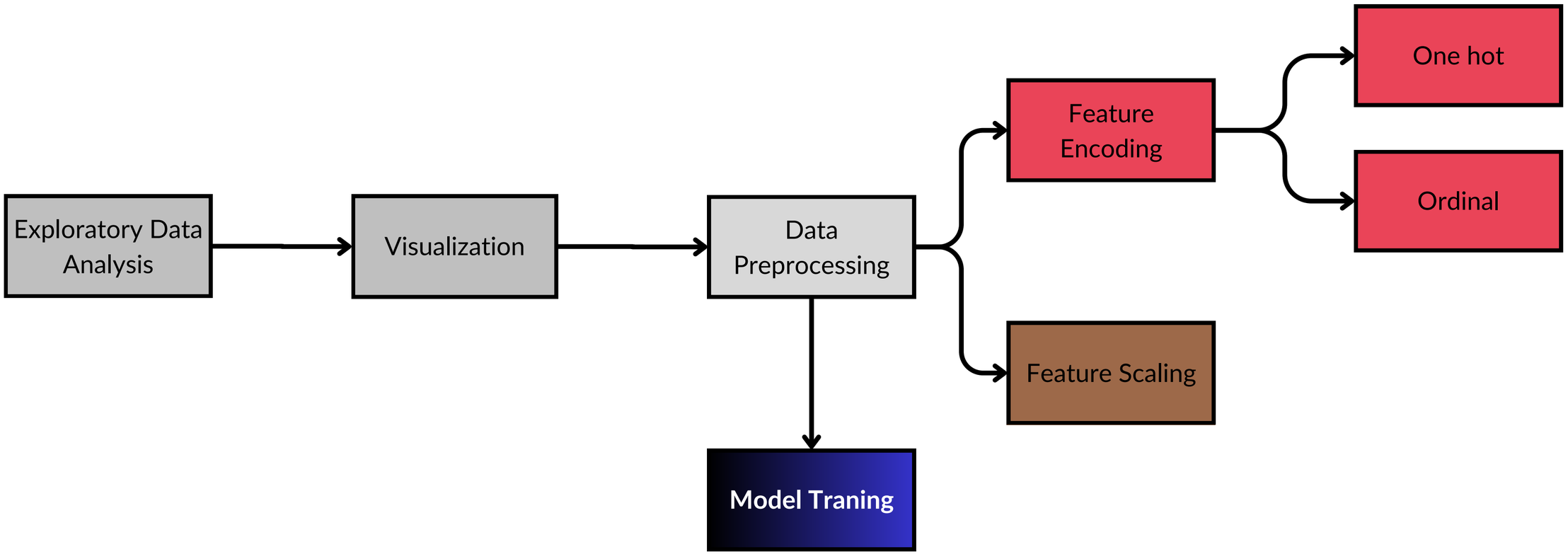
FEATURE SCALING

What is Feature Scaling

Feature scaling refer to the methods used to normalize the range of values of independent variables.

- In other word, the methods to set the feature value range within a similar scale.



Why Feature Scaling

Regression Coefficient and Scale

The size of the regression coefficient depends on the scale of the variable. Larger numbers in data can overpower smaller ones when building a regression model.

Bigger Numbers Dominate

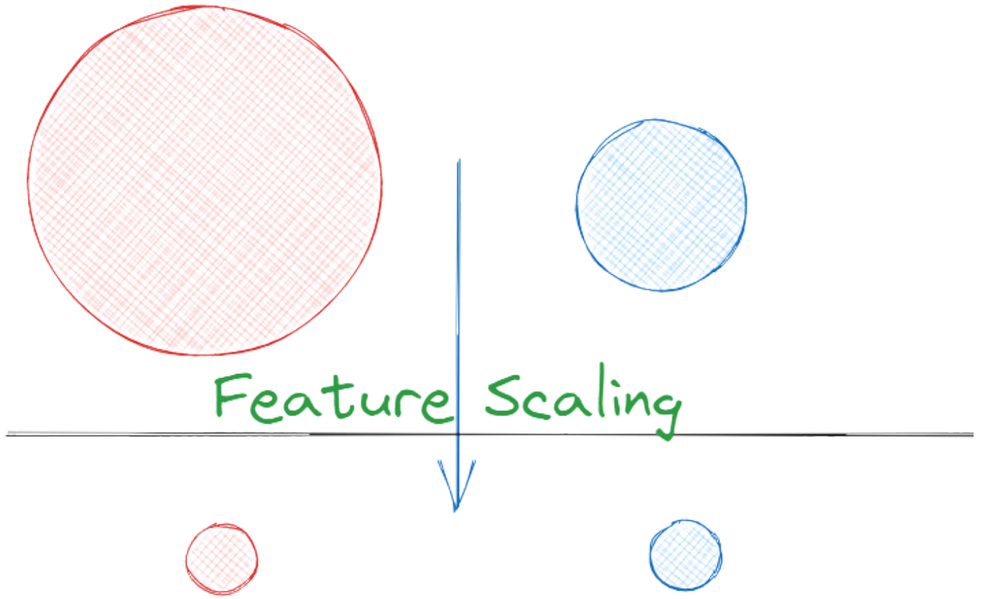
If one variable has values in thousands and another has values in single digits, the larger numbers will have more influence in the model, even if the smaller numbers are just as important.

Gradient Descent and Scaling

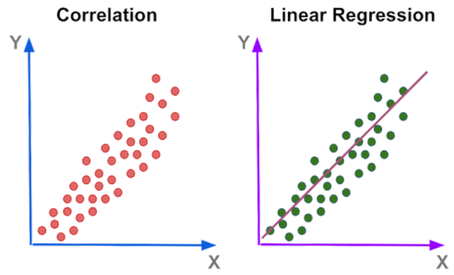
When all features are on a similar scale (e.g., 0–1), gradient descent, a method to train models, works faster and finds the best solution more quickly.

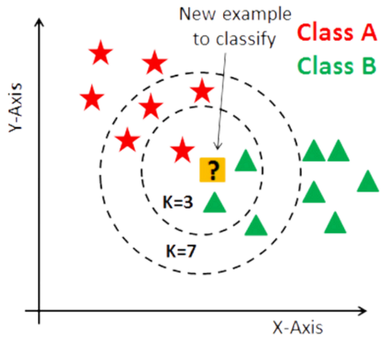
Scaling for SVM

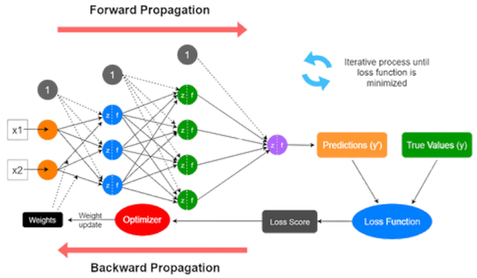
Scaling features also speeds up the process of finding the support vectors in Support Vector Machines (SVMs), making the algorithm more efficient.



Why Feature Scaling







The Machine Learning models affected by the magnitude of the feature:

- **Linear and Logistics Regression**
- **Neural Networks**
- **KNN**
- **K-means clustering**
- **SVMs**
- **Linear Discriminant Analysis (LDA)**
- **Principal Component Analysis (PCA)**

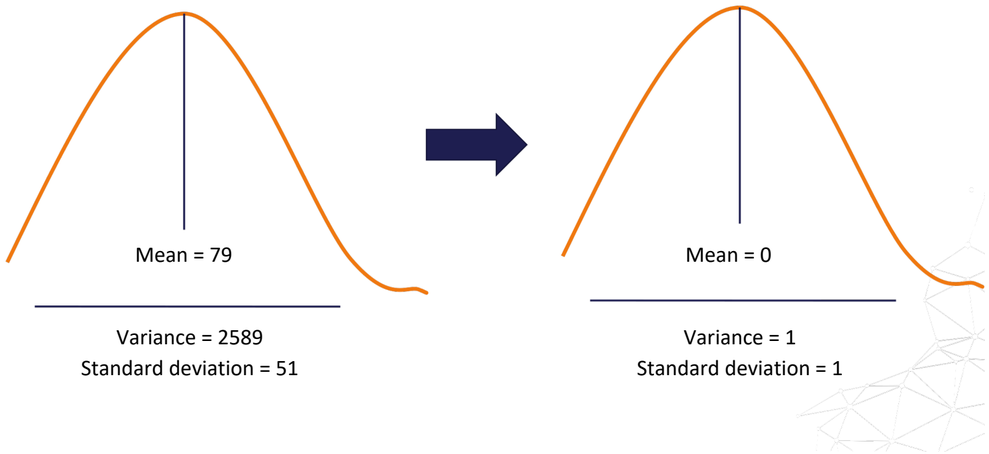
Machine Learning model insensitive to feature magnitude are the are the ones based on tree:

- **Classification and Regression Trees**
- **Random Forests**
- **Gradient Boosted Trees**

Standardisation

Mean the variables at 0 and sets the variance to 1

$$z = \frac{x - \text{mean}(x)}{\text{std}(x)}$$



Standardisation: example

Price
100
90
50
40
20
100
50
60
120
40
200

Mean = 79
Standard dev = 51

Obs. -Mean

Standard dev

Price
0.41
0.22
-0.57
-0.76
-1.16
0.41
-0.57
-0.37
0.80
-0.76
2.37

Min-Max Scaing

Scales the variable between 0 and 1

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

MinMaxScaling: example

Price
100
90
50
40
20
100
50
60
120
40
200

Max = 79
Min = 20
Range = 180

Price
0.44
0.39
0.17
0.11
0.00
0.44
0.17
0.22
0.56
0.11
1.00

Min-Max vs.Standardisation

ទីវិទ្យាស្ថានសាន់ប៊ុយស៍
SUNRISE INSTITUTE

	Standardisation	Min-Max
Range	**Mean = 0, Std. Dev = 1**	**Less sensitive**
Sensitivity to outliers	**Scaled to [0, 1] (or other specified range)**	**Highly sensitive**
Preferred for	**Distance-based algorithms (e.g., PCA, SVM)**	**Neural Networks, kNN, K-Means**