+

https://github.com/pvateekul/ieat2026

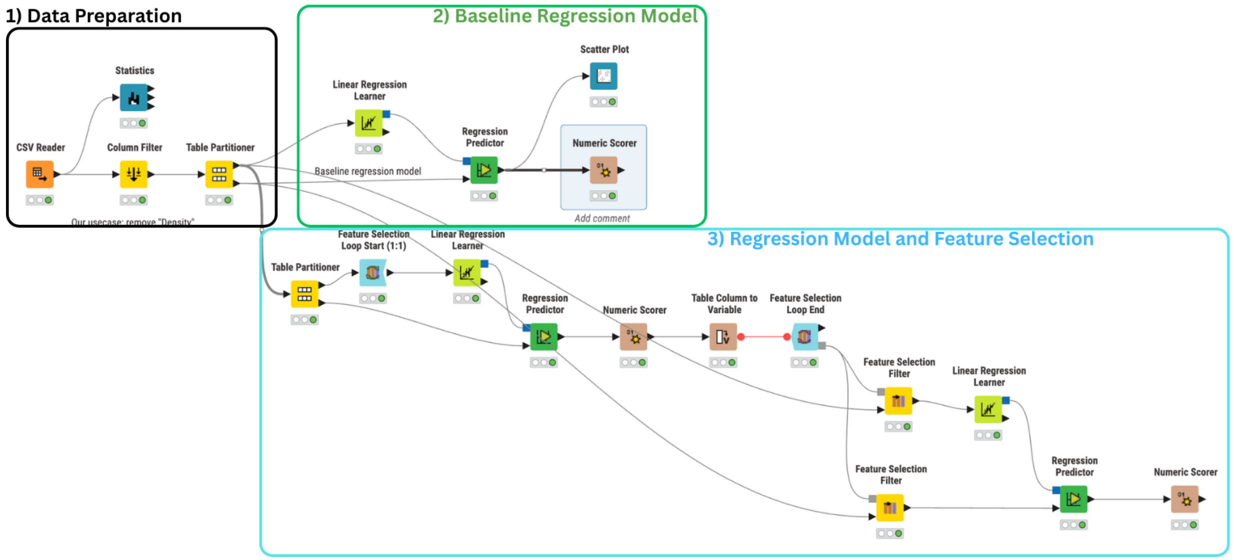
การนิคมอุตสาหกรรมแห่งประเทศไทย

Regression

Prof. Peerapon Vateekul, Ph.D.
Peerapon.v@chula.ac.th

Outlines

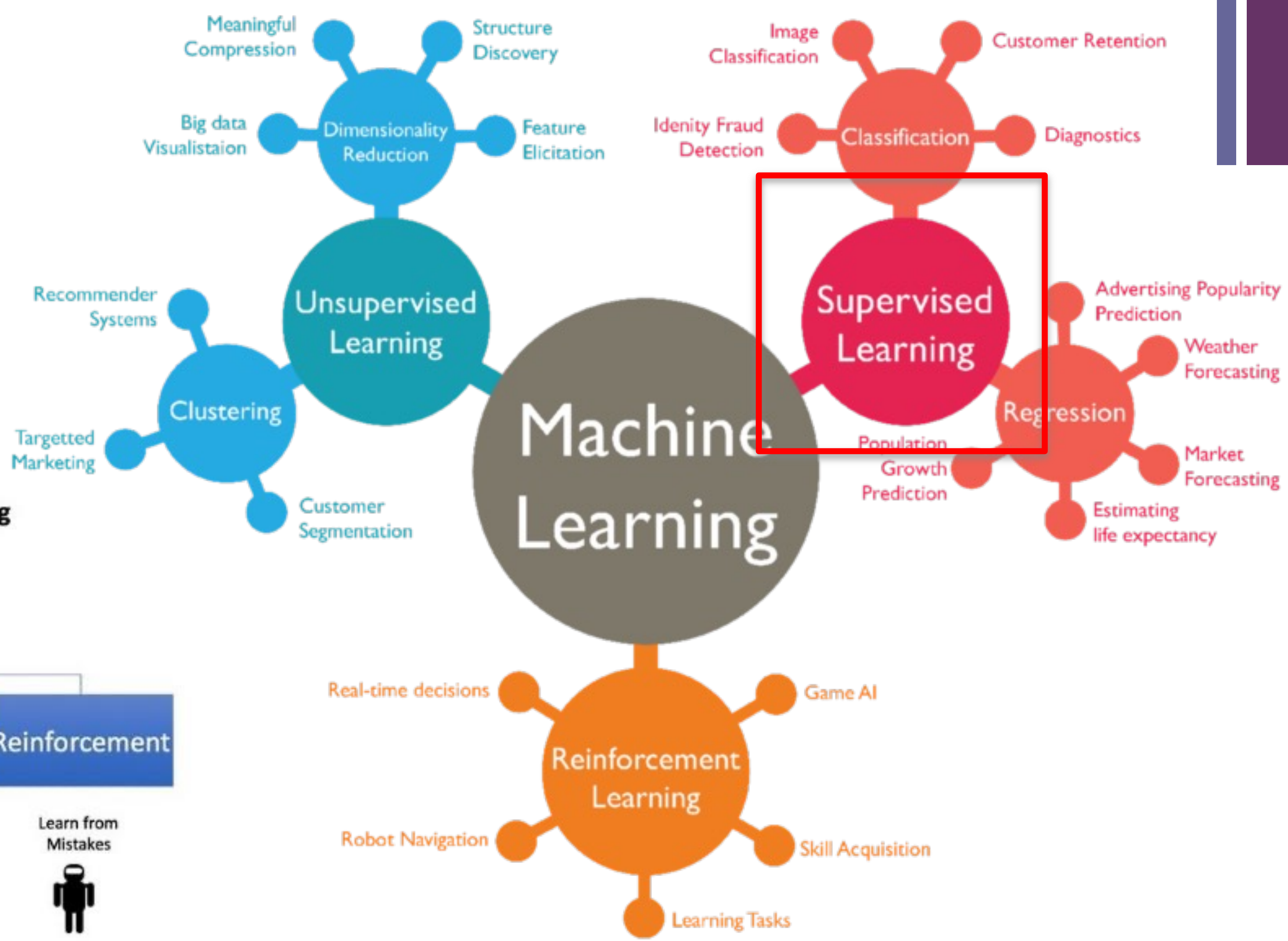
- Linear Regression
- Handle Categorical Variables
- Feature Selection
- Regression Performance
- LAB 2: Regression
- LAB 3: AutoML

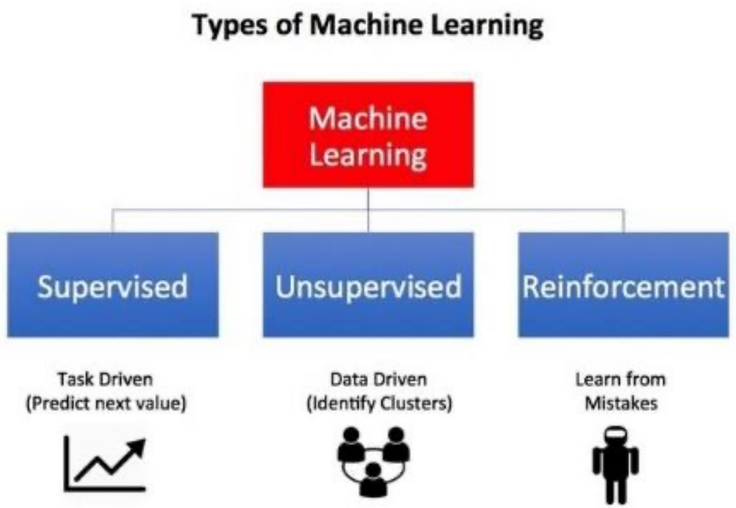


+

Supervised Learning (recap)

+ Machine Learning





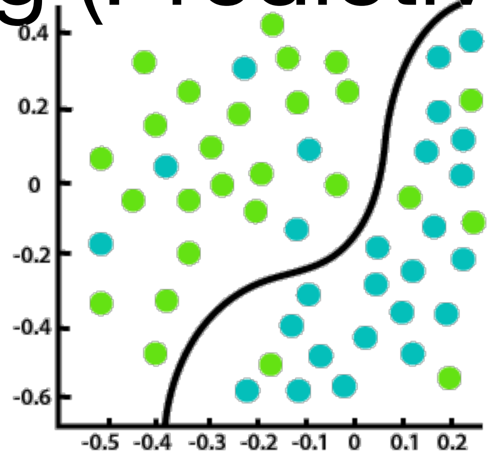
Supervised Learning (Predictive Task)

inputs | target

Age	Temp	Gender	Smell	Covid
25	39.0	Female	No	Yes
35	38.9	Female	No	Yes
32	36.5	Male	Yes	No

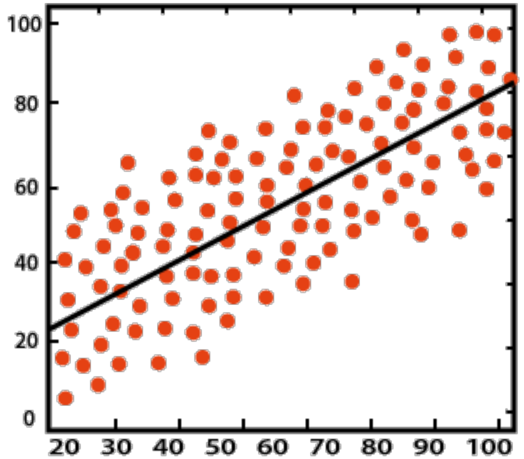
- Goal: To learn a prediction model mapping from inputs to output.

- Data without label (answer) is meaningless!
- Label should be provided by experts!



Classification

- Target is categorical variable.
- Example
- Covid diagnosis (yes/no)
- Disease diagnosis from gait information:
  - 1) Normal,
  - 2) Sick/Knee OA
  - 3) Sick/Parkinson



Regression

- Target is numeric variable.
- Example
- PD's state diagnosis from movement data.
- Glucose level prediction from breath particles.

There are two main processes: Train/Test

1) Training Phase: Model Construction

Training Data

inputs

target

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

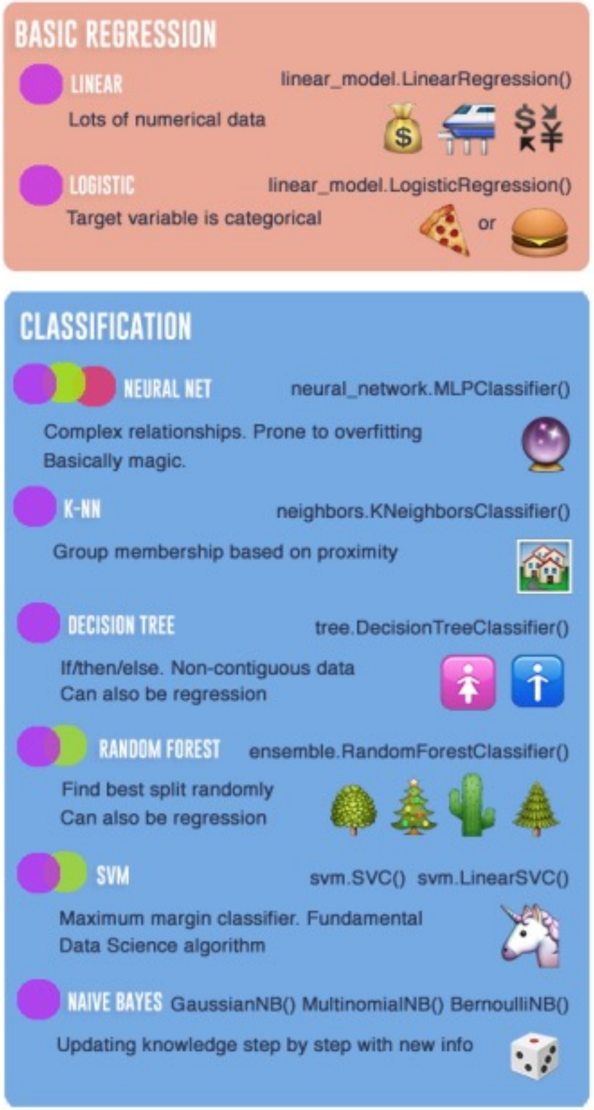
2) Testing Phase: Model Evaluation, Model Assessment
Also called "prediction, inference, scoring"

Testing Data

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	?

+ Prediction Algorithms

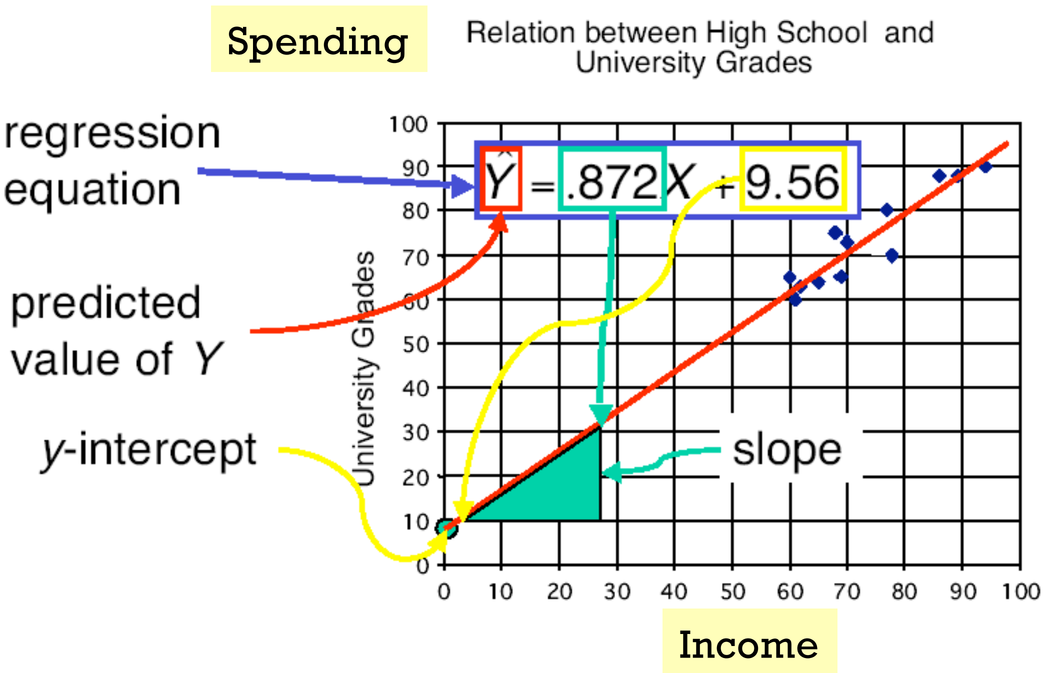
- Decision Tree
- (Logistic) Regression
- Neural Networks (NN)
- kNN
- Support Vector Machine
- Deep Learning



CHULA ENGINEERING
Foundation toward Innovation
COMPUTER

Linear Regression

Linear Regression



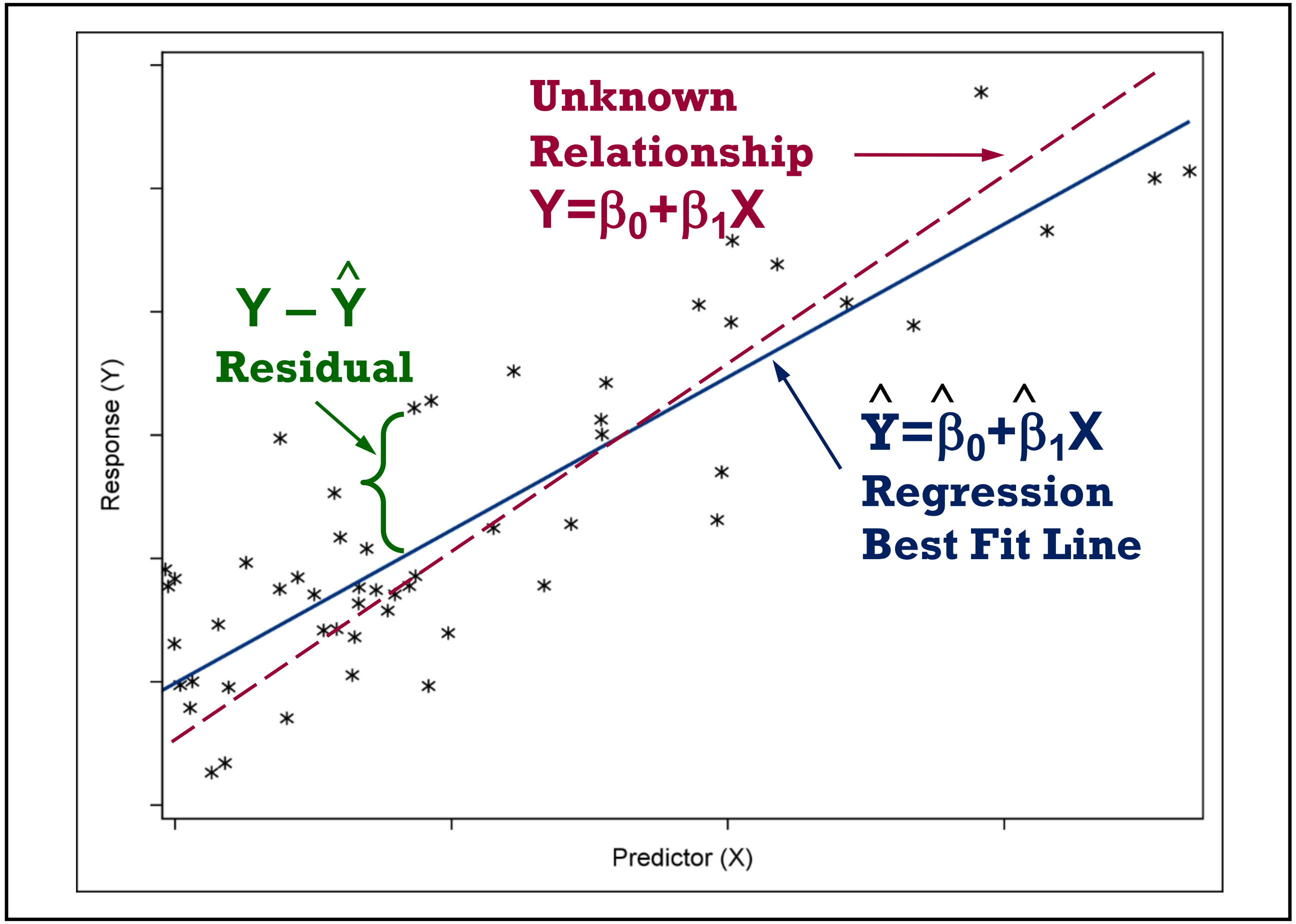
weight, coefficient

$$\underbrace{\hat{y}}_{\text{target}} = \underbrace{\hat{w}_0}_{\text{intercept}} + \hat{w}_1 x_1 + \hat{w}_2 \underbrace{x_2}_{\text{input}}$$

- **The least square method** aims to minimize the following term

$$\sum_{training\ data} (y_i - \hat{y}_i)^2$$

Ordinary Least Squares (OLS) Regression



Multiple Linear Regression Model
Matrix Multiplication Approach

	inputs	target
Age	**Income in k$**	**Spending**
25	25	400
35	50	500
32	35	550

$Y=\beta_0(1)+\beta_1 X_1+\beta_2 X_2+\varepsilon$

$$[Y] = [X][\beta]$$
$$[\beta]=[X]^{-1}[Y]$$

General least-squares solution:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{y} = \begin{bmatrix} 400 \\ 500 \\ 550 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 25 & 25 \\ 1 & 35 & 50 \\ 1 & 32 & 35 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

In this toy example, $\mathbf{X}$ is 3×3 and invertible, so it fits **exactly** and you can also use:

$$\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{y}$$

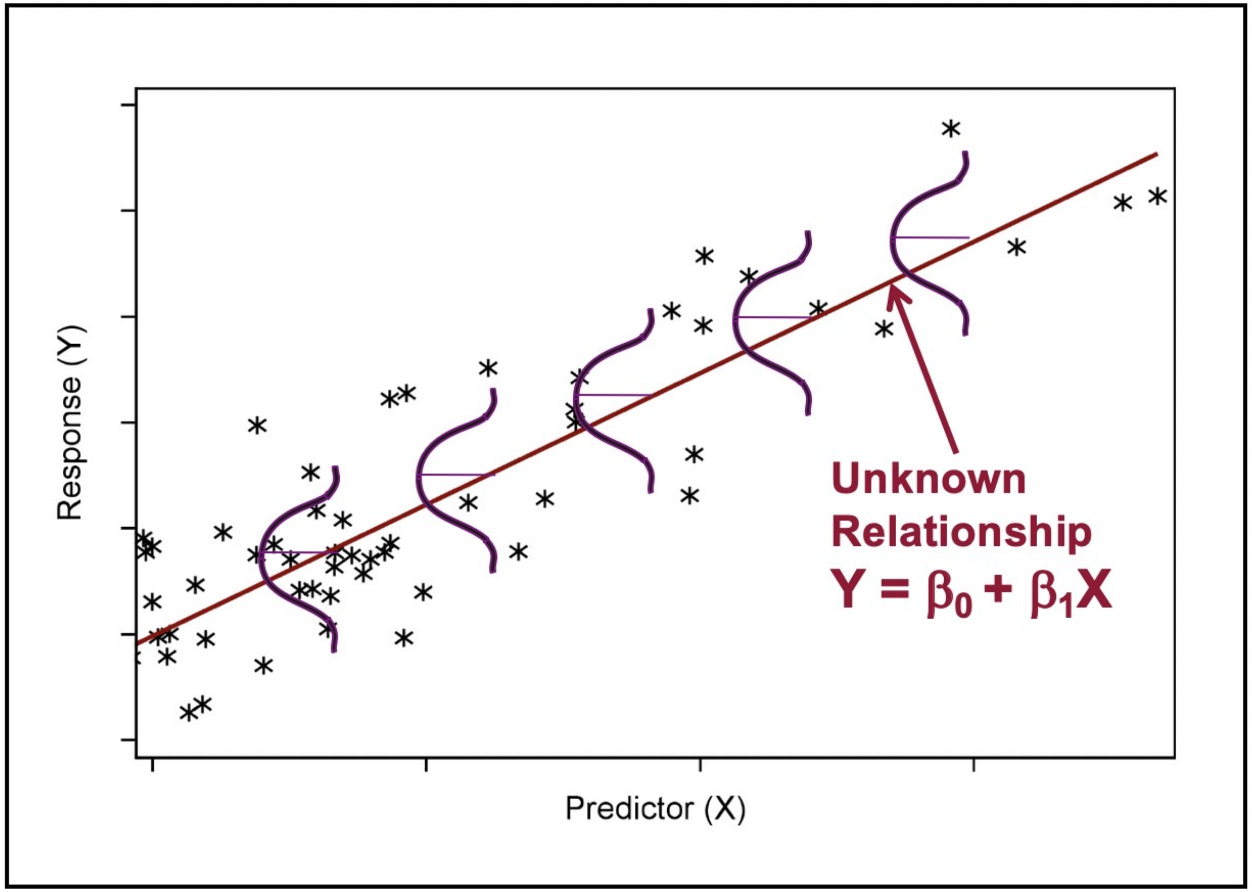
$$\boldsymbol{\beta} = \begin{bmatrix} -250 \\ \frac{110}{3} \\ -\frac{32}{3} \end{bmatrix} \approx \begin{bmatrix} -250 \\ 36.6667 \\ -10.6667 \end{bmatrix}$$

$$\widehat{Spending} = -250 + 36.6667(\text{Age}) - 10.6667(\text{Income in k\$})$$

https://sites.google.com/site/mystatistics01/regression-correlation-analysis/multiple-linear-regression-step

Linear Regression Assumption

Spending = 500 + 10*Income10K + 2*Age



- Linear relationship between (y, x_i). [Pearson correlation]
- Error is normal distributed. [remove outliers, log-transformation]
- Error has equal variance (homoscedasticity) [remove outliers, log-transformation]
- Errors are independent from each other. [design new data correction]

Linear Regression Limitation

- Manage missing value
- Handle outliers (skewness)
- Handle categorical variables (dummy codes)
- Variable selection:
  - Wrapper (all combinations)
  - Forward, Backward, Stepwise
- Accounted for nonlinearities

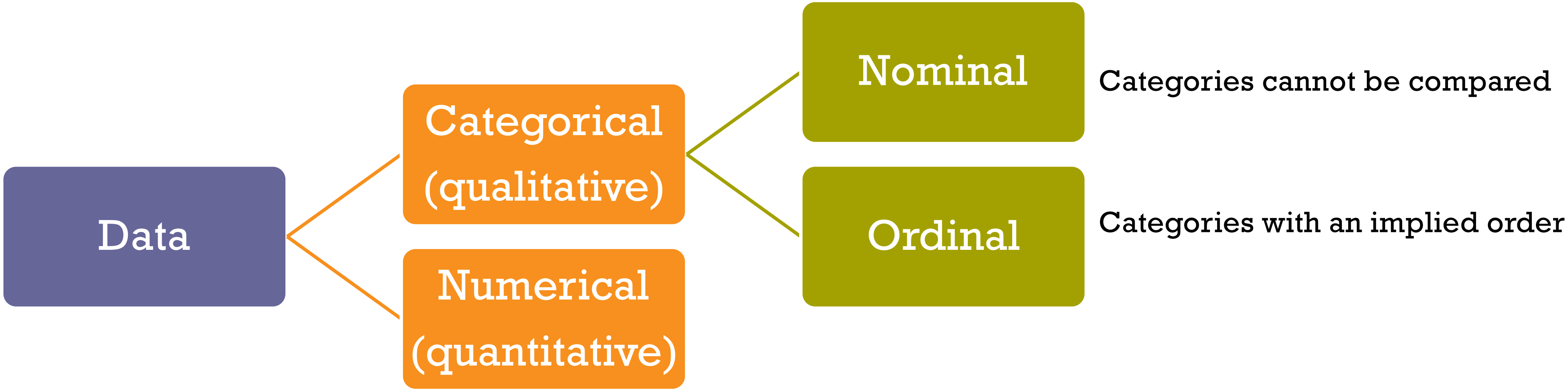
Remarks:
- Require good data preparation
- Variable selection

CHULA ΣNGINEERING Foundation toward Innovation COMPUTER

Handle Categorical Variables

Dummy coding, OneHotEncoder

Terminology: Kinds of data (recap)



Ordinal: Recode

Grade	GradeNum
A	4.00
B+	3.50
B	3.00
C+	2.50
C	2.00
D+	1.50
D	1.00
F	0.00

Size	SizeNum
XL	5
L	4
M	3
S	2
XS	1

Categorical

- Dummy coding = (n-1) dummy codes

Branch	BranchNum	D_BKK	B_Patum	D_Non	D_BKK	B_Patum	
BKK	1	1	0	0	1	0	
Patumtani	2	0	1	0	0	1	
Nontaburi	3	0	0	1	0	0	reference

+

Feature Selection

Feature Selection Approach

- Sequetial Feature Selection:
  - Sequential Feature Selection
  - Backward Feature Elimination
- Best Subet Selection (create all combinations & pick the best one)

CHULA ΣNGINEERING Foundation toward Innovation COMPUTER

Sequential Feature Selection

Start with no features, then add the feature that improves performance the most at each step.

Goal: Predict *Spending*

Candidate features:

`Age, Income, Education, Gender`

Step 0: No features

- Model: intercept only
- RMSE = 120

Step 1: Try adding one feature at a time

Feature added	RMSE
Age	80
Income	**60** ✅
Education	95
Gender	110

➡️ Select **Income**

Step 2: Try adding one more feature (given Income)

Features	RMSE
Income + Age	**45** ✅
Income + Education	58
Income + Gender	62

➡️ Add **Age**

Step 3: Try adding another feature

Features	RMSE
Income + Age + Education	46
Income + Age + Gender	47

➡️ No improvement → **STOP**

✅ **Final selected features**

```code
Income, Age
```

+ Backward Feature Elimination

Backward elimination starts with all features and iteratively removes the least useful one based on performance impact.

Goal: Predict *Spending*

Candidate features:

`Age, Income, Education, Gender`

Step 0 — Start with ALL features

```code
Age, Income, Education, Gender
```

Baseline: **RMSE = 60**

Step 1 — Try removing ONE feature at a time

Feature removed	Remaining features	Val RMSE
Age	Income, Education, Gender	95 ❌
Income	Age, Education, Gender	110 ❌
Education	Age, Income, Gender	62 ❌
Gender	**Age, Income, Education**	**58** ✅ (better)

✅ **Decision:** remove **Gender** (RMSE improves: 60 → 58)

Step 2 — Current set

```code
Age, Income, Education
```

Baseline: **RMSE = 58**

Try removing one:

Feature removed	Remaining features	Val RMSE
Age	Income, Education	88 ❌
Income	Age, Education	105 ❌
Education	**Age, Income**	**57** ✅ (better)

✅ **Decision:** remove **Education** (58 → 57)

Step 3 — Current set

```code
Age, Income
```

Baseline: **RMSE = 57**

Try removing one:

Feature removed	Remaining features	Val RMSE
Age	Income	80 ❌
Income	Age	92 ❌

🔴 **Decision:** STOP

✅ **Final selected features**

```code
Income, Age
```

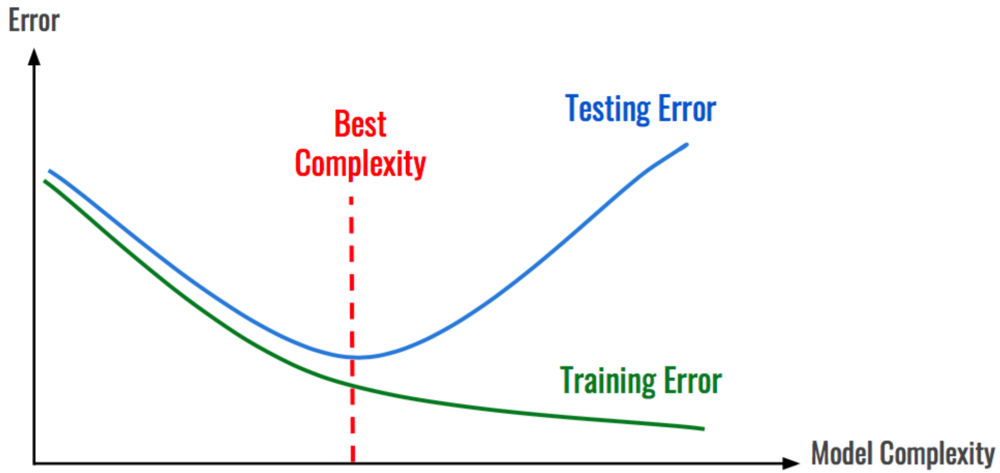
Evaluation (Train/Test Split)

Training Data

Age	Income	Purchase
25	25,000	Yes
35	50,000	Yes
32	35,000	No

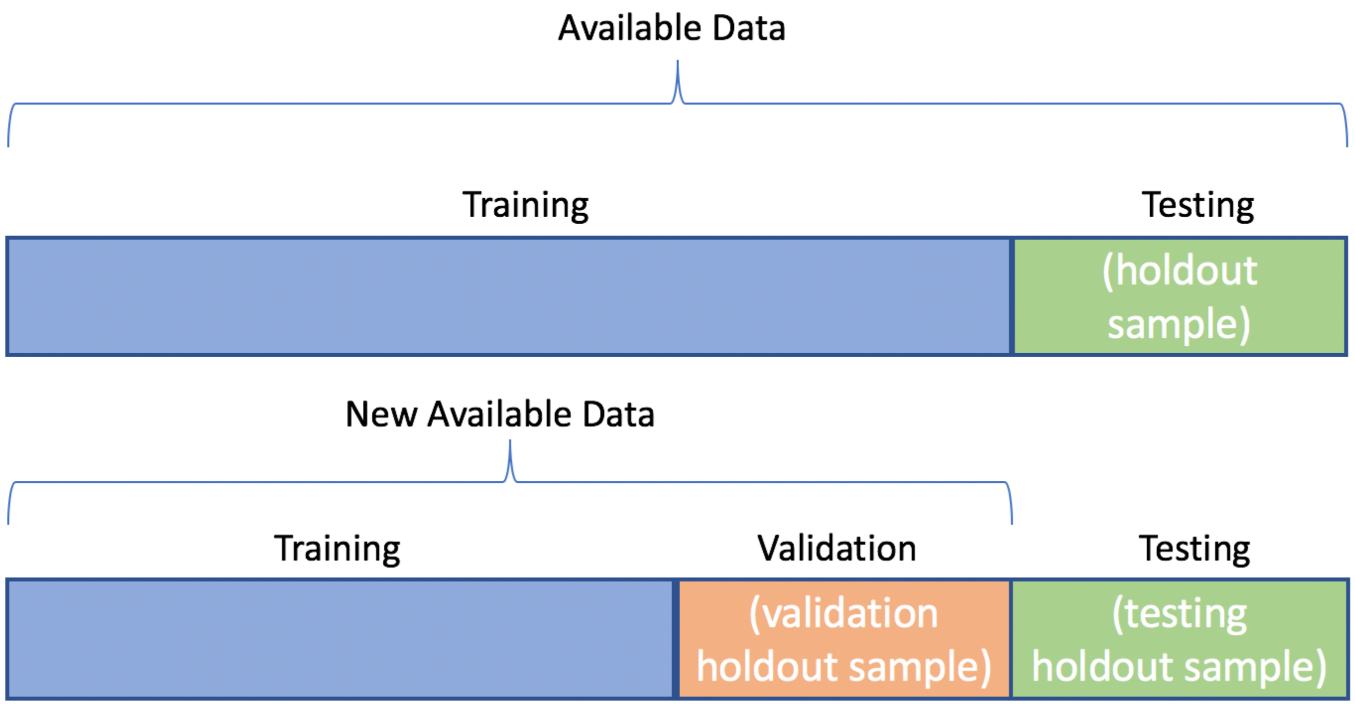
Testing Data

Age	Income	Purchase
27	35,000	Yes
23	20,000	No
45	34,000	No



+ Train (Validation) & Test

- Training data = Textbook
- Validation data = Exercise
- Testing data = Final exam

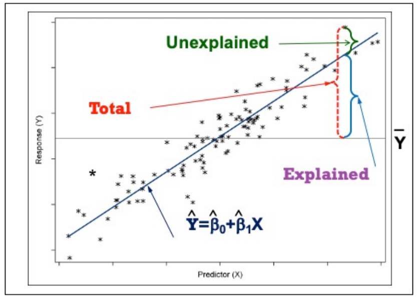


Regression Performance

Regression	
‘mean_absolute_error’	`sklearn.metrics.mean_absolute_error`
‘mean_squared_error’	`sklearn.metrics.mean_squared_error`
‘r2’	`sklearn.metrics.r2_score`

Coefficient of Determination

Regression	
'mean_absolute_error'	`sklearn.metrics.mean_absolute_error`
'mean_squared_error'	`sklearn.metrics.mean_squared_error`
'r2'	`sklearn.metrics.r2_score`



$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

id	chol (x)	bp (y)	predict	error	squred error (SE)	guess	(y - y_bar)	squred total (ST)
1	437	194	196.1897	(2.1897)	4.7948	143.4286	50.5714	2,557.4694
2	264	121	141.4179	(20.4179)	416.8906	143.4286	(22.4286)	503.0408
3	249	131	136.6689	(5.6689)	32.1364	143.4286	(12.4286)	154.4694
4	297	159	151.8657	7.1343	50.8982	143.4286	15.5714	242.4694
5	243	123	134.7693	(11.7693)	138.5164	143.4286	(20.4286)	417.3265
6	272	161	143.9507	17.0493	290.6786	143.4286	17.5714	308.7551
7	161	115	108.8081	6.1919	38.3396	143.4286	(28.4286)	808.1837
average	274.7143	143.4286		SSE	972.2548		SST	4,991.7143
				MSE	138.8935			
				RMSE	**11.7853**			
	R^2	**1 - (SSE/SST)**	**0.8052**					

CHULA ΣNGINEERING
Foundation toward Innovation
COMPUTER

LAB 2: Regression

LAB 2: Regression

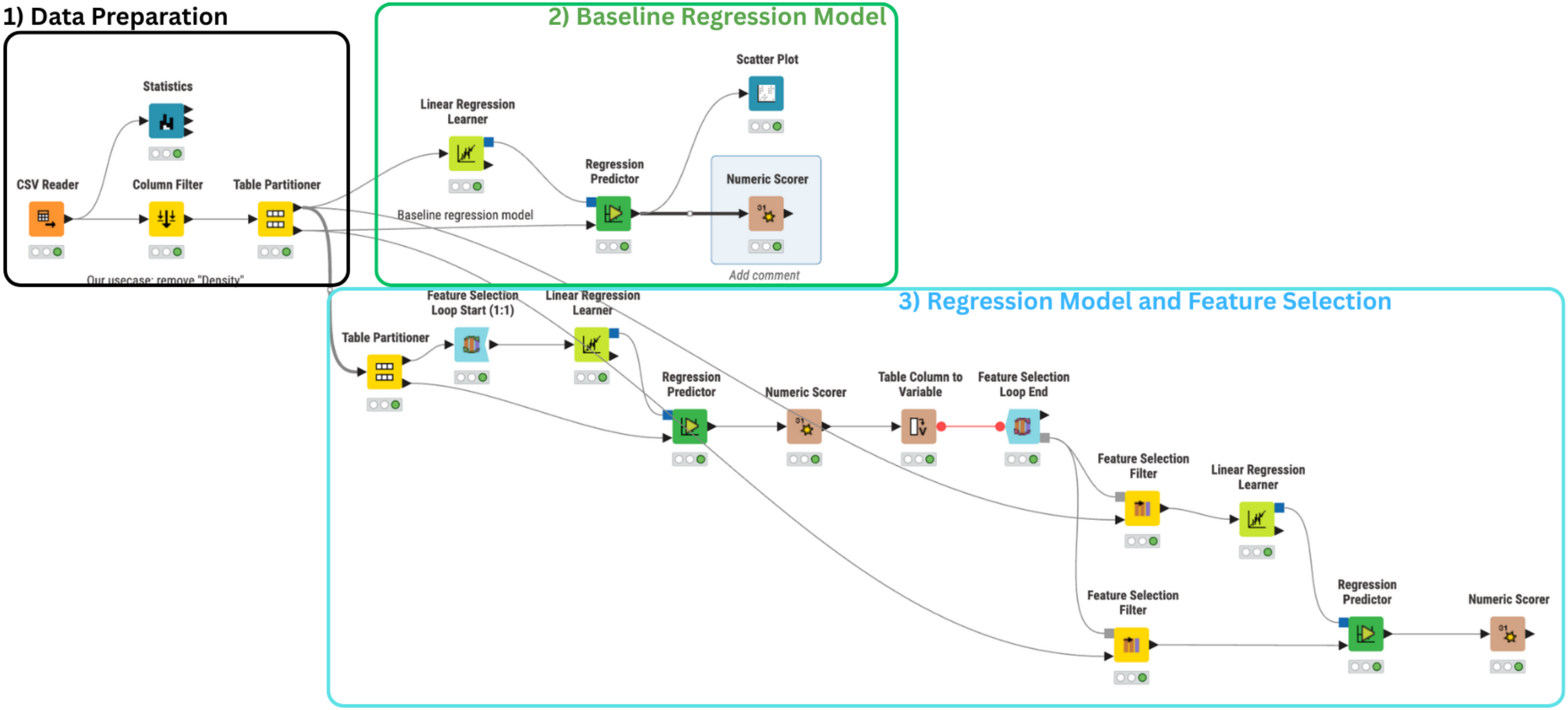
Data Description

- The BodyFat dataset contains 252 adult male subjects. Each row is one subject. The goal is to predict BodyFat (%) from body measurements.
- Target Variable: Body fat percentage (%)

Input:

- Age: Age (years)
- Weight: Body weight (lbs)
- Height: Height (inches)
- Neck: Neck circumference (cm)
- Chest: Chest circumference (cm)
- Abdomen: Abdomen/waist circumference (cm)
- Hip: Hip circumference (cm)
- Thigh: Thigh circumference (cm)
- Knee: Knee circumference (cm)
- Ankle: Ankle circumference (cm)
- Biceps: Biceps circumference (cm)
- Forearm: Forearm circumference (cm)
- Wrist: Wrist circumference (cm)
- Density: Body density estimate (typically in g/cm^3).
- Lab note: We remove Density to simulate a realistic scenario where this measurement is not available.

LAB 2: Regression



CHULA ΣNGINEERING
Foundation toward Innovation
COMPUTER

LAB 3: AutoML

LAB 3: AutoML

Data Description

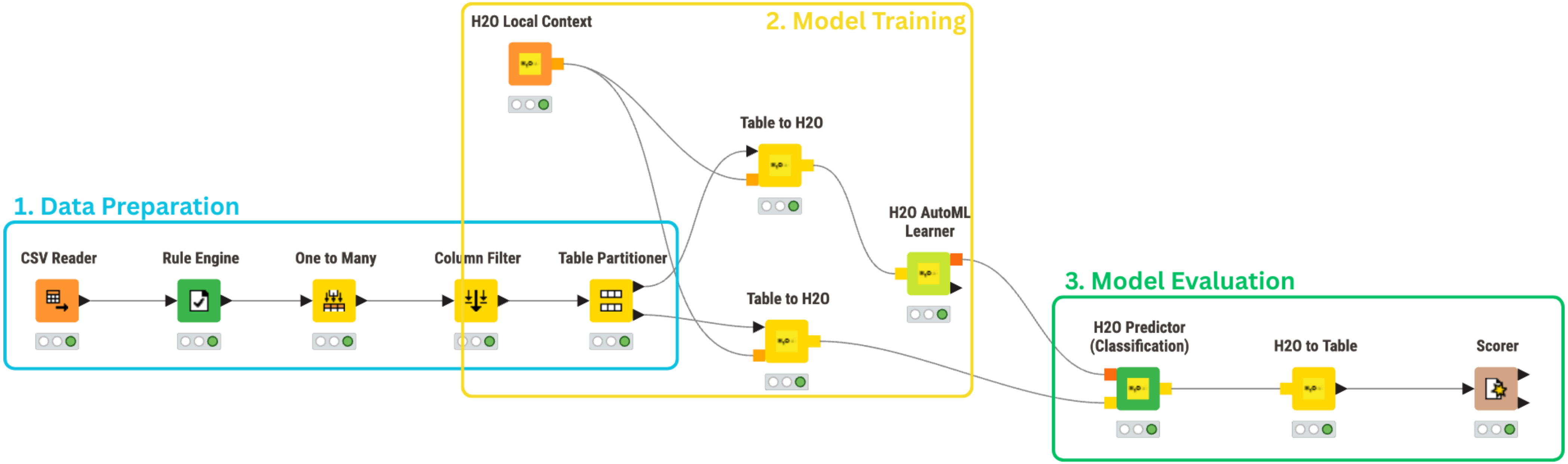
- The Mall Customer dataset contains 200 customers. Each row is one subject. The goal is to predict the Spending Class (High or Low Spender) of the customer.

- Target Variable: High Spender: “yes” if spending score >= 60; else “no”

Input:

- CutomerID
- Gender: Male/Female
- Age: (years)
- Annual Income: (k$)
- Spending Score: Derived Metric (0-100)

LAB 3: AutoML


1. Data Preparation
CSV Reader
Rule Engine
One to Many
Column Filter
Table Partitioner
2. Model Training
H2O Local Context
Table to H2O
H2O AutoML Learner
Table to H2O
3. Model Evaluation
H2O Predictor (Classification)
H2O to Table
Scorer

Thank you & any questions