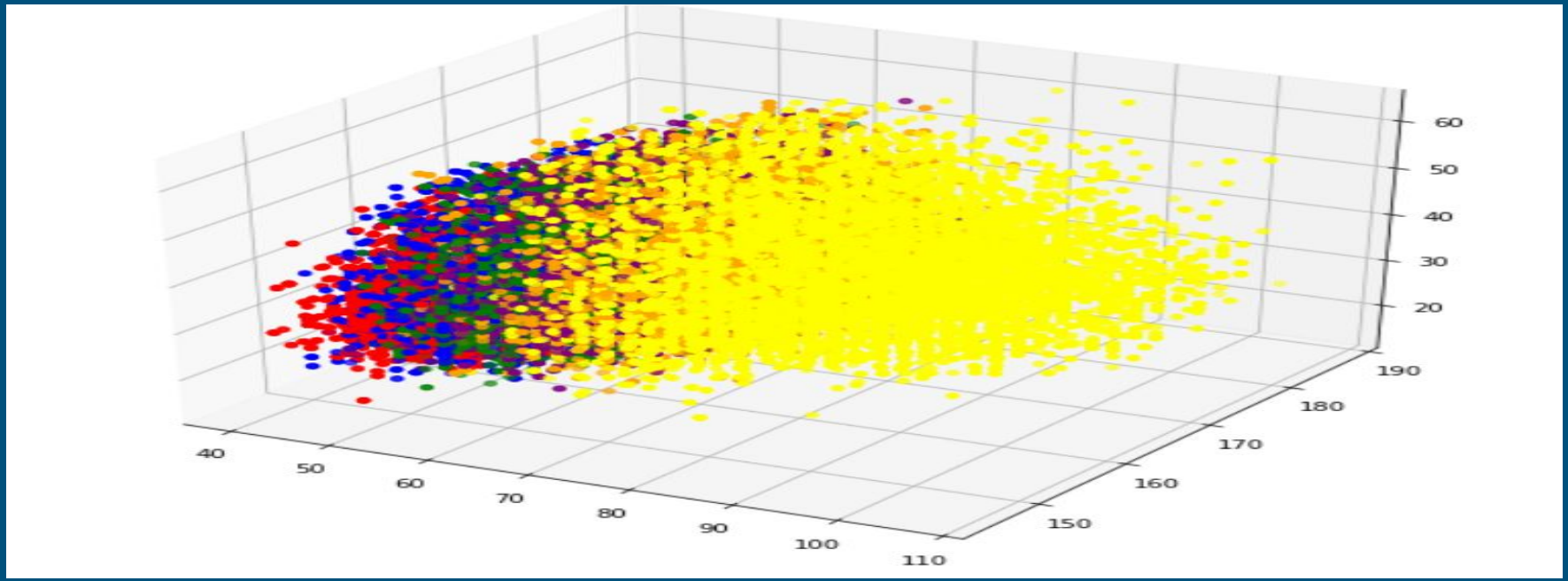
02. DATA EXPLORATION & CLEANING



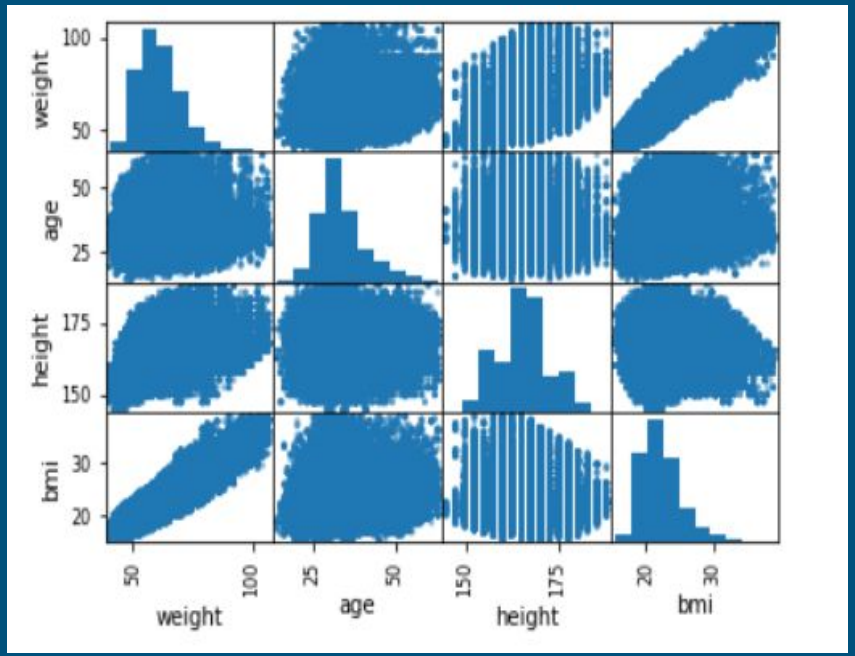
02. DATA EXPLORATION & CLEANING

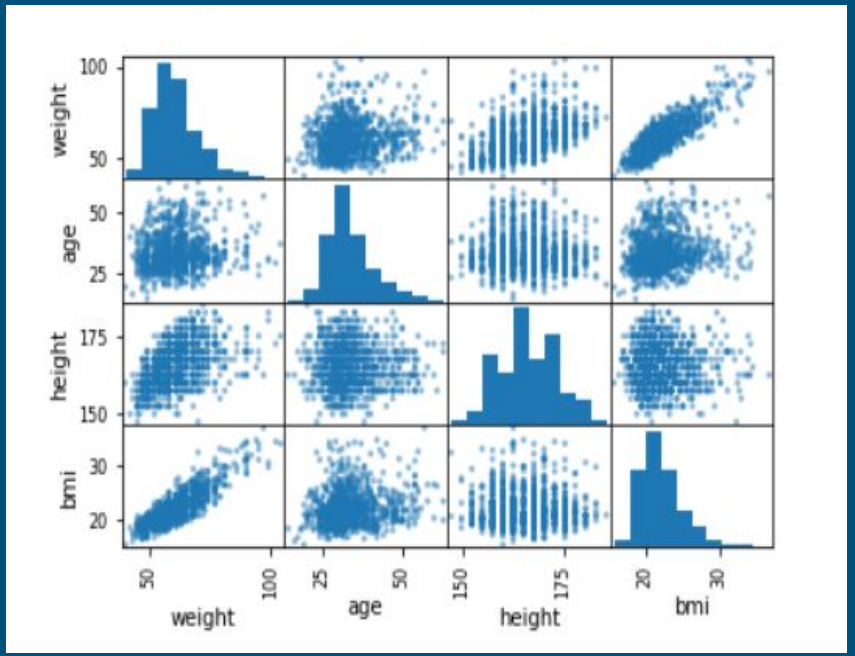


02. DATA EXPLORATION & CLEANING

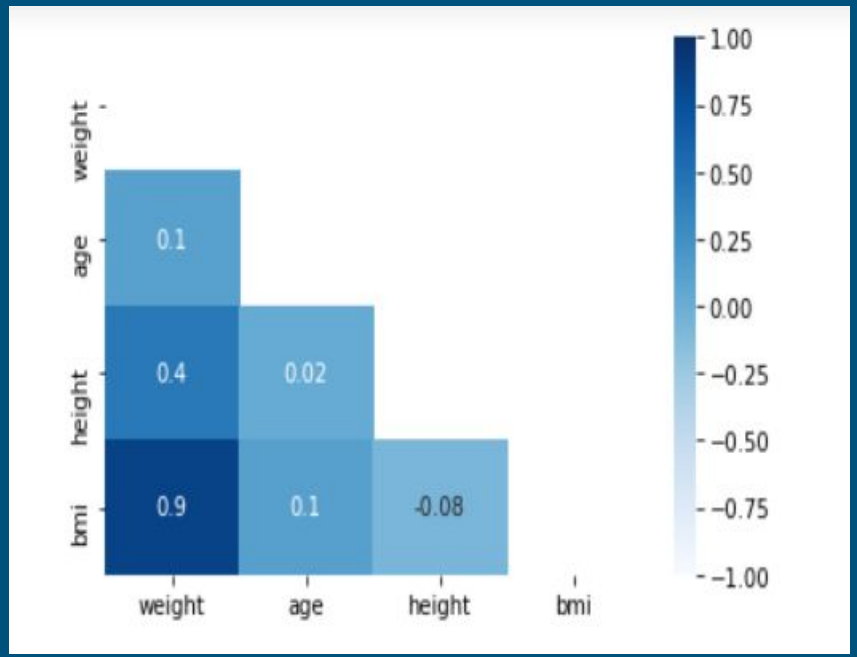


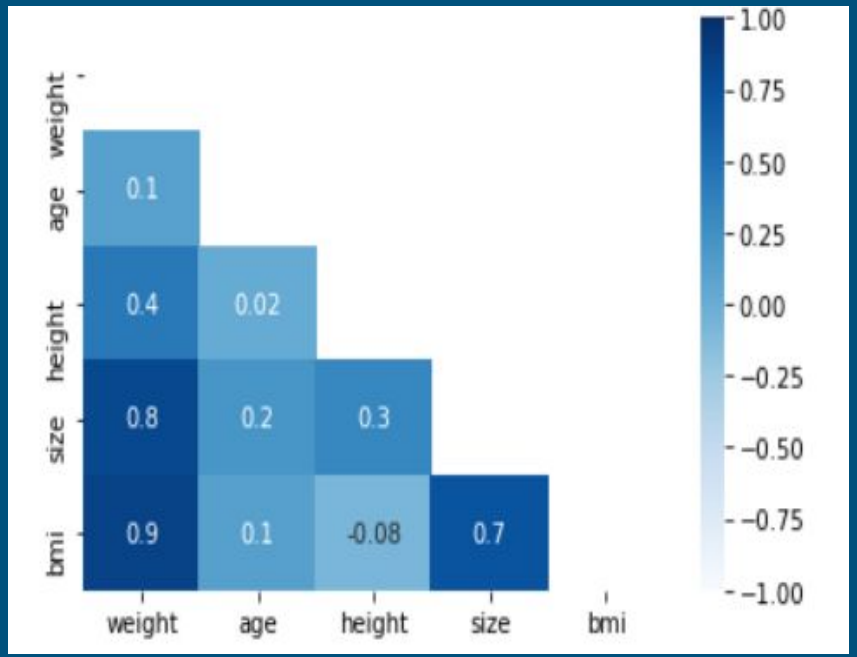
02. DATA EXPLORATION & CLEANING



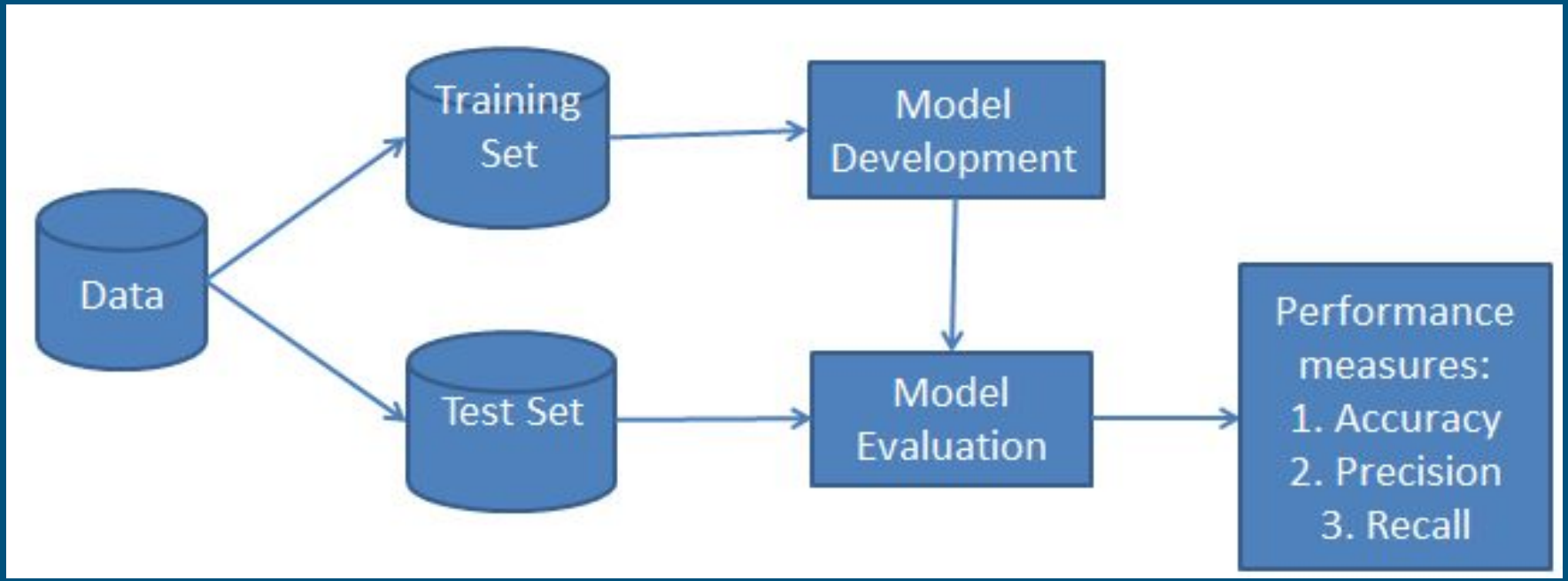


02. DATA EXPLORATION & CLEANING

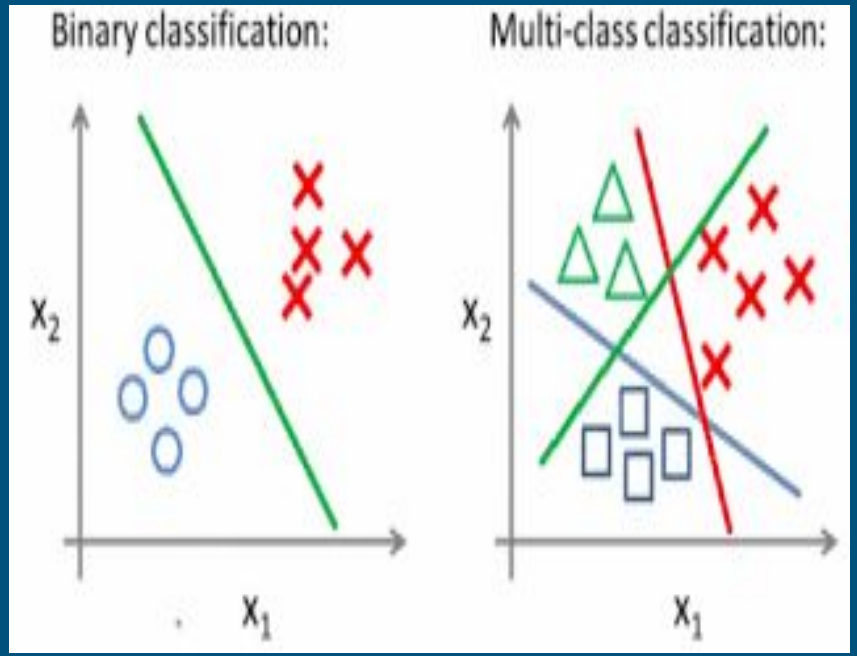




03. MACHINE LEARNING ALGORITHMS



03. MACHINE LEARNING ALGORITHMS



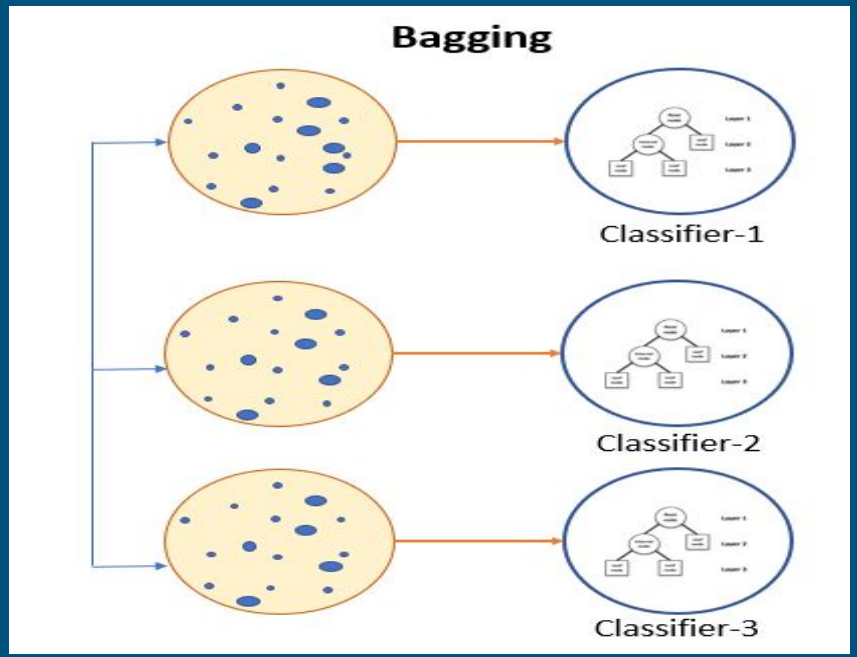
1) k-Nearest Neighbors
2) Decision Trees
3) Random Forest
4) Gradient Boosting
5) Logistic Regression
6) Support Vector Machine

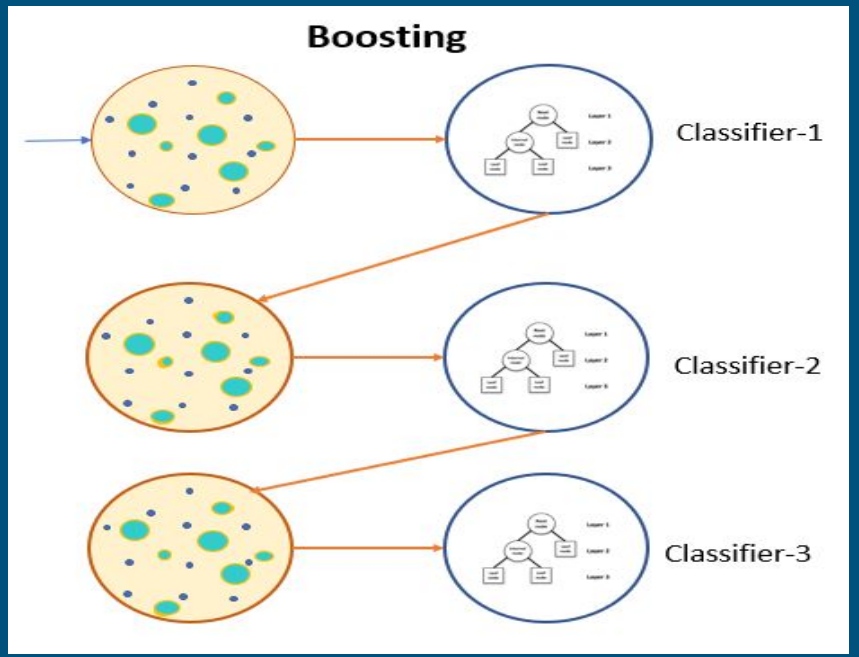
* Base

* Ensemble

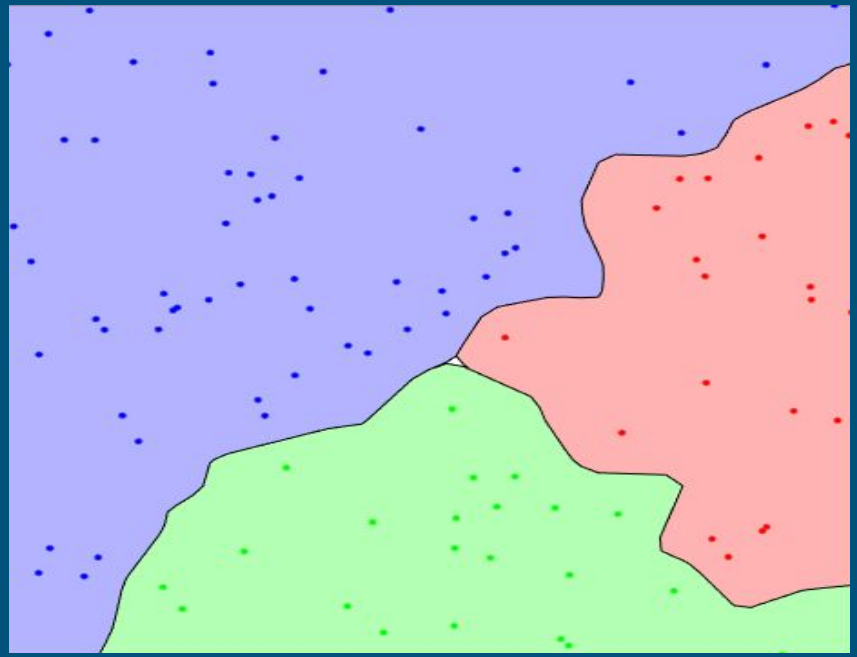
* Binary

03. MACHINE LEARNING ALGORITHMS





03. MACHINE LEARNING ALGORITHMS



01. k-Nearest Neighbors

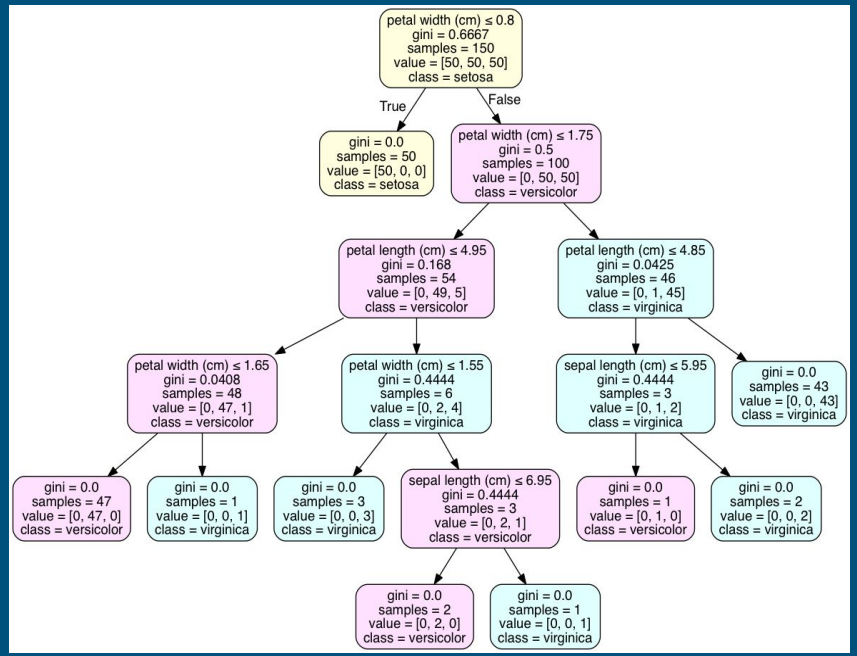
Advantages

- Easy to use
- Memory-based approach
- A variety of distance metrics

Disadvantages

- Computational complexity
- Poor performance on imbalance data
- k_neigbors tuning

03. MACHINE LEARNING ALGORITHM



02. Decision Trees

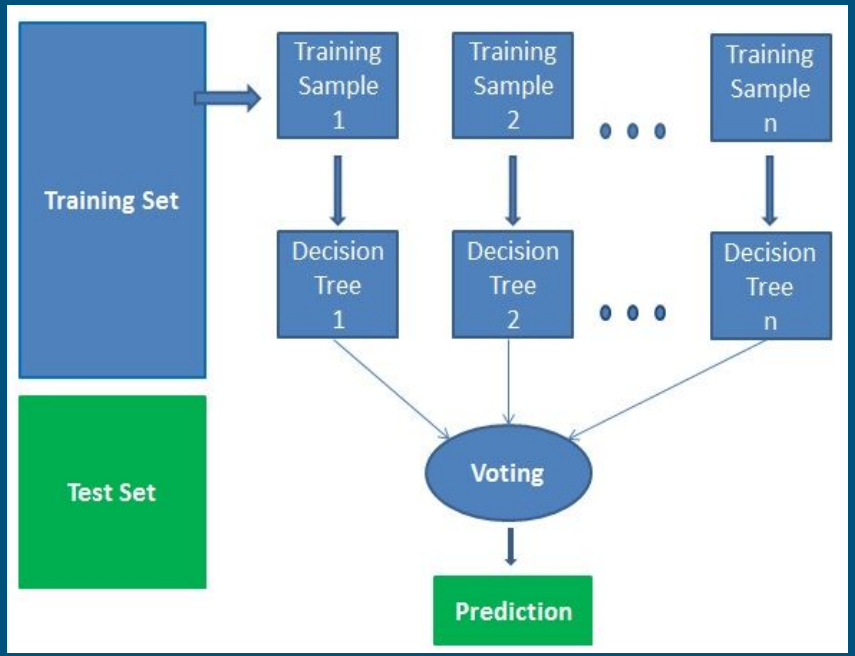
Advantages

- Easy to use and visualize
- Automatic data preprocessing
- Not much hyperparameter tuning

Disadvantages

- Memory-and time-consuming complexity
- Sensitive reproducibility
- Prone to overfitting

03. MACHINE LEARNING ALGORITHMS



03. Random Forest

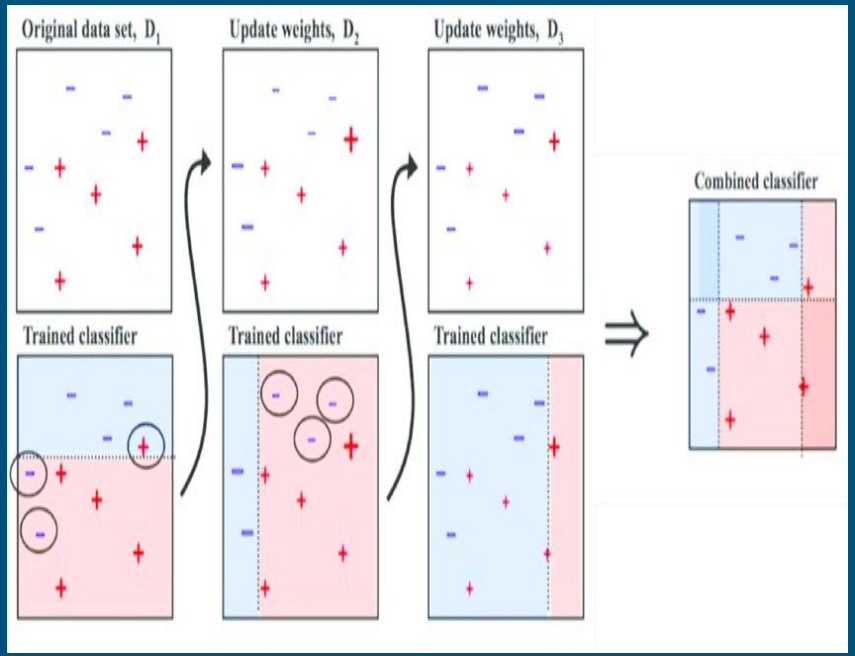
Advantages

- Reduce overfitting
- Handle large & high dimensional dataset
- Handle missing values & imbalanced data

Disadvantages

- Computational & time complexity
- Not consider each variable's significance
- Incomprehensible prediction processes

03. MACHINE LEARNING ALGORITHMS



04. Gradient Boosting

Advantages

- Perform well with hyperparameter tuning
- No data preprocessing required

Disadvantages

- Prone to overfitting
- Sensitive to extreme values and noises

04. MODEL EVALUATION

	precision	recall	f1-score	support
2XL	1.00	1.00	1.00	6
L	1.00	1.00	1.00	8
M	1.00	1.00	1.00	9
S	1.00	1.00	1.00	8
XL	1.00	1.00	1.00	6
XS	1.00	1.00	1.00	8
accuracy			1.00	45
macro avg	1.00	1.00	1.00	45
weighted avg	1.00	1.00	1.00	45

1.0

```
[[3739  268   50   12  756    8]
 [ 207 2644  890  223  945   67]
 [  26 1033 2136  978  311  335]
 [   1  274  891 2276   38 1362]
 [ 707 1237  420  100 2335   34]
 [   0   19  195 1011    2 3666]]
```

	precision	recall	f1-score	support
2XL	0.80	0.77	0.79	4833
L	0.48	0.53	0.51	4976
M	0.47	0.44	0.45	4819
S	0.49	0.47	0.48	4842
XL	0.53	0.48	0.51	4833
XS	0.67	0.75	0.71	4893
accuracy			0.58	29196
macro avg	0.57	0.58	0.57	29196
weighted avg	0.57	0.58	0.57	29196

04. MODEL EVALUATION

05. CONCLUSION

- CROSS VALIDATION
- HYPERPARAMETER TUNING
- TRANSFORMATION & SCALING TECHNIQUES
- USING Z-SCORE
- RESAMPLE
- BINARY CLASSIFICATION ALGORITHMS

05. CONCLUSION

- BENEFITS OVER SIZE CHARTS
- FEATURES
- FEATURE ENGINEERING
- FEATURE SELECTION

THANK YOU!

Q&A