KHOA HỌC DỮ LIỆU

WATER QUALITY CLASSIFICATION

NHÓM 9

GIỚI THIỆU ĐỀ TÀI

WATER QUALITY CLASSIFICATION - KAGGLE

LÝ DO CHỌN ĐỀ TÀI

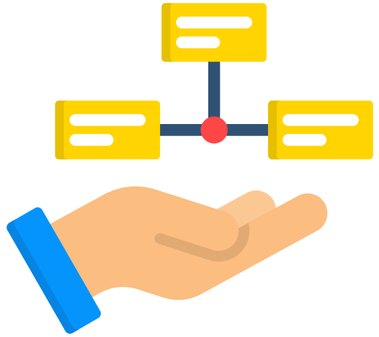
NƯỚC SẠCH QUAN TRỌNG VỚI SỰ SỐNG CON NGƯỜI

KIỂM ĐỊNH CHẤT LƯỢNG NƯỚC LÀ QUAN TRỌNG

ÁP DỤNG CÁC THUẬT TOÁN HỌC MÁY (ML)

HỆ THỐNG KIỂM ĐỊNH TỐN KÉM, TỐN THỜI GIAN, KHÔNG HIỆU QUẢ TRONG MỘT SỐ TÌNH HUỐNG

GIỚI THIỆU
9 THUỘC TÍNH



pH

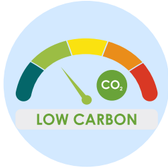
Hardness

Solids

Chloramines

Sulfate

Conductivity



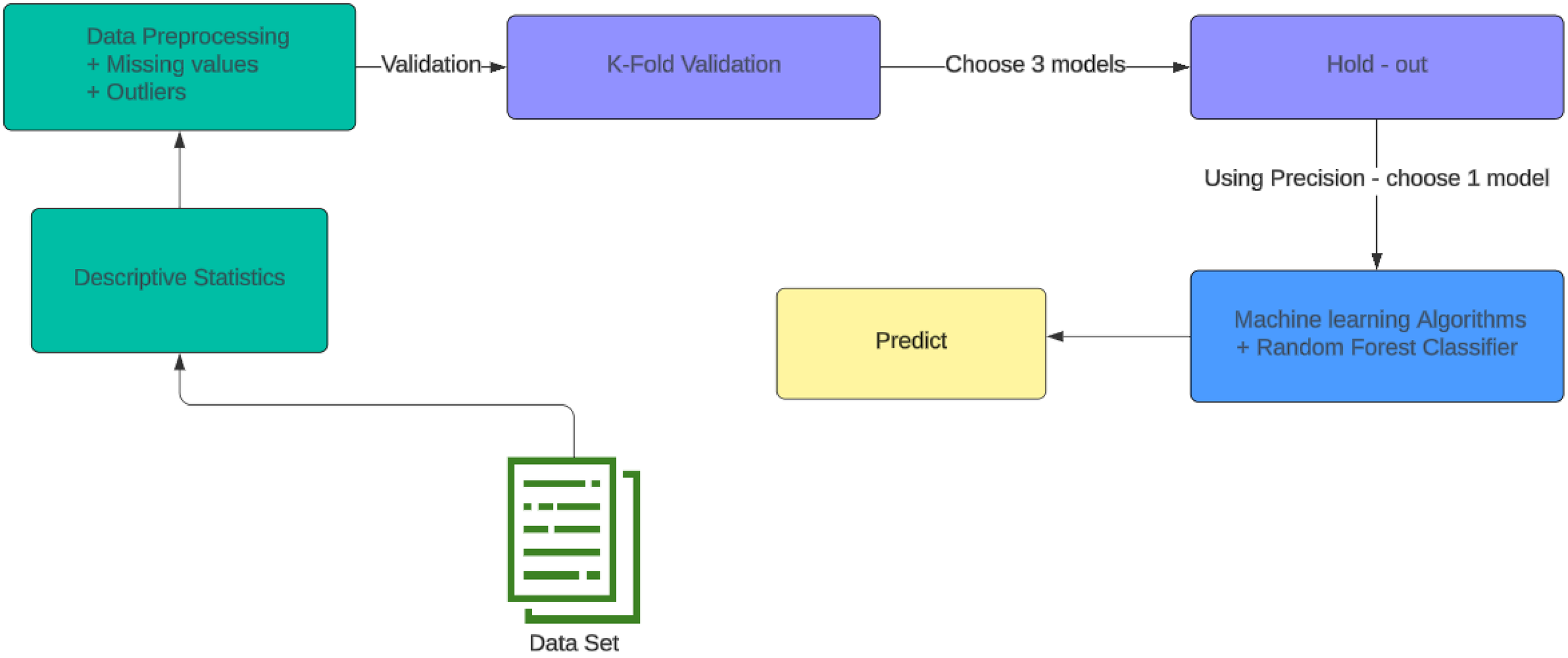
Organic carbon

Trihalomethanes

Turbidity

QUY TRÌNH PHÂN TÍCH

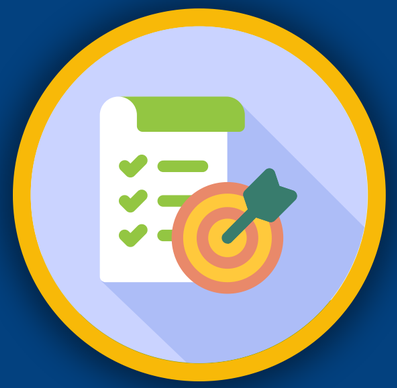
WATER QUALITY CLASSIFICATION - KAGGLE



ROOF &
PLUMBING

TIÊU CHÍ ĐÁNH GIÁ TÍNH HIỆU QUẢ

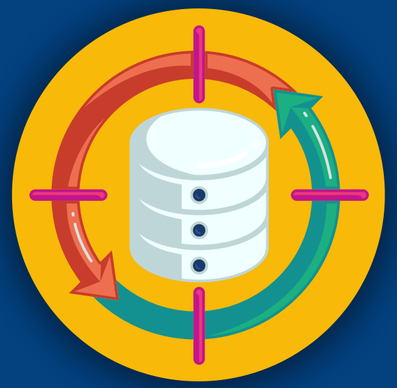
Sử dụng mô hình học máy dựa vào **ma trận nhầm lẫn (confusion matrix)** để đưa ra những đánh giá từ đó đưa ra mô hình có độ chính xác cao nhất.



PRECISION

RECALL

ACCURACY



F1 - SCORE

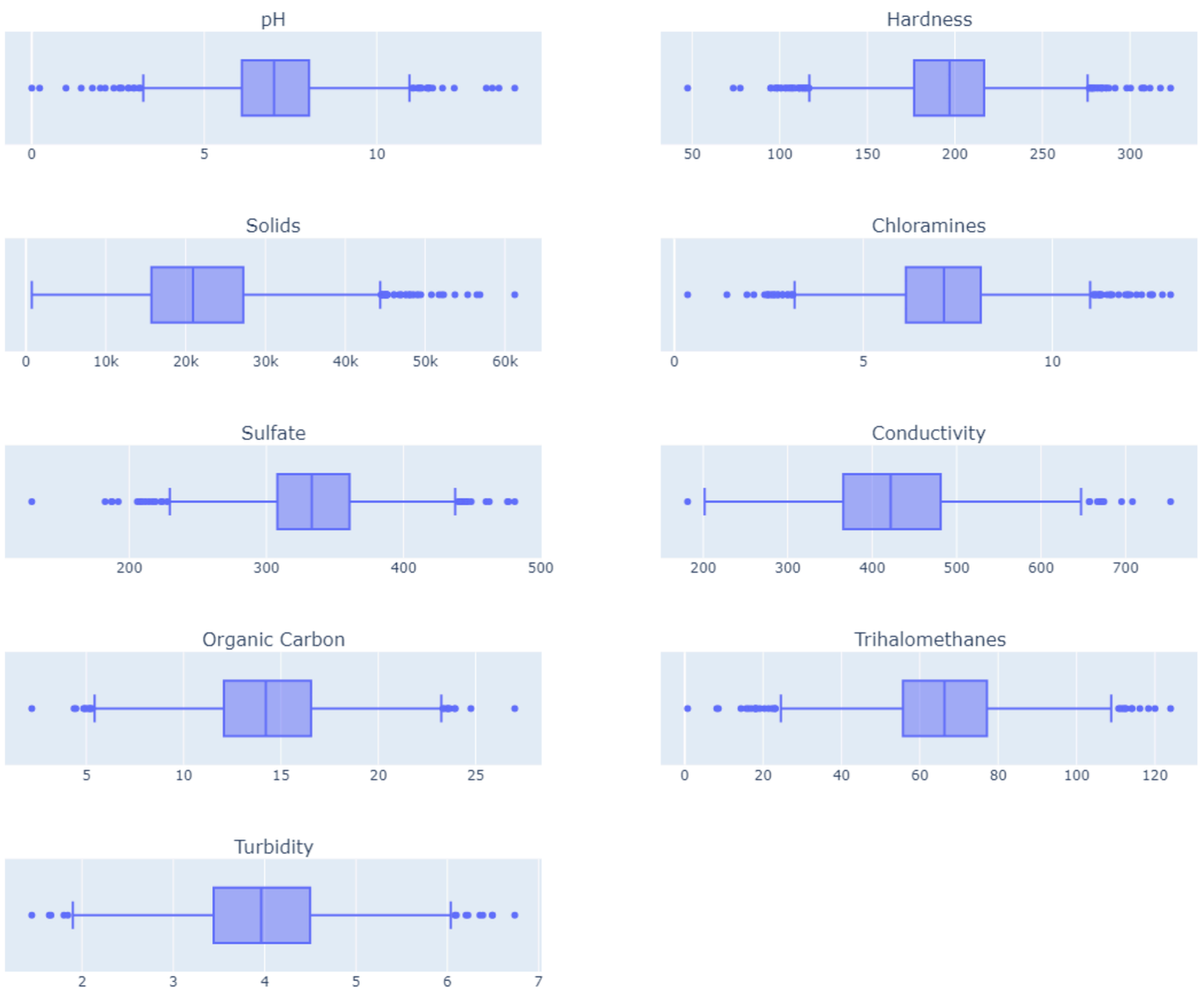
TỔNG QUAN DỮ LIỆU

ĐÁNH GIÁ SƠ BỘ

- **BẢNG THỐNG KÊ MÔ TẢ**

Sử dụng **Python** để tạo "Bảng thống kê mô tả" của các biến trong dataset.

	count	mean	std	min	25%	50%	75%	max
ph	2234.000000	7.071024	1.588089	0.000000	6.089577	7.022285	8.035837	14.000000
Hardness	2620.000000	196.200911	33.137494	47.432000	176.734160	196.926487	216.657847	323.124000
Solids	2620.000000	21973.377381	8702.501618	728.750830	15715.740730	20926.882155	27216.632550	61227.196008
Chloramines	2620.000000	7.130793	1.584109	0.352000	6.125502	7.131972	8.105680	13.127000
Sulfate	1999.000000	334.040923	41.573027	129.000000	307.992545	333.073546	360.601377	481.030642
Conductivity	2620.000000	425.749573	81.038498	181.483754	365.842780	421.884968	481.089353	753.342620
Organic_carbon	2620.000000	14.292469	3.272972	2.200000	12.067417	14.220645	16.541731	27.006707
Trihalomethanes	2495.000000	66.194111	16.186110	0.738000	55.711220	66.299162	77.127254	124.000000
Turbidity	2620.000000	3.965649	0.783198	1.450000	3.441806	3.962234	4.497554	6.739000
Potability	2620.000000	0.394275	0.488788	0.000000	0.000000	0.000000	1.000000	1.000000



TỔNG QUAN DỮ LIỆU

ĐÁNH GIÁ SƠ BỘ

- **DISTRIBUTION OF FEATURES**

  Sử dụng **Python** để tạo bảng biểu thị phân phối của các biến quan sát trong dataset.

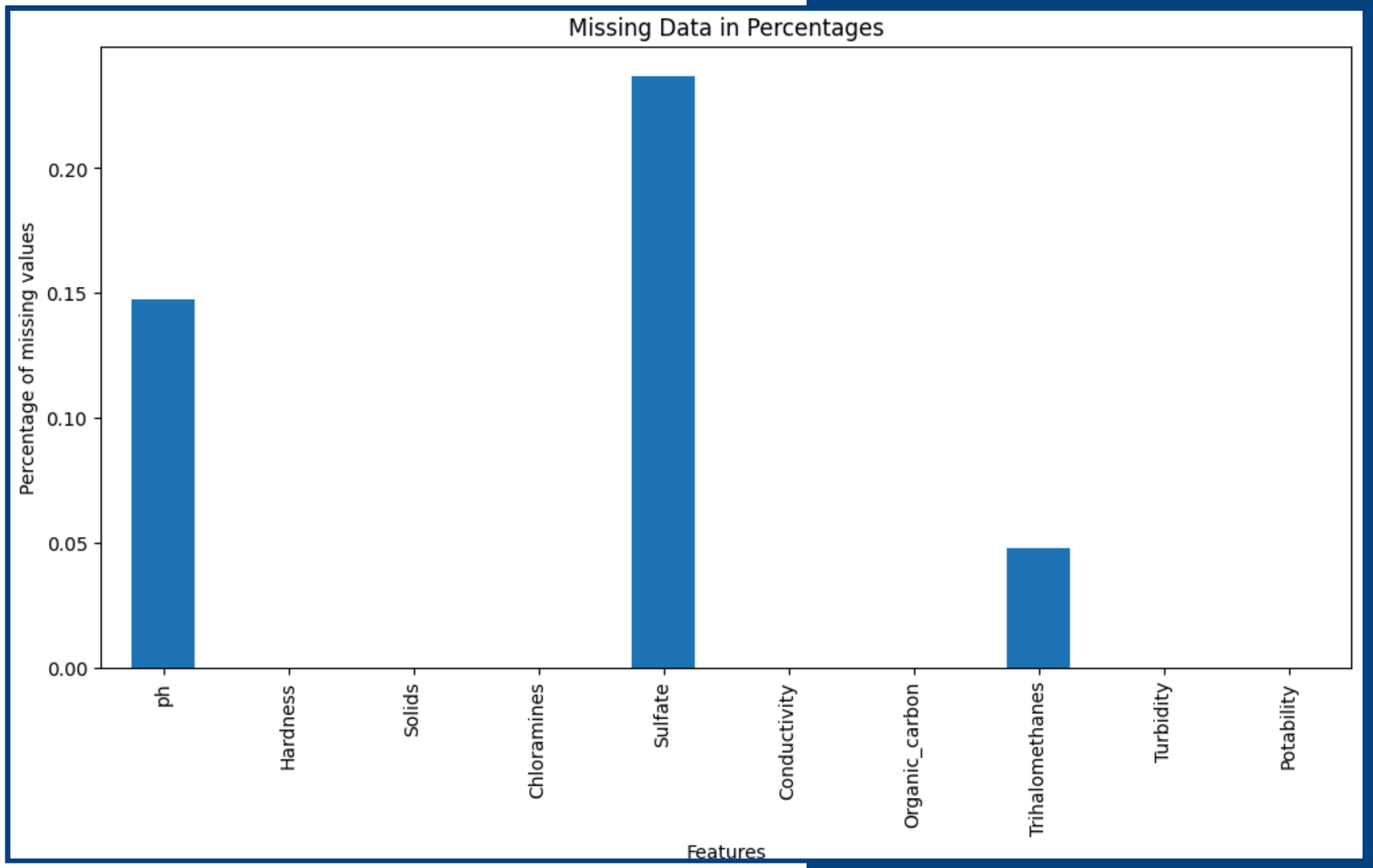
TIỀN XỬ LÝ
MISSING VALUES

14.7% **pH**

23.7% **Sulfate**

4.8% **Trihalomethanes**

Phương pháp giải quyết: Thay thế missing values của mỗi quan sát bằng trung vị (median) của chúng.



(*) Missing values: những giá trị bị thiếu/không được điền/không được cập nhật vào dataset

TIỀN XỬ LÝ OUTLIERS

QUY TRÌNH XỬ LÝ

(*)Outliers là những điểm dữ liệu khác biệt đáng kể so với các quan sát khác trong tập dữ liệu.

TÍNH GIÁ TRỊ IQR = Q3 - Q1 → **TÍNH GIỚI HẠN SÀN VÀ GIỚI HẠN TRẦN**

lower = Q1 - 1.5*IQR
upper = Q3 + 1.5*IQR

Bình thường ~ [lower; upper]
Outlier: còn lại

QUY TRÌNH ĐÁNH GIÁ MÔ HÌNH

K-FOLD CROSS VALIDATION

CHIA TẬP DỮ LIỆU THÀNH K PHẦN (K=5) → TRAIN MÔ HÌNH K LẦN → DỰA VÀO KẾT QUẢ ĐÁNH GIÁ VÀ LỰA CHỌN MÔ HÌNH

KẾT QUẢ ĐÁNH GIÁ:

TRUNG BÌNH CỘNG KẾT QUẢ ĐÁNH GIÁ CỦA K LẦN TRAIN

LỰA CHỌN MÔ HÌNH:

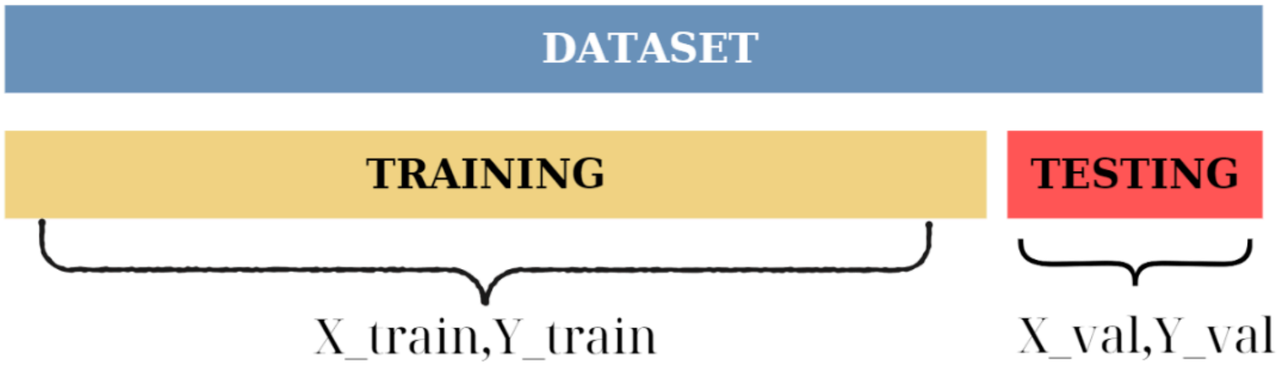
CHỌN 3 MÔ HÌNH CÓ KẾT QUẢ CAO NHẤT

CHẠY MÔ HÌNH
TRAINING DATA

Chia Dataset thành 2 phần:

- Training data (80%)
- Test data (20%)

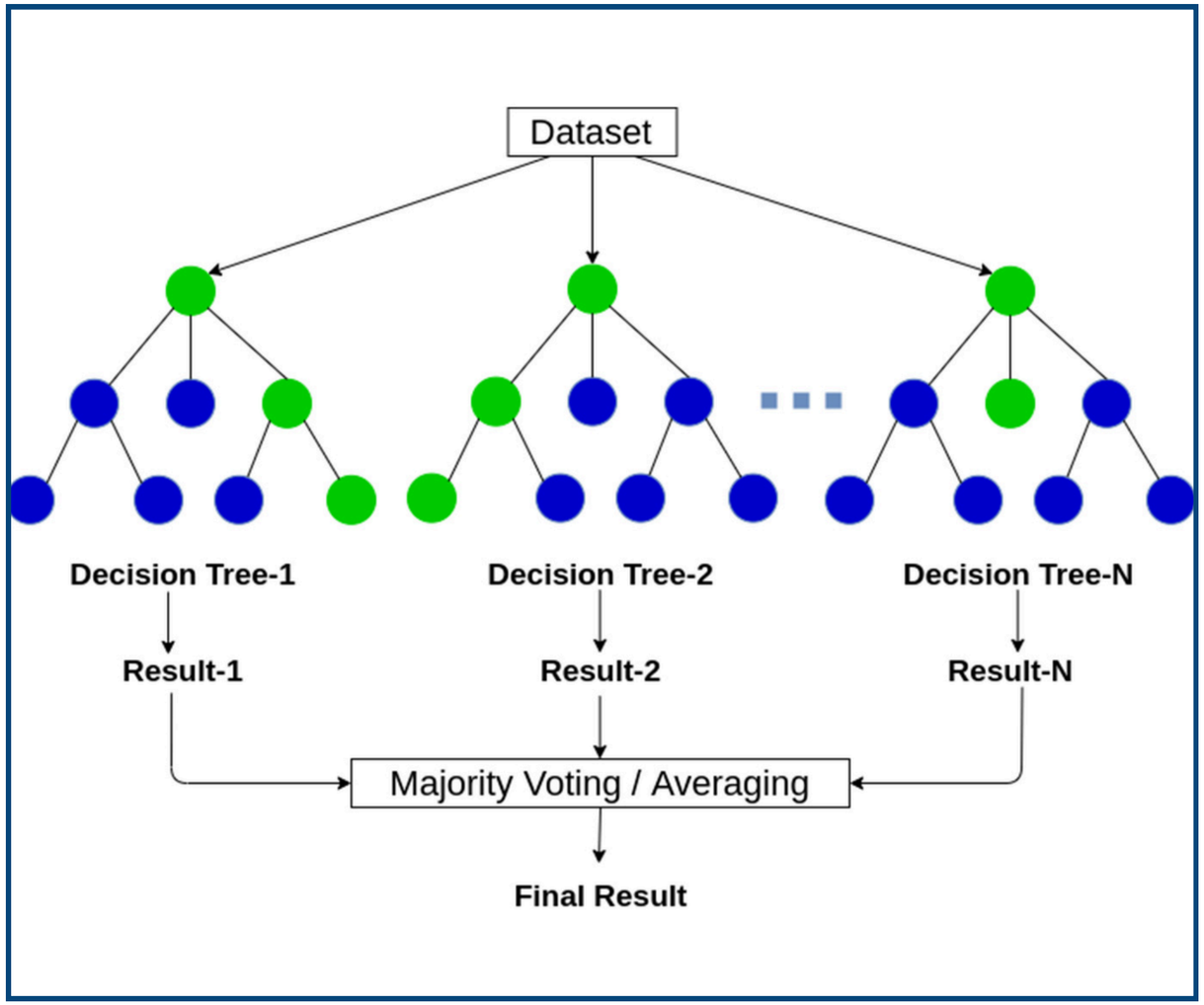
Mục đích: nếu lấy 100% dữ liệu vào train thì có thể khiến mô hình bị overfitting hoặc underfitting.



Để đánh giá xem nên lựa chọn mô hình nào, ta xem xét chỉ số precision và recall trong bảng sau:

	Precision	Recall
RandomForestClassifier	0.75	0.88
DecisionTreeClassifier	0.73	0.88
XGBClassifier	0.66	0.81

MÔ HÌNH MÁY HỌC
DECISION TREE - RANDOM FOREST



KHÁI NIỆM

(*) DECISION TREE:

- Thuật toán học có giám sát
- Cấu trúc dạng cây:
  - + Mỗi nút tương ứng một thuộc tính
  - + Mỗi nhánh tương ứng một giá trị thuộc tính
  - + Mỗi lá đại diện một quyết định/dự đoán

(**)RANDOM FOREST:

- Bộ phân loại chứa một số cây quyết định trên các tập con khác nhau
- Lấy dự đoán từ mỗi cây, dựa trên đa số phiếu dự đoán => kết quả cuối cùng.

FOR ALL YOUR ATTENTION

THANK YOU

CẢM ƠN THẦY VÀ CÁC BẠN ĐÃ LẮNG NGHE
PHẦN THUYẾT TRÌNH CỦA NHÓM 9.

www.reallygreatsite.com