

# DATA SCIENCE FOR ALL

---

Taweesak Samanchuen

21/1/2023

# ข้อมูล



# Data Analytics

**Data Analytic** คือ ศาสตร์ของการใช้ข้อมูลต่าง ๆ จากที่ต่าง ๆ มาวิเคราะห์ร่วมกันเพื่อวัตถุประสงค์บางประการ เช่น เพิ่มความสามารถในการแข่งขัน เพิ่มยอดขาย และเกิดความเข้าใจลูกค้า โดยออกมาในรูปแบบของรายงานผลการวิเคราะห์

- ระดับของ Data Analytics

- Descriptive Analytics วิเคราะห์ให้รู้ว่าเกิดอะไรขึ้น
- Diagnostic Analytics วิเคราะห์ต่อให้รู้ว่าสิ่งนั้นเกิดขึ้นเพราะอะไร
- Predictive Analytics แล้วอีกหน่อยจะเกิดอะไรขึ้นได้อีกบ้าง
- Prescriptive Analytics ถ้าเราทำแบบนี้แล้วจะเกิดอะไรขึ้นได้อีกบ้าง

# Data Science



# DATA SCIENCE

---

# Data Science

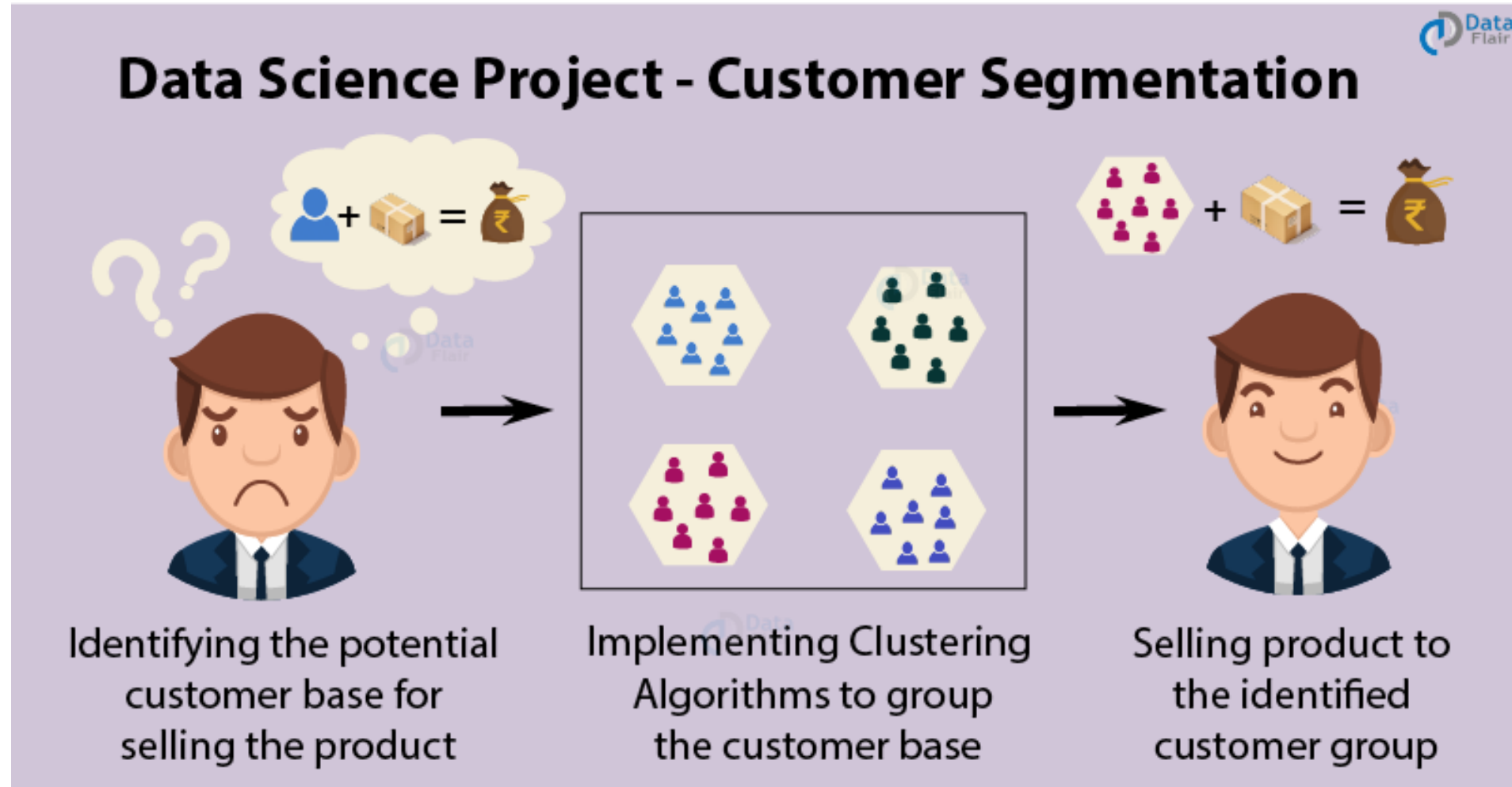
Data Science (วิทยาการข้อมูล) คือศาสตร์ที่ว่าด้วยการนำข้อมูลที่มีมาหาองค์ความรู้ใหม่ ด้วยวิธีการต่าง ๆ เพื่อเพิ่มความสามารถในการแข่งขันให้กับองค์กร โดยมีผลลัพธ์เป็น รายงานหรือระบบการทำงานอัตโนมัติ

# Data Science Project: Movie Recommendation System



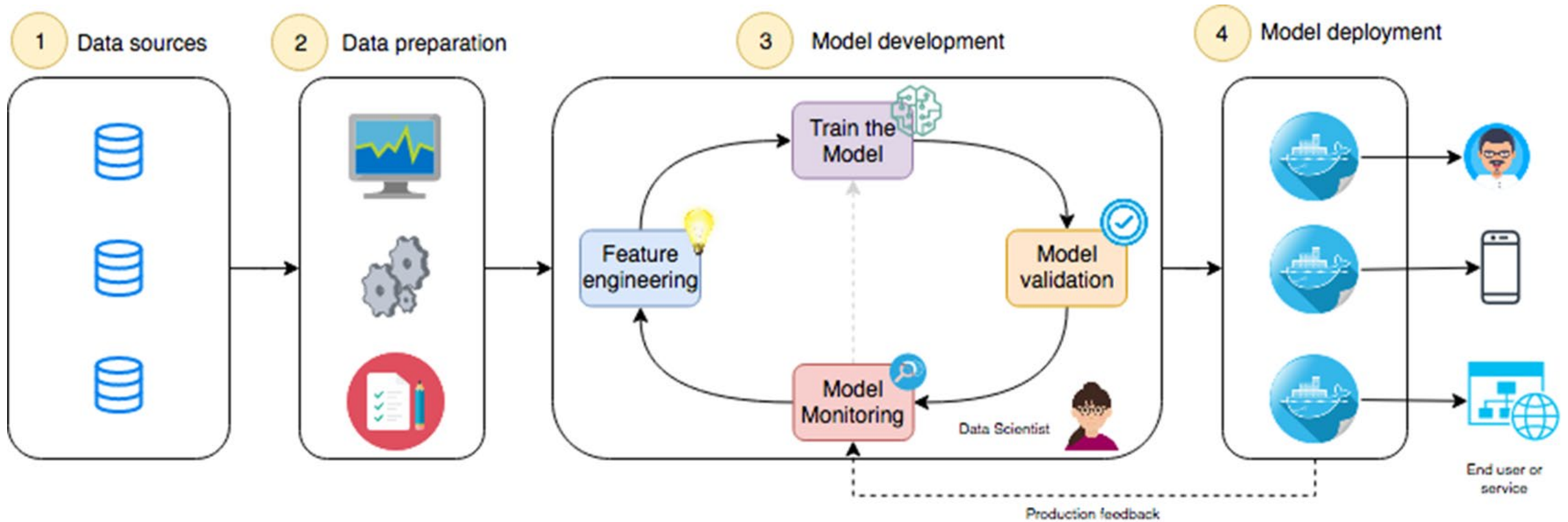


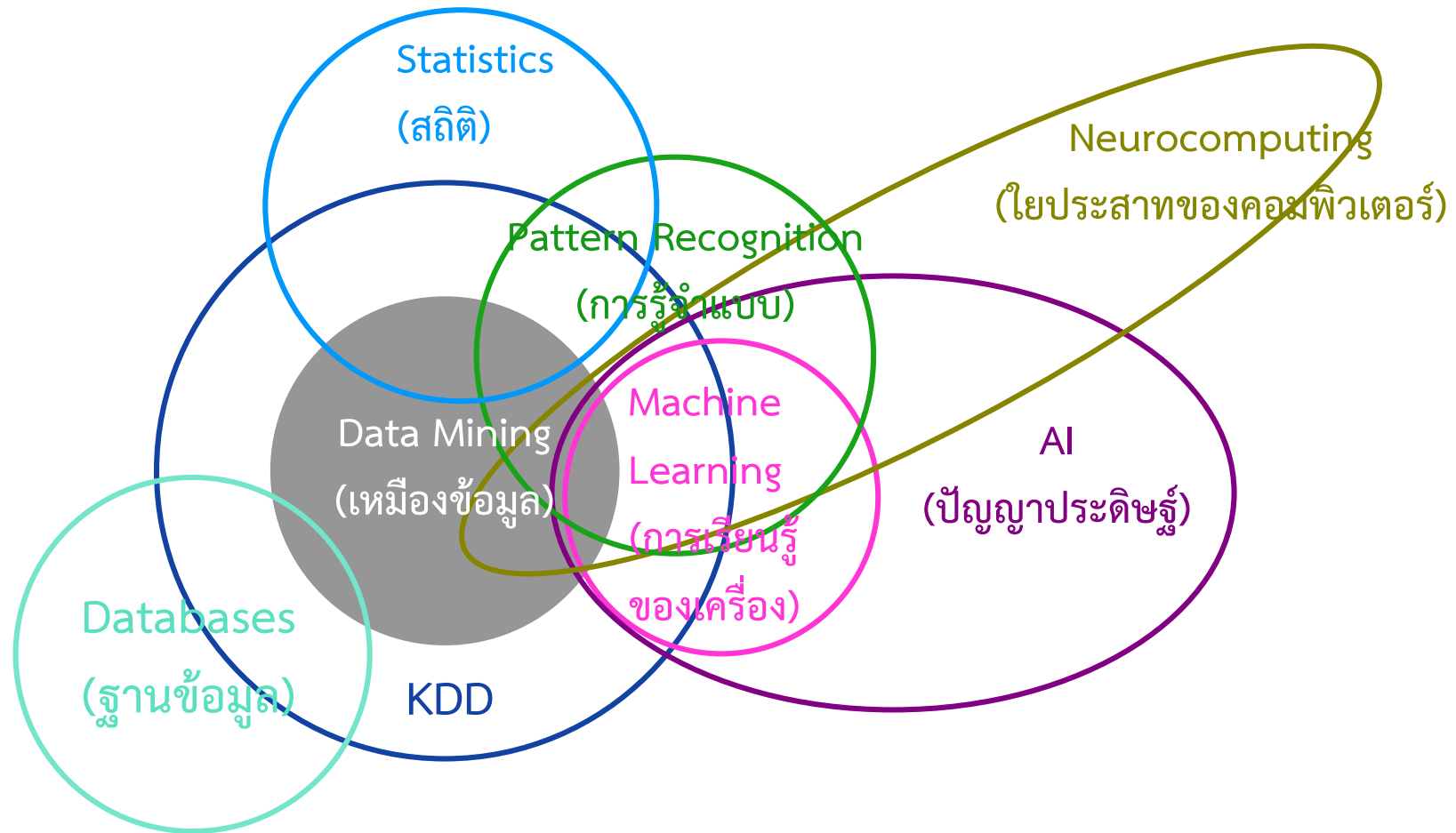
# Data Science Project : Customer Segmentation








# System Overview



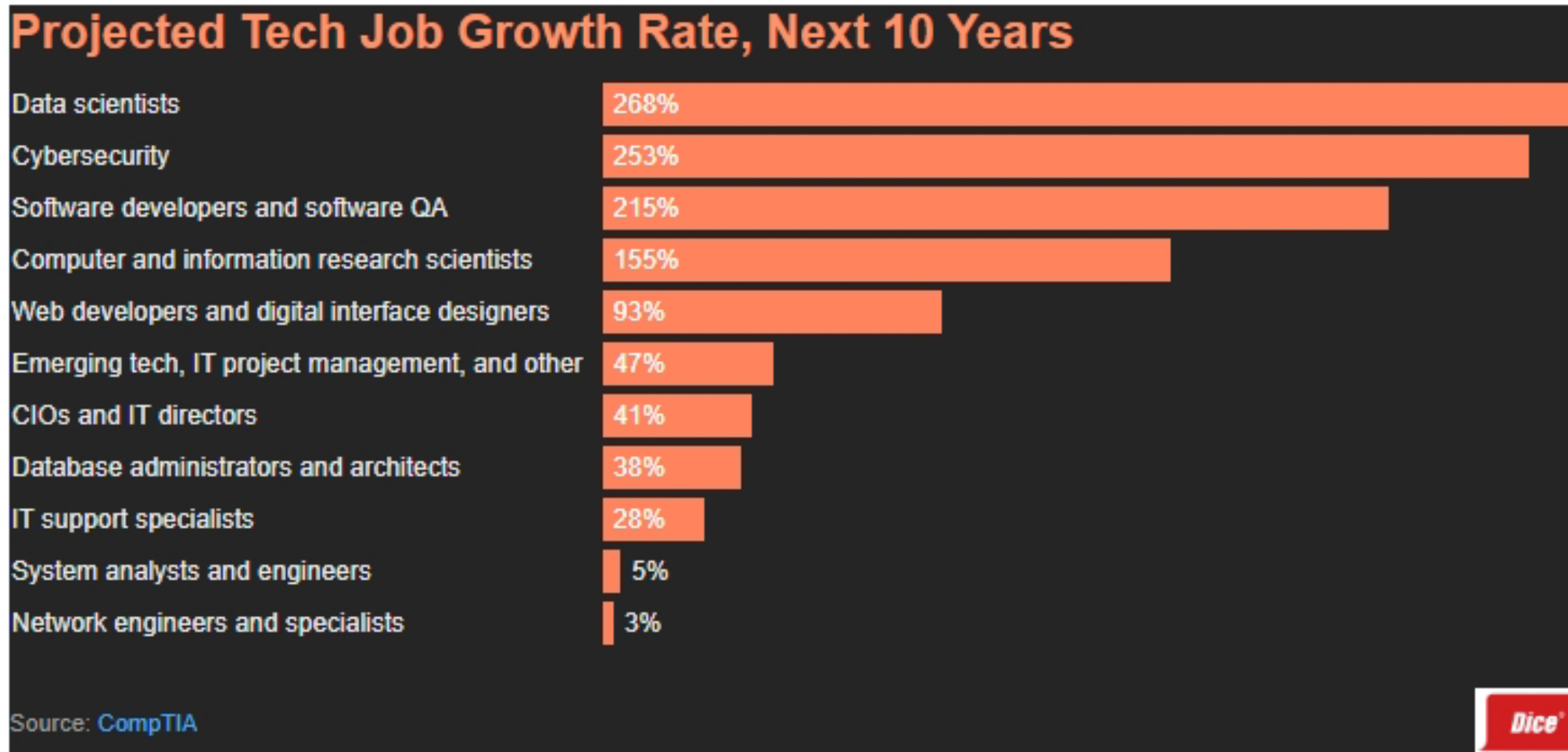


- KDD (Knowledge Discovery in Database Process) : กระบวนการการค้นพบความรู้ในการวิเคราะห์ข้อมูล
- AI (Artificial Intelligence) : ปัญญาประดิษฐ์

# อาชีพใน Data Science

<p>Data Engineer</p> 	<p>Data Scientist</p> 	<p>Business and Data Analyst</p> 
<p>หน้าที่ (Responsibilities)</p>	<p>ออกแบบโมเดลจากข้อมูล เพื่อหาช่องทางใหม่ๆ ให้องค์กร</p>	<p>วิเคราะห์ และออกแบบการนำเสนอข้อมูล เพื่อแก้ปัญหาในส่วนต่างๆ ของธุรกิจ</p>
<p>ออกแบบช่องทางของข้อมูล และวิธีการจัดเก็บ และใช้งาน</p>		

# คาดการณ์อาชีพ ปี 10 ข้างหน้า

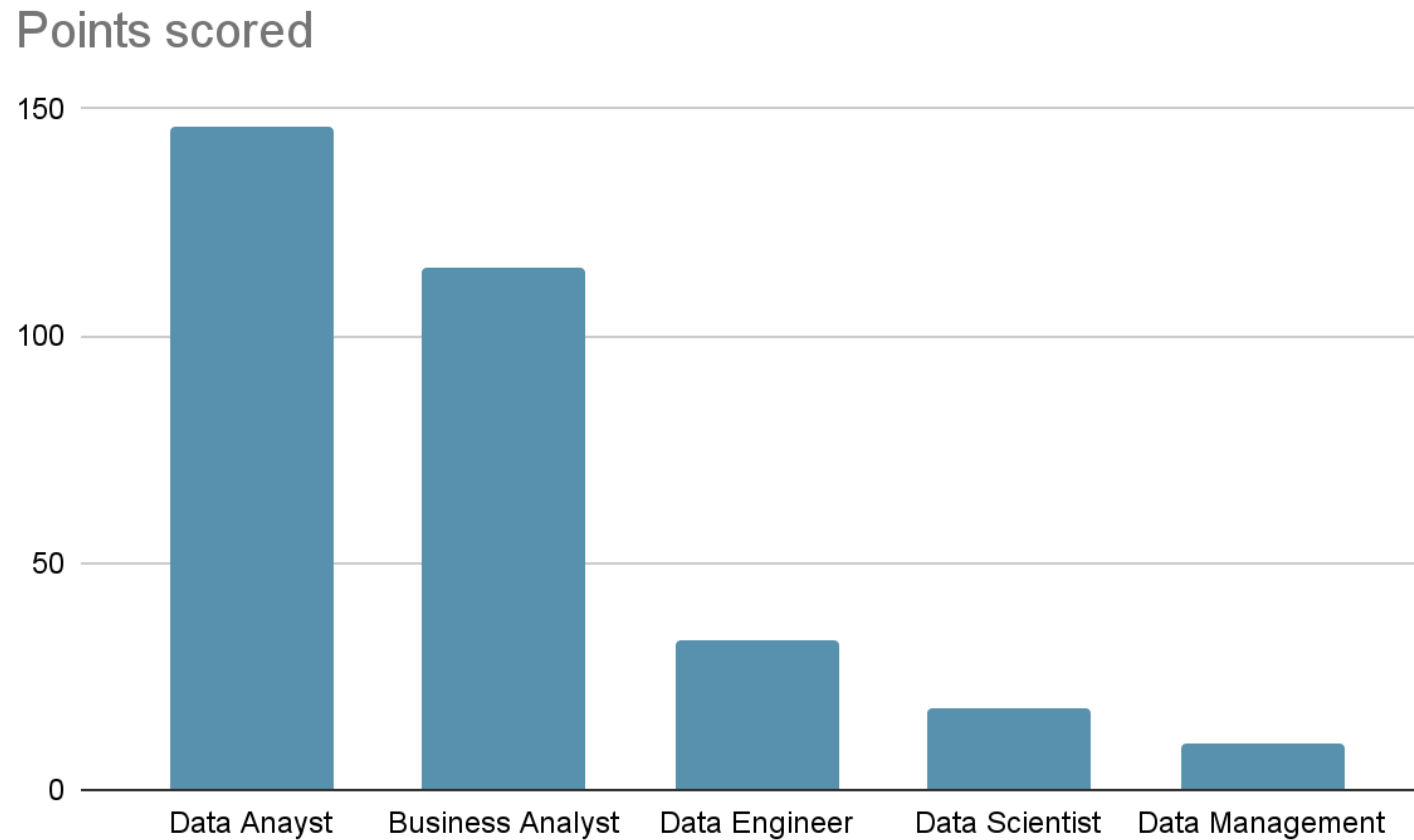


# ค่าตอบแทนของอาชีพใน USA

- Software Engineering Manager \$134,156.
- Mobile Applications Developer - \$111,468.
- Information Systems Security Manager - \$153,677.
- Database Manager - \$58,161.
- Data Security Analyst -\$71,226.
- Product Manager - \$100,000.
- Artificial Intelligence (AI) Engineer - \$110,000.
- Full-Stack Developer - \$106,000.
- Cloud Architect - \$107,000.
- DevOps Engineer - \$140,000.
- Blockchain Engineer - \$150,000.
- Software Architect - \$114,000.
- Big Data Engineer - \$140,000.
- Internet of Things (IoT) Solutions Architect - \$130,000.
- **Data Scientist - \$150,000.**

<https://www.simplilearn.com/highest-paying-tech-jobs-article>

# อาชีพที่รับสมัคร วันที่ 21/12/22



<https://docs.google.com/spreadsheets/d/1S8Vs4-j6DChPJkpkM48gJRUV4HB-yPrTytu2AcAgq1Y/edit?usp=sharing>

# จำนวนหลักสูตรด้าน DATA Science

- ปริณญาตรี 24 หลักสูตร
- ปริญาโท 14 หลักสูตร
- ระยะสั้น 39 หลักสูตร

[https://docs.google.com/spreadsheets/d/13JUftMrQIxp2MUZj-I\\_OeCV\\_QHaBPk1d/edit?usp=sharing&ouid=110135164357156798011&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/13JUftMrQIxp2MUZj-I_OeCV_QHaBPk1d/edit?usp=sharing&ouid=110135164357156798011&rtpof=true&sd=true)

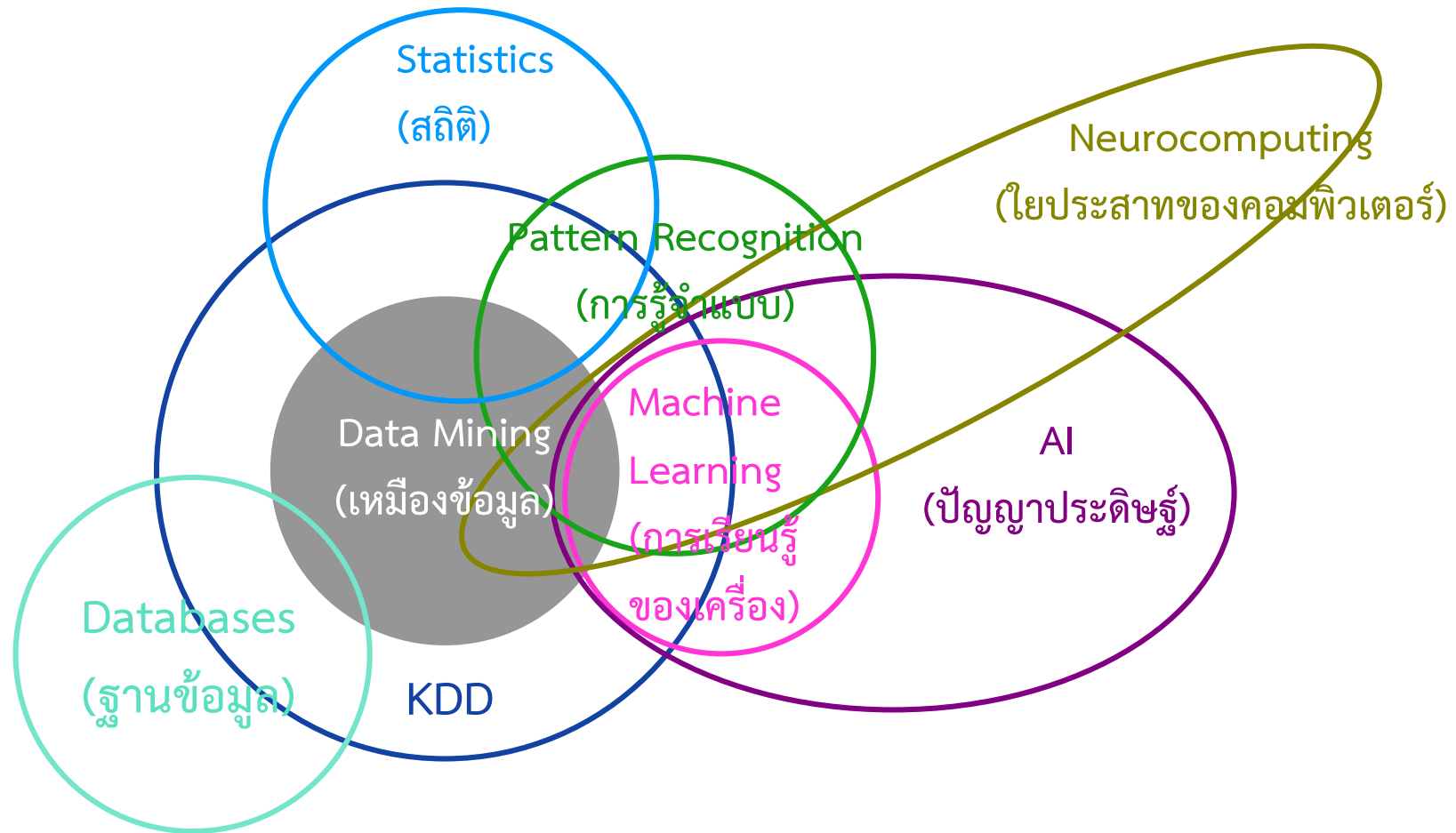


# MACHINE LEARNING

FOR DATA SCIENCE

---

Taweesak Samanchuen



- KDD (Knowledge Discovery in Database Process) : กระบวนการการค้นพบความรู้ในการวิเคราะห์ข้อมูล
- AI (Artificial Intelligence) : ปัญญาประดิษฐ์

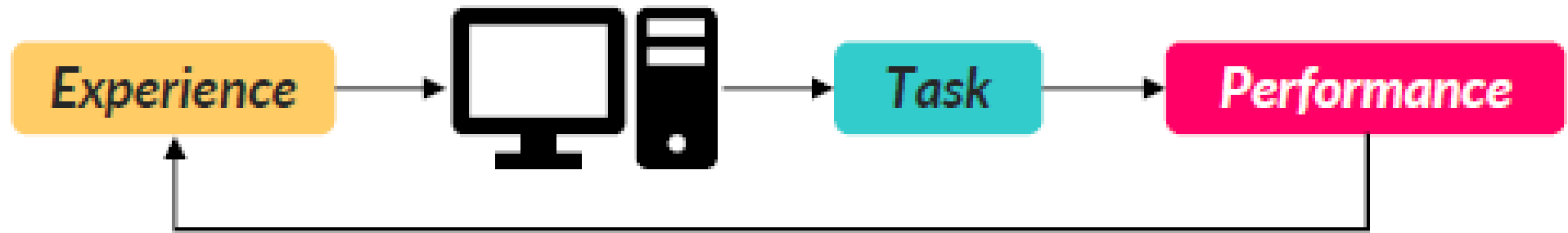
# ความหมายของ ML

- **Arthur Samuel (1959).** Machine Learning:  
Field of study that gives computers the ability to learn without being explicitly programmed.
- **Tom Mitchell (1998)** Well-posed Learning Problem:  
A computer program is said to learn from experience  $E$  with respect to some task  $T$  and some performance measure  $P$ , if its performance on  $T$ , as measured by  $P$ , improves with experience  $E$ .

## ความหมายของ ML (ต่อ)

- Machine Learning คือ กระบวนการที่ระบบคอมพิวเตอร์เรียนรู้การทำงานใด ๆ โดยใช้การฝึกจากข้อมูลที่อยู่ในอดีตและนำผลลัพธ์ของการฝึกมาปรับปรุงการทำงานให้ดีขึ้นไปเรื่อย ๆ

## ความหมายของ ML (ต่อ)



*Can we improve computer's ability to perform task T over time (without being explicitly programmed)?*

# ประเภทของ ML

- 1) Supervised learning
- 2) Unsupervised learning
- 3) Reinforcement learning
- 4) Recommender systems

# SUPERVISED LEARNING

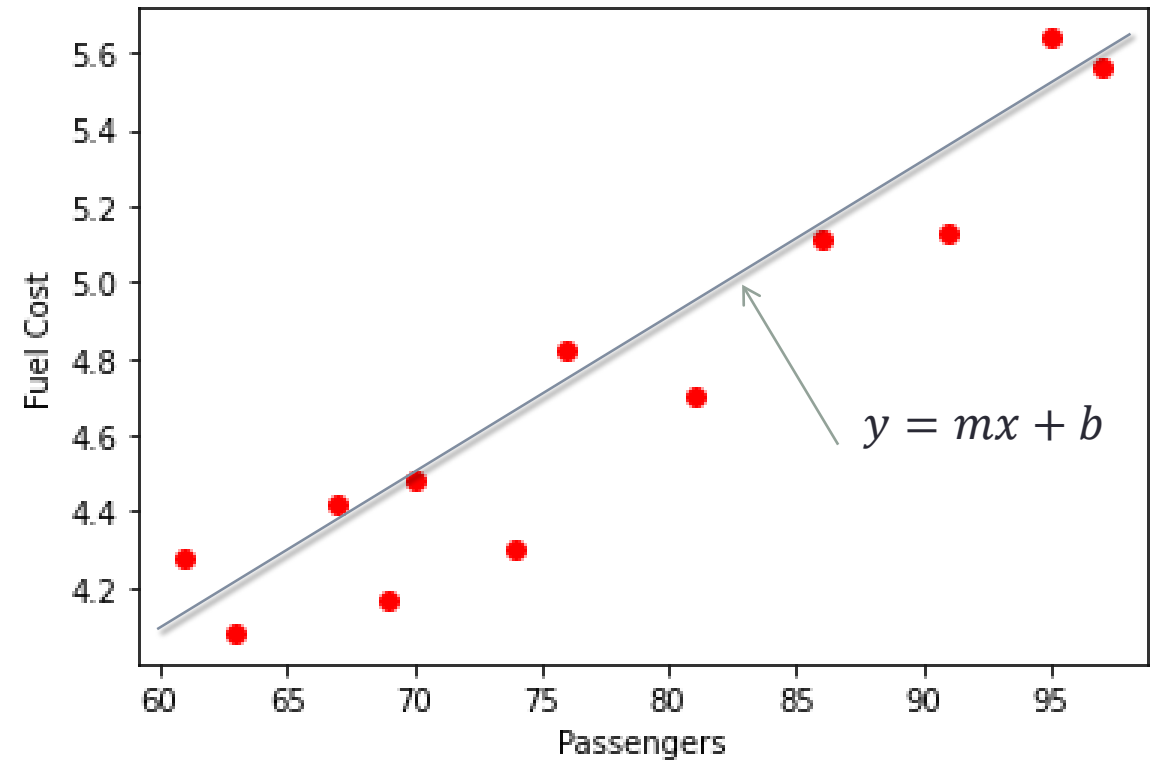
---



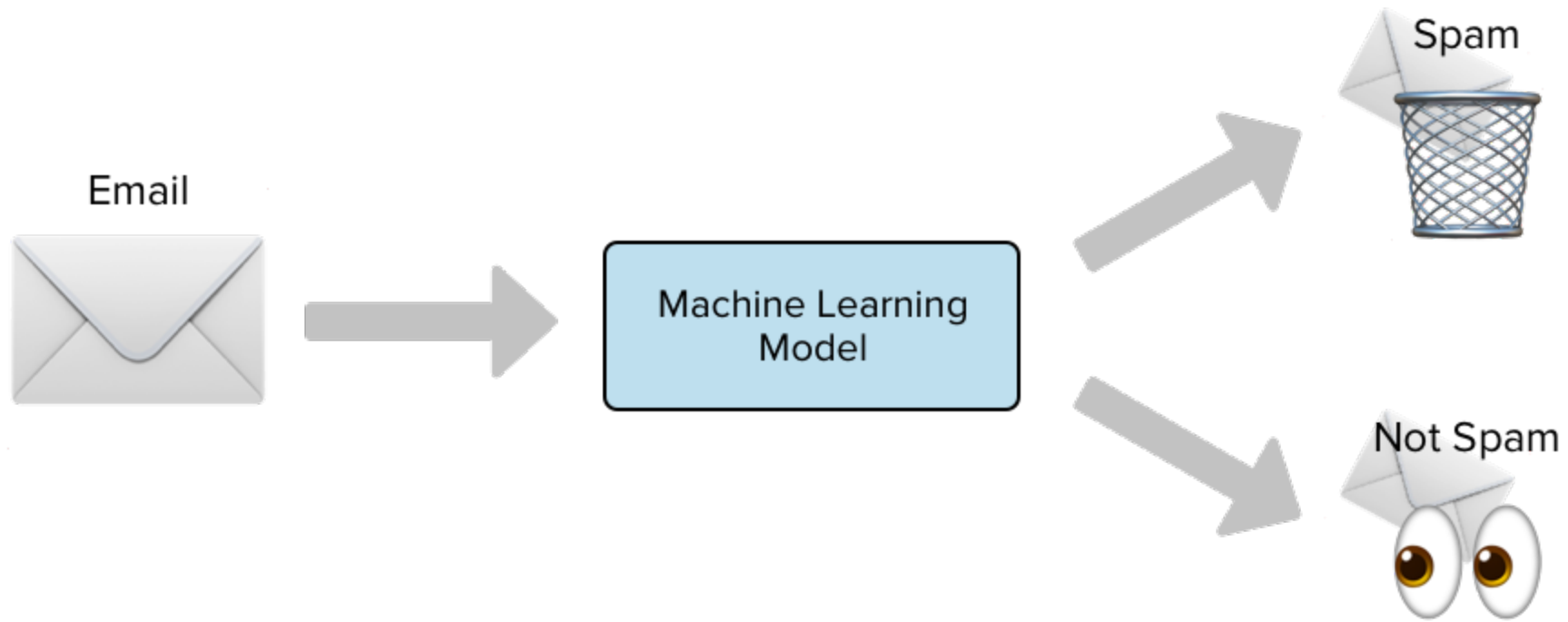
# Supervised Learning

- เป็นกลุ่ม ML ที่มีการระบุคำตอบของข้อมูลที่นำมาสร้างแบบจำลอง เช่น
  - กรณีการทำนายค่าเชื้อเพลิงจากจำนวนผู้โดยสาร
  - กรณีการคัดแยก E-mail ที่รับมานั้นเป็น Spam หรือไม่

# Airline Cost Data



# Spam Prediction



# Supervised Learning Types

- Regression

เป็น ML ที่มีการระบุคำตอบของข้อมูลที่นำมาสร้างแบบจำลอง (Model) เป็นแบบค่าต่อเนื่อง

เช่น การทำนายค่าเฉลี่ยพึงพอใจจากจำนวนผู้โดยสาร

- Classification เป็น ML ที่มีการระบุคำตอบของข้อมูลที่นำมาสร้างแบบจำลอง (Model)

เป็นแบบเซตคำตอบที่รู้สมาชิกของคำตอบอย่างแน่นอน

เช่น การคัดแยก E-mail ที่เป็น Spam

# UNSUPERVISED LEARNING

---

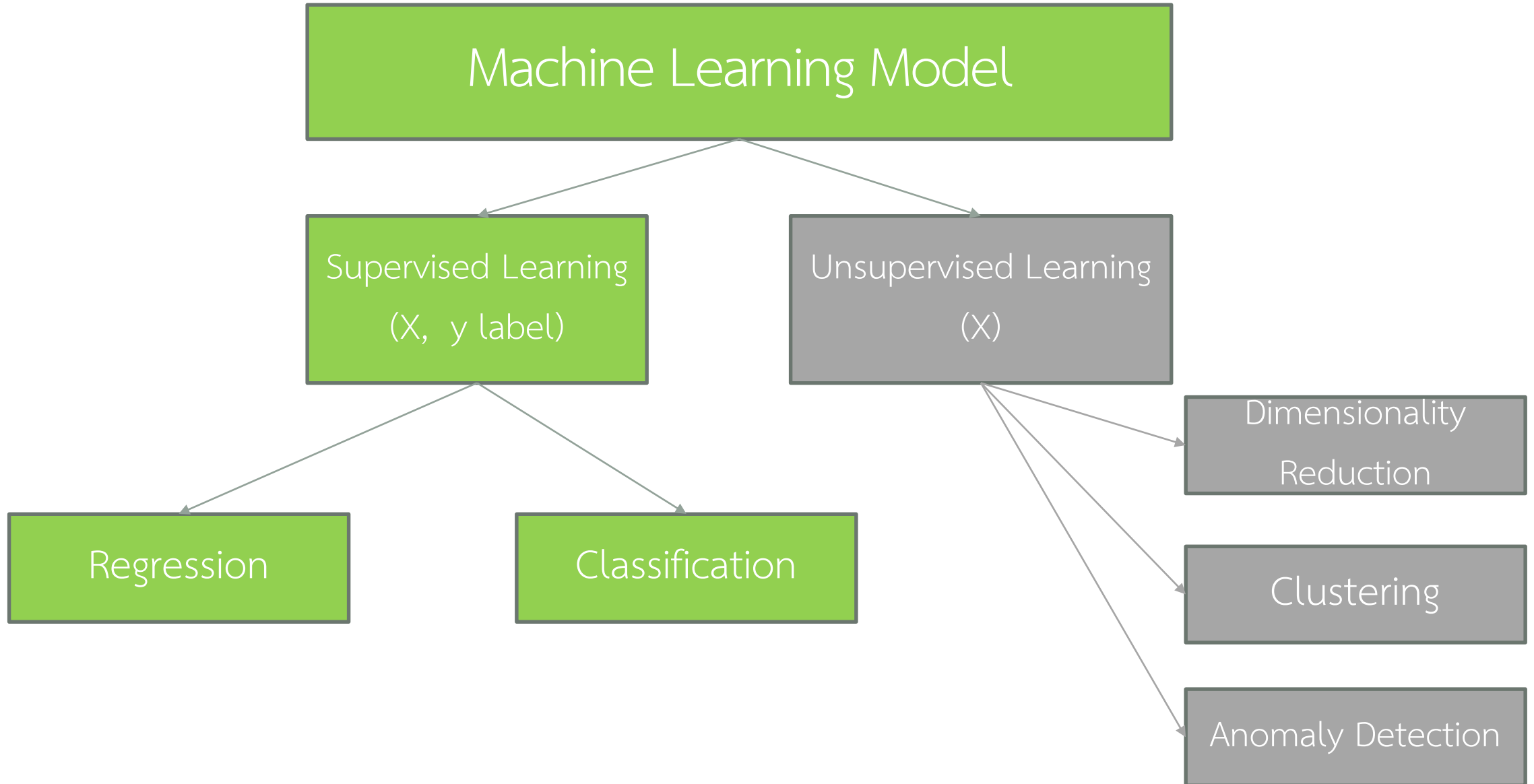
# Unsupervised Learning

- เป็นกลุ่ม ML ที่ไม่มีการระบุคำตอบของข้อมูลที่นำมาสร้างแบบจำลอง เช่น
  - การแบ่งกลุ่มลูกค้า
  - การแบ่งกลุ่ม E-mail
  - การลดจำนวนมิติข้อมูล
  - การตรวจจับธุรกรรมที่ผิดปกติ

# ประเภทของ Unsupervised Learning

- Clustering (K-Means)
- Anomaly Detection
- Dimensionality Reduction





# แบบฝึกหัด

ปัญหาต่อไปนี้เป็น ML แบบใด

1. การทำนายความสูงของคนจากน้ำหนักตัว
2. การทำนายชนิดของผลไม้ (Apple & Orange) จากขนาดและน้ำหนัก
3. แบ่งกลุ่มอีเมล
4. ตรวจสอบข้อความว่าเป็น hate speech ด้วยคอมพิวเตอร์

# INTRODUCTION TO RAPIDMINER STUDIO

---



# Gartner 2018 & 2020 Magic Quadrant for Data Science Platforms



# RapidMiner Products

The screenshot displays the RapidMiner Studio interface. On the left, a 'Process' tree lists various operators like 'Modeling (4/6)', 'Predictive (5/1)', 'Lury (2)', 'Default Model', 'k-NN', 'Random (2)', 'Naive Bayes (4/6)', 'Trees (5)', 'Decision Tree', 'Random Forest', 'Gradient Boosted Trees', 'CHAD', 'G2', 'Decision Dump', 'Decision Tree (Single-Seed)', 'Random Tree', 'Rules (1)', 'Rule Induction', 'Single Rule Induction', 'Single Rule Induction (Single Attribute)', 'Subgroup Discovery', 'Tree to Rules', 'Neural Nets (4)', 'Deep Learning', 'Neural Net', 'AutoML', 'Perceptron', 'Functions (3)', 'Logistic Regression (3)', 'Logistic Regression', 'Logistic Regression (SVM)', 'Logistic Regression (Evolutionary)', 'Support Vector Machines (3)', 'Support Vector Machines (2)', 'Select Attributes', 'Create Model', 'P-Create Model', 'Apply Model', and 'Filter Examples'. The main workspace shows a workflow with operators: 'Retrieve Customer...', 'Set Rule', 'Numerical to String...', and 'Cross Validation'. The 'Parameters' panel on the right shows settings for 'logit' and 'logit'. A 'Help' panel at the bottom right provides information about 'Gradient Boosted Trees'.

**Robust machine learning library**

RapidMiner Studio

The screenshot displays the RapidMiner AI Hub interface. It features a 'Hello auth0! Welcome to the RapidMiner AI Hub.' message. The interface is divided into several sections: 'APPS' (Projects and repositories, Executions and schedules, Grafana, Platform Administration), 'ADMIN TOOLS' (Platform Administration), 'RESOURCES' (RM Documentation, RM Community), and 'Identity and Security'. The 'RapidMiner Go' section highlights 'AutoML, built for domain experts, business users, and analysts.' and 'JupyterHub' (Notebooks at hand, integrated with RapidMiner repositories).

RapidMiner AI-Hub

The screenshot displays the RapidMiner Go interface, specifically the 'Select your inputs' step. It shows a table of column quality metrics for 18 columns. The table includes columns for Name, Quality, Correlation, ID-ness, Stability, and Missing. The 'Quality' column uses color-coded tags: 'High ID-ness' (red), 'Highly stable' (red), 'Low correlation' (yellow), and 'Low correlation' (yellow). The 'ID-ness' column shows percentages, and the 'Stability' column shows percentages. The 'Missing' column shows percentages.

Name	Quality	Correlation	ID-ness	Stability	Missing
loan_id	High ID-ness	0.07%	99.00%	0.10%	0.00%
emp_title	High ID-ness	38.55%	89.50%	0.33%	6.20%
handship_flag	Highly stable	?	0.05%	100.00%	0.00%
total_acc	Low correlation	0.00%	3.05%	4.65%	0.00%
issue_d	Low correlation	0.00%	?	0.16%	0.00%
loan_amnt		1.67%	16.29%	6.63%	0.00%
int_rate		17.17%	?	8.79%	0.00%
installment		0.55%	?	0.88%	0.00%
grade		18.88%	0.36%	37.23%	0.00%

RapidMiner Go

# RAPIDMINER STUDIO

---



## 1. Main Panel

## 2. Views Panel

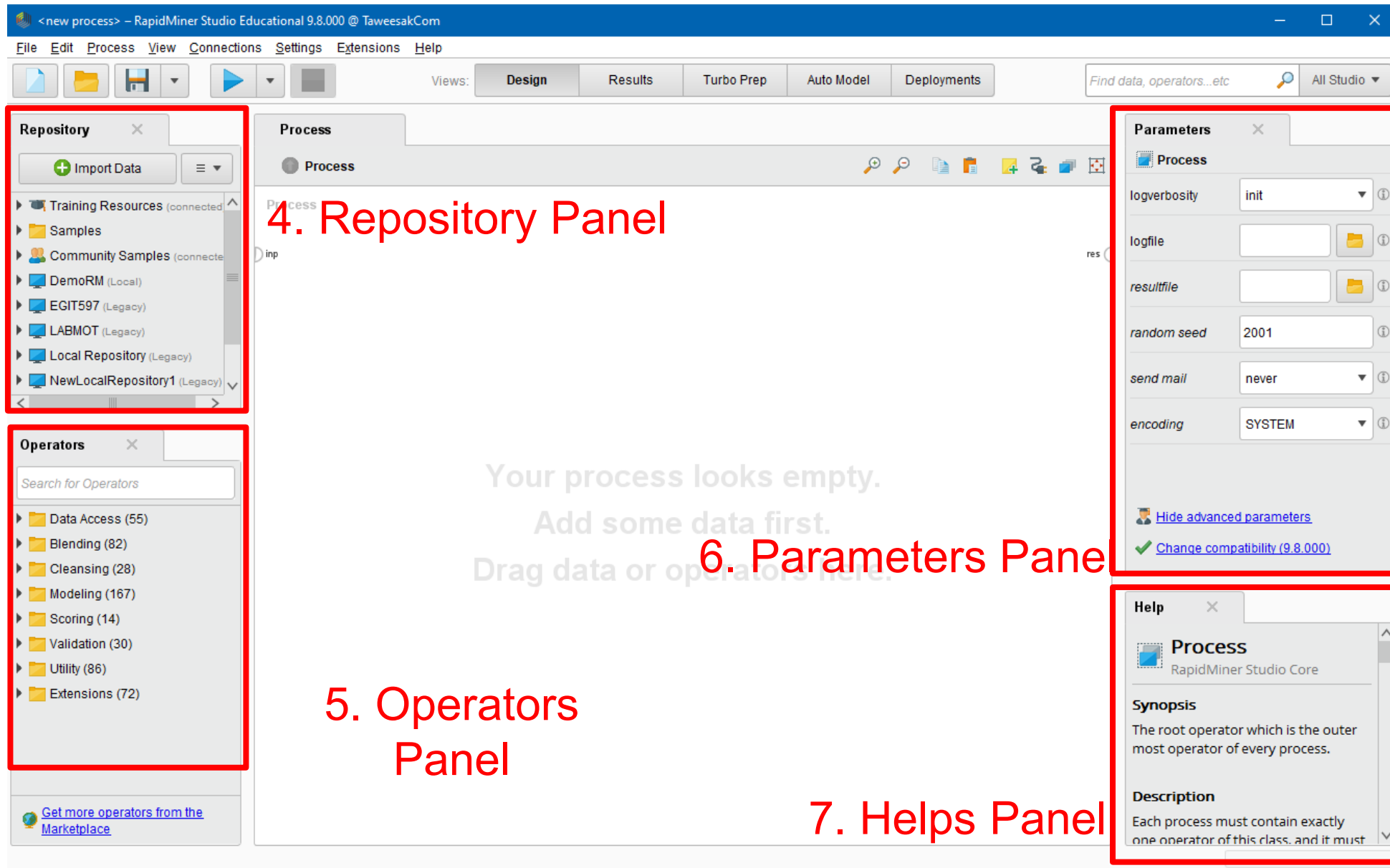
The screenshot shows the RapidMiner Studio interface. The top menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, and Help. Below the menu bar is a toolbar with icons for file operations and a 'View' dropdown menu. The 'View' dropdown menu is highlighted with a red box and contains the following options: Design, Results, Turbo Prep, Auto Model, and Deployments. The main workspace is divided into three panels:

- Main Panel (Left):** Contains the 'Repository' and 'Operators' panels. The 'Repository' panel shows a tree view of data sources, including 'Training Resources', 'Samples', 'Community Samples', 'DemoRM', 'EGIT597', 'LABMOT', 'Local Repository', and 'NewLocalRepository1'. The 'Operators' panel shows a search bar and a list of operator categories: Data Access (55), Blending (82), Cleansing (28), Modeling (167), Scoring (14), Validation (30), Utility (86), and Extensions (72).
- Views Panel (Top Right):** Contains the 'Parameters' panel, which shows settings for the 'Process' operator, including 'logverbosity', 'logfile', 'resultfile', 'random seed', 'send mail', and 'encoding'. It also includes links for 'Hide advanced parameters' and 'Change compatibility (9.8.000)'.
- Process Panel (Center):** Contains the 'Process' panel, which is currently empty. It displays the text: 'Your process looks empty. Add some data first. Drag data or operators here.'

The 'Process' panel is highlighted with a red box. The 'Parameters' panel is also highlighted with a red box. The 'Main Panel' is highlighted with a red box.

## 3. Process Panel





- ▶ **Data Access (58)**
- ▶ Blending (77)
- ▶ Cleansing (26)
- ▶ Modeling (129)
- ▶ Scoring (9)
- ▶ Validation (28)
- ▶ Utility (87)
- ▶ Extensions (796)

- ▼ **Blending (77)**
  - ▶ Attributes (44)
  - ▶ Examples (11)
  - ▶ Table (11)
  - ▶ Values (11)
- ▼ **Cleansing (26)**
  - ▶ Normalization (3)
  - ▶ Binning (5)
  - ▶ Missing (6)
  - ▶ Duplicates (1)
  - ▶ Outliers (4)
  - ▶ Dimensionality Reduction (7)

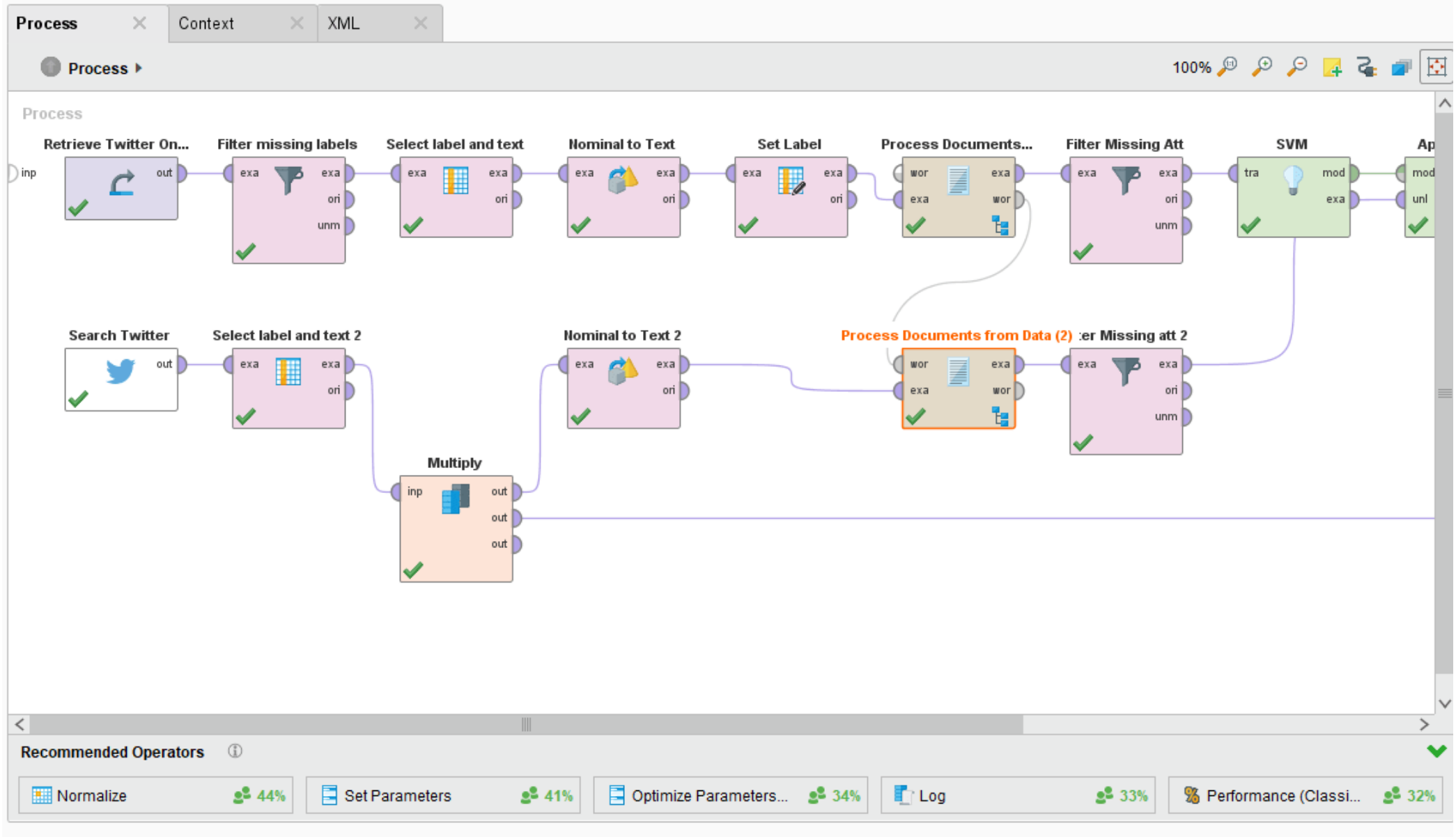
- ▼ **Data Access (58)**
  - ▶ Files (21)
  - ▶ Database (3)
  - ▶ NoSQL (9)
  - ▶ Hadoop (1)
  - ▶ Applications (10)
  - ▶ Cloud Storage (8)
  - ▶ Retrieve
  - ▶ Store
  - ▶ Rename Repository Entry
  - ▶ Copy Repository Entry
  - ▶ Move Repository Entry
  - ▶ Delete Repository Entry

- ▼ **Modeling (129)**
  - ▶ Predictive (61)
  - ▶ Segmentation (13)
  - ▶ Associations (6)
  - ▶ Correlations (8)
  - ▶ Similarities (4)
  - ▶ Feature Weights (17)
  - ▶ Optimization (20)

- ▼ **Scoring (9)**
  - ▶ Confidences (8)
  - ▶ Apply Model
- ▼ **Validation (28)**
  - ▶ Performance (20)
  - ▶ Visual (3)
  - ▶ Cross Validation
  - ▶ Split Validation
  - ▶ Bootstrapping Validation
  - ▶ Wrapper Split Validation
  - ▶ Wrapper-X-Validation

- ▼ **Predictive (61)**
  - ▶ Lazy (2)
  - ▶ Bayesian (2)
  - ▶ Trees (9)
  - ▶ Rules (5)
  - ▶ Neural Nets (4)
  - ▶ Functions (8)
  - ▶ Logistic Regression (3)
  - ▶ Support Vector Machines (7)
  - ▶ Discriminant Analysis (3)
  - ▶ Ensembles (14)
  - ▶ Update Model
  - ▶ Group Models
  - ▶ Ungroup Models
  - ▶ Create Formula

- ▼ **Utility (87)**
  - ▼ **Scripting (5)**
    - ▶ Execute Script
    - ▶ Execute SQL
    - ▶ Execute Program
    - ▶ Execute Python
    - ▶ Execute R
  - ▶ Process Control (30)
  - ▶ Macros (5)
  - ▶ Files (11)
  - ▶ Annotations (4)
  - ▶ Logging (7)
  - ▶ Data Anonymization (2)
  - ▶ Random Data Generation (13)
  - ▶ Misc (6)
  - ▶ Multiply
  - ▶ Subprocess
  - ▶ Schedule Process
  - ▶ Execute Process



# LAB 1: แนะนำ PROGRAM RM

---

## LAB 2: การสร้าง REPOSITORY ใน RM

---

# LAB 3: การนำเข้าข้อมูลใน RM

---

# LAB 4: การสำรวจข้อมูลเบื้องต้น

---

# ประเด็นการตรวจสอบ

- ความหมายของข้อมูลโดยรวม
- ความของข้อมูลแต่ละ attribute
- คุณลักษณะ
- ความสมบูรณ์
- ความทันสมัย
- ความจำเป็นในการปรับปรุงข้อมูล
- ความเพียงพอของข้อมูล



# คุณสมบัติของข้อมูลใน RM

---

# ระดับการวัดของข้อมูลในสถิติ

- Nominal
  - Examples: ID numbers, eye color, zip codes
- Ordinal
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio
  - Examples: temperature in Kelvin, length, time, counts

# คุณสมบัติของข้อมูลใน RM

คุณสมบัติของข้อมูลใน RM แบ่งได้เป็น 2 ประเภท

- **Type** (ชนิด) หมายถึง คุณสมบัติที่บอกถึงความหมายข้อมูลที่ได้จัดเก็บเช่น เป็นตัวหนังสือ เป็นตัวเลข หรือเป็นวันที่
- **Role** (บทบาท) หมายถึง บทบาทของข้อมูลนั้นในการใช้งานใน RM ว่าทำบทบาทหรือหน้าที่ในสถานะใดเช่น ข้อมูลทั่วไป id หรือ ข้อมูลเป้าหมายในการฝึก

# Type (ชนิด) ของข้อมูลใน RM

- **Text** (ตัวหนังสือ)

ข้อมูลถูกจัดเก็บเป็นข้อความ เป็นประโยค มีความหลากหลายของคำ

- **Nominal** ข้อมูลถูกตีความเป็นคำนาม

- **Binominal** ข้อมูลถูกตีความเป็นคำนาม และมีแค่สองคำ เช่น yes/no, ขาว/ดำ,

**Polynominal** ข้อมูลถูกตีความเป็นคำนามเช่นเดียวกับ Binominal แต่มีมากกว่าสองคำเช่นสี ดำ ขาว แดง

- **Numeric** (ตัวเลข) ถูกตีความเป็นตัวเลข

- **Integer** ตัวเลขจำนวนเต็มเช่น 4 -10

- **Real** ตัวเลขจำนวนจริง เช่น 3.2 4.5 -5

- **Date\_Time** ข้อมูลที่ถูกจัดเก็บหมายถึงวันที่และเวลาโดยสามารถนำไปดำเนินการทางคณิตศาสตร์ เช่น 23.12.2014 17:59

- **Date** หมายถึงวันที่จัดเก็บเป็น วันเดือนปี เช่น 23.12.2014

- **Time** หมายถึงเวลาที่จัดเก็บเป็น ชั่วโมง นาที เช่น 17:59

# Role (บทบาท) ของข้อมูลใน RM

- **Attributes, Regular Attribute** หมายถึง Attributes ทั่วไปโดยจะถูกใช้ในฝึกโมเดล ข้อมูลที่ถูกนำเข้าในระบบ RM จะถูกตั้ง default ไว้เป็นค่านี้
- **Label** หมายถึง Attribute ที่ถูกกำหนดเป็น Target Attribute หรือค่าเป้าหมายในการฝึก
- **Id** หมายถึง Attribute ที่ทำหน้าที่เป็นดัชนีข้อ (index) โดยจะไม่มีการซ้ำกันของข้อมูลเลย
- **Cluster** หมายถึง Attribute ที่กำหนดว่าข้อแต่ละตัวถูกกำหนดเป็น Cluster ไດ โดยบทบาทนี้จะใช้หรือเกิดขึ้นในกรณีการทำงานของแบบจำลองประเภท Clustering
- **Prediction** หมายถึง Attribute ที่ถูกสร้างขึ้นมาเพื่อเป็นผลการทำงานของ Model ประเภท Supervised Learning

## Attributes, Feature

Examples



<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Type : Binominal,  
Nominal  
Role: Regular Attr.

# การสร้าง MODEL

---

# กระบวนการทางสร้าง Model

- Training Process

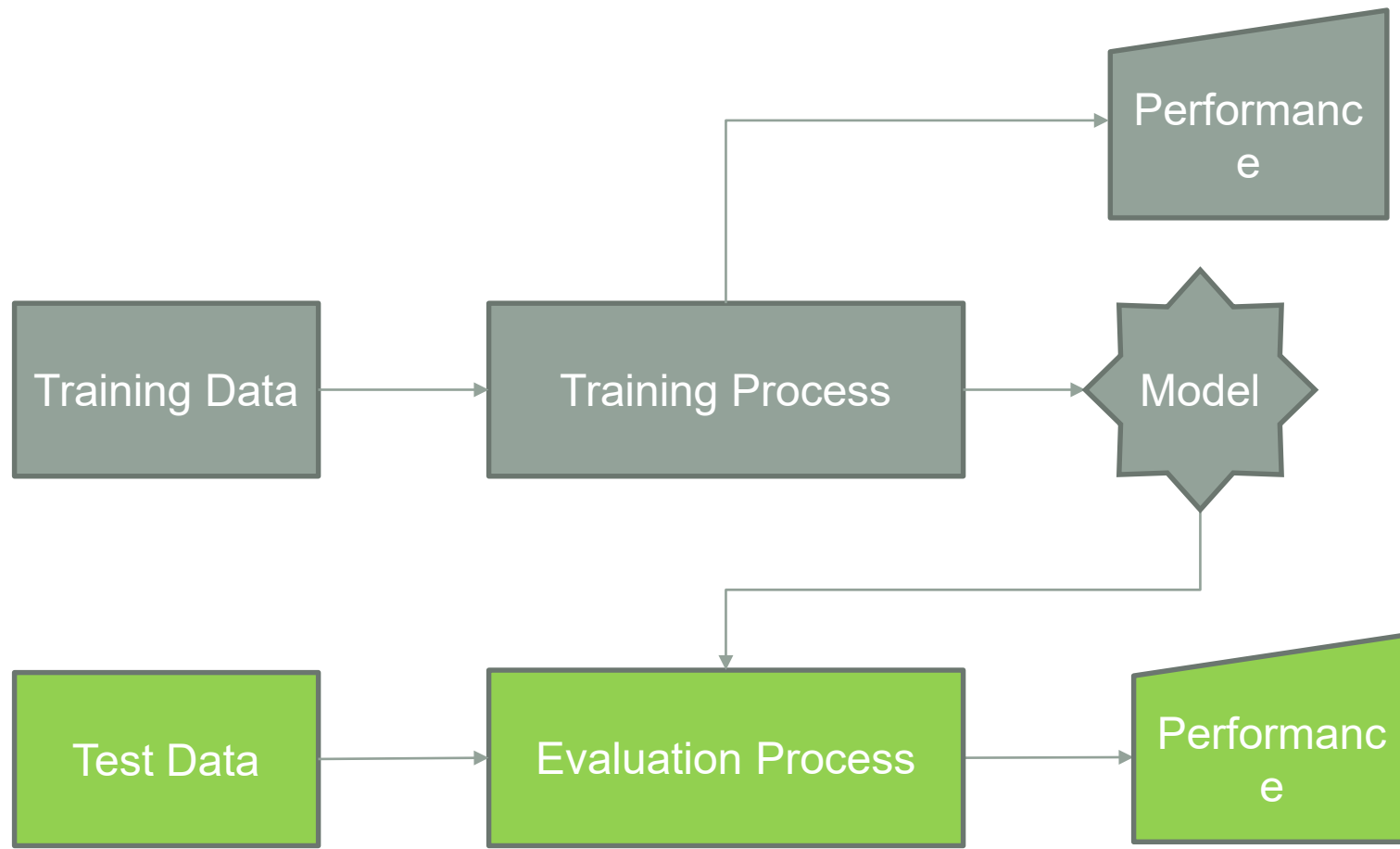
- เป็นกระบวนการฝึกแบบจำลอง
- ใช้ข้อมูลจำนวนหนึ่งเพื่อนำมาฝึกแบบจำลองให้มีประสิทธิภาพที่ยอมรับได้
- ข้อมูลที่ใช้ฝึกนี้จะถูกเรียกว่า Training Set

- Evaluation Process

- เป็นกระบวนการประเมินแบบจำลอง
- ใช้ข้อมูลจำนวนหนึ่งซึ่งปกติขนาดเล็กกว่า Training Set มาใช้ประเมิน
- ข้อมูลชุดนี้จะถูกเรียกว่า Test Set



## กระบวนการทางสร้าง Model (ต่อ)



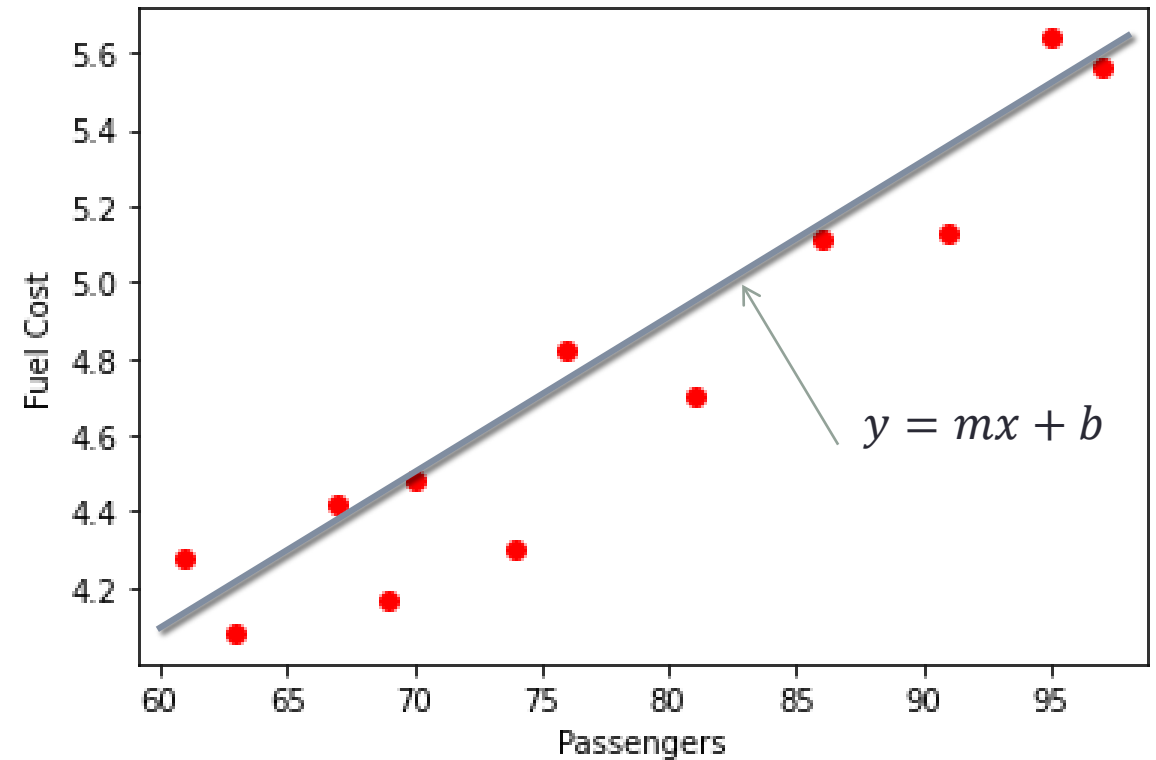
# SUPERVISED LEARNING:

## REGRESSION

---

# Regression Model

## Airline Cost Data

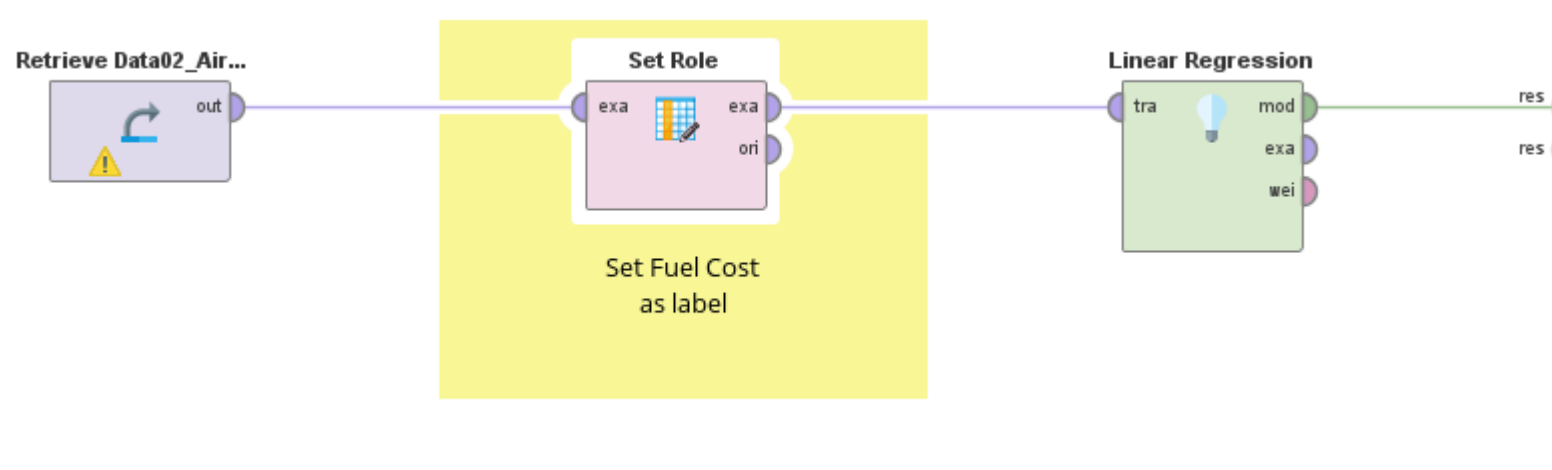


# LAB 5: LINEAR REGRESSION

---

# Lab 5: Process

- Use data set DATA2\_Airplan\_Fuel\_Cost.xls
- Create model Linear Regression



# Lab 5: Model Result

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance
Passenger	0.041	0.004	0.948	?
(Intercept)	1.570	0.338	?	?

Regression equation

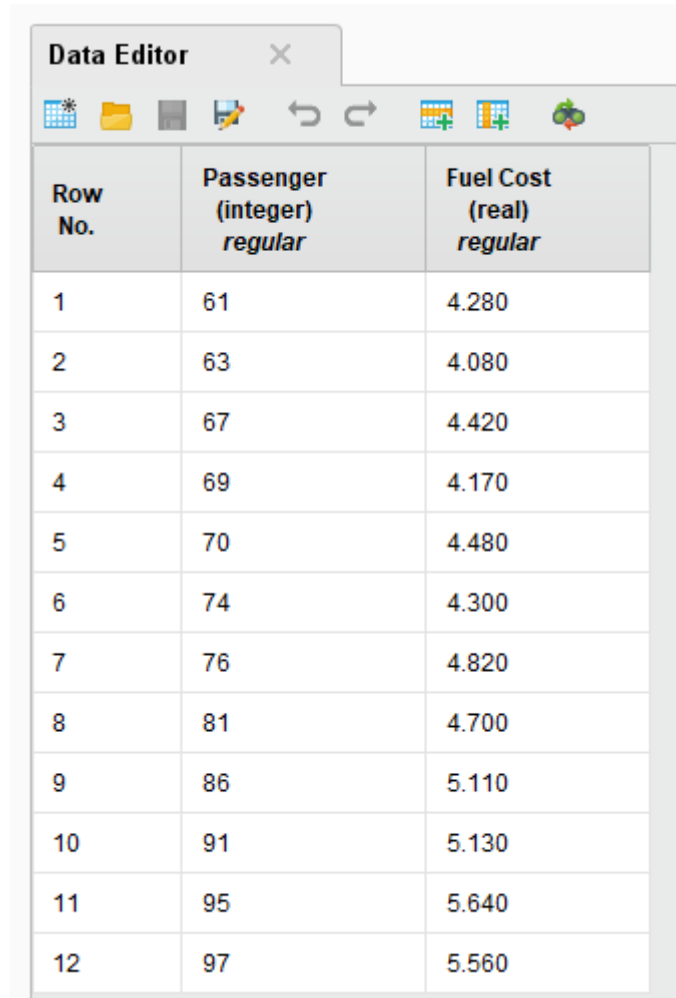
$$\hat{Y} = 1.57 + 0.041X$$

# LAB 6: LINEAR REGRESSION

---

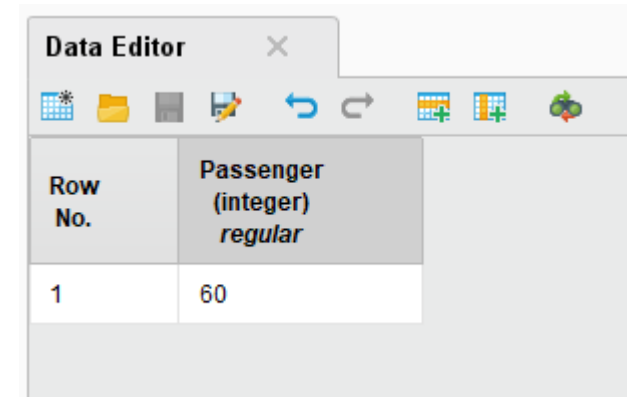
Prediction Value

# Lab 6: Create Test Data for Prediction



A screenshot of a 'Data Editor' window. It contains a table with three columns: 'Row No.', 'Passenger (integer) regular', and 'Fuel Cost (real) regular'. The table lists 12 rows of data, representing training data.

Row No.	Passenger (integer) regular	Fuel Cost (real) regular
1	61	4.280
2	63	4.080
3	67	4.420
4	69	4.170
5	70	4.480
6	74	4.300
7	76	4.820
8	81	4.700
9	86	5.110
10	91	5.130
11	95	5.640
12	97	5.560



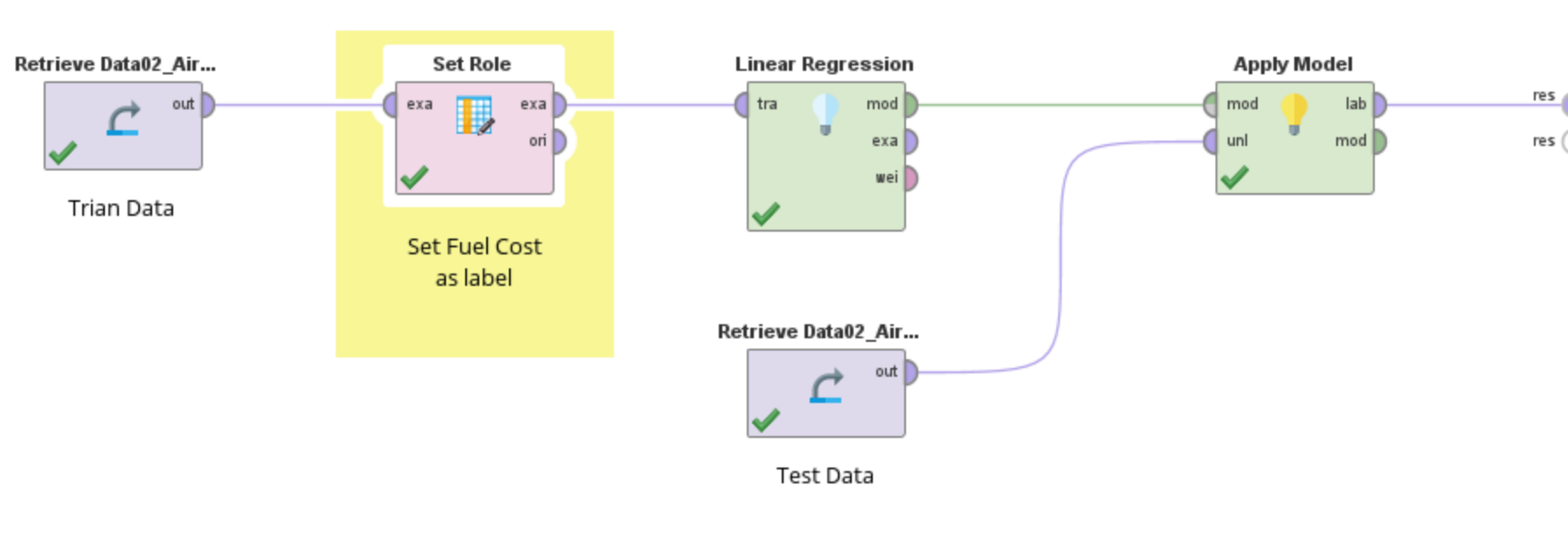
A screenshot of a 'Data Editor' window. It contains a table with two columns: 'Row No.' and 'Passenger (integer) regular'. The table lists 1 row of data, representing test data.

Row No.	Passenger (integer) regular
1	60

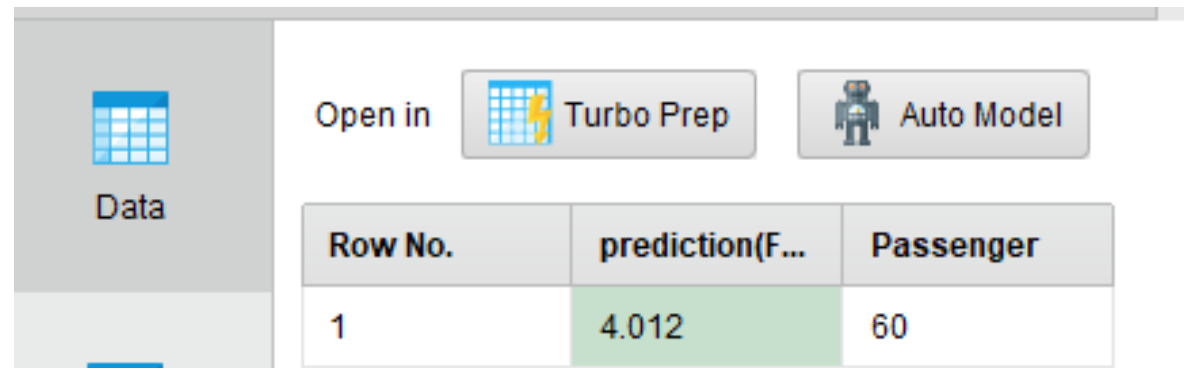
Modified the training  
to the test data



# Lab 6: Process



# Lab 6: Prediction Value



The screenshot shows a software interface with a sidebar on the left containing a 'Data' icon and label. The main area has two buttons: 'Open in Turbo Prep' (with a lightning bolt icon) and 'Auto Model' (with a robot icon). Below these buttons is a table with three columns: 'Row No.', 'prediction(F...', and 'Passenger'. The first row of data shows '1' for Row No., '4.012' for prediction(F... (highlighted in green), and '60' for Passenger.

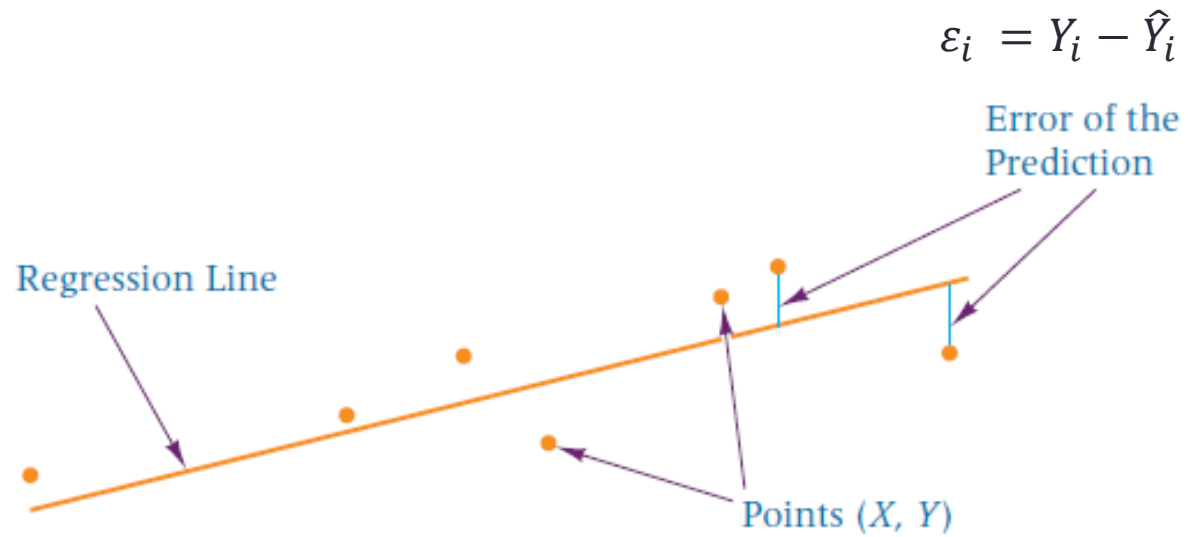
Row No.	prediction(F...	Passenger
1	4.012	60

# LAB 7: LINEAR REGRESSION

---

## Performance Evaluation

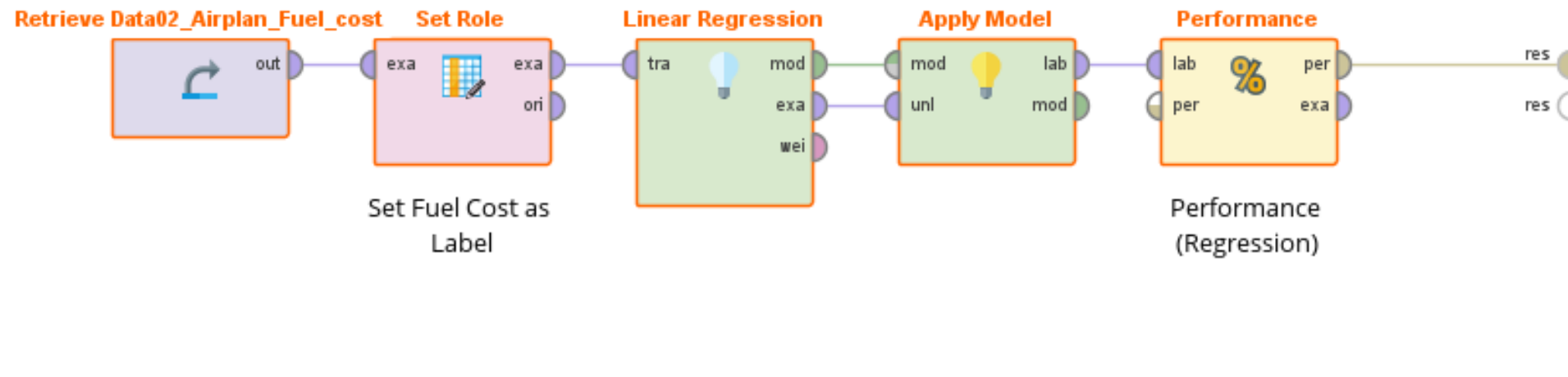
## Root Mean Square Error: Performance of Regression



$$\varepsilon_i = Y_i - \hat{Y}_i$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}$$

# Lab 7: Process



# Lab 7: Performance Result

The screenshot shows a software interface with a 'Result History' tab. A sub-tab labeled 'Performance' is active, displaying a yellow percentage icon. To the right, a panel titled 'PerformanceVector (Performance)' shows the selected criterion 'root mean squared error' in a blue box. The main display area shows the criterion name 'root\_mean\_squared\_error' in a large font, followed by the value 'root\_mean\_squared\_error: 0.162 +/- 0.000' in a monospaced font. A third tab labeled 'ExampleSet (//NPR\_...' is partially visible on the right.

Result History

PerformanceVector (Performance)

ExampleSet (//NPR\_...

Criterion

root mean squared error

root\_mean\_squared\_error

root\_mean\_squared\_error: 0.162 +/- 0.000

# SUPERVISED LEARNING: CLASSIFICATION

---

# DECISION TREE

---



# Sample Dataset (was Tennis played?)

Day	Outlook	Temp	Humidity	Wind	Tennis?
<i>D1</i>	Sunny	Hot	High	Weak	<i>No</i>
<i>D2</i>	Sunny	Hot	High	Strong	<i>No</i>
<i>D3</i>	Overcast	Hot	High	Weak	<i>Yes</i>
<i>D4</i>	Rain	Mild	High	Weak	<i>Yes</i>
<i>D5</i>	Rain	Cool	Normal	Weak	<i>Yes</i>
<i>D6</i>	Rain	Cool	Normal	Strong	<i>No</i>
<i>D7</i>	Overcast	Cool	Normal	Strong	<i>Yes</i>
<i>D8</i>	Sunny	Mild	High	Weak	<i>No</i>
<i>D9</i>	Sunny	Cool	Normal	Weak	<i>Yes</i>
<i>D10</i>	Rain	Mild	Normal	Weak	<i>Yes</i>
<i>D11</i>	Sunny	Mild	Normal	Strong	<i>Yes</i>
<i>D12</i>	Overcast	Mild	High	Strong	<i>Yes</i>
<i>D13</i>	Overcast	Hot	Normal	Weak	<i>Yes</i>
<i>D14</i>	Rain	Mild	High	Strong	<i>No</i>

Sunny = มีแสงแดดมาก

Overcast = มีเมฆมาก

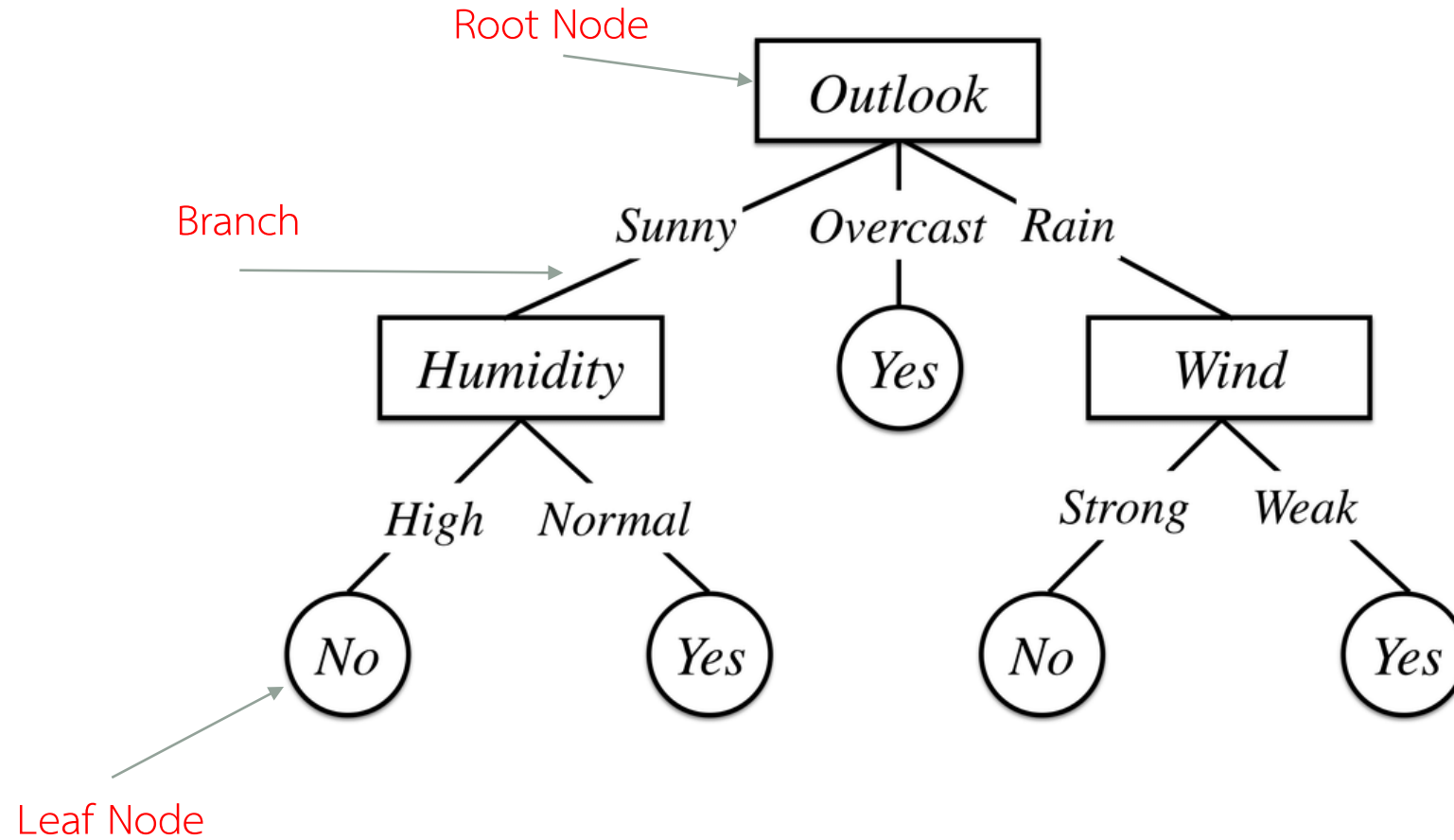
Rain = มีฝน

Hot = ร้อน

Mild = อบอุ่น

Cool = เย็น

# Terminology



# Representation in decision trees

■ Example of representing rule in DT's:

*if* outlook = sunny AND humidity = normal

OR

*if* outlook = overcast

OR

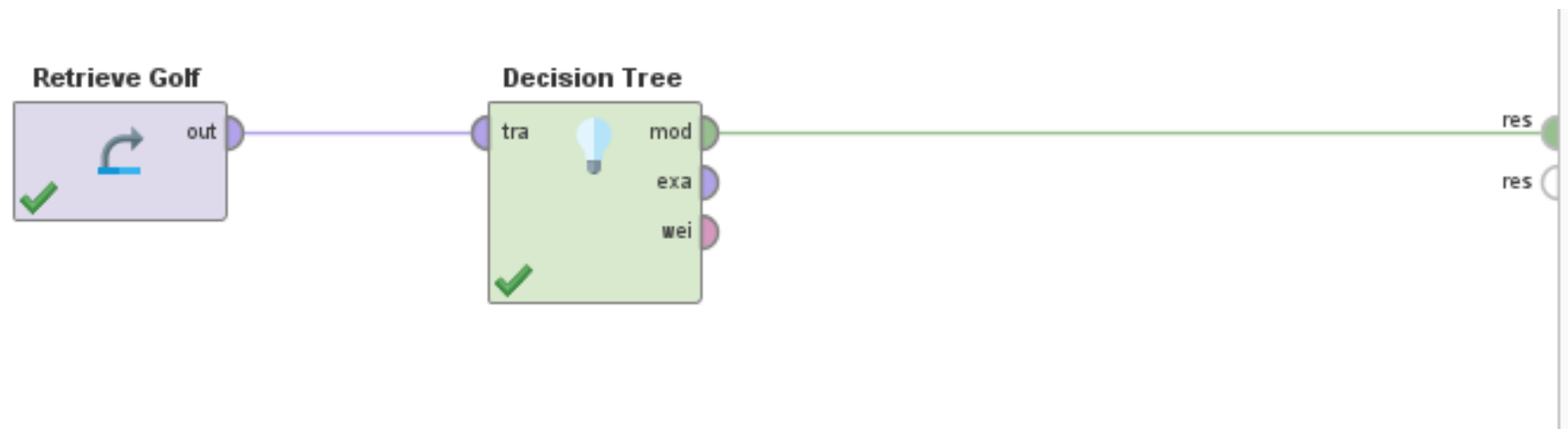
*if* outlook = rain AND wind = weak

*then* playtennis

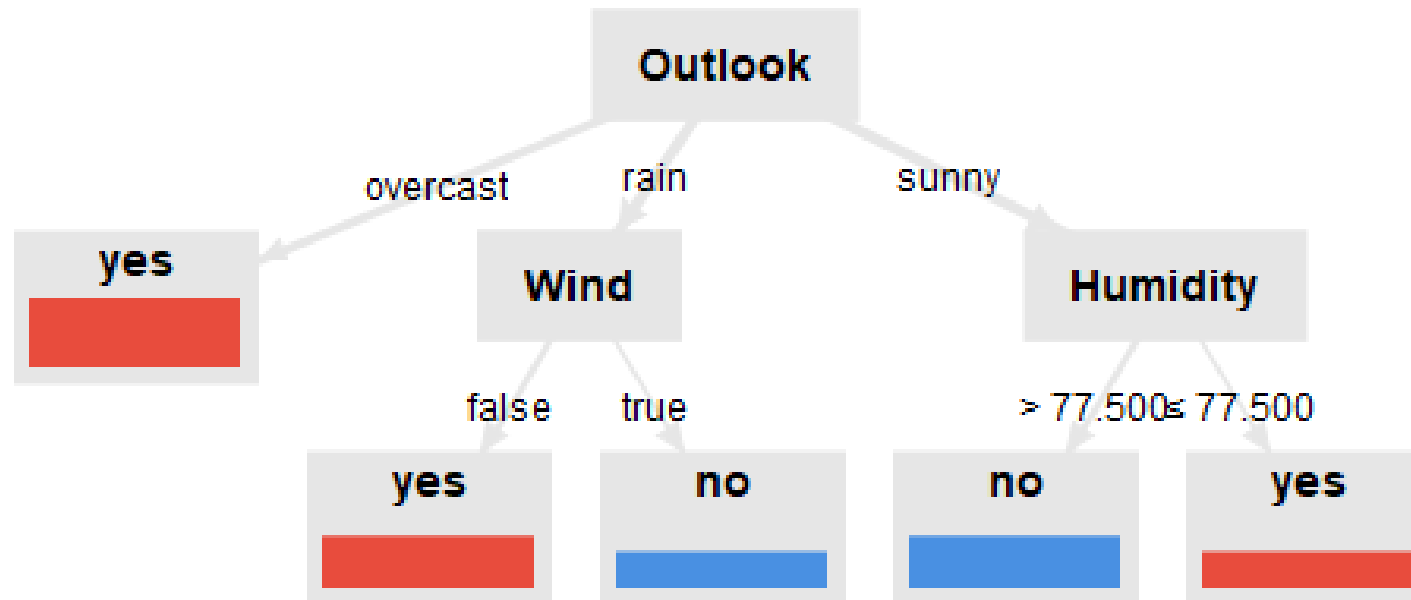
# LAB 8: DECISION TREE

---

# Lab 8



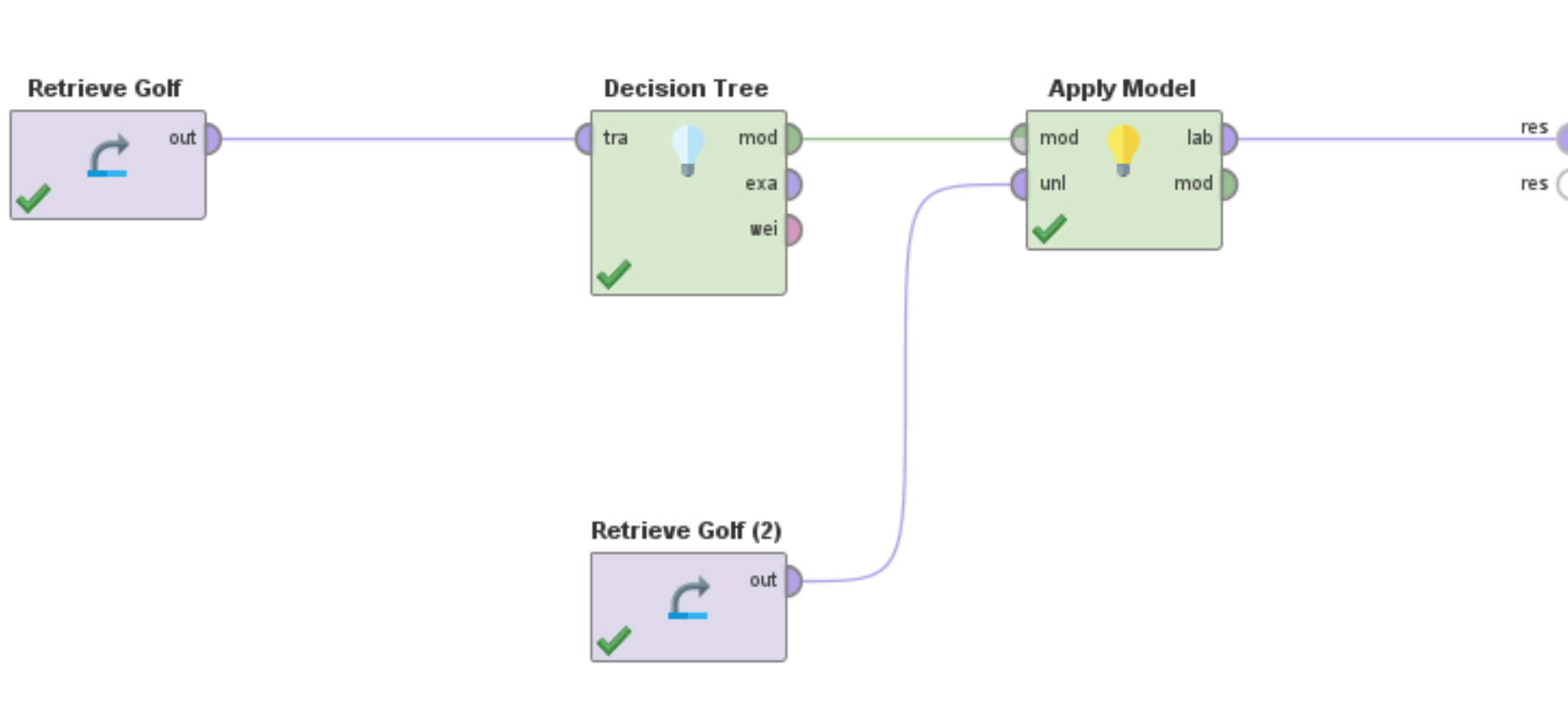
# Lab 8



# LAB 9: DECISION TREE WITH APPLY MODEL

---

# Lab 9





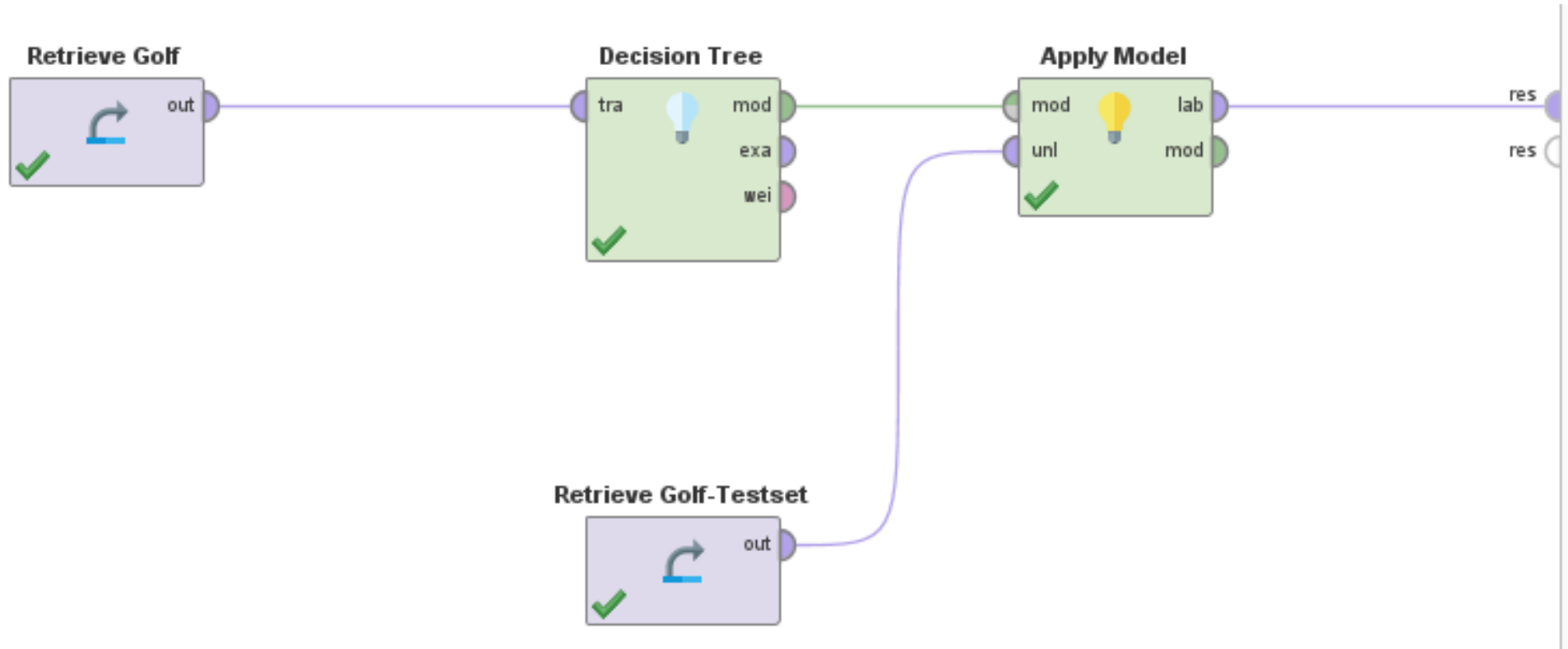
# Lab 9

Row No.	Play	prediction(P...	confidence(...	confidence(...	Outlook	Temperature	Humidity	Wind
1	no	no	1	0	sunny	85	85	false
2	no	no	1	0	sunny	80	90	true
3	yes	yes	0	1	overcast	83	78	false
4	yes	yes	0	1	rain	70	96	false
5	yes	yes	0	1	rain	68	80	false
6	no	no	1	0	rain	65	70	true
7	yes	yes	0	1	overcast	64	65	true
8	no	no	1	0	sunny	72	95	false
9	yes	yes	0	1	sunny	69	70	false
10	yes	yes	0	1	rain	75	80	false
11	yes	yes	0	1	sunny	75	70	true
12	yes	yes	0	1	overcast	72	90	true
13	yes	yes	0	1	overcast	81	75	false
14	no	no	1	0	rain	71	80	true

# LAB 10: DECISION TREE WITH TEST SET

---

# Lab 10



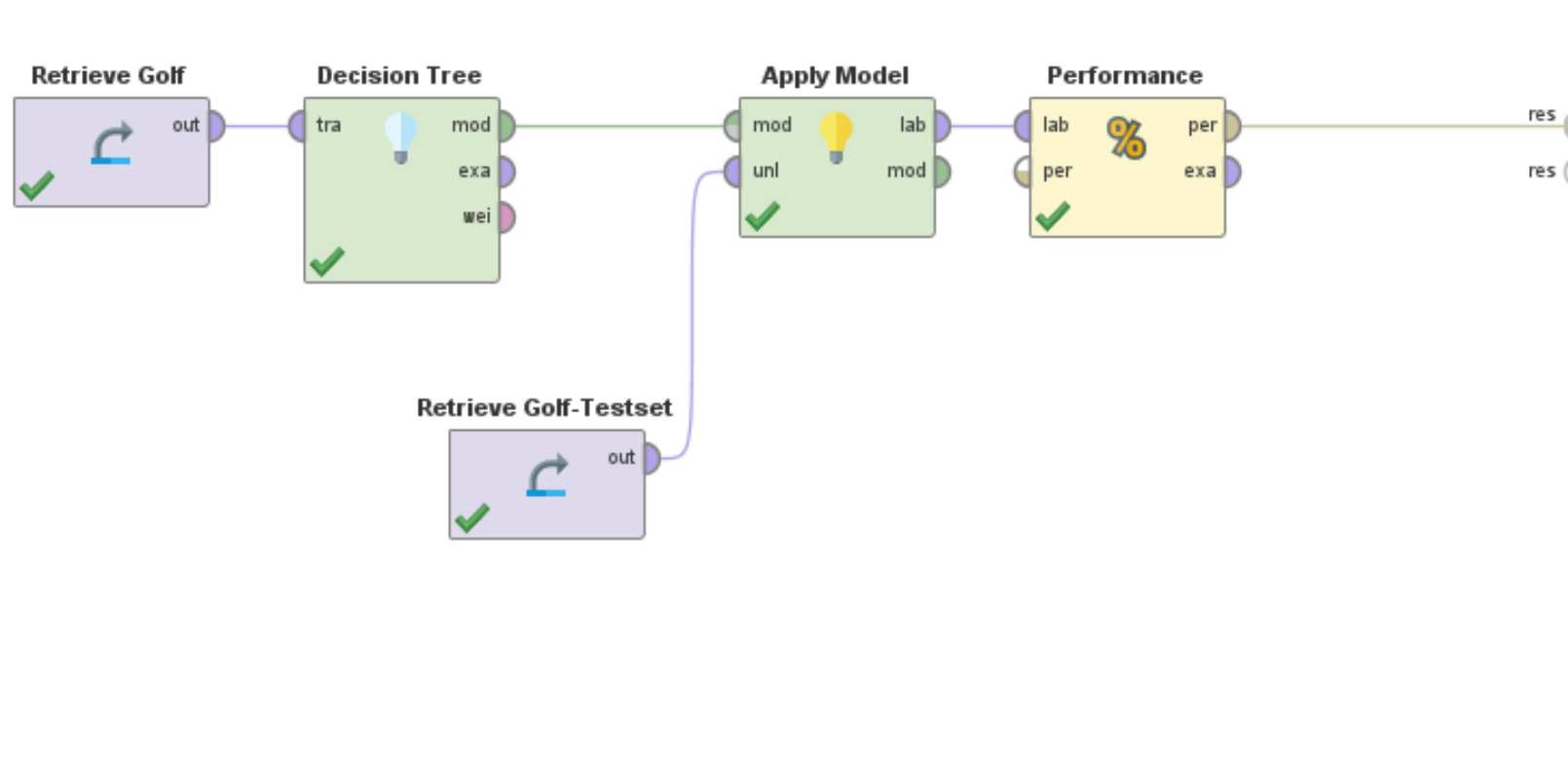
# Lab 10

Row No.	Play	prediction(P...	confidence(...	confidence(...	Outlook	Temperature	Humidity	Wind
1	yes	no	1	0	sunny	85	85	false
2	no	yes	0	1	overcast	80	90	true
3	yes	yes	0	1	overcast	83	78	false
4	yes	yes	0	1	rain	70	96	false
5	yes	no	1	0	rain	68	80	true
6	no	no	1	0	rain	65	70	true
7	yes	yes	0	1	overcast	64	65	true
8	no	no	1	0	sunny	72	95	false
9	yes	yes	0	1	sunny	69	70	false
10	no	no	1	0	sunny	75	80	false
11	yes	yes	0	1	sunny	68	70	true
12	yes	yes	0	1	overcast	72	90	true
13	no	yes	0	1	overcast	81	75	true
14	yes	no	1	0	rain	71	80	true

# LAB 11: DECISION TREE PERFORMANCE EVALUATION

---

# Lab 11



# Lab 11

accuracy: 64.29%

	true no	true yes	class precision
pred. no	3	3	50.00%
pred. yes	2	6	75.00%
class recall	60.00%	66.67%	

# MODEL EVALUATION MATRICES FOR CLASSIFICATION PROBLEM

---



# Model Evaluation Metrics

- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1-Score
- Area under curve (AUC)
- Kappa

More and more..

# Confusion Matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

**True Positive (TP):** หมายถึงจำนวนการทำนายว่าเป็น**บวก**เมื่อคลาสทำนายนั้นเป็น**บวก**

**True Negative (TN):** หมายถึงจำนวนการทำนายว่าเป็น**ลบ**เมื่อคลาสทำนายนั้นเป็น**ลบ**

**False Positive (FP):** หมายถึงจำนวนการทำนายว่าเป็น**บวก**เมื่อคลาสทำนายนั้นเป็น**ลบ**

**False Negative (FN):** หมายถึงจำนวนการทำนายว่าเป็น**ลบ**เมื่อคลาสทำนายนั้นเป็น**บวก**

# Example of confusion matrix

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

# Accuracy

		True Class	
		Positive	Negative
Predicted Class	Positive	TP 30	FP 10
	Negative	FN 20	TN 40

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Predictions}}$$

$$\text{Accuracy} = \frac{30 + 40}{100} = 0.7$$

# ข้อจำกัดของ Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

		True Class	
		Positive	Negative
Predicted Class	Positive	TP 0	FP 0
	Negative	FN 10	TN 9990

- If model predicts everything to be class 0, accuracy is  $9990/10000 = 99.9\%$
- Accuracy is misleading because model does not detect any class 1 example

# Precision

		True Class	
		Positive	Negative
Predicted Class	Positive	TP 0	FP 0
	Negative	FN 10	TN 9990

$$\text{Precision Class 1} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

$$\text{Precision Class 1} = \frac{0}{0} = \infty$$

$$\text{Precision Class 0} = \frac{9990}{10000} = 0.99$$

# Recall

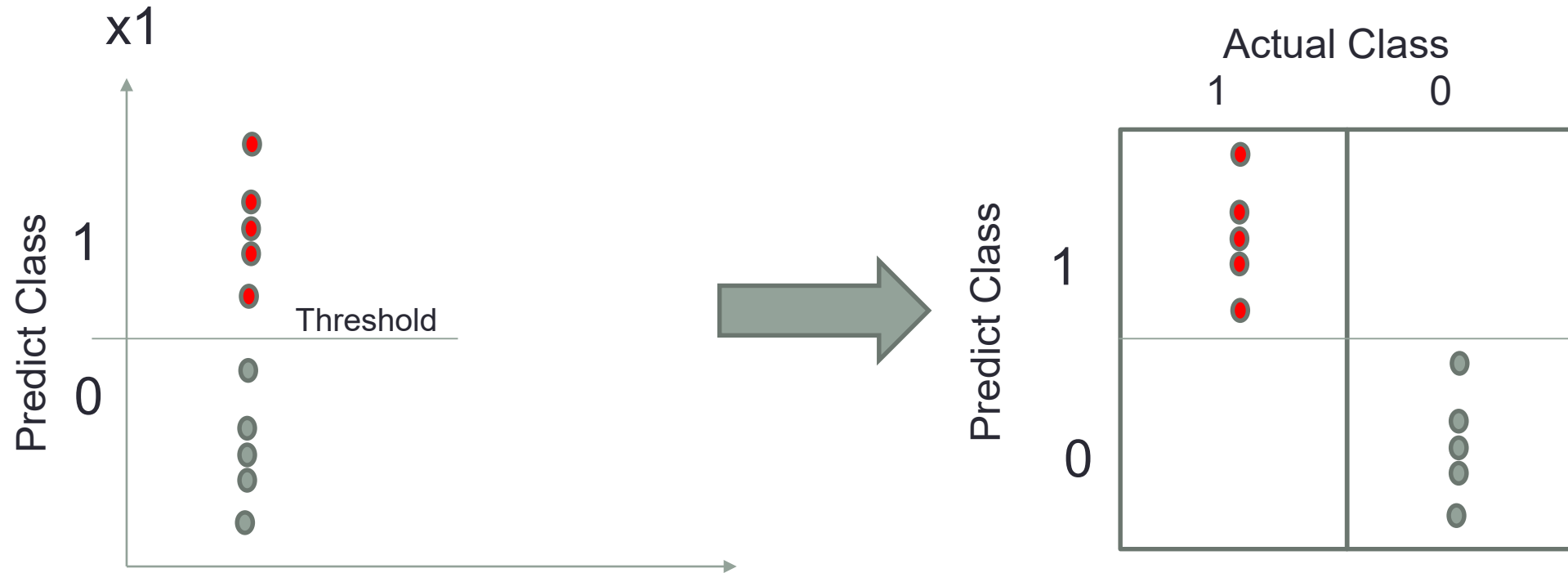
		True Class	
		Positive	Negative
Predicted Class	Positive	TP 0	FP 0
	Negative	FN 10	TN 9990

$$\text{Recall Class 1} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{Recall Class 1} = \frac{0}{10} = 0$$

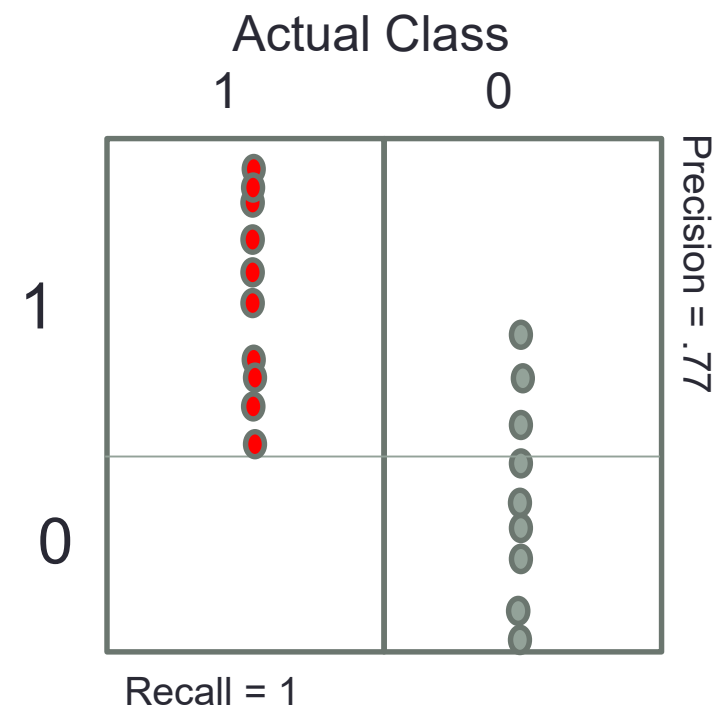
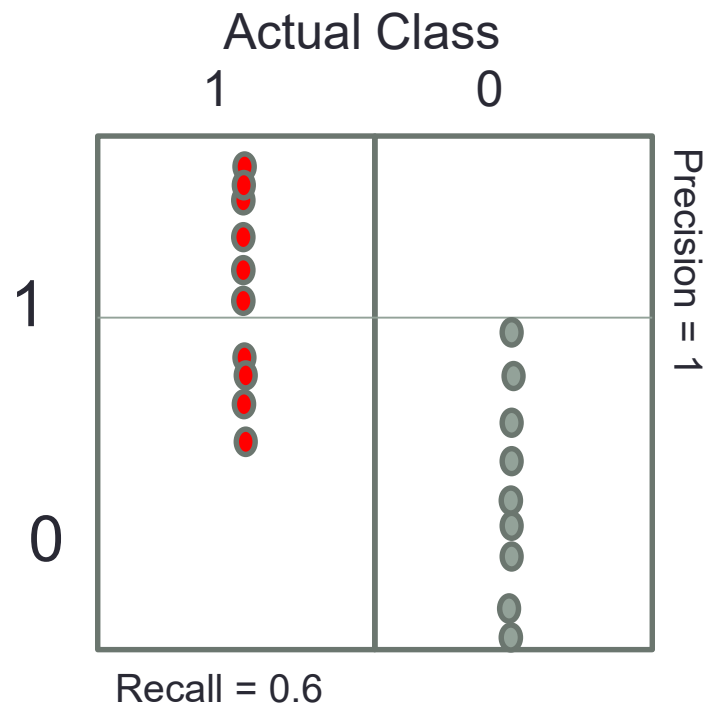
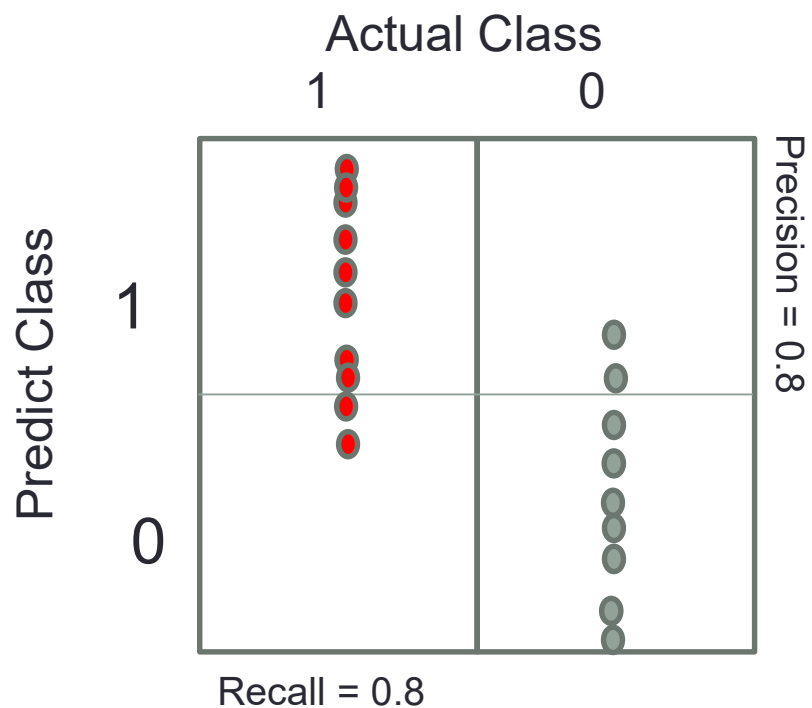
$$\text{Recall Class 0} = \frac{9990}{9990} = 1$$

# Confusion Matrix Intuition





# Confusion Matrix Intuition



# F1-Score

## F-Score

$$F_{\beta} = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

where  $\beta$  is chosen such that recall is considered  $\beta$  times as important as precision

For  $\beta = 1$ , it becomes F1-Score

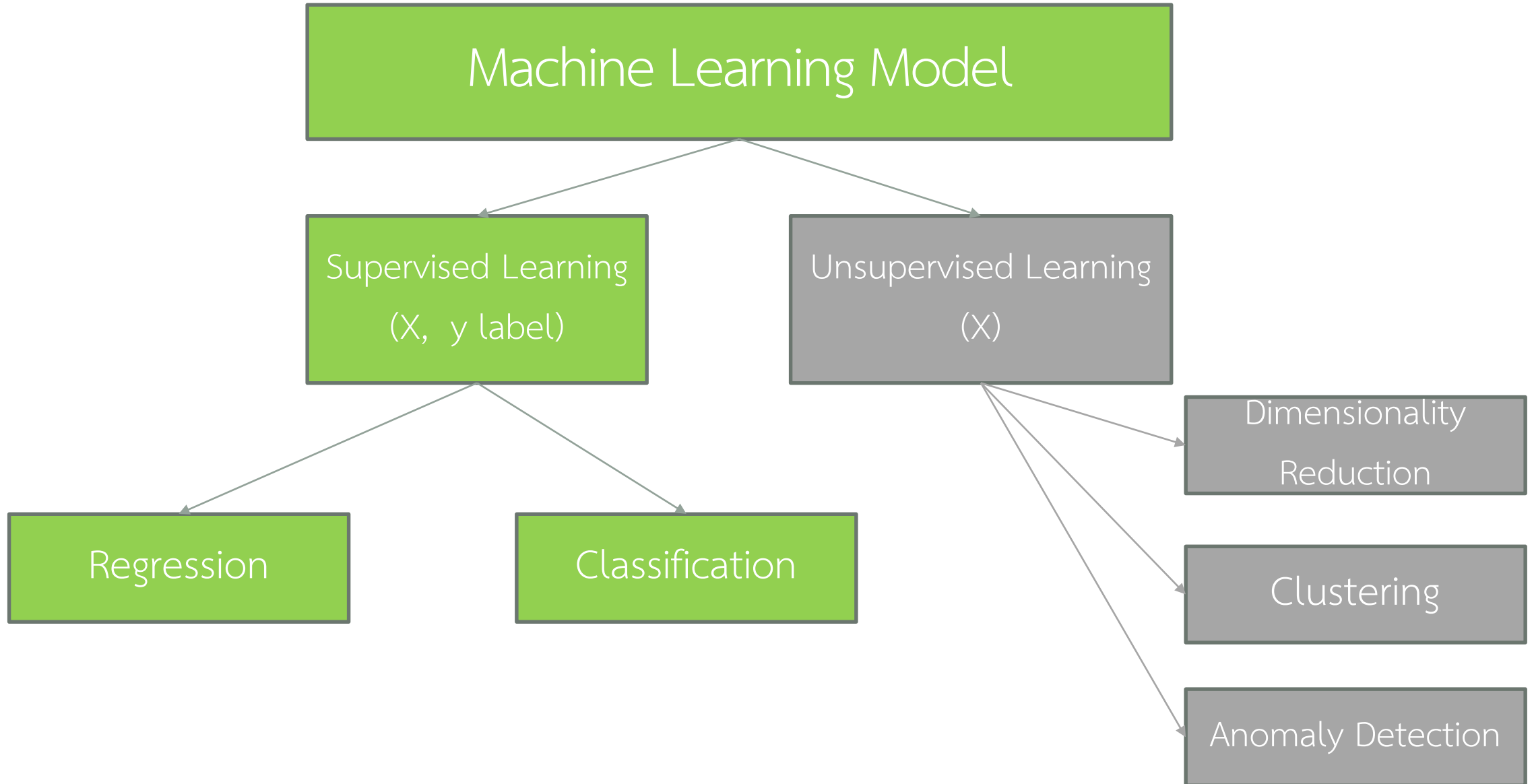
$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

# F1-Score

How to compare precision/recall numbers?

	Precision(P)	Recall (R)
Algorithm 1	0.5	0.4
Algorithm 2	0.7	0.1
Algorithm 3	0.02	1.0

$$F_1 \text{ Score: } 2 \frac{PR}{P+R}$$

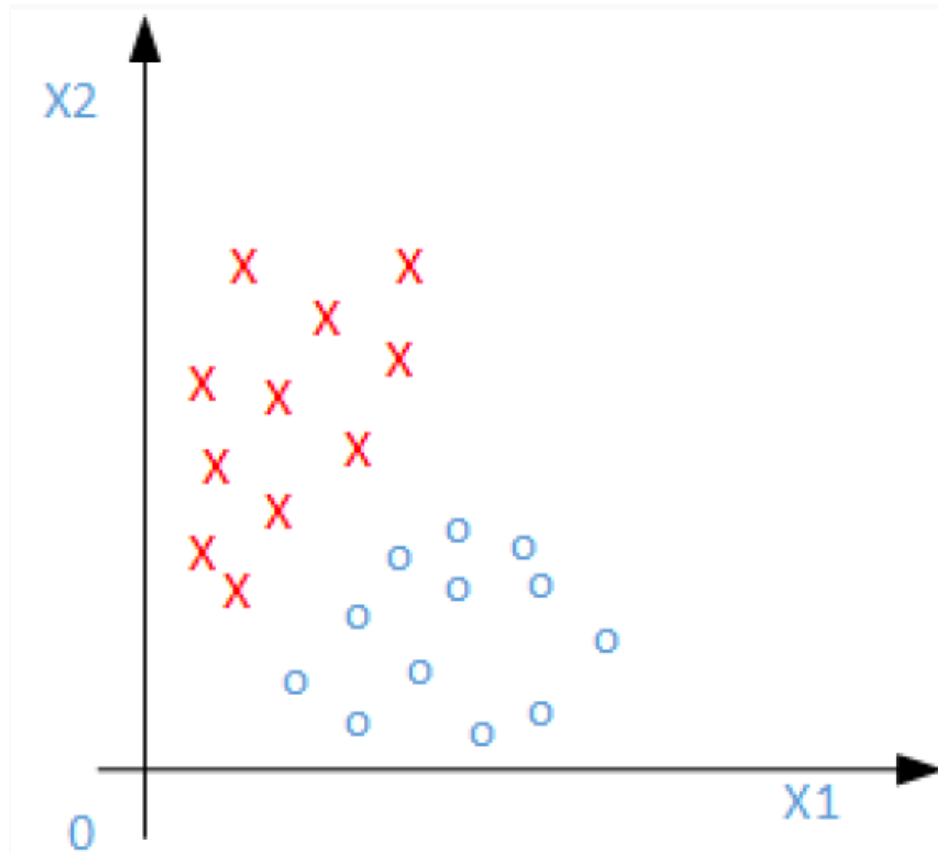


# LOGISTIC REGRESSION

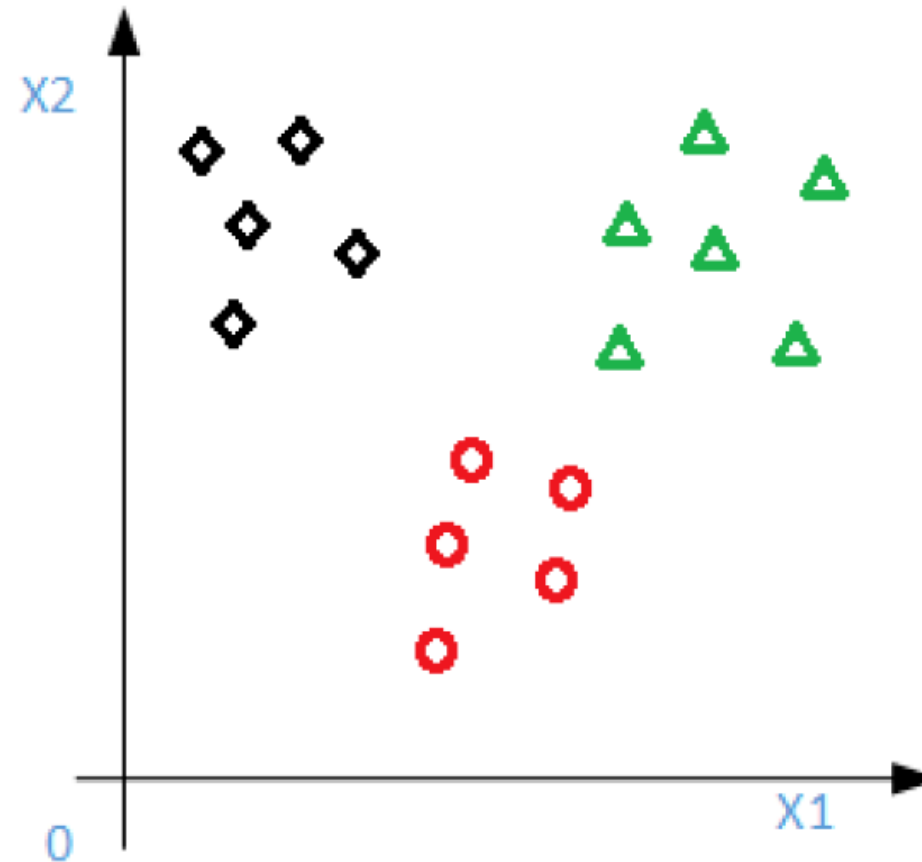
---

# Logistic Regression

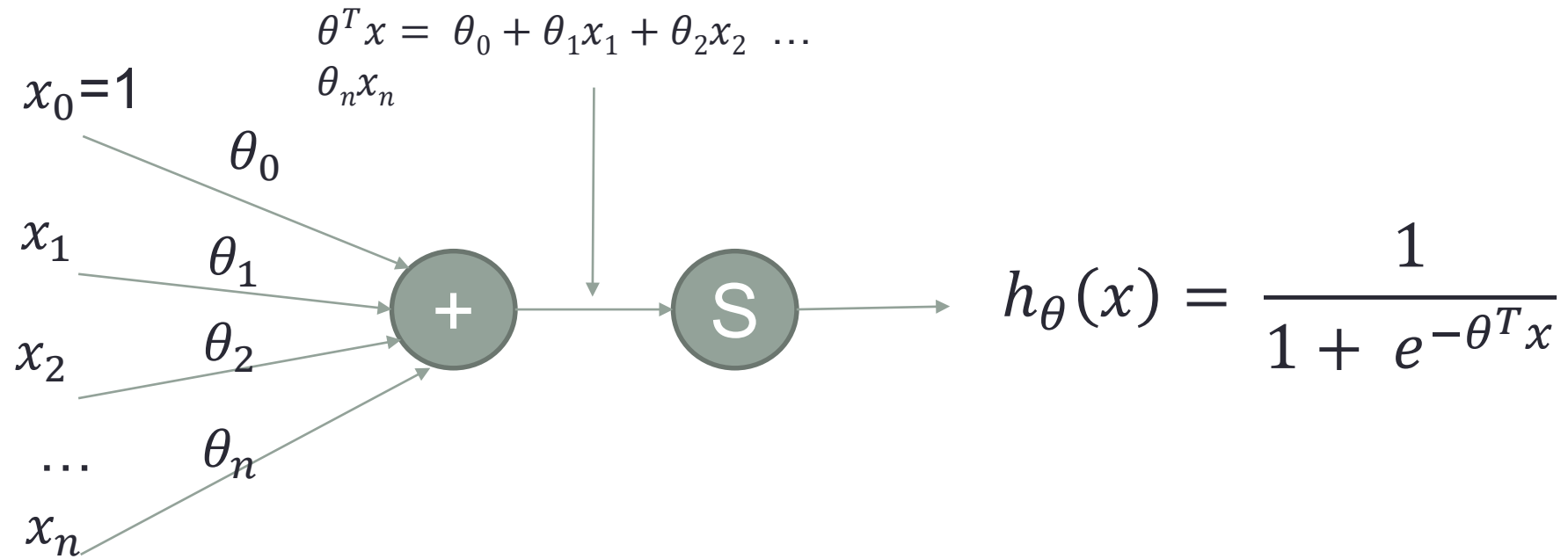
Binary Classification :



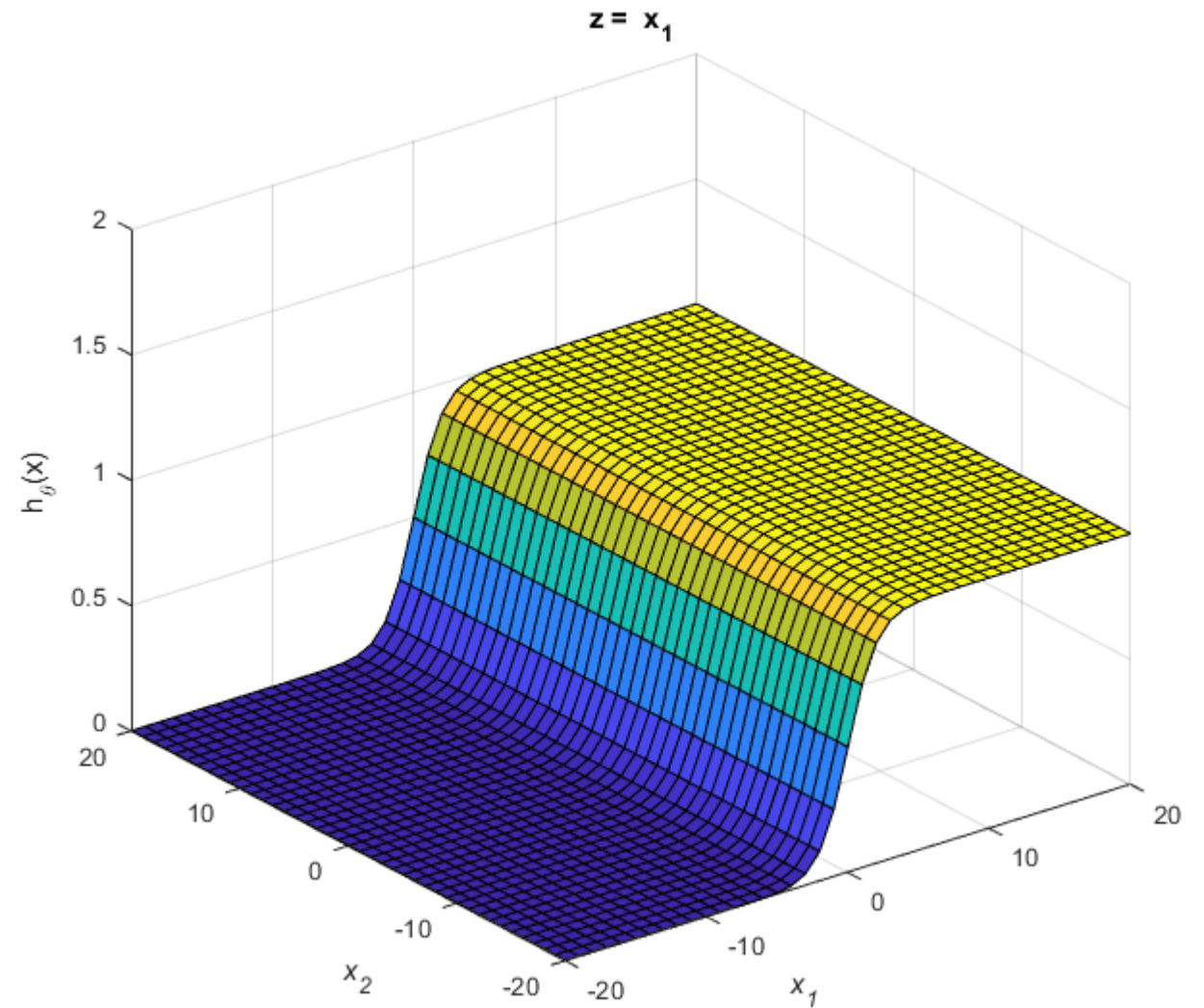
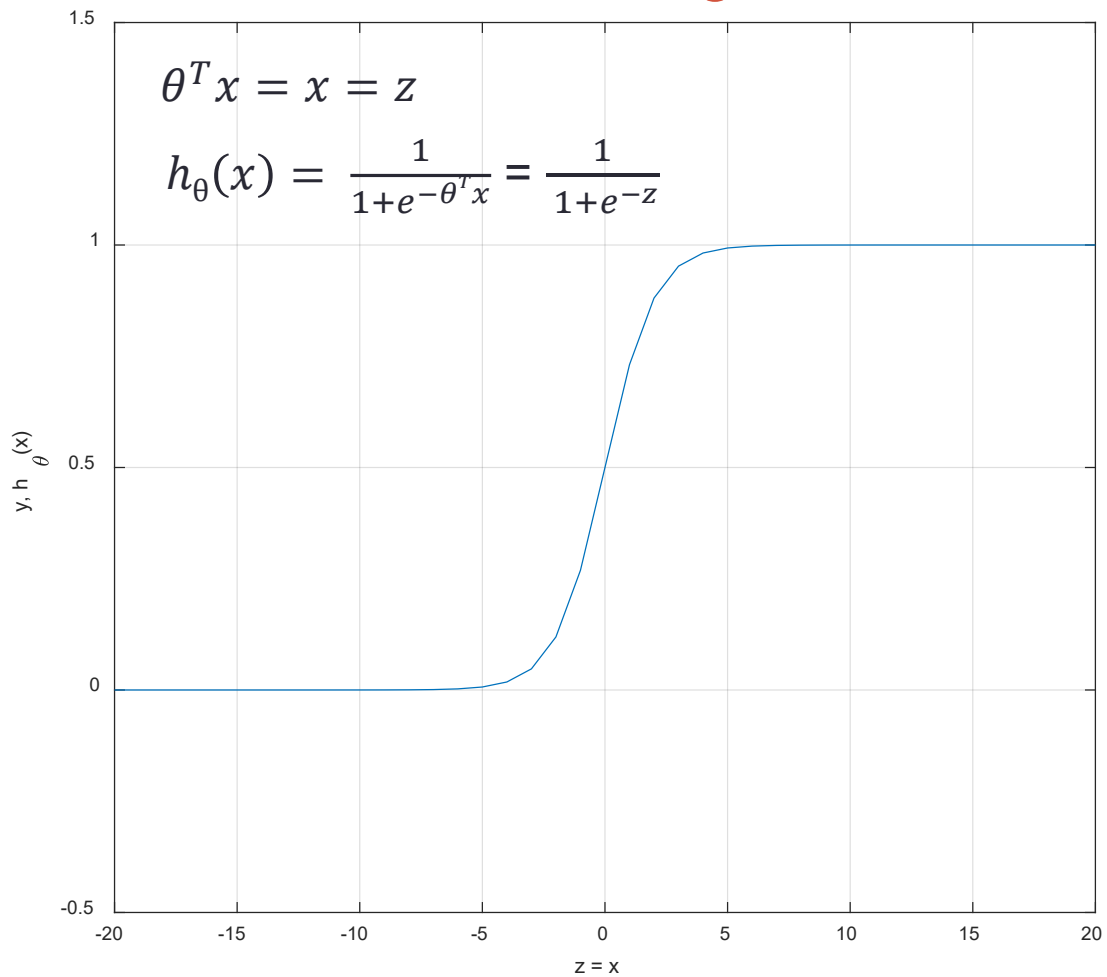
Multi-Class  
Classification :



# Logistic Regression Model



# Example of $h_{\theta}(x)$



$$z = \theta^T x = x_1$$

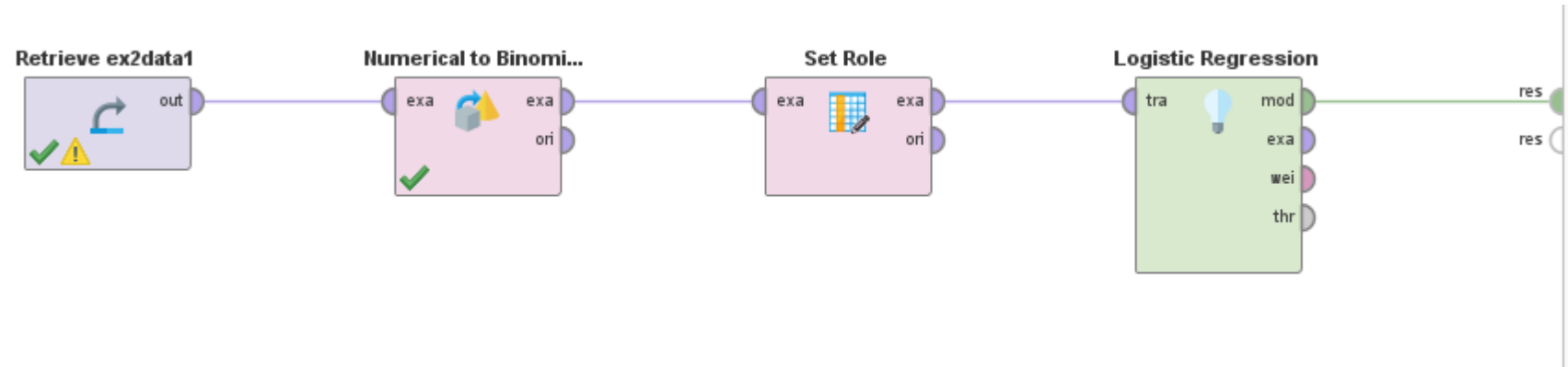


# LAB 12:

---

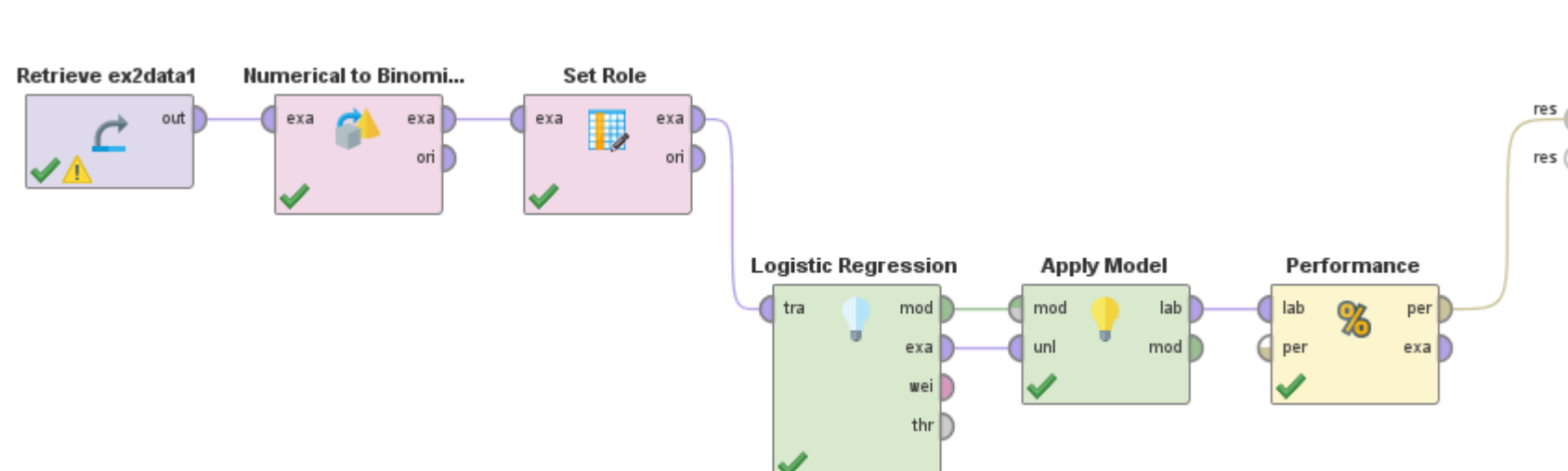
## Logistic Regression

# Lab 12: Process



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
att1	0.206	4.013	0.048	4.296	0.000
att2	0.201	3.744	0.049	4.143	0.000
Intercept	-25.161	1.718	5.799	-4.339	0.000

# Lab 12: Performance



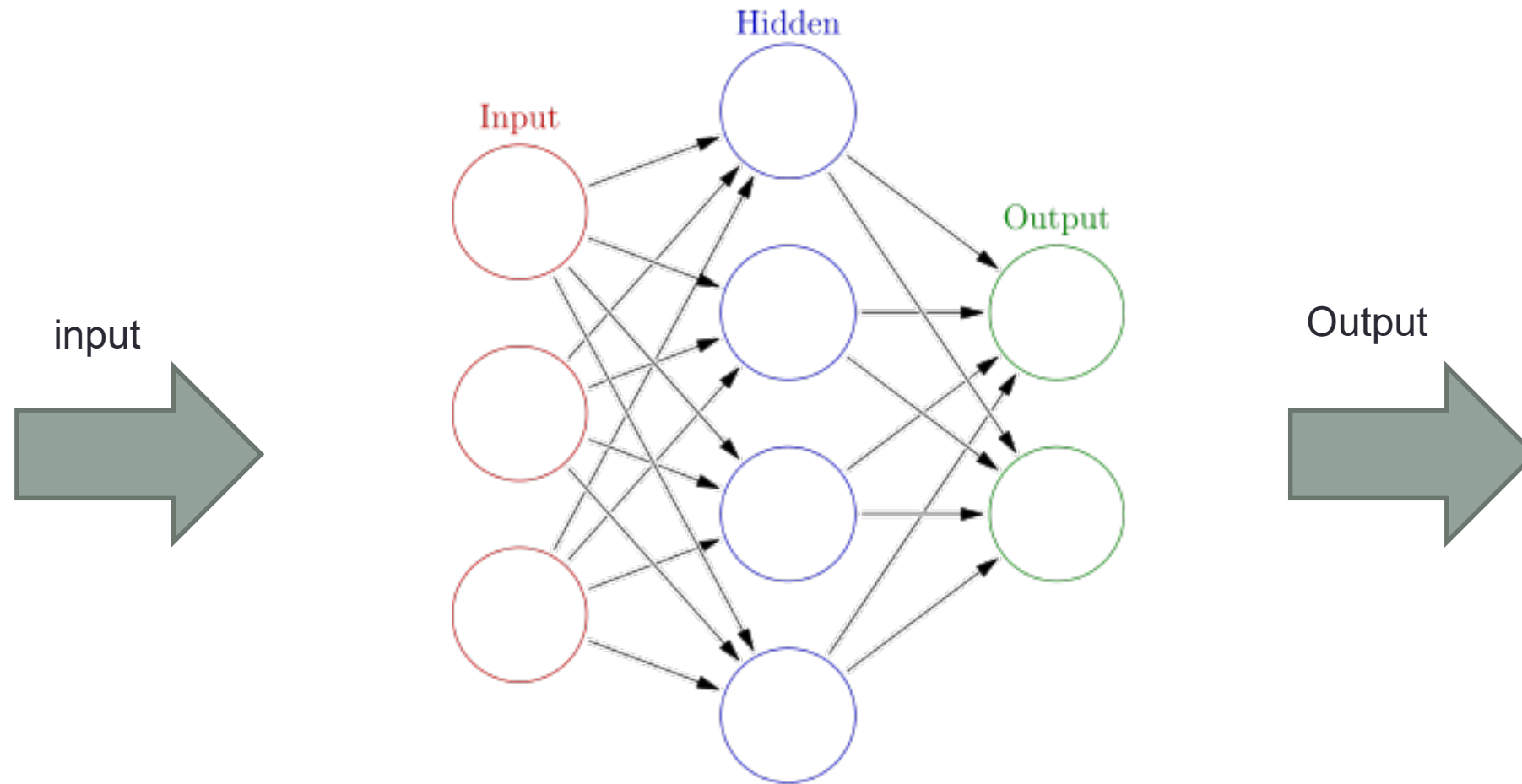
accuracy: 89.00%

	true false	true true	class precision
pred. false	34	5	87.18%
pred. true	6	55	90.16%
class recall	85.00%	91.67%	

# NEURAL NETWORK

---

# Artificial Neural Network

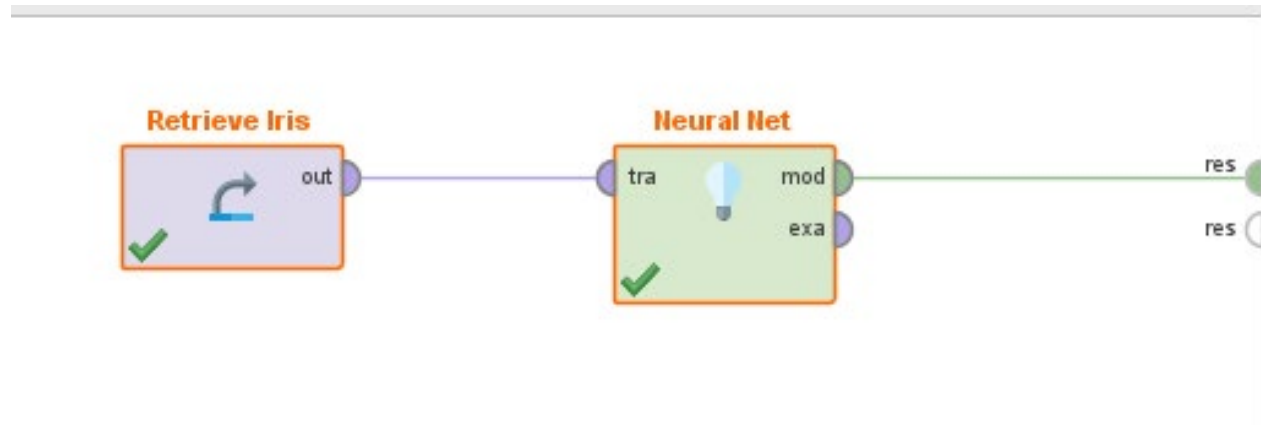


# LAB 13:

---

Neural Network for Iris

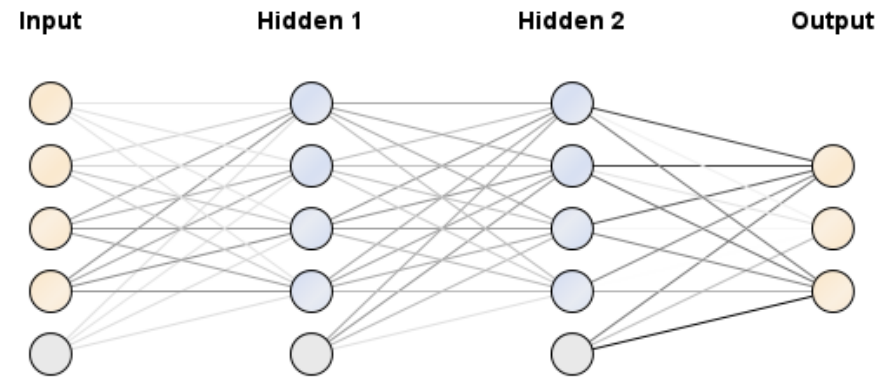
# Lab 13: Model



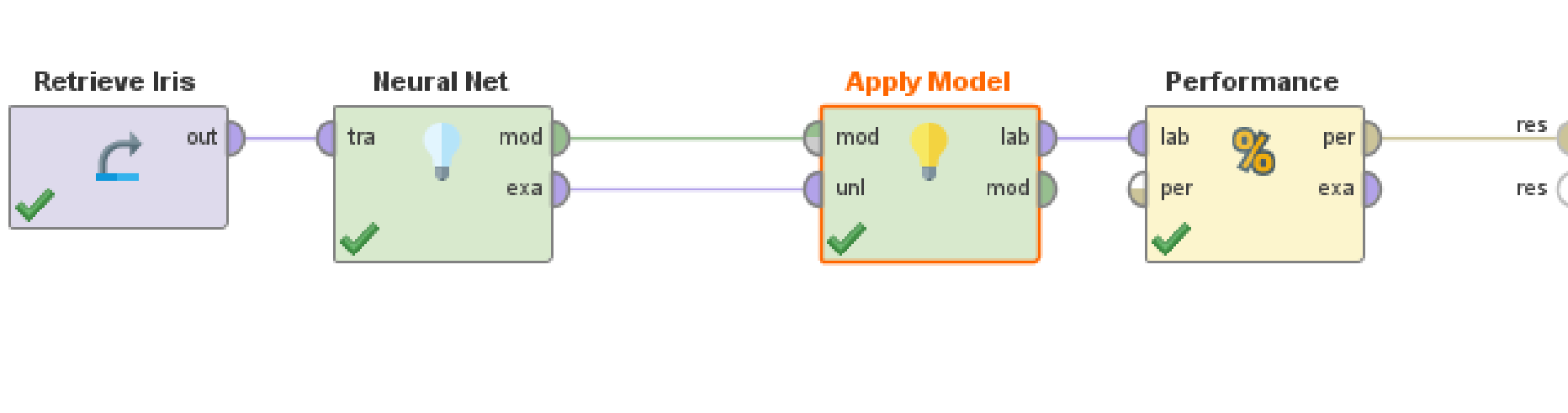
Edit Parameter List: hidden layers

Edit Parameter List: **hidden layers**  
Describes the name and the size of all hidden layers.

hidden layer name	hidden layer sizes
L1	4
L2	4



# Lab 13: Performance





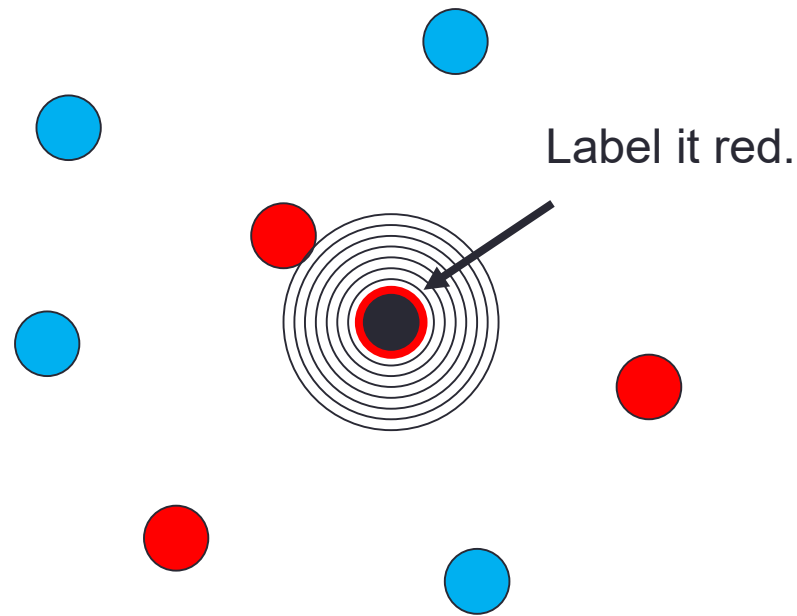
# NEAREST NEIGHBOR CLASSIFIERS

---

Adopt from slides by  
Carla P. Gomes

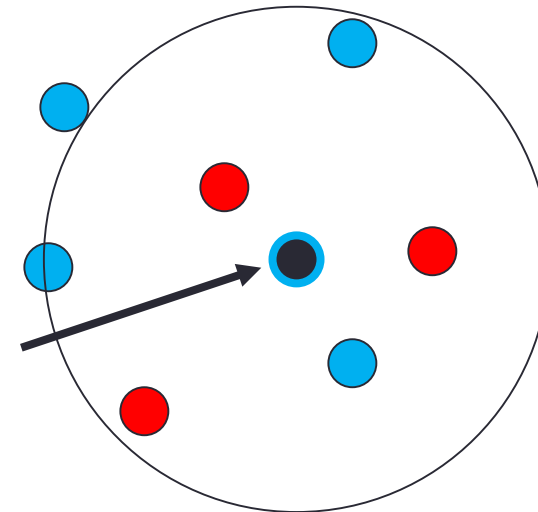
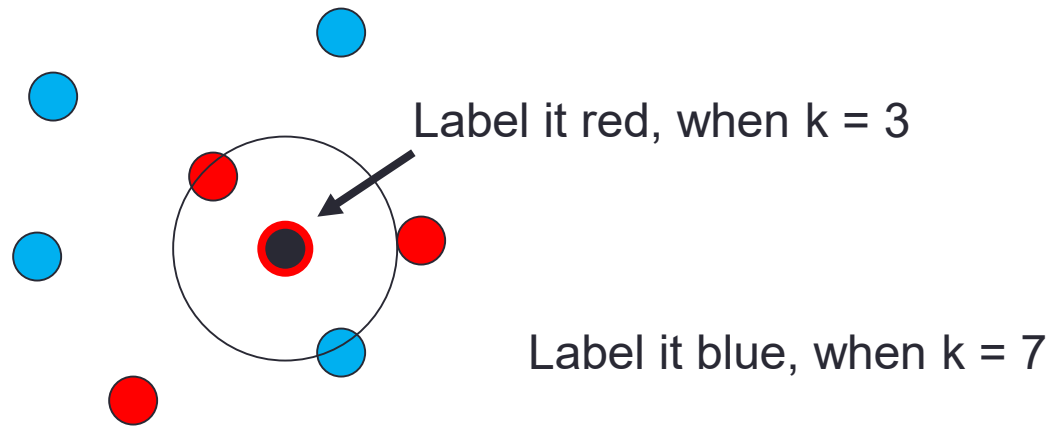
# 1-Nearest Neighbor

- เป็นโมเดลที่มีการทำงานที่ง่ายที่สุดโมเดลหนึ่ง
- **Simple idea:** การทำนาย class จะพิจารณาจากค่าข้อมูลที่อยู่ใกล้ที่สุดของข้อมูลที่ทำนาย



# k – Nearest Neighbor

- เพื่อเป็นการลด noise ที่จะเกิดขึ้นในการทำนาย จึงพิจารณาทำนายด้วยข้อมูลมากขึ้น
- การทำนายข้อมูลจะดูจากข้อมูลที่ใกล้เคียงจำนวน  $k$  ตัว โดยจะทำนายเป็น class ที่มากที่สุด ในจำนวน  $k$  ตัว



# Selecting the Number of Neighbors

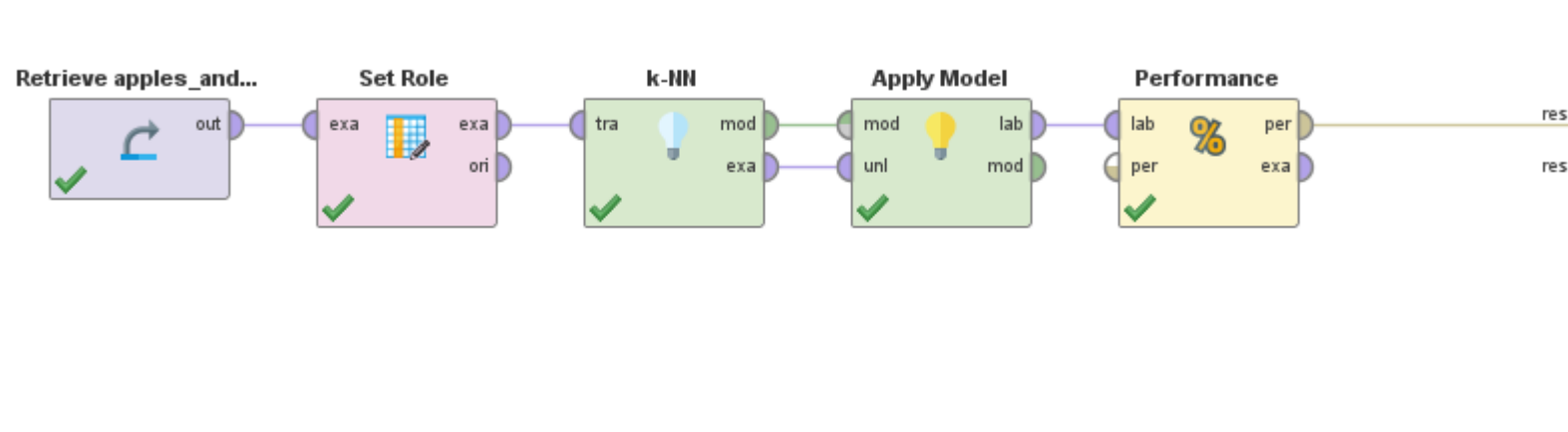
- Increase k:
  - k ขนาดใหญ่ทำให้ การทำนายลดความไวต่อ noise
- Decrease k:
  - k ขนาดเล็กจะทำให้การทำนายละเอียดขึ้น
- ➔ การเลือกค่า k จะต้องเลือกให้เหมาะสมไม่เล็กหรือใหญ่เกินไป (depends on data) นิยมทำเป็นเลขคี่

# LAB 14: APPLE AND ORANGE

---

KNN Class Prediction

# Lab 14 : Process

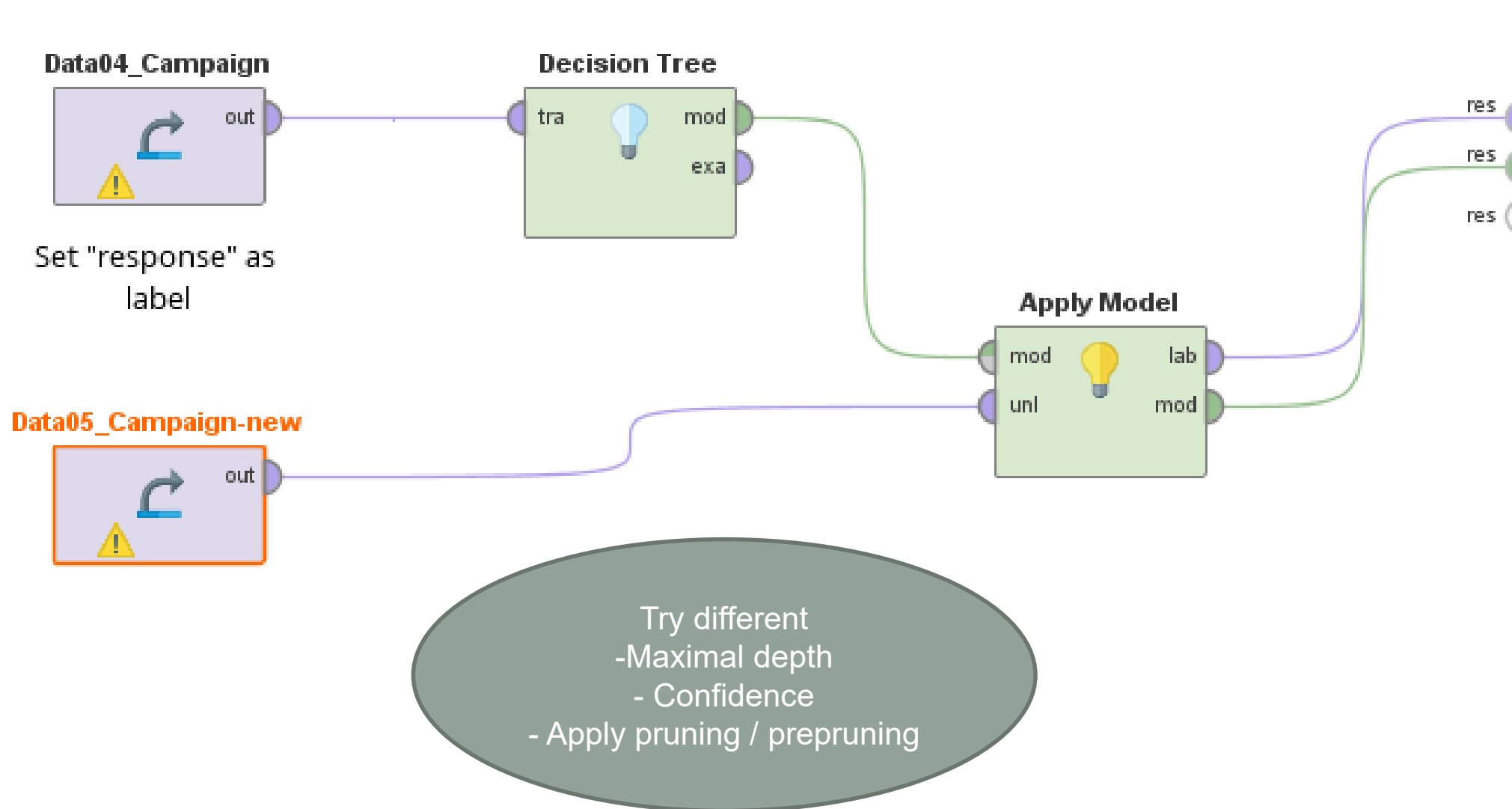


# LAB 15: APPLICATION EXAMPLE

## CAMPAIGN

---

# Lab 15 : Process





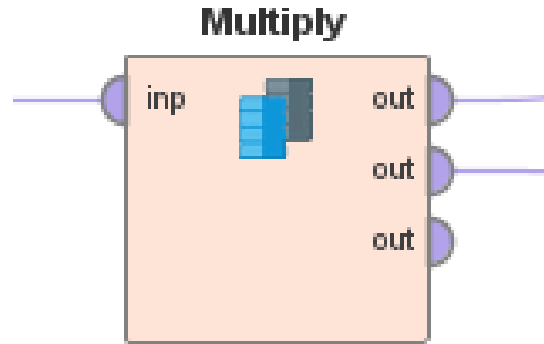
# LAB 16:

---

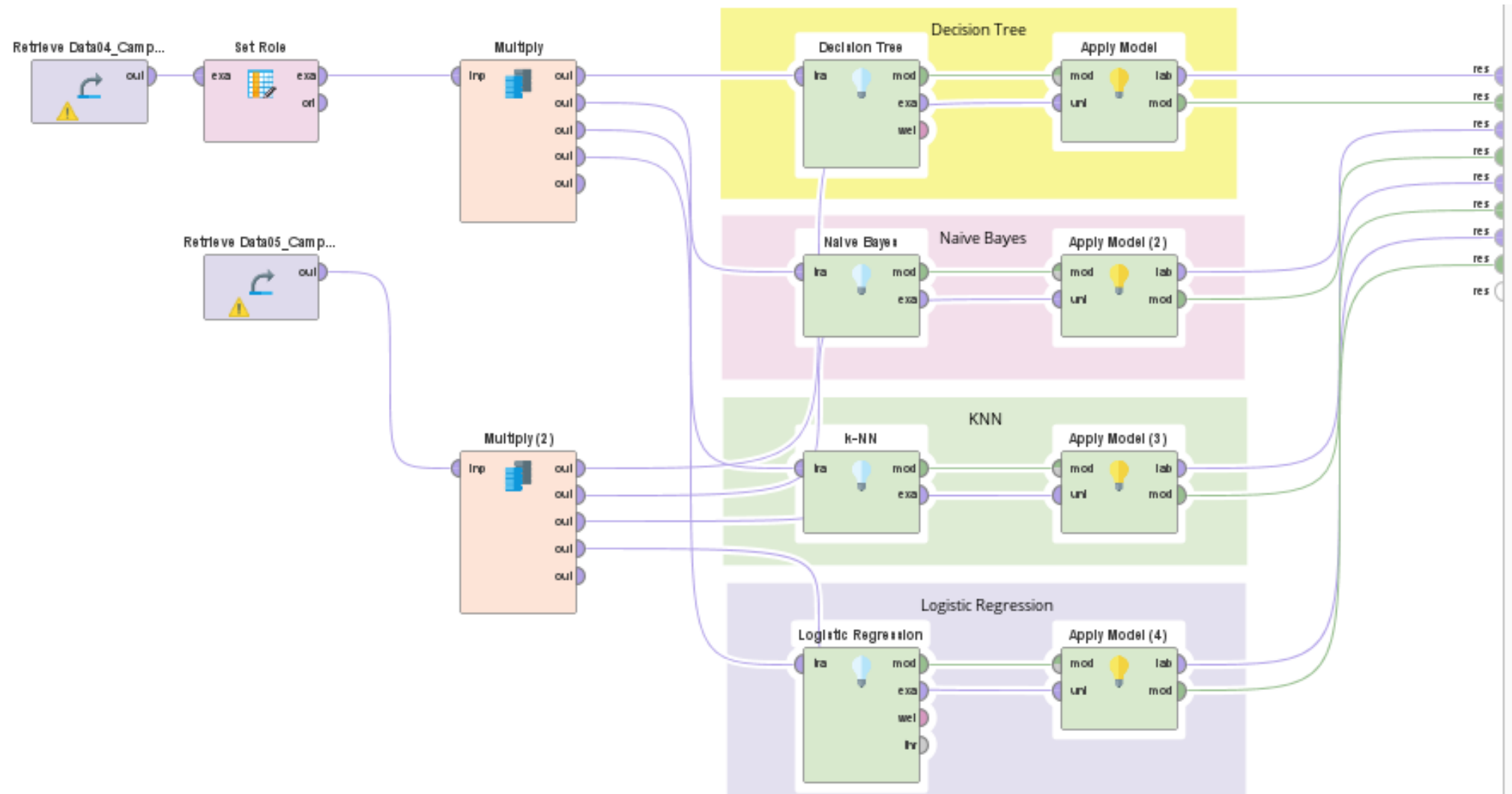
Decision Tree, Naïve Bayes, k-Nearest Neighbors and Logistic Regression

## Lab 16:

Use this operator if you want to  
uses the output more than once



# Lab 16: Process



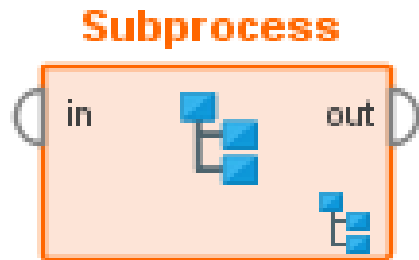
# LAB 17:

---

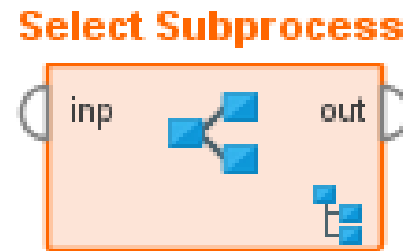
Decision Tree, Naïve Bayes, and k-Nearest Neighbors with Sub Processes

# Lab 17 :

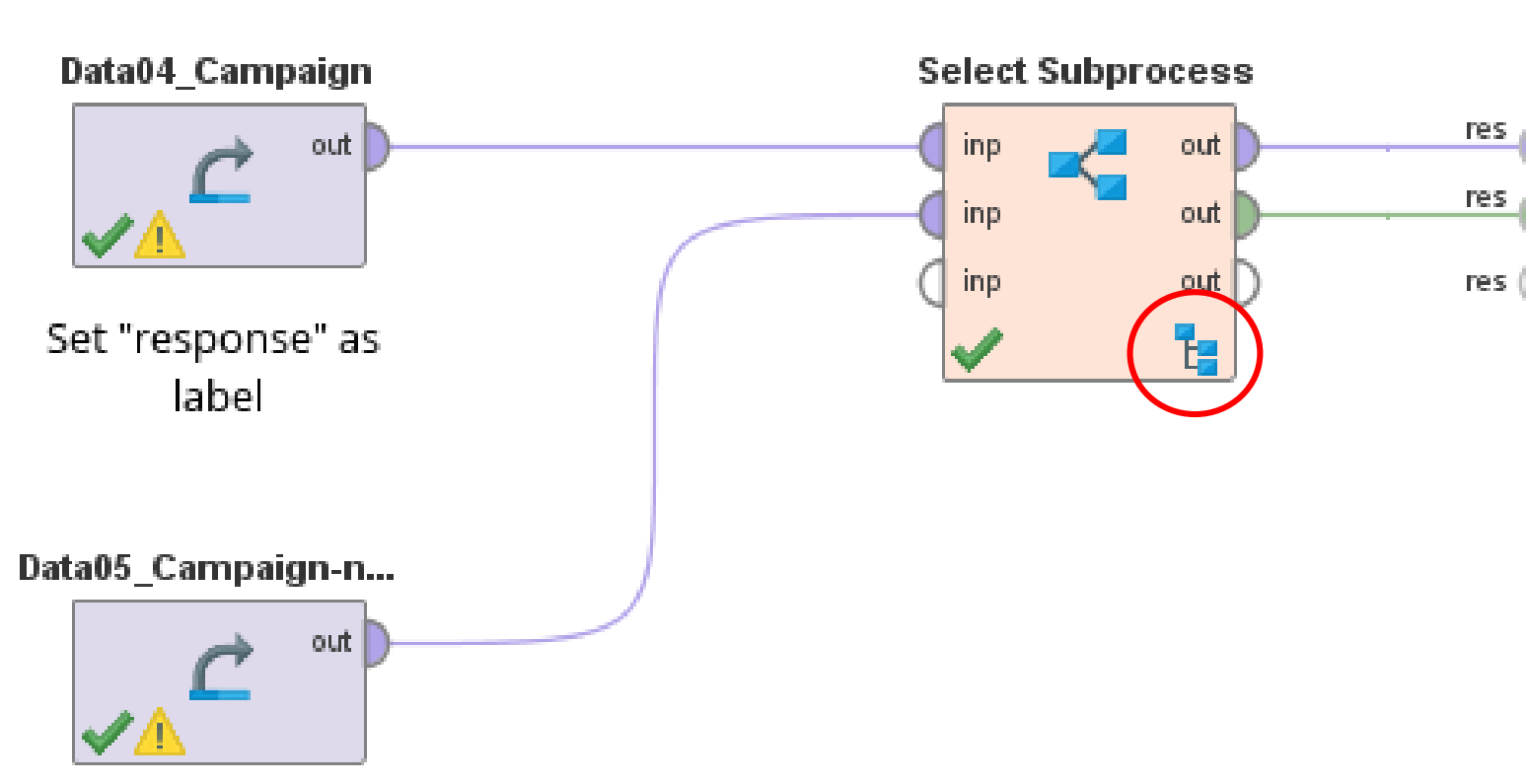
Use this operators to  
group some operator  
in one block



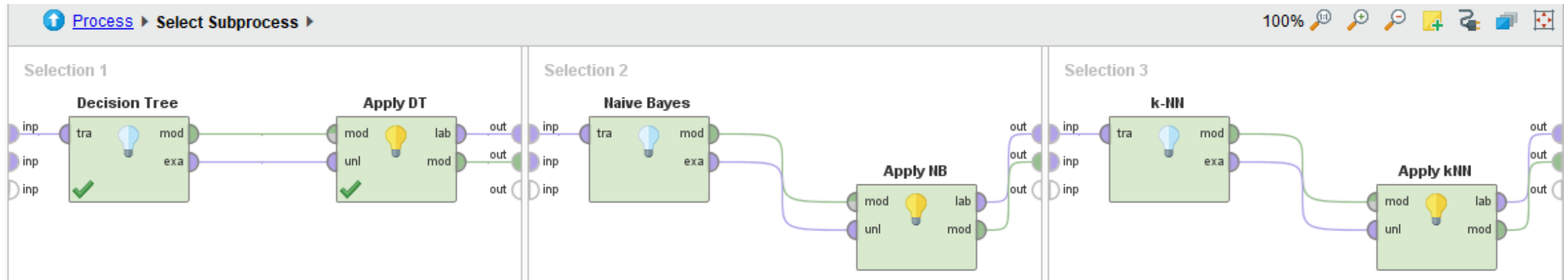
Use this operators to  
select subprocesses  
(like switch-case)



# Lab 17: Process



# Lab 17: Process



## Parameters

Select Subprocess

select which

1

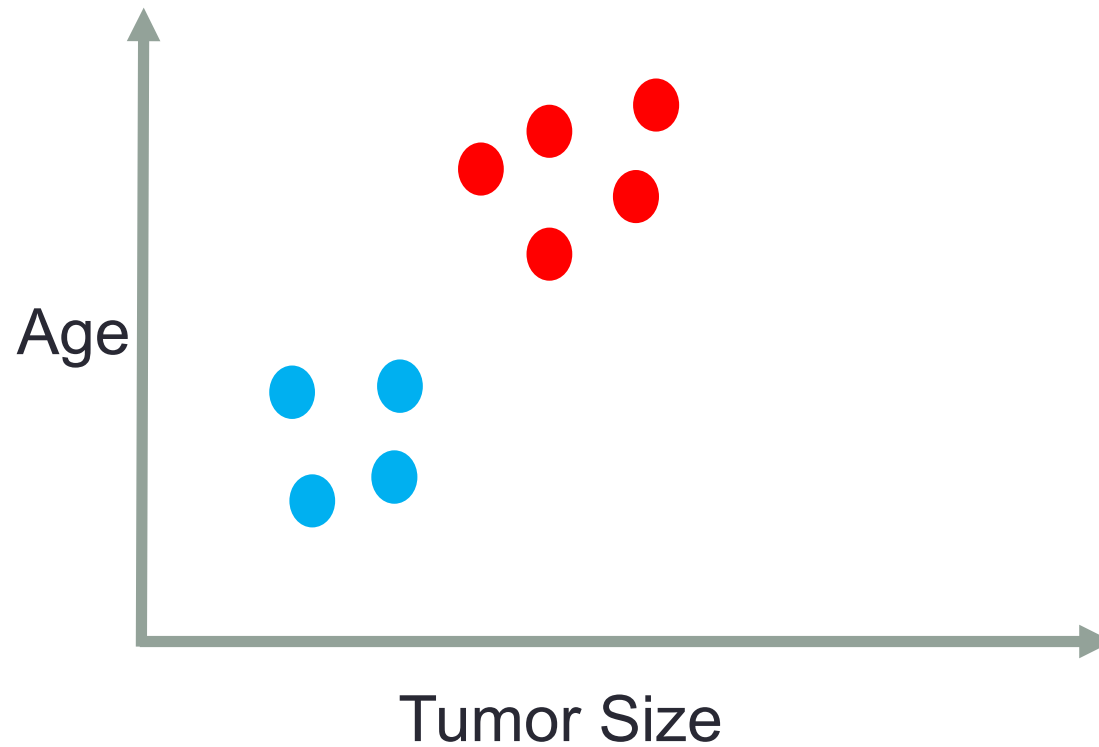
# UNSUPERVISED LEARNING:

## K-MEANS

---

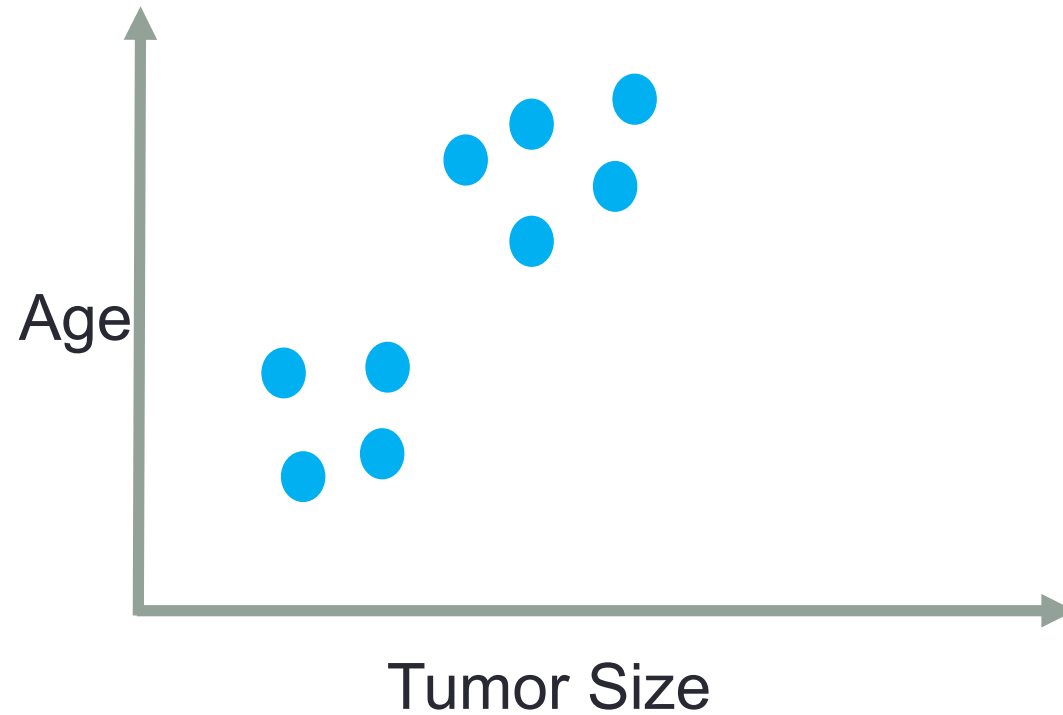


# Supervised Learning : Classification



$$\text{Training Set} = \{(x_1^{(1)}, x_2^{(1)}, y^{(1)}), (x_1^{(2)}, x_2^{(2)}, y^{(2)}), \dots, (x_1^{(n)}, x_2^{(n)}, y^{(n)})\}$$

# Unsupervised Learning



$$\text{Training Set} = \{(x_1^{(1)}, x_2^{(1)}), (x_1^{(2)}, x_2^{(2)}), \dots, (x_1^{(n)}, x_2^{(n)})\}$$

# ตัวอย่างการใช้งาน Clustering



Market segmentation

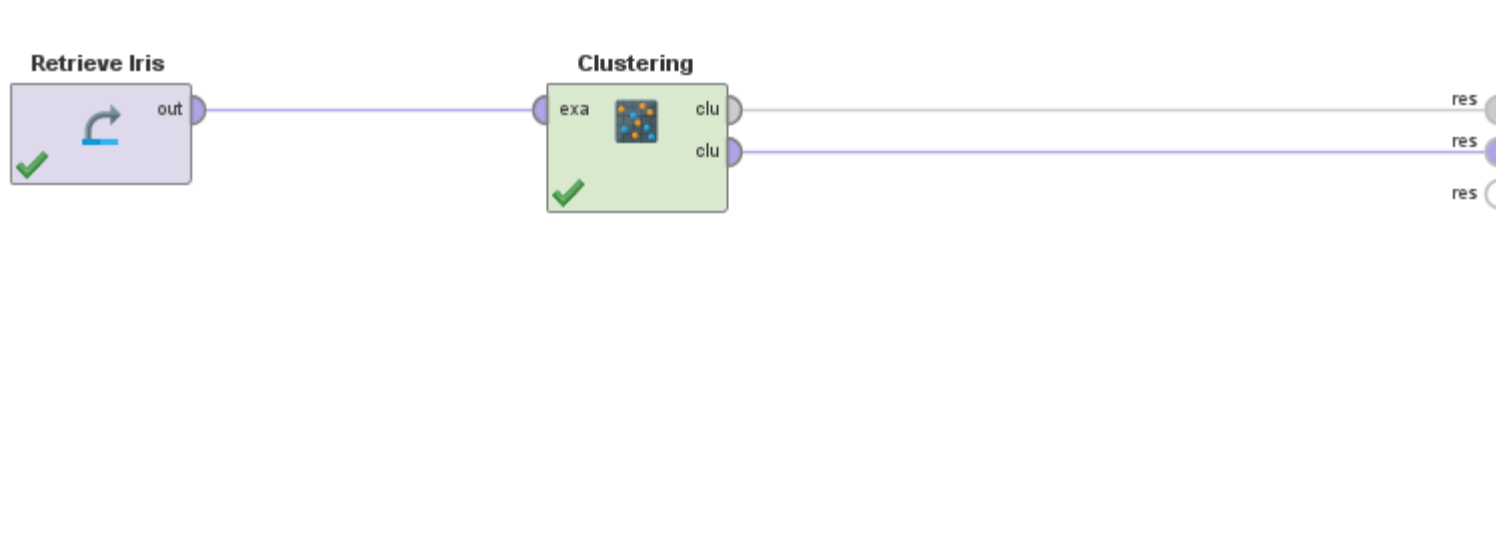


Social Network Analysis

# LAB 18: K-MEAN CLUSTERING

---

# Lab 18: Process



# Further Study

- Preprocessing
- Model Deployment
- Feature Selection
- Hyperparameter turning
- ...

# Resource

- [www.kaggle.com](http://www.kaggle.com)
- <https://archive.ics.uci.edu/ml/index.php>
- <https://academy.rapidminer.com/>