داده ها و انواع الگوریتم ها در یادگیری ماشین

Mohammad Jahanbakhsh

انواع داده

- عددی (پیوسته یا گسسته) Numeric
- ترتیبی Order
- اسمی Nominalیا طبقه ای Categorical
- باینری Binary

وزن ، تعداد بچه

عالی / خوب / متوسط

رنگ چشم

جنسیت ، تب داشتن

Binary data

	Beer	Wine	Juice	Coffee	Tea
Leia	*no*	*yes*	*yes*	*no*	*yes*
Luke	*yes*	*yes*	*yes*	*no*	*no*
Han	*yes*	*no*	*yes*	*yes*	*no*

Possible values: yes, no

Binary data

	Beer	Wine	Juice	Coffee	Tea
Leia	0	1	1	0	1
Luke	1	1	1	0	0
Han	1	0	1	1	0

1 = yes, 0 = no

Another possible codification could be “yes” = 1, and “no” = 0

Or also with logical values: “yes” = TRUE, and “no” = FALSE.

Ordinal encoding

How often do you drink these beverages?

	Beer	Wine	Juice	Coffee	Tea
Leia	*never*	*some*	*always*	*some*	*always*
Luke	*always*	*some*	*always*	*never*	*never*
Han	*always*	*some*	*some*	*always*	*never*

Possible values: never, sometimes, always

How often do you drink these beverages?

	Beer	Wine	Juice	Coffee	Tea
Leia	1	2	3	2	3
Luke	3	2	3	1	1
Han	3	2	2	3	1

1 = never, 2 = sometimes, 3 = always

you will very likely have to transform the values of the categories to some numeric coding. For instance, you can assign values 1 = “never”, 2 = “sometimes”, and 3 = “always.”

One-hot encoding

Color
Red
Red
Yellow
Green
Yellow

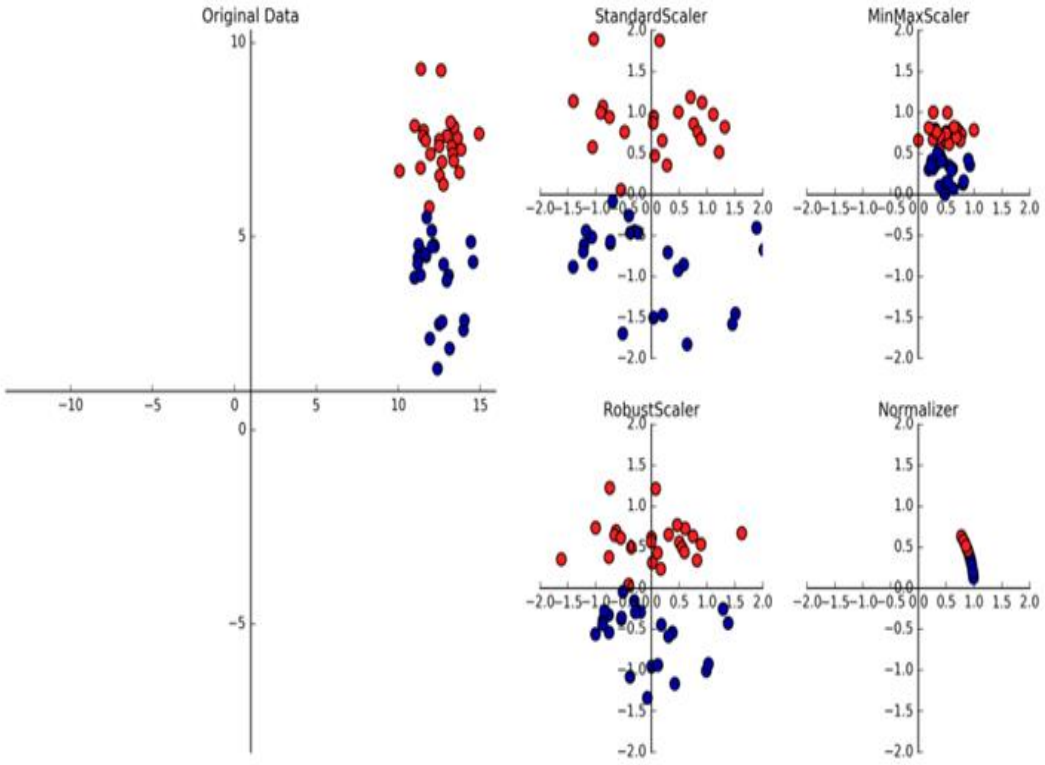
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

?

?

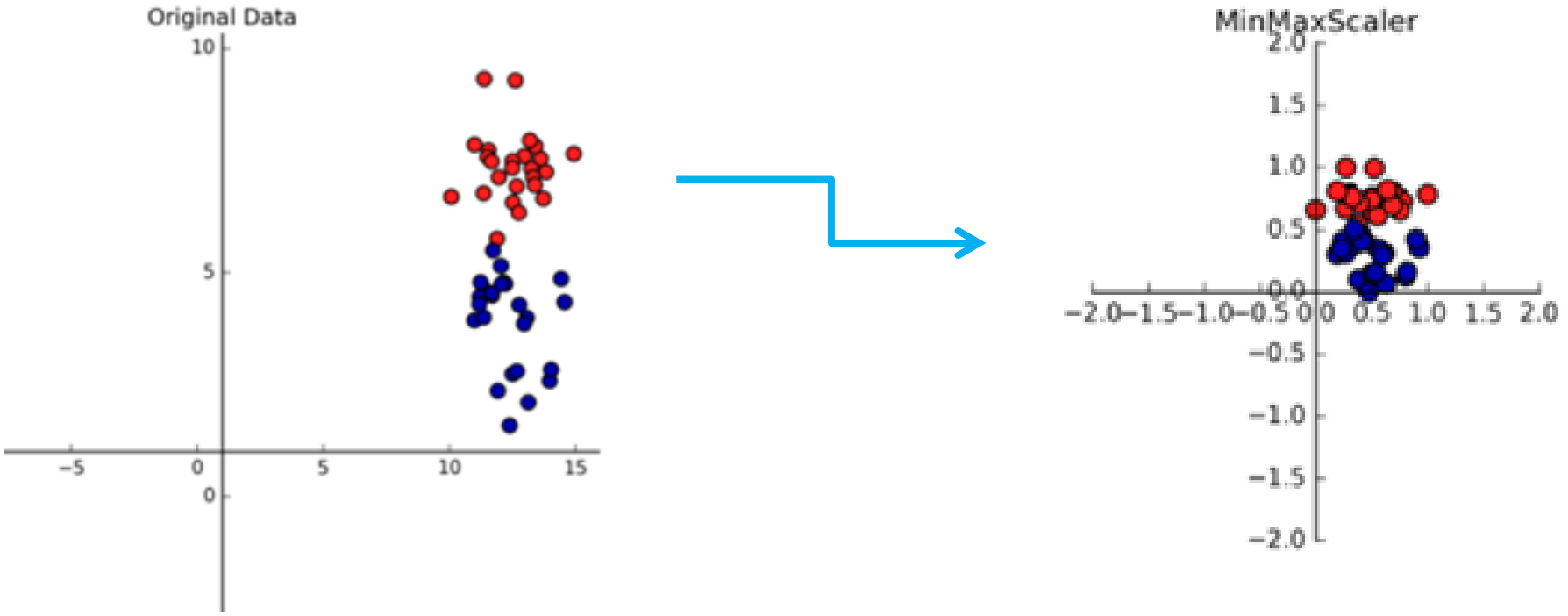
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Feature Normalization



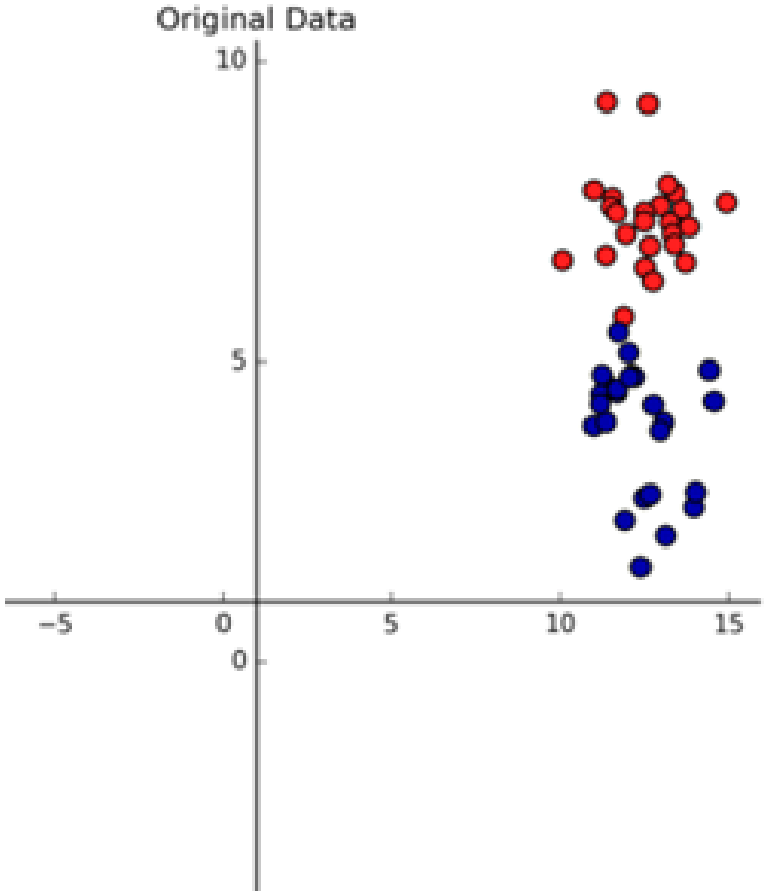
Min Max Scaler

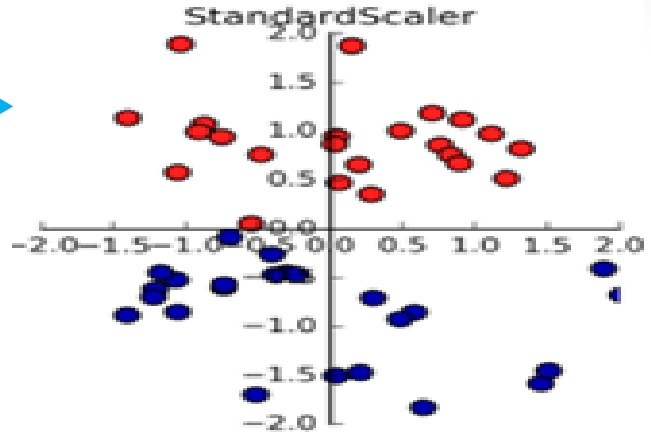
$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$



Z(0,1)

Standardization





$$z = \frac{x - \mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

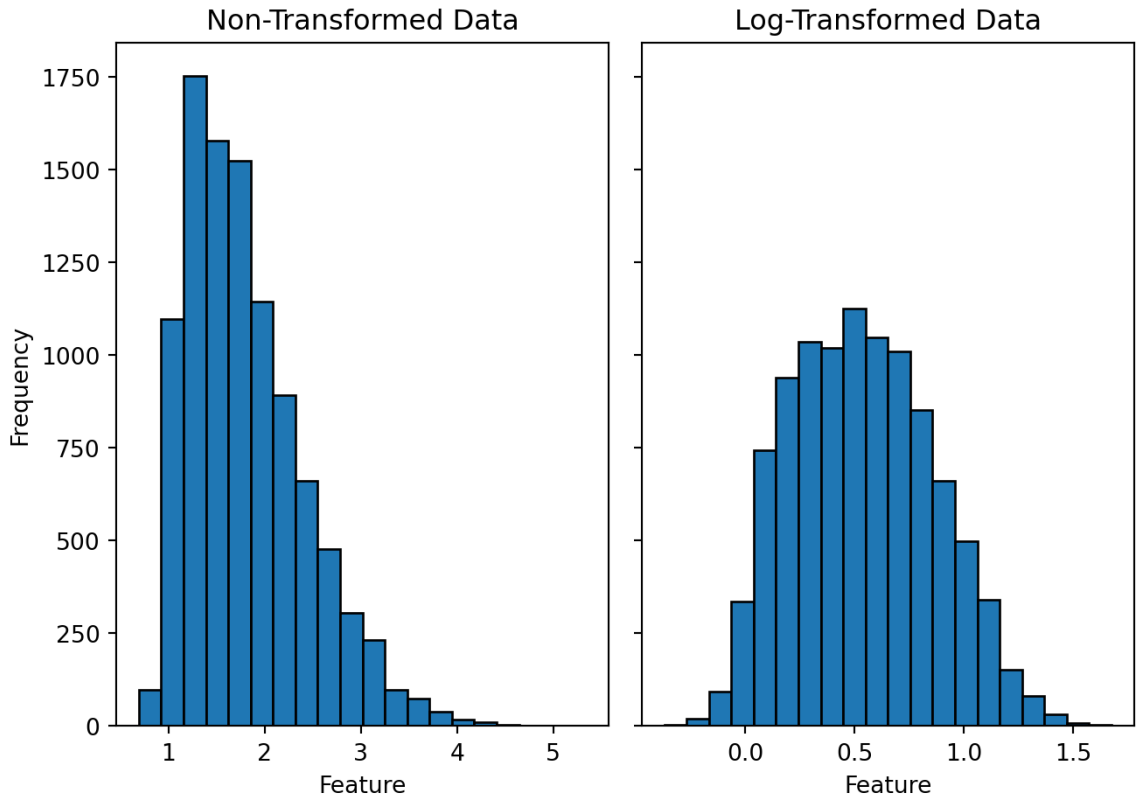
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Log normalization

Natural log using the constant e (≈ 2.718)

$e^{3.4} = 30$

useful when we have features with high variance.



Number	Log
30	3.4
300	5.7
3000	8

انواع الگوریتم های یادگیری

یادگیری با ناظر	**Supervised Learning**
یادگیری بی ناظر	**Unsupervised Learning**
یادگیری نیمه نظارتی	**Semi_supervisedLearning**
یادگیری تقویتی	**Reinforcement Learning**

CLASSICAL MACHINE LEARNING

SUPERVISED

Predict a category

CLASSIFICATION

Predict a number

REGRESSION

nDivide the ties by length»

UNSUPERVISED

Data is not labeled in any way

Divide by similarity

CLUSTERING

Find hidden dependencies

DIMENSION REDUCTION

Identify sequences

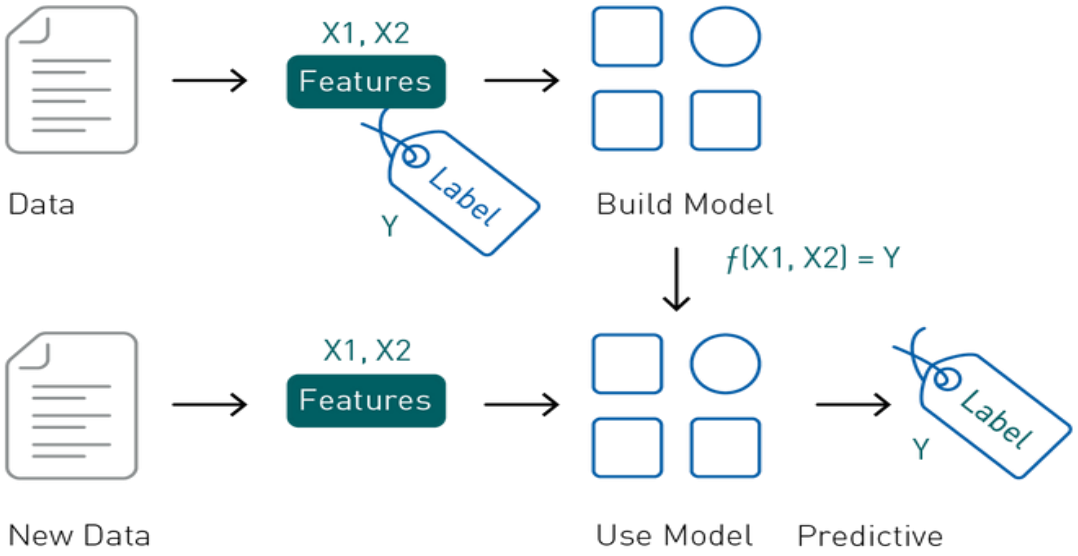
ASSOCIATION

Supervised Learning یادگیری با ناظر

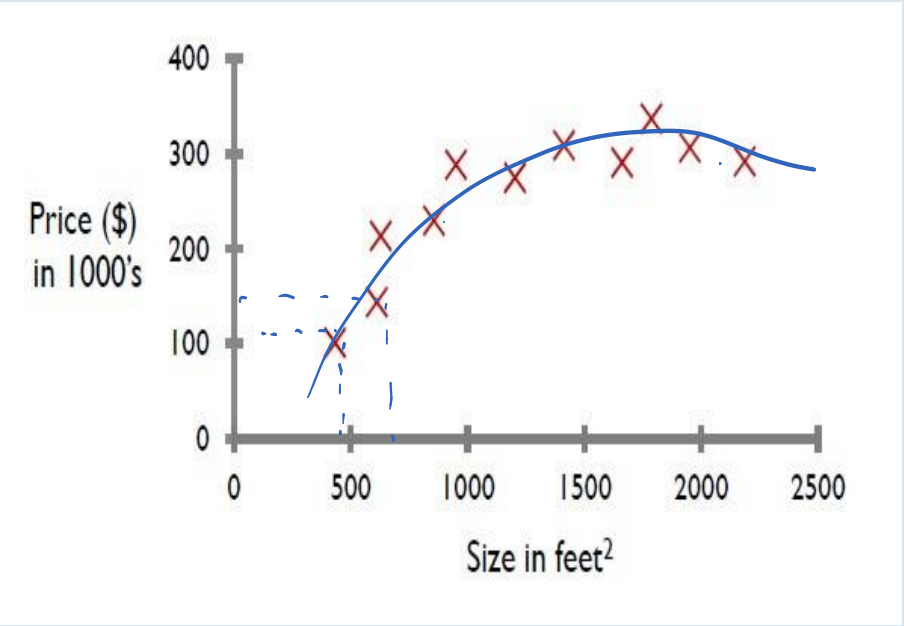
features

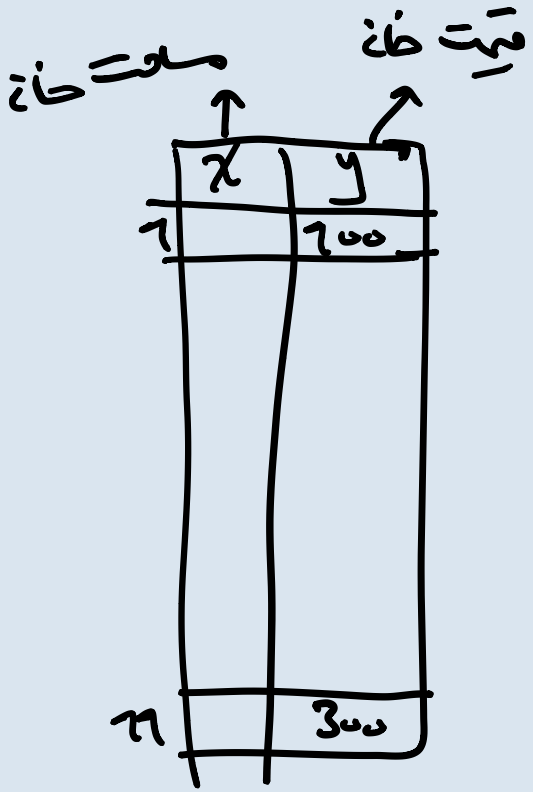
labels

	x_1	x_2	...	x_d	y (Target)
Sample 1					
Sample 2					
...					
Sample n-1					
Sample n					



	AGE	SEX	BMI	BP	⋯	Serum Measurements				⋯	Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

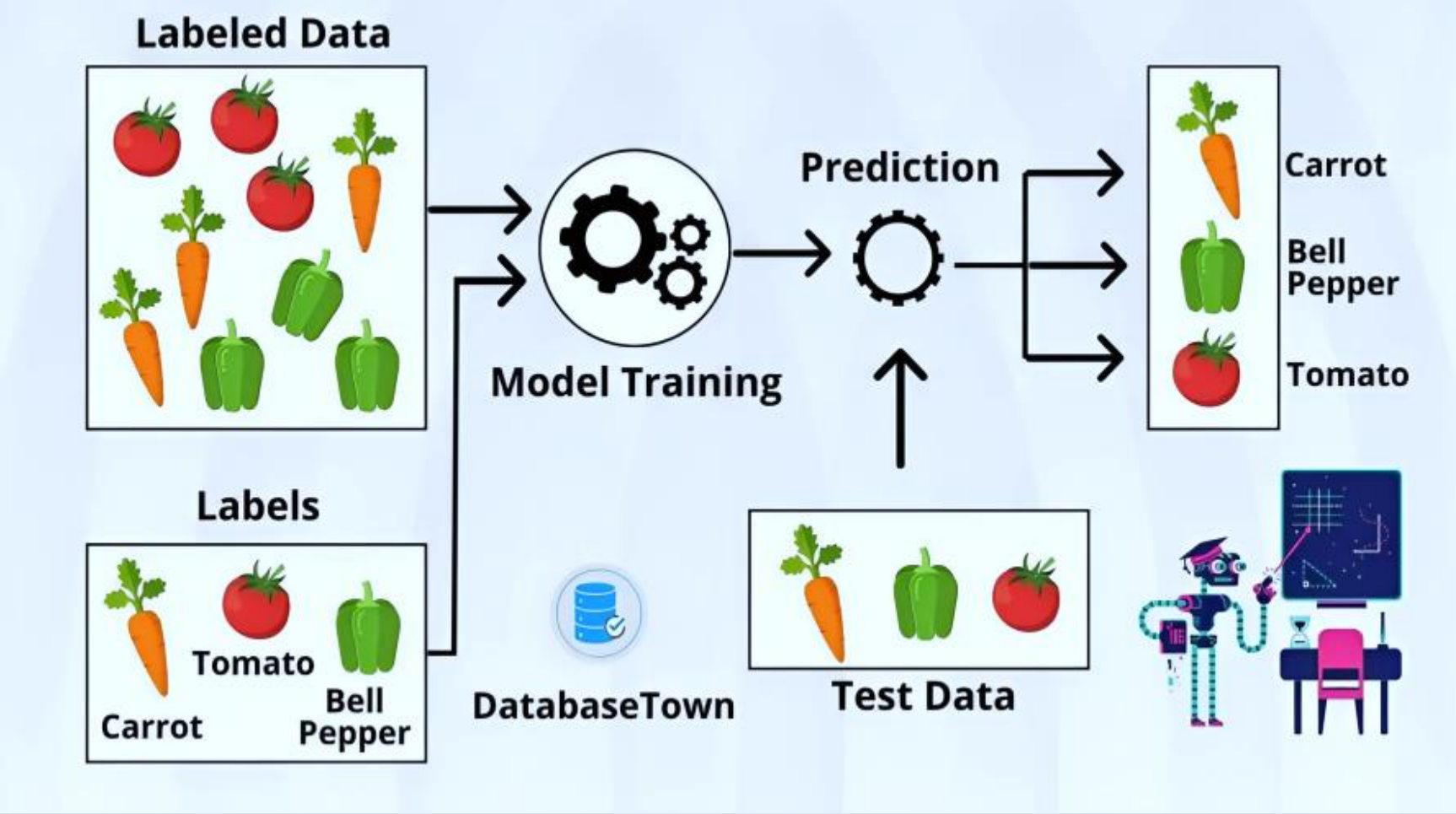


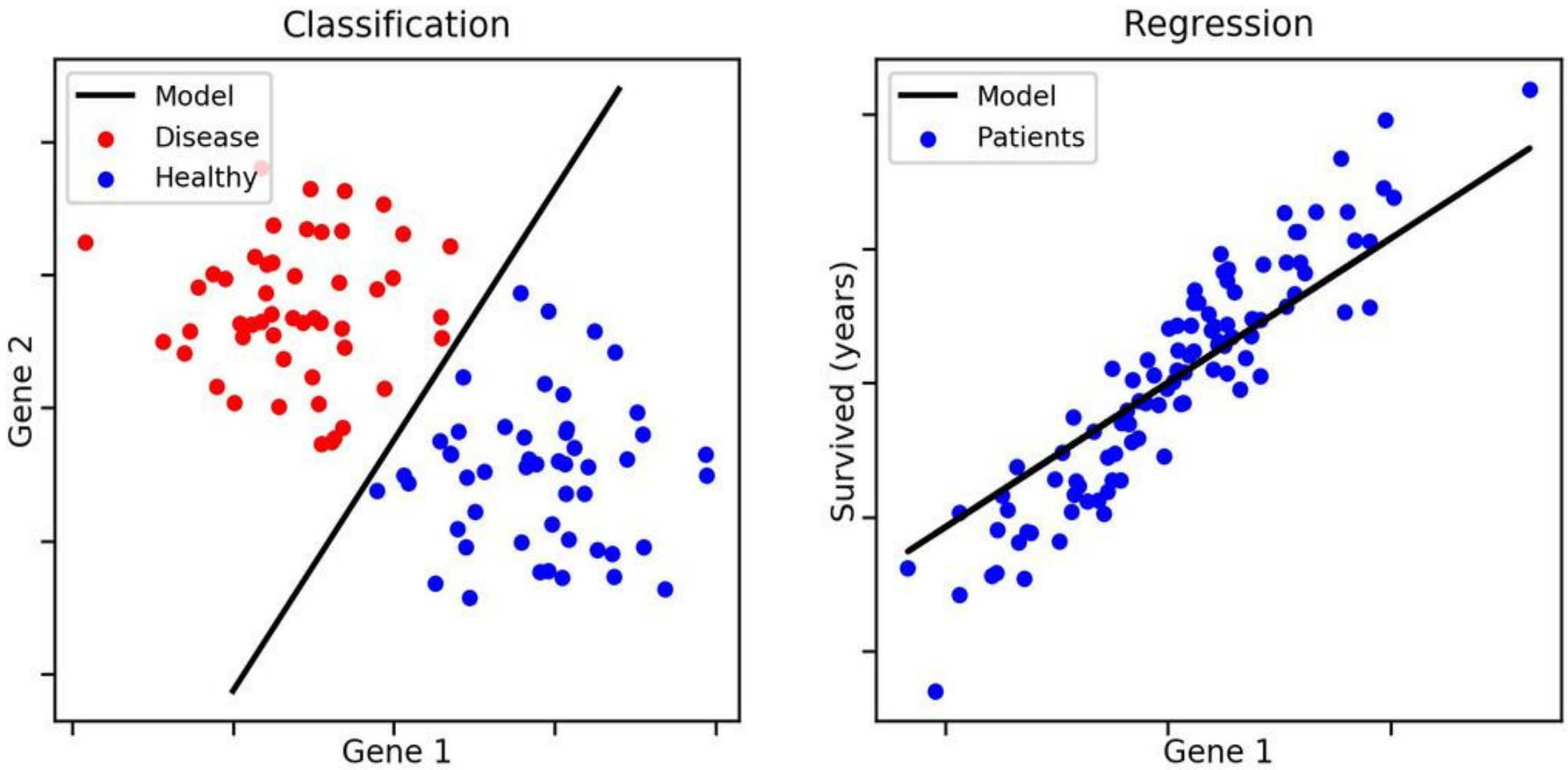


Boston Housing

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	PRICE
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

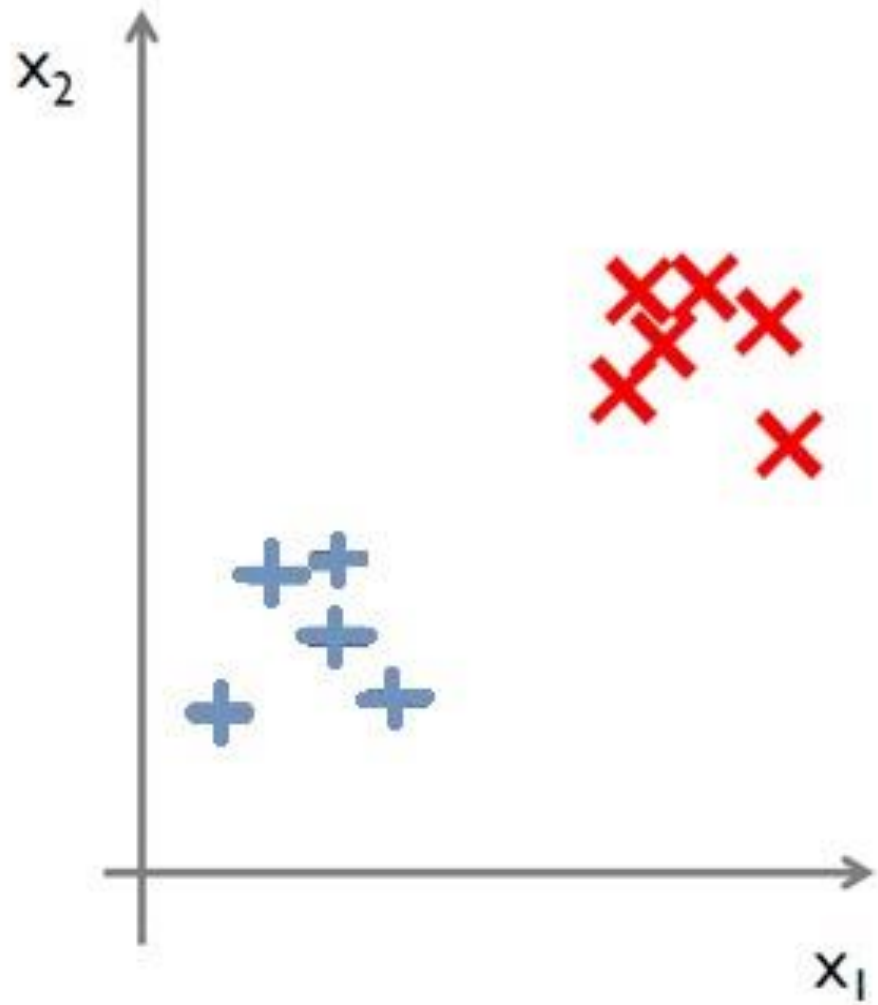
طبقه بندی classification




Classification
Model
Disease
Healthy
Gene 2
Gene 1
Regression
Model
Patients
Survived (years)
Gene 1

Training data

x_1	x_2	y
0.9	2.3	+1
3.5	2.6	+1
2.6	3.3	+1
2.7	4.1	+1
1.8	3.9	+1
6.5	6.8	-1
7.2	7.5	-1
7.9	8.3	-1
6.9	8.3	-1
8.8	7.9	-1
9.1	6.2	-1

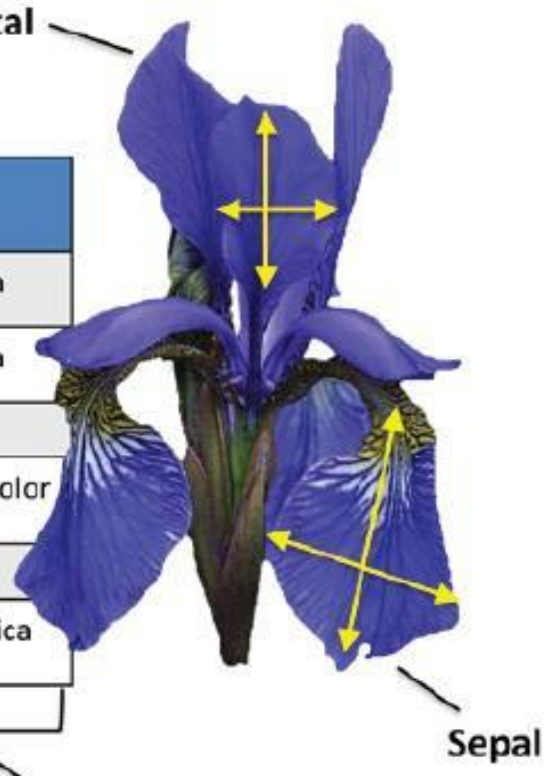


Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes , dimensions)

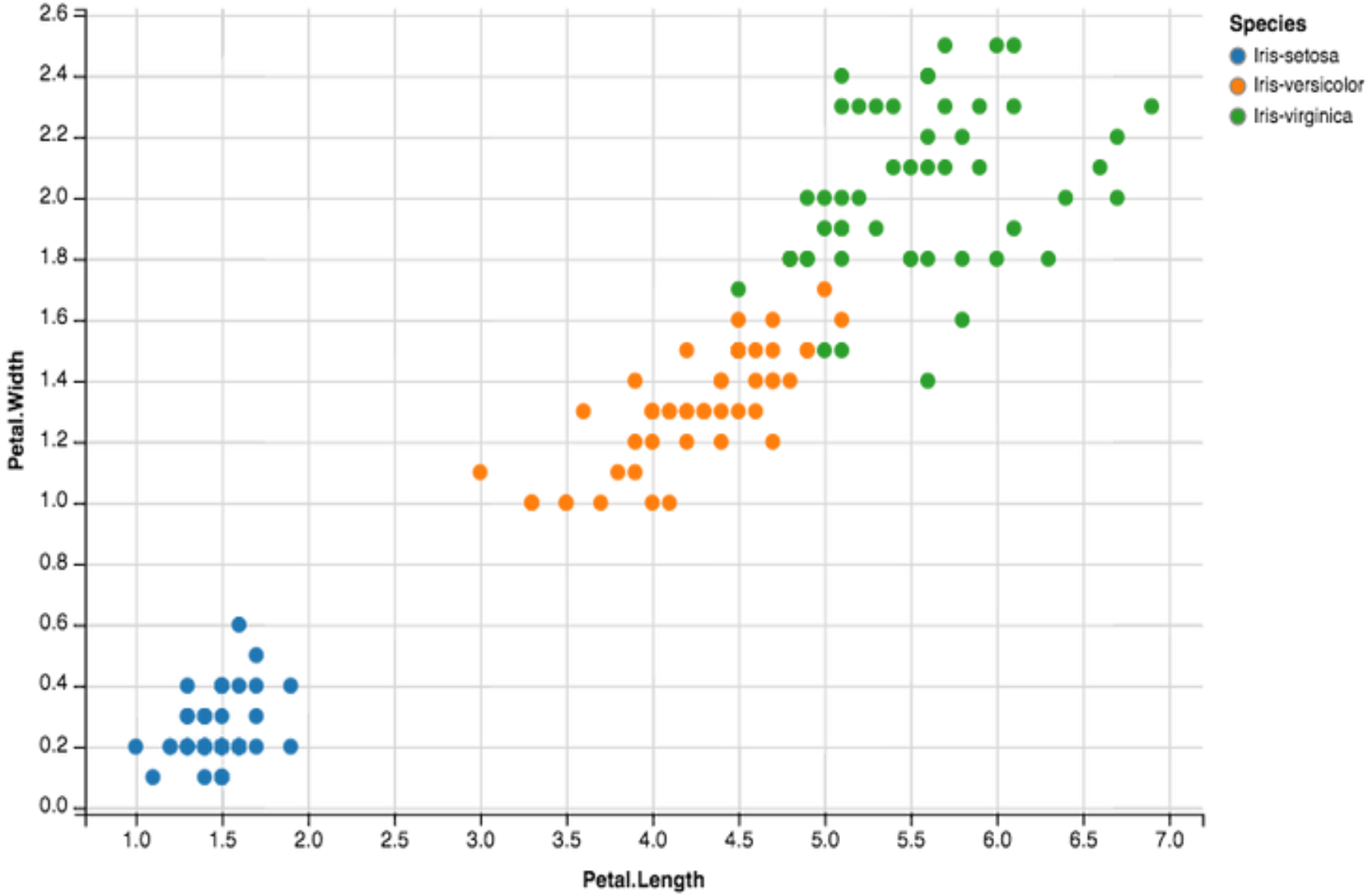
Class labels
(targets)



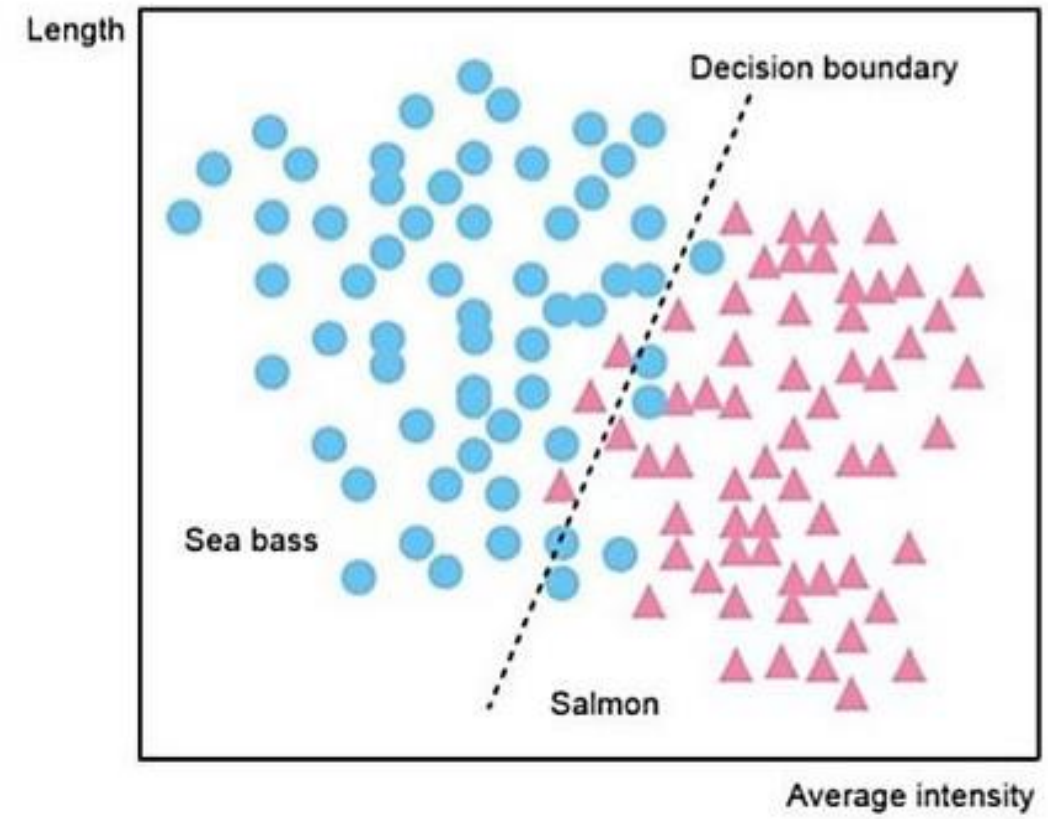
Virginica

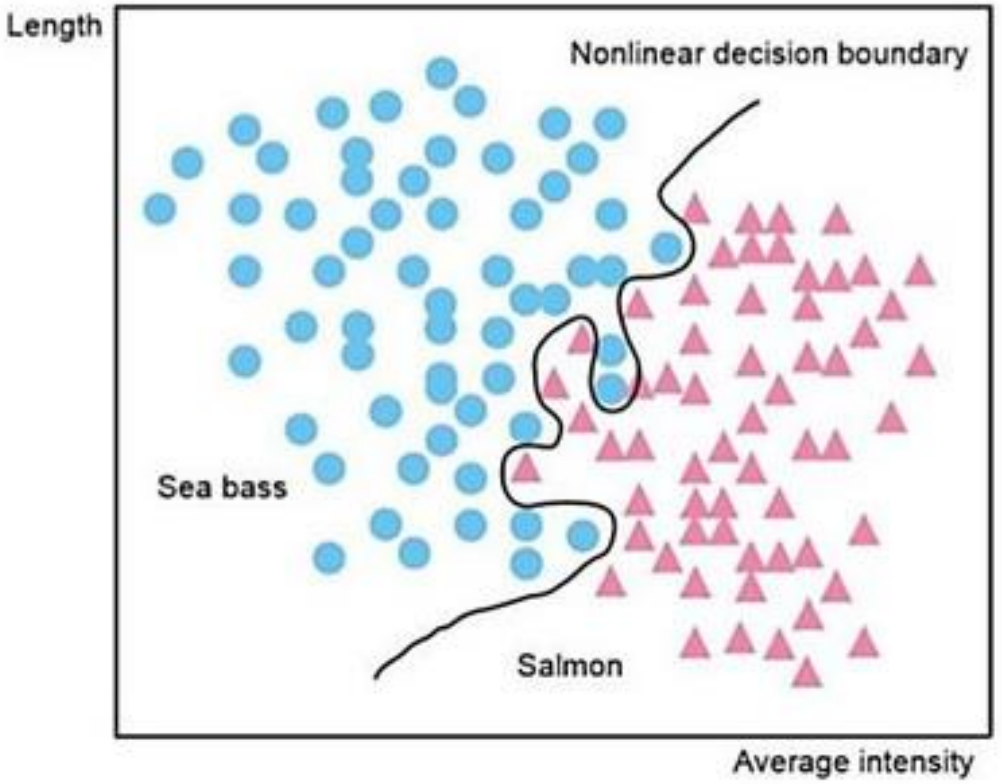
Setosa

Versicolor


Species
Iris-setosa
Iris-versicolor
Iris-virginica
Petal.Width
Petal.Length
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0

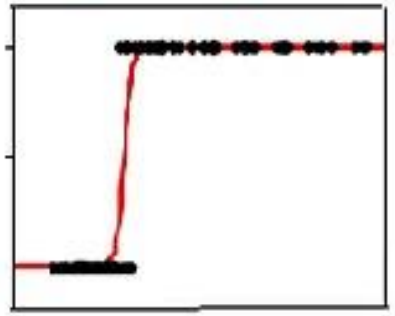
مرز تصمیم گیری



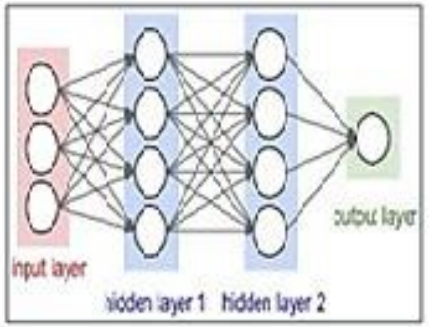


معروف ترین دسته بند ها

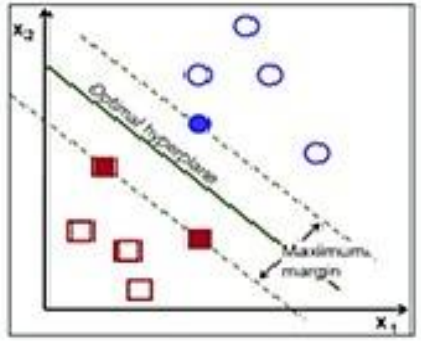
Logistic regression



Neural Network



SVM

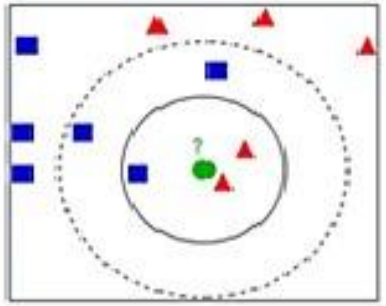


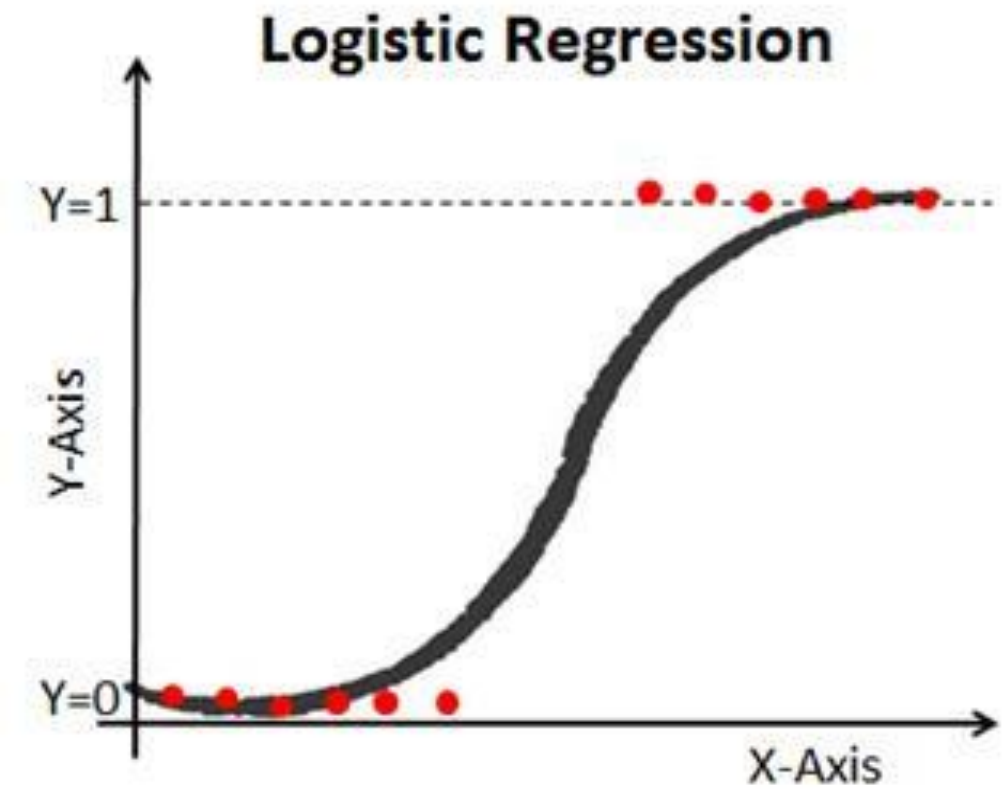
Decition Tree

Naive Bayesian

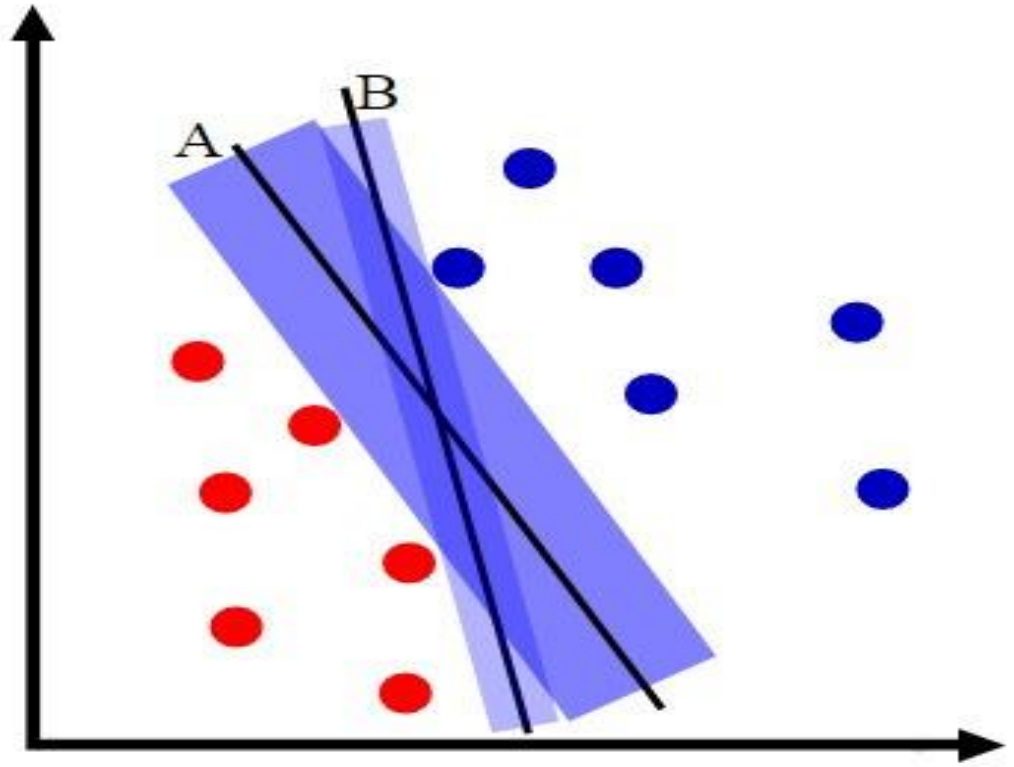
$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

KNN

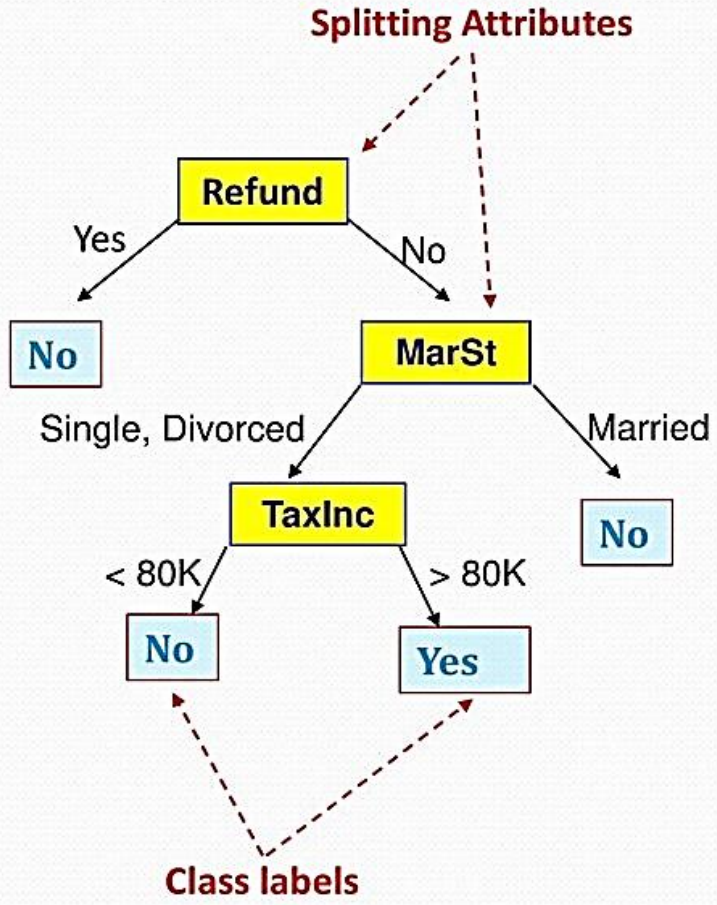



Logistic Regression
Y=1
Y=0
Y-Axis
X-Axis

ماشین بردار پشتیبان



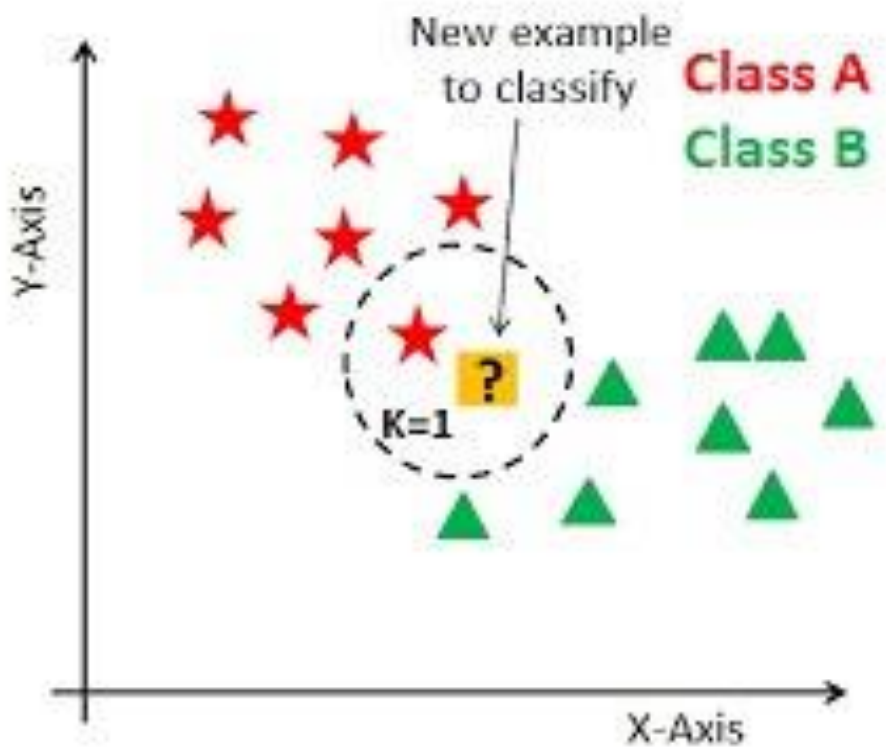
درخت تصمیم

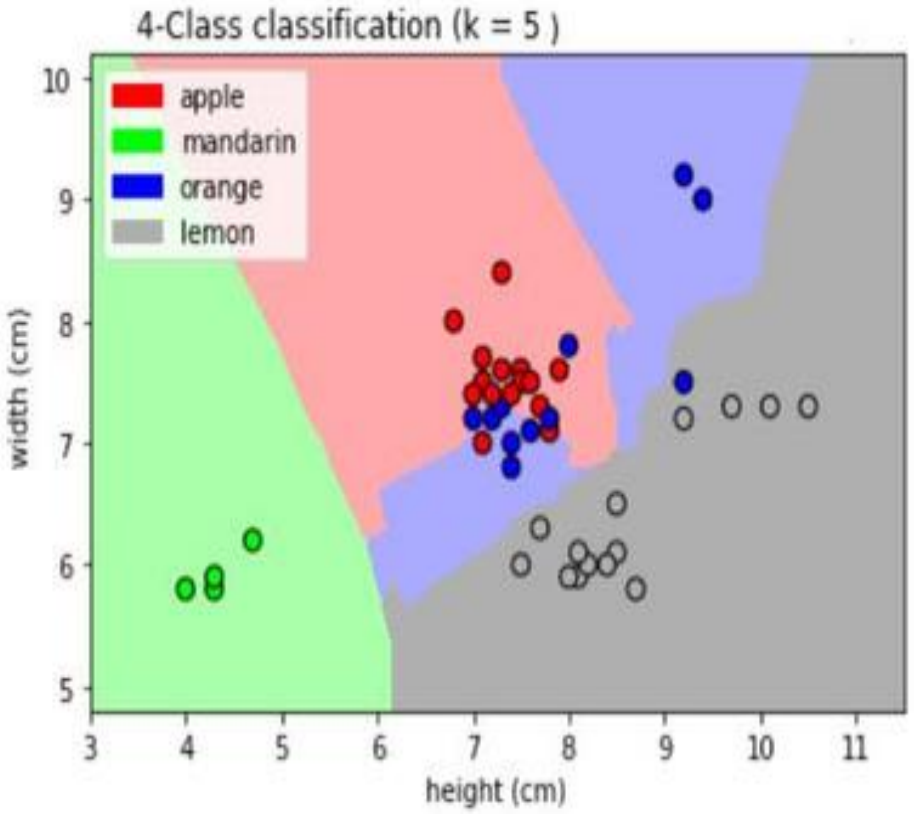


	categorical	categorical	Integer	class
T id	Refund	Marital status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

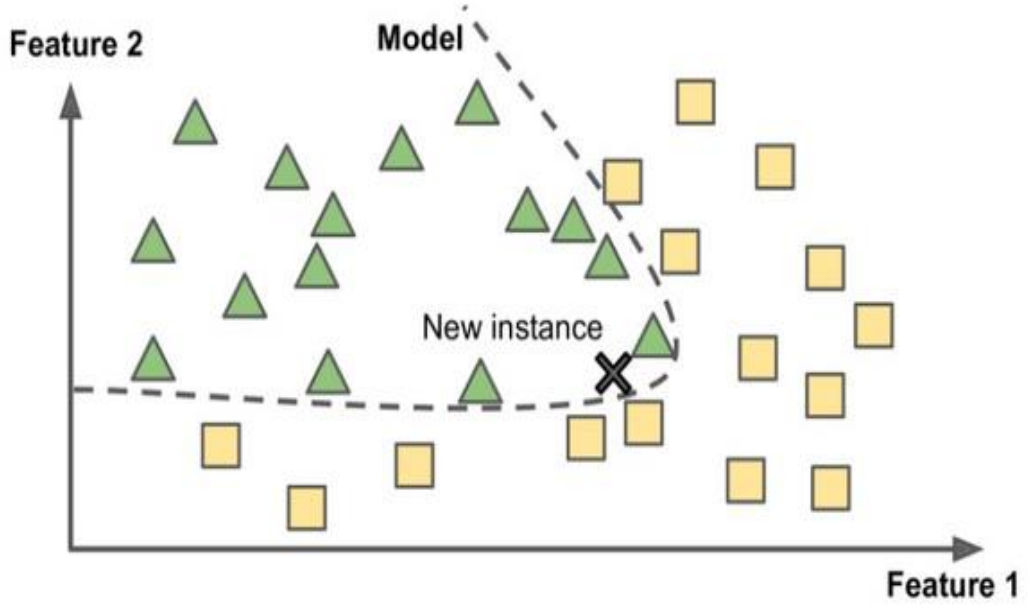
Kنزدیک ترین همسایه



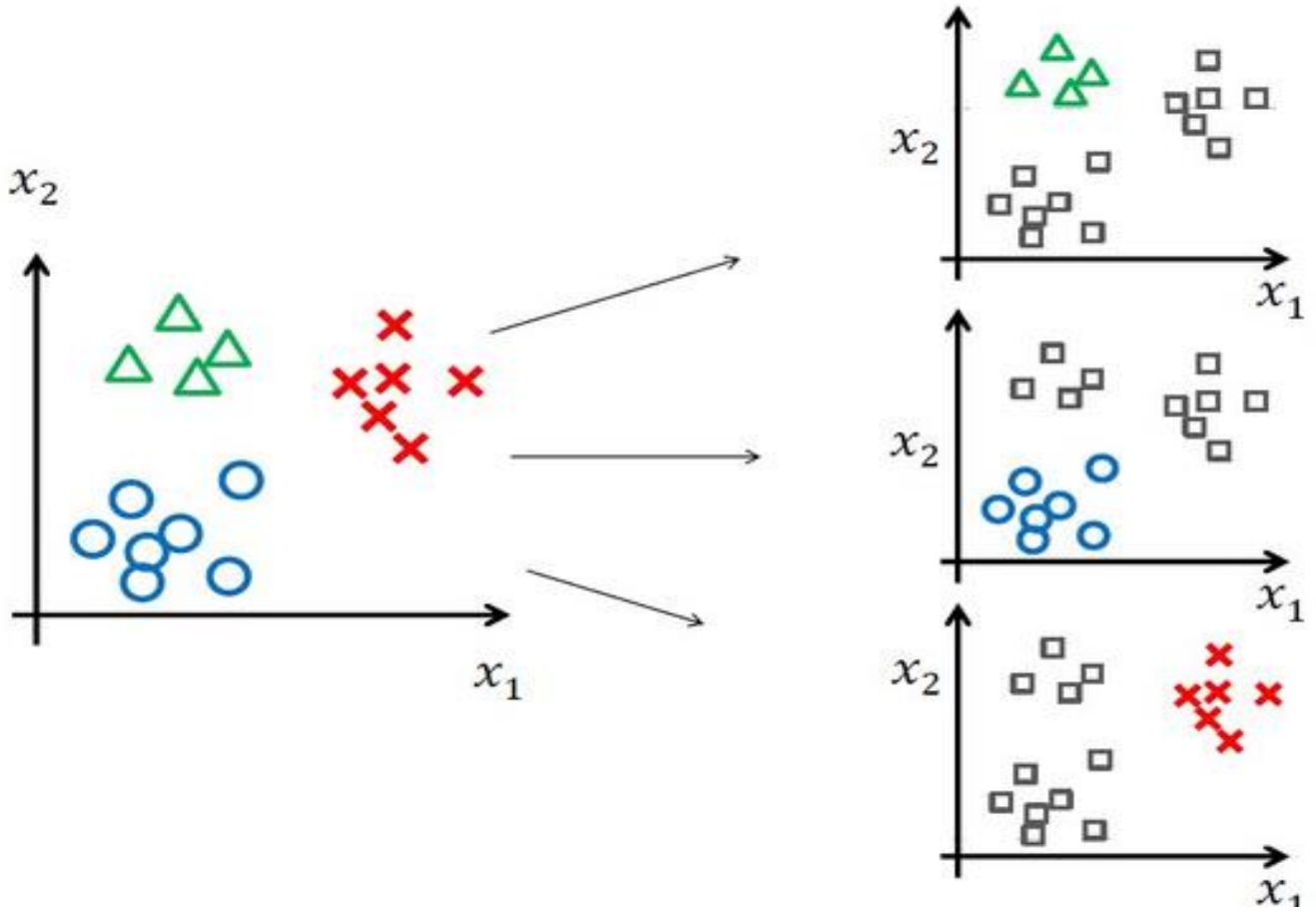


Instance-based learning

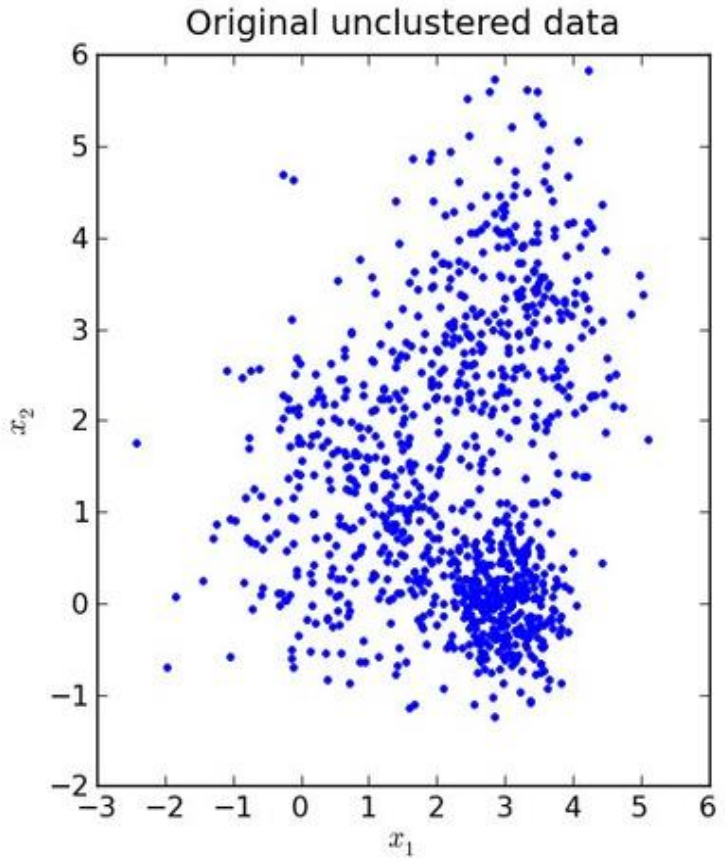
Model-based learning

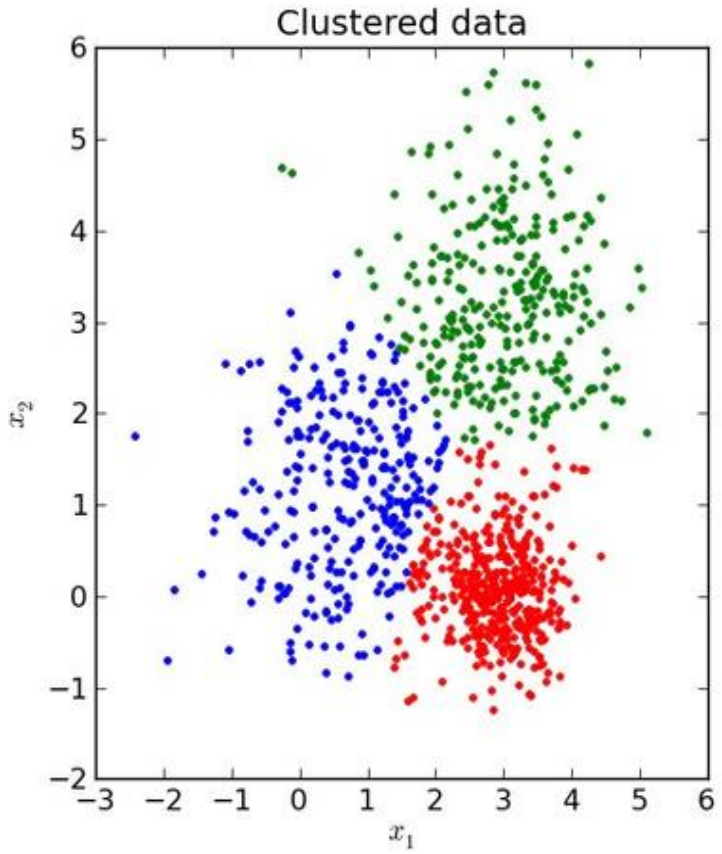


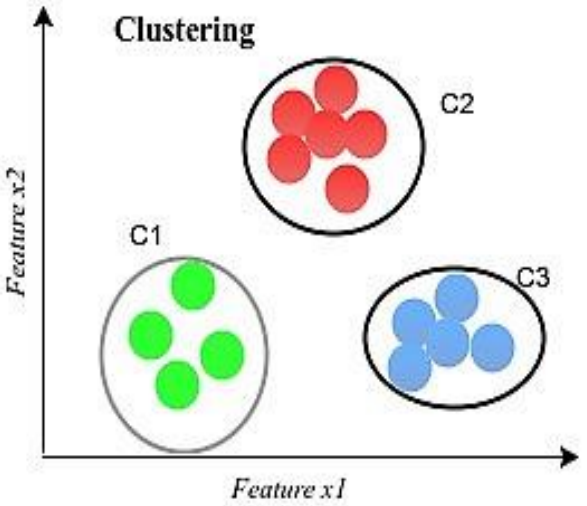
طبقه بندی چندکلاسه خطی



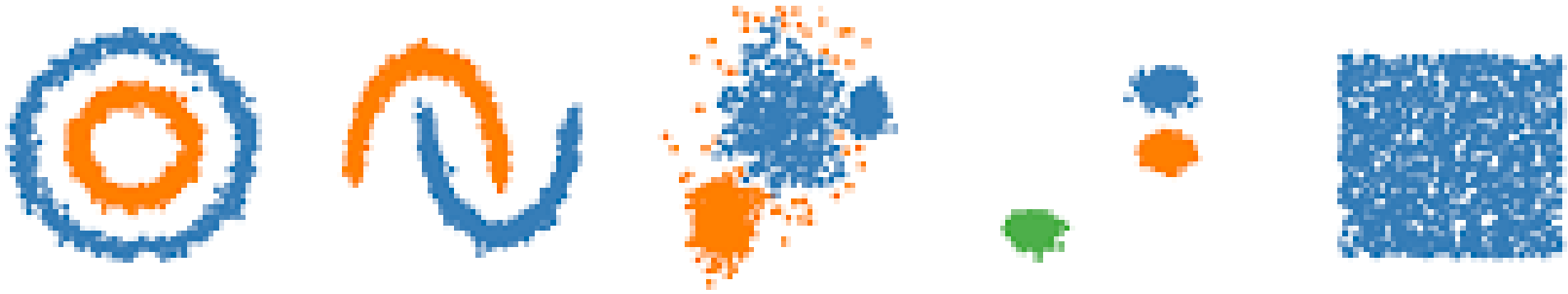
خوشه بندی


Original unclustered data
x_2
x_1

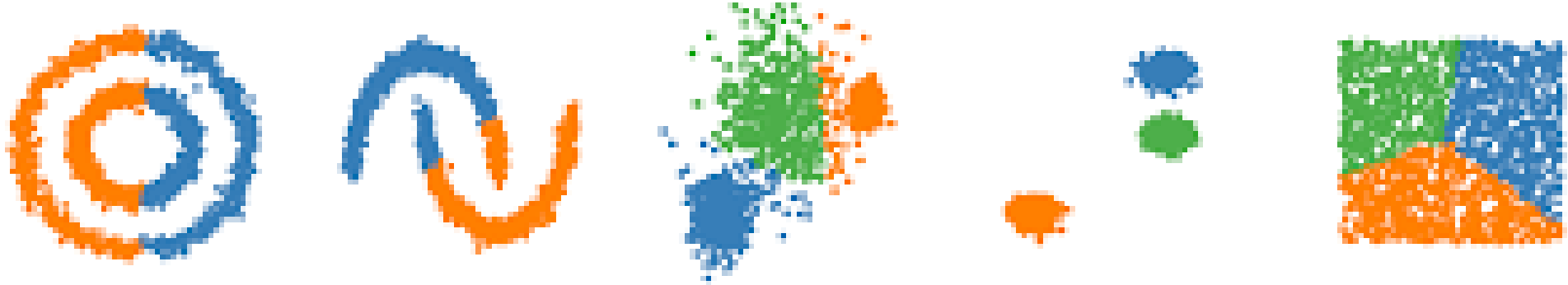

Clustered data
x_2
x_1


Clustering
C1
C2
C3
Feature x2
Feature x1

DBSCAN



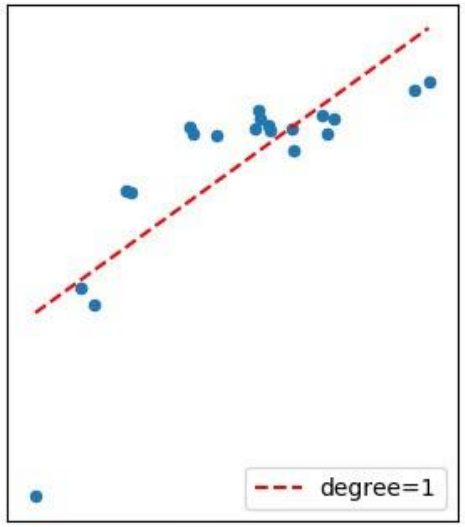
K-means



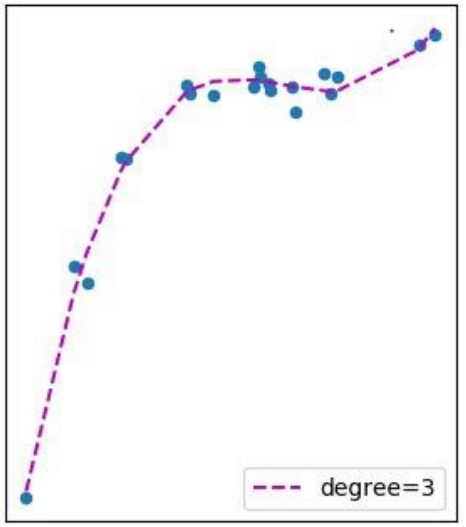
کاربرد های یادگیری ماشین کلاسیک

تقسیم کتاب ها به موضوعات مختلف
دسته‌بندی ژن‌هایی با عملکردهای مشابه
میزان فروش محصولات
تشخیص خبر جعلی
شناسایی تقلب(داده پرت)
پیش بینی قیمت سهام
پیشبینی بارندگی
پیش بینی میزان فروش
پیش بینی اهدای خون
تشخیص چهره

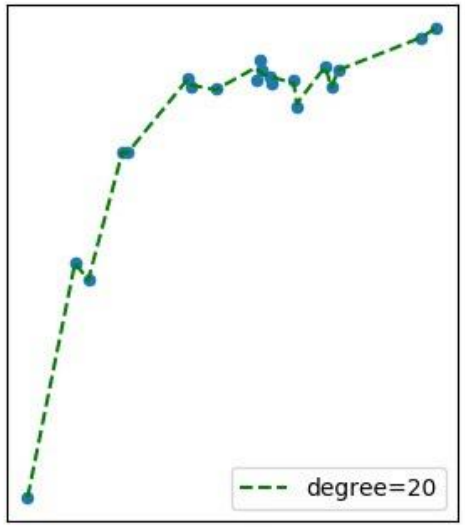
Overfit & under fit



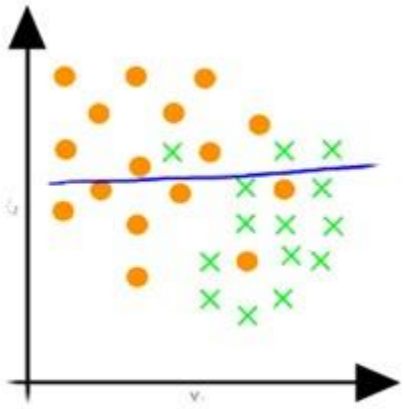
Underfit
High Bias
Low Variance

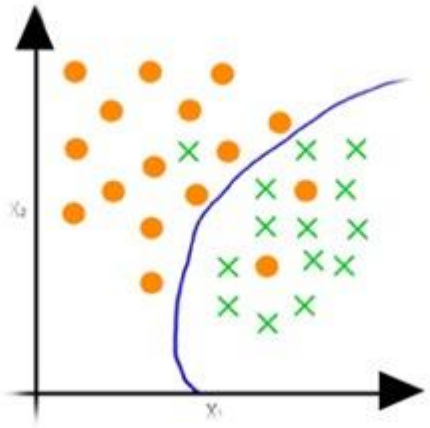


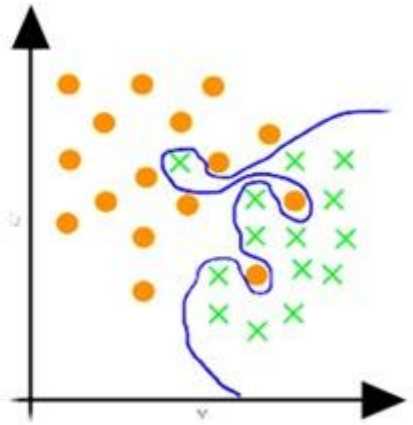
Correct Fit
Low Bias
Low Variance



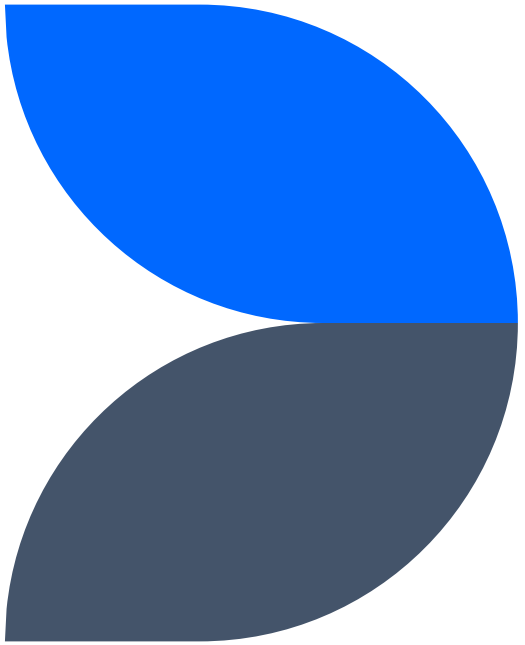
Overfit
Low Bias
High Variance

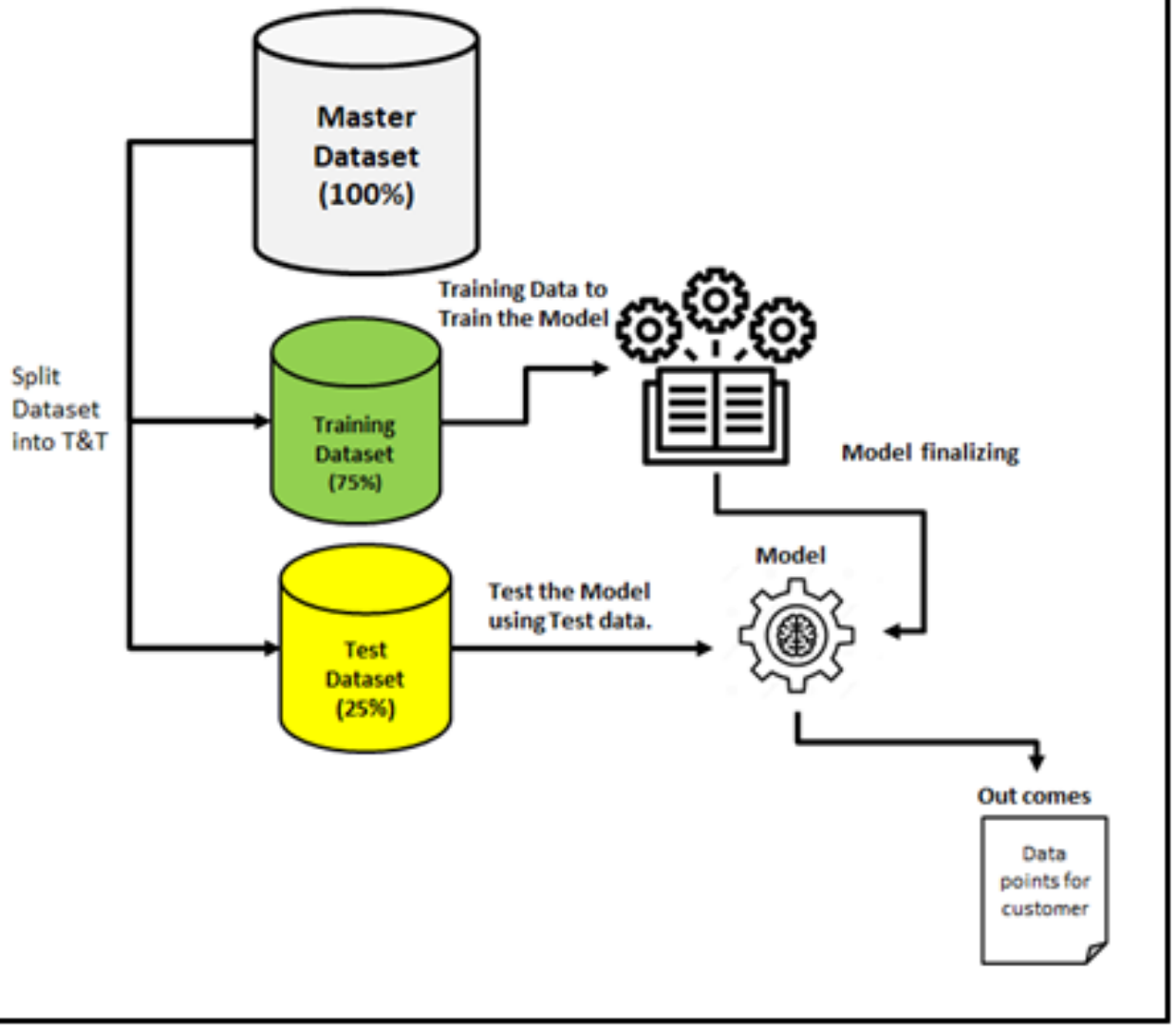


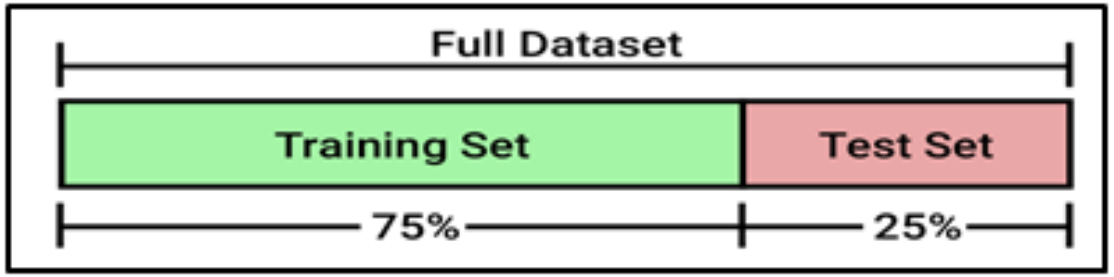


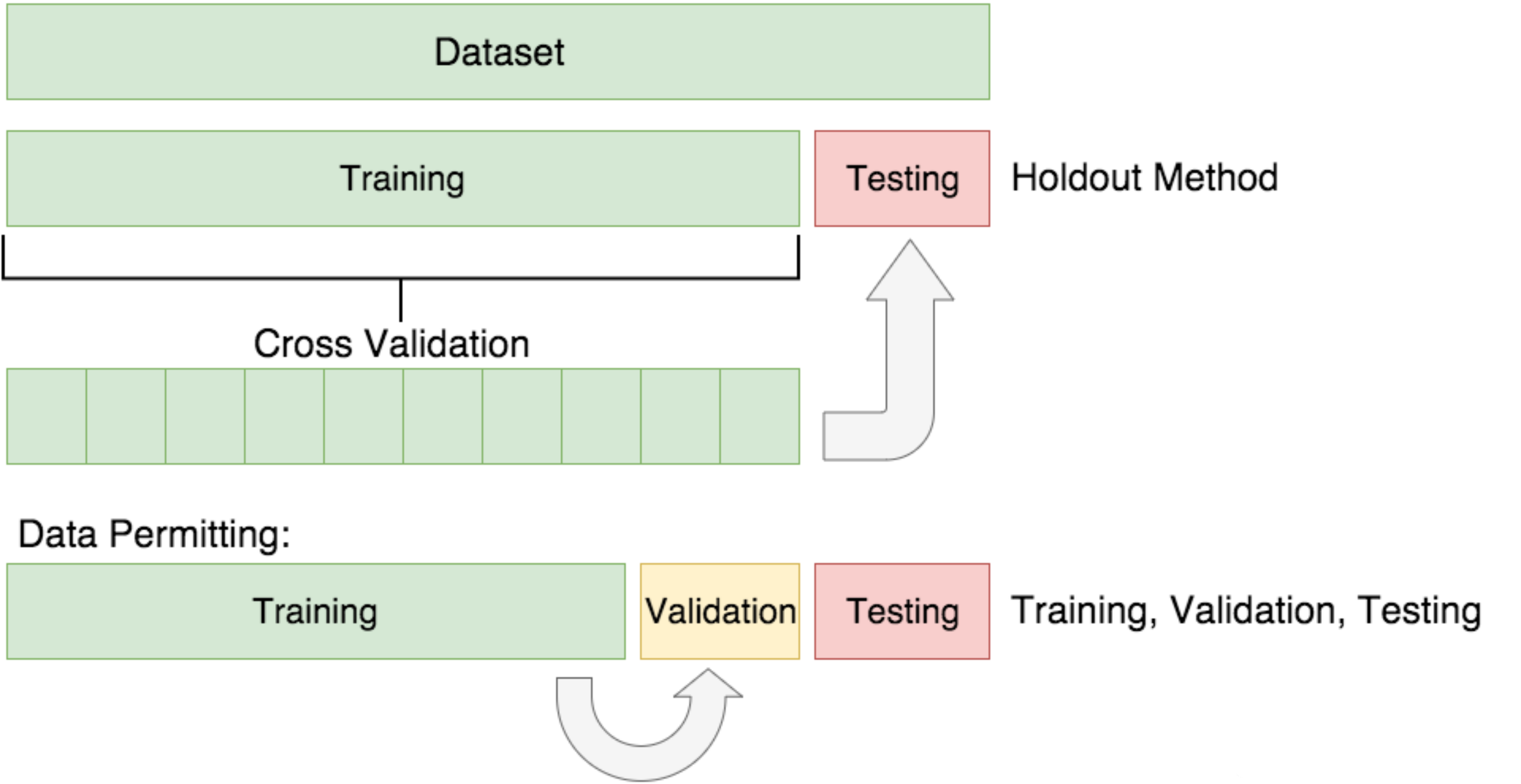


Model Evaluation




Master Dataset (100%)
Training Data to Train the Model
Split Dataset into T&T
Training Dataset (75%)
Model finalizing
Model
Test the Model using Test data.
Test Dataset (25%)
Out comes
Data points for customer


Full Dataset
Training Set
Test Set
75%
25%


Dataset
Training
Testing
Holdout Method
Cross Validation
Data Permitting:
Training
Validation
Testing
Training, Validation, Testing

Confusion matrix

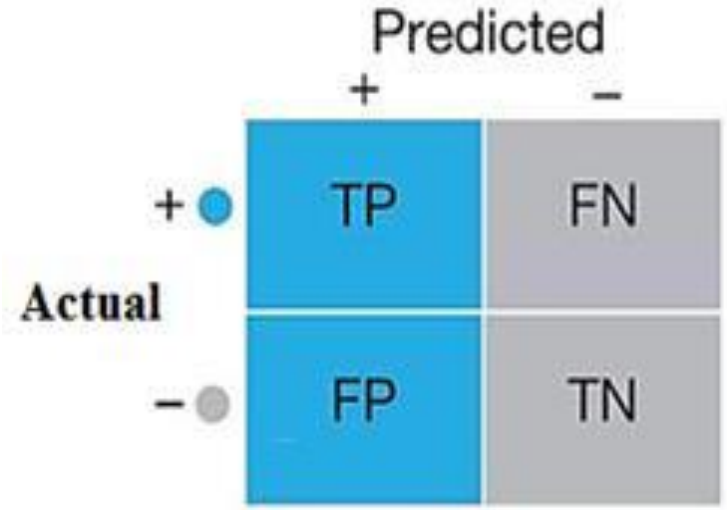
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN}$$



(True Positive) :the number of correct classifications of positive examples.

(True Negative) : the number of correct classifications of negative examples.

(False Positive):the number of incorrect classifications of negative examples.

(False Negative) : : the number of incorrect classifications of positive examples.

Model Evaluation

Regression

- (MAE) mean absolute error
- (MSE) mean squared error • (RMSE) root mean squared error; interpretable in the same units as the response vector or y units
- (RAE) Relative absolute error, also known as residual sum of square
- (RSE) Relative squared error
- (R2) ; Popular metric for the accuracy of your model. represents how close the data values are to the fitted regression line. The higher the better

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

$$\text{RAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

$$\text{RSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$