ICT Academy of Kerala

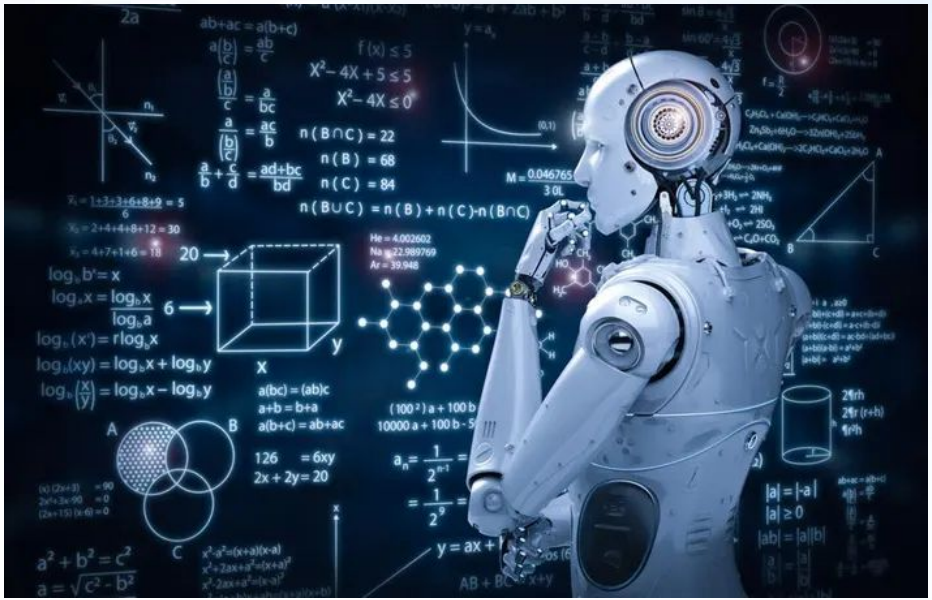
Building the Nation's Future

Certified Specialist in Machine Learning and Artificial Intelligence

Supervised Learning & Regression

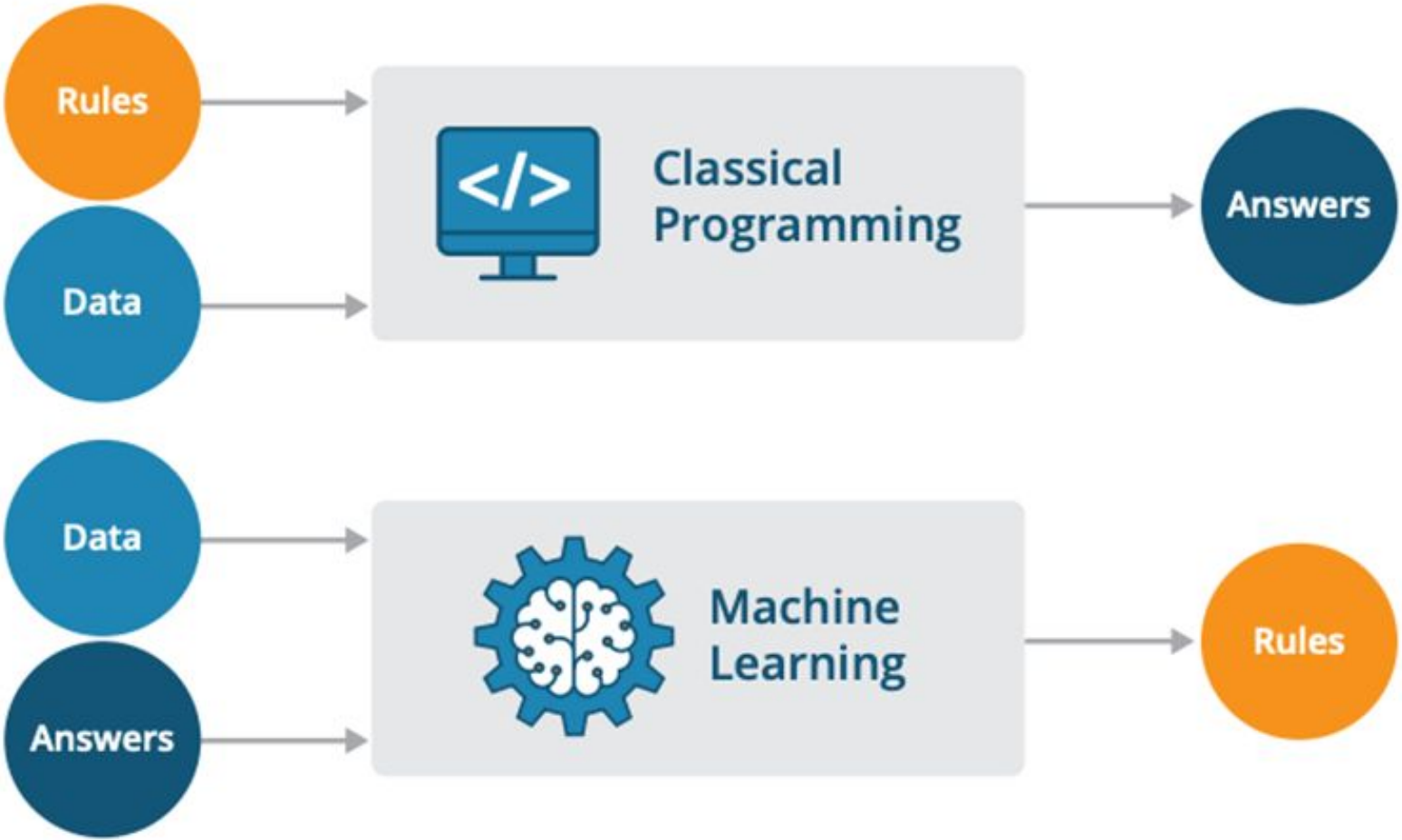
A **GOVT. OF INDIA** SUPPORTED, **GOVT. OF KERALA** PARTNERED SOCIAL ENTERPRISE.

Machine Learning



Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed.

Arthur Samuel



A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

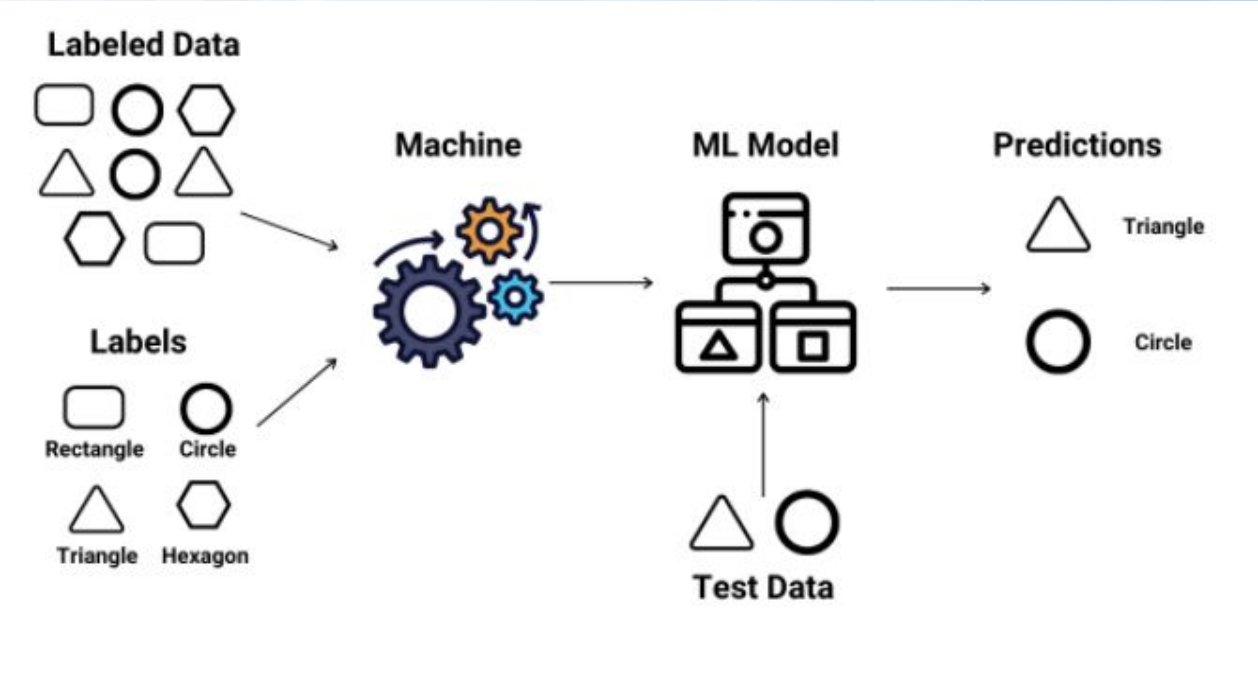
Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning

- Machine learns under guidance
- Class labels will be given in the data
- We teach the machine with labelled data
  - Regression
  - Classification

Supervised Learning



Regression

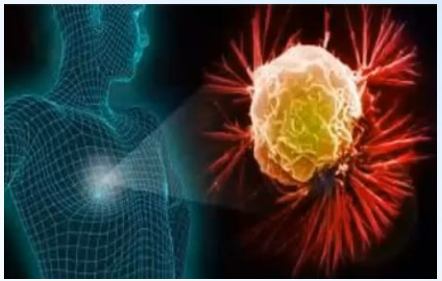
- Can you predict what is the amount of rainfall that will take place in the year 2025? What will be the data that will be used for your prediction?
- Can you predict by seeing data for sales of Apple of the past 9 years to predict whether the sales will rise or lower by what amount the next year?

Classification

- Can you predict an email is spam or not?

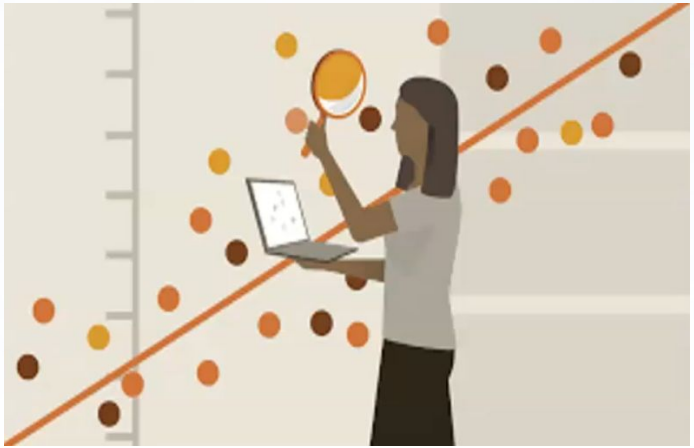
  But, how?

- If a given tissue is cancerous or not? And, how?

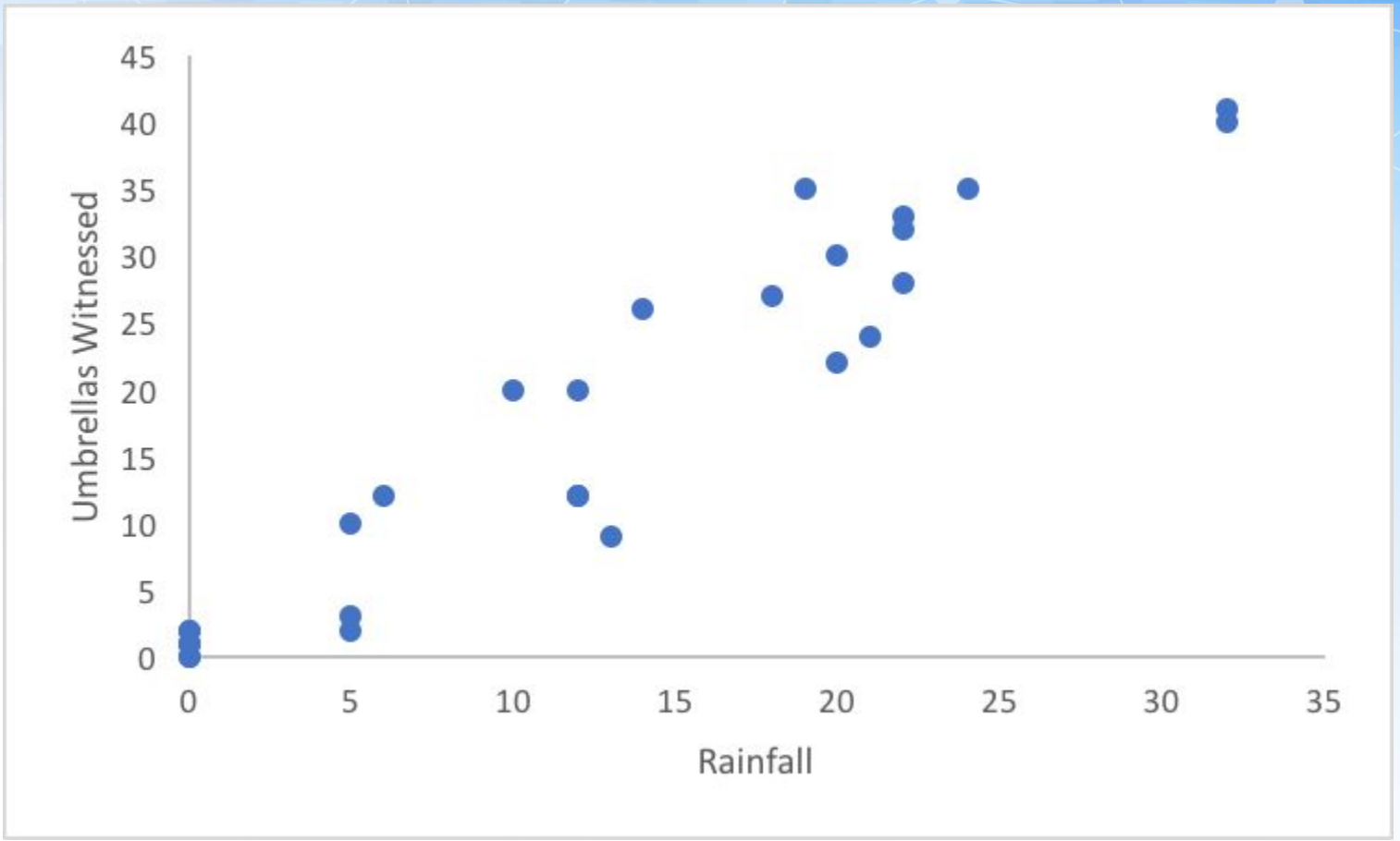


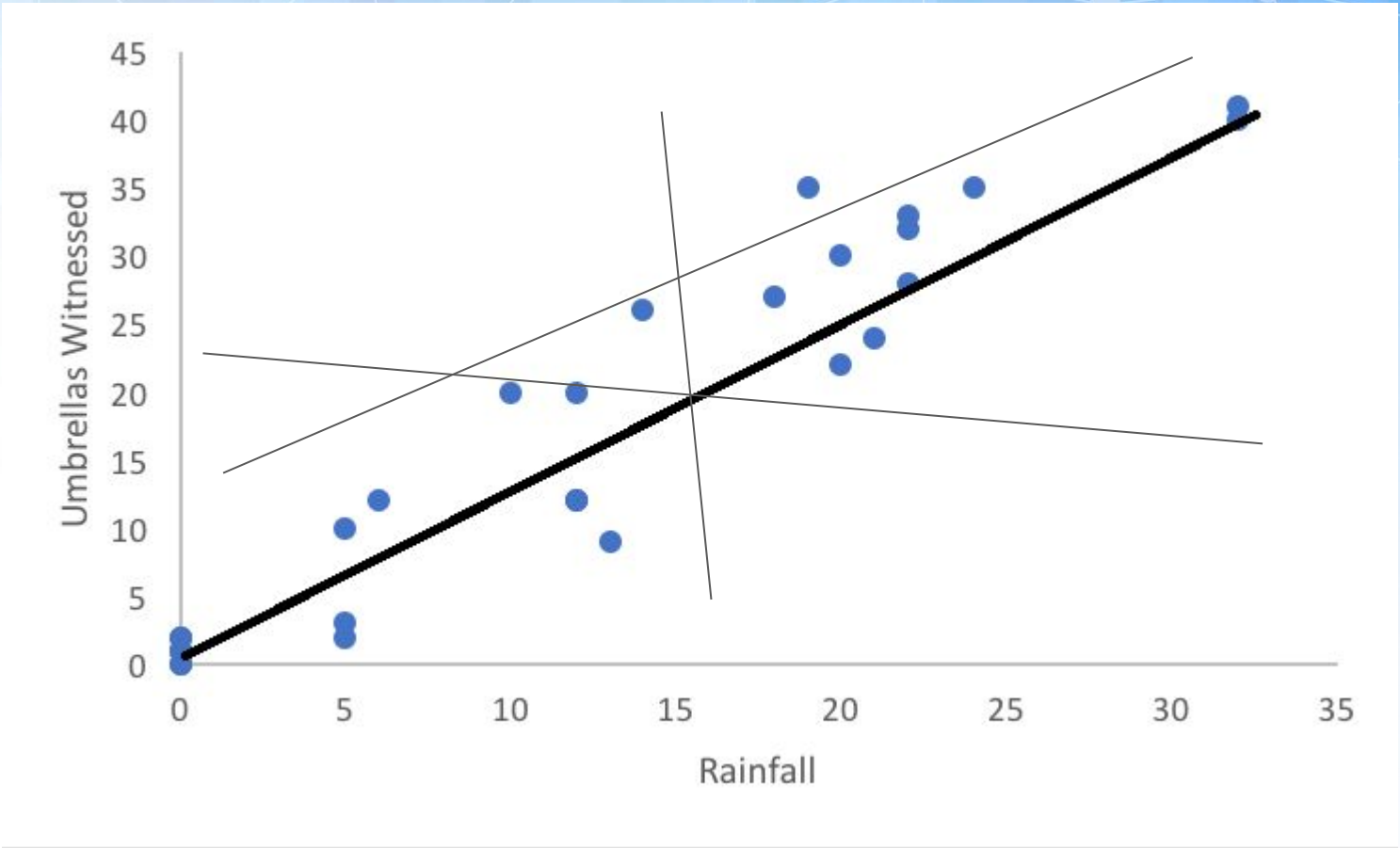
- Can you predict if a customer will default on a loan or not? If so, by how much?

What is Regression?



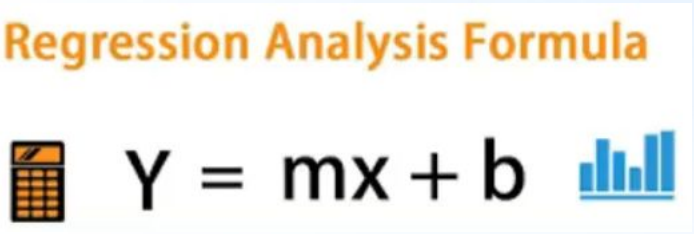
- Regression Analysis is a predictive modelling technique.
- It estimates the relationship between an independent variable(predictor) and a dependent variable(target)

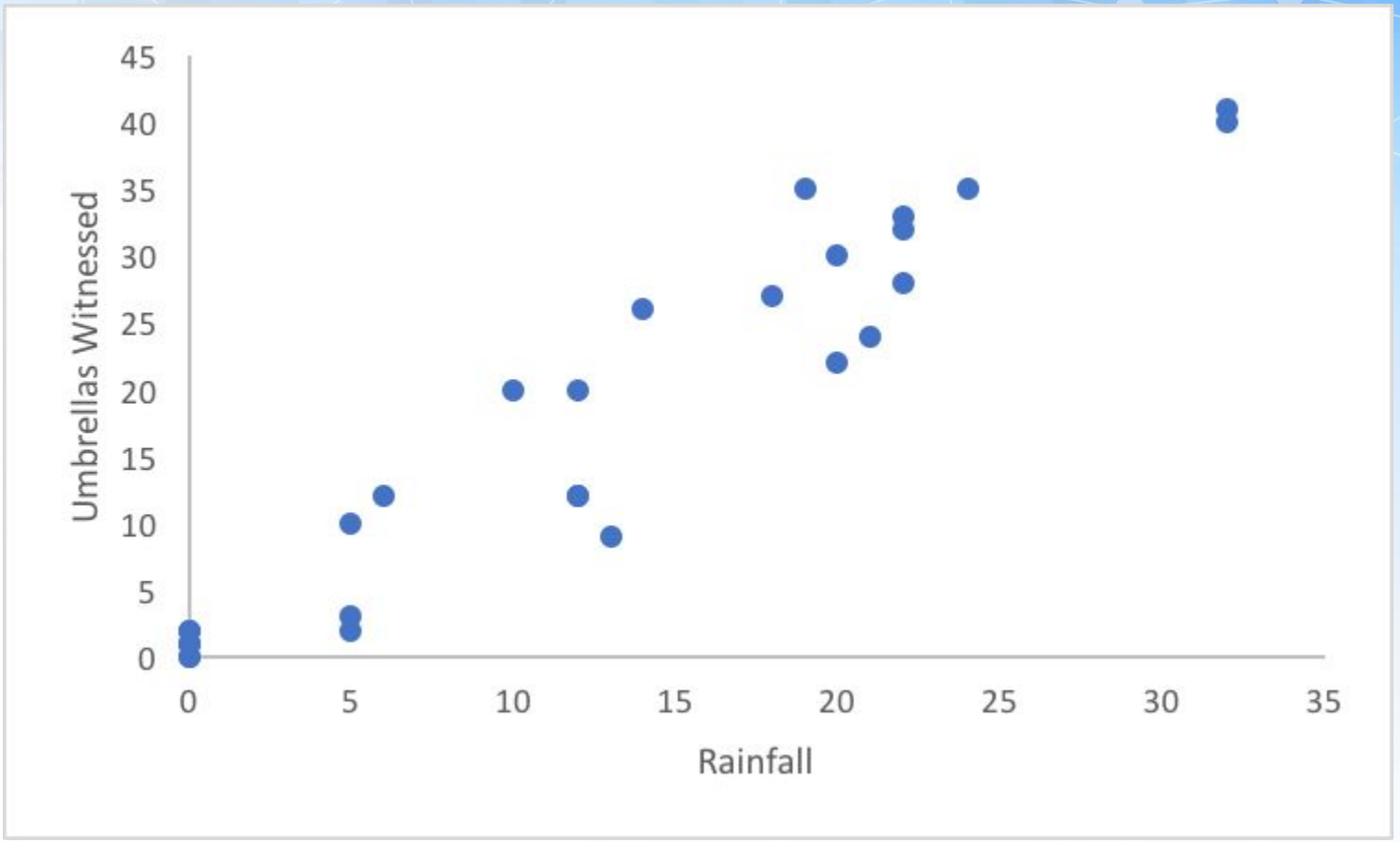


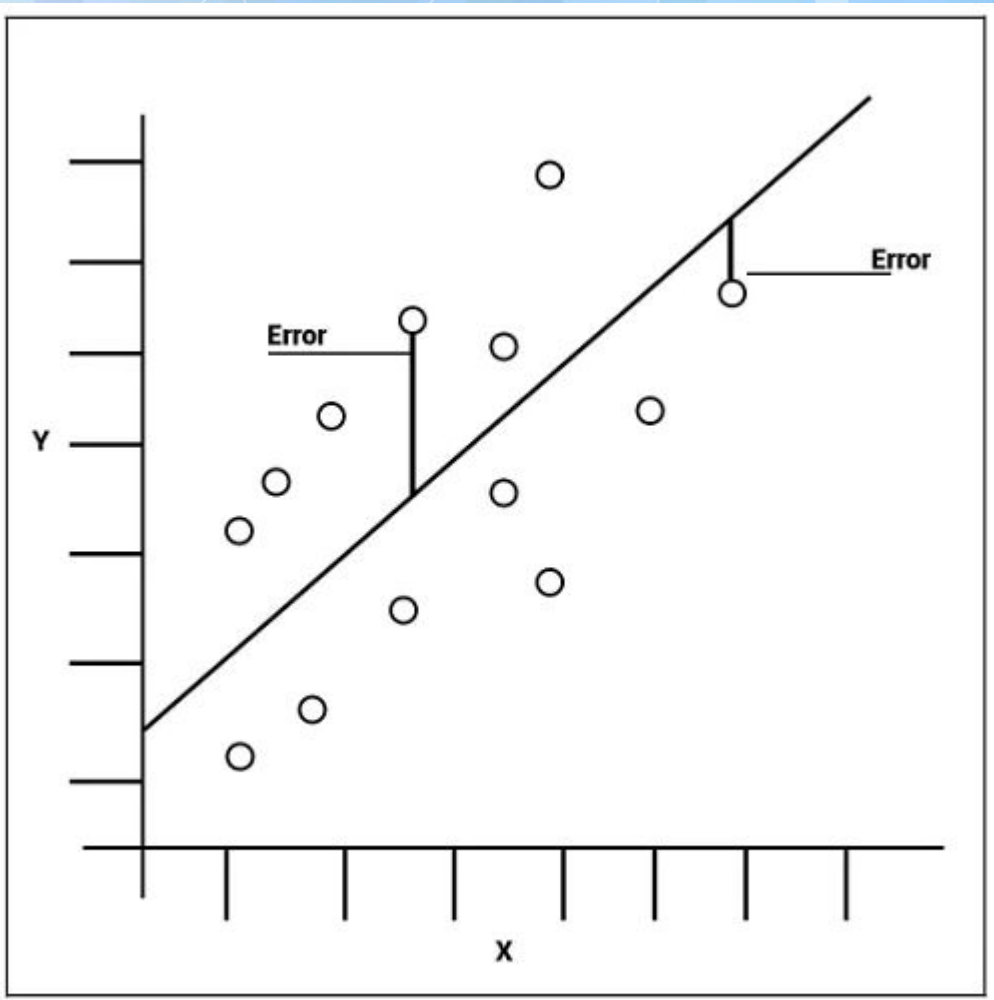


Linear Regression

- Linear regression is used to predict the relationship between 2 variables or factors







Cost Function

The cost function (loss function) is the mathematical expression that measures how well the model fits the training data. It quantifies the error between the predicted values and the actual target values.

Mean Squared Error

- Estimator which measures the average of error squares
- Average squared distance between the estimated value and the true value
- It is always non-negative
- Values close to zero are better

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Why Should We Square The Residuals?

1. It makes them positive

1. Shows larger deviations

Mean Absolute Error

- It is the difference between the measured value and “true” value.
- Absolute value of the error is taken
- It is always non-negative
- Values close to zero are better

$$(\Delta x) = x_i - x,$$

R squared value

- It is known as coefficient of determination
- It is a goodness-of-fit measure for regression models
- indicates the percentage of the variance in the dependent variable that the independent variables explain collectively
- Values close to one are better

Metrics – Points to Remember

- Never compare different metrics with each other
- Try to use more than 1 loss function
- Always calculate evaluation metrics (loss functions) for both testing and training data set.
- If you have outlier in the data and you want to ignore them, MAE is a better option
- If you want to account the outliers in your loss function, go for MSE/RMSE.

Assumptions of Linear Regression

- The relationship between X and the mean of Y is linear.
- The variance of residual is the same for any value of X.
- Observations are independent of each other.
- For any fixed value of X, Y is normally distributed.

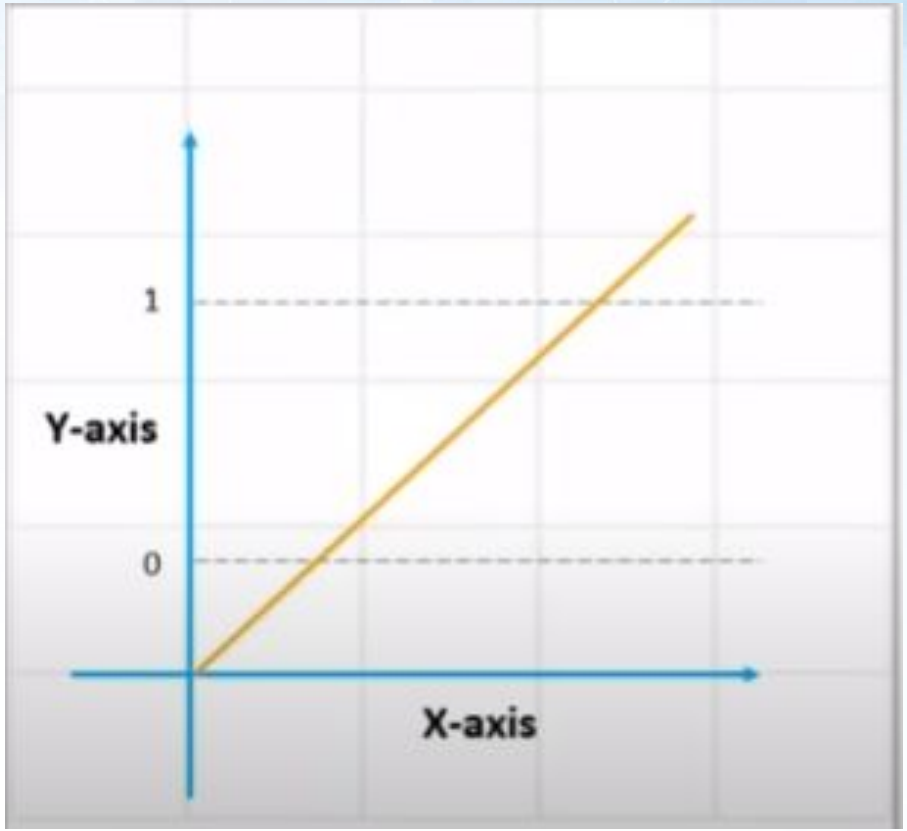
Classification

- Unlike regression where you predict a continuous number, you use classification to predict a category.
- There is a wide variety of classification applications from medicine to marketing.
- Classification models include linear models like Logistic Regression, SVM, and nonlinear ones like K-NN, Kernel SVM and Random Forests.

Logistic Regression

- Produces results in the form of binary format
- The outcome should be discrete/categorical such as:
  - 0 or 1
  - Yes or No
  - True or False
  - High and Low

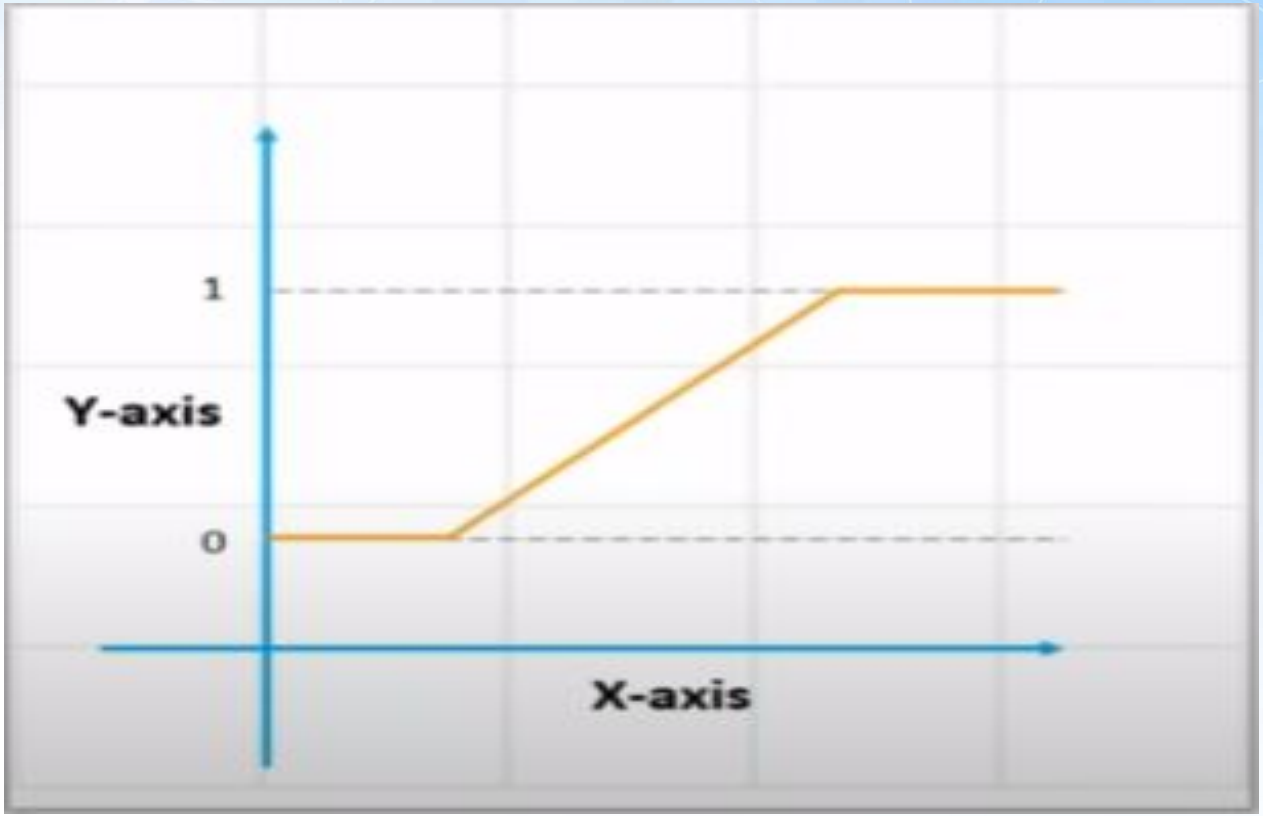
Logistic Regression



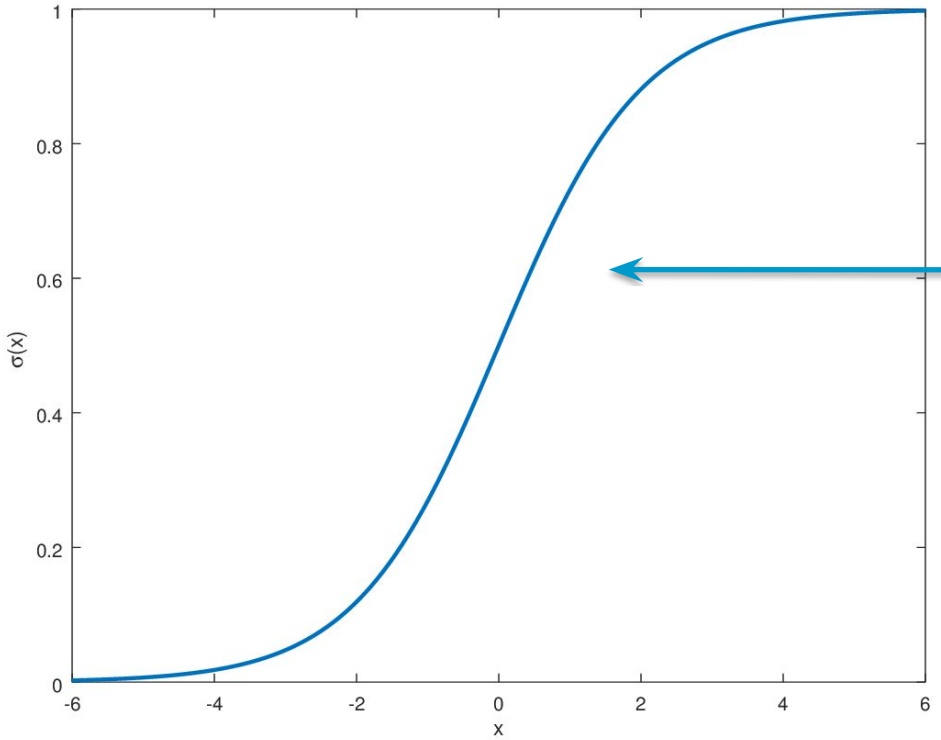
$y = mx + c$

$y = 0 \text{ or } 1$

Logistic Regression



Logistic Regression



Sigmoid curve

Logistic Regression

- y = c + m1 x1 + m2 x2 + … where y ranges from – infinity to infinity
- y = c + m1 x1 + m2 x2 + … where y ranges from 0 to 1
- y /(y-1) = c + m1 x1 + m2 x2 + … + where y ranges from 0 to infinity
- log(y/(y-1)) = c + m1 x1 + m2 x2 + … + where y ranges from 0 to 1

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

Logistic Regression

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Where, y is dependent variable and x1, x2 ...

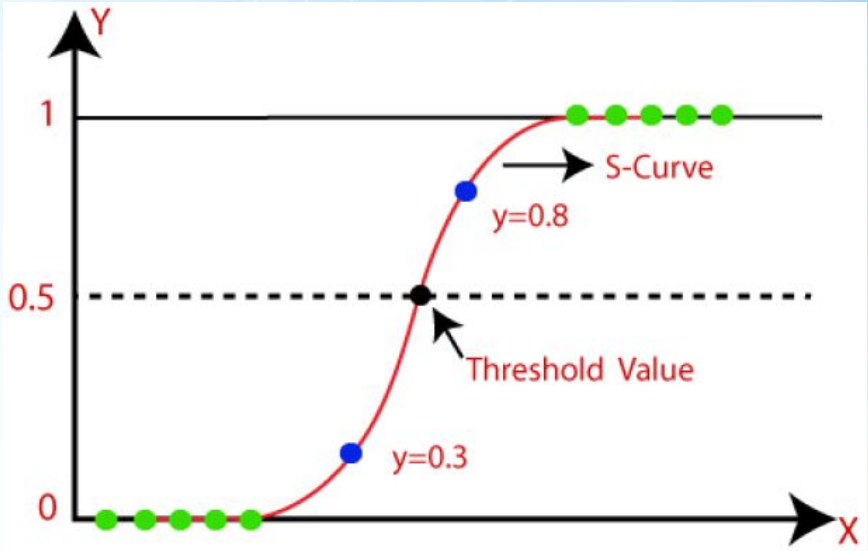
and Xn are explanatory variables.

Sigmoid Function:

$$p = 1/1 + e^{-y}$$

Apply Sigmoid function on linear regression:

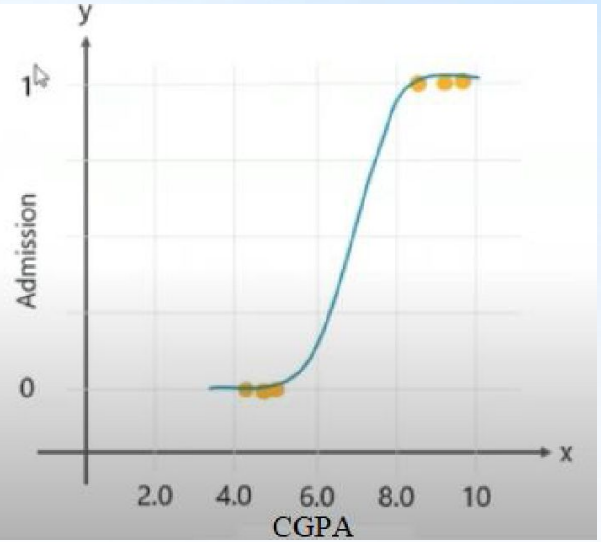
$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots\ldots \beta nXn)}$$



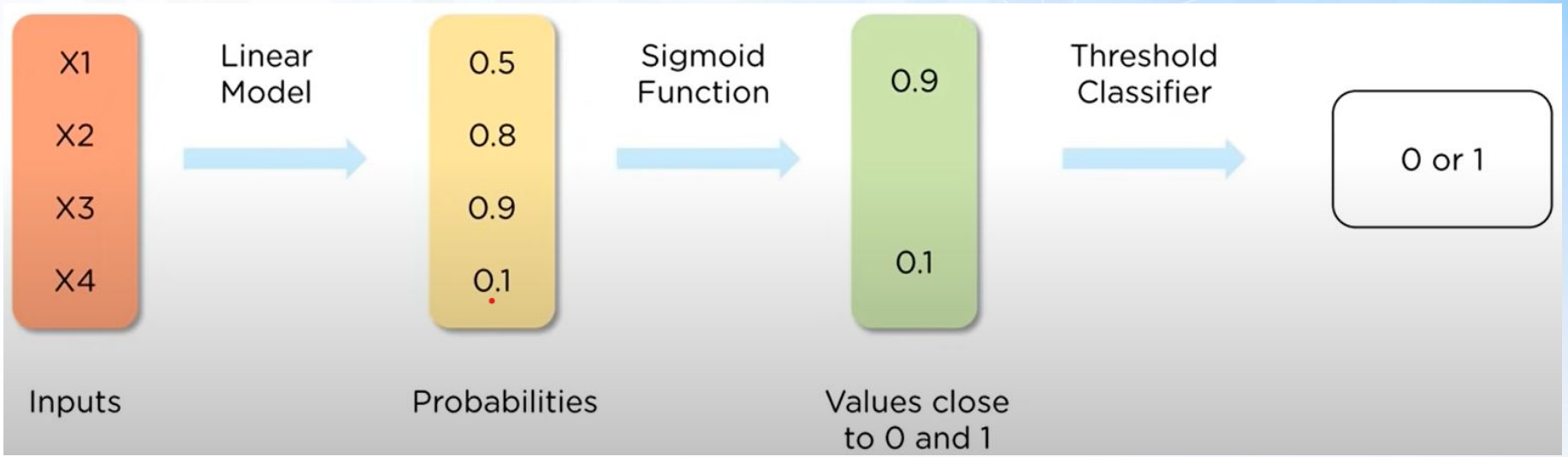
Logistic Regression

- To predict if a student will get admitted to a school based on CGPA

Admission	CGPA
0	4.2
0	5.1
0	5.5
1	8.2
1	9.0
1	9.1



Logistic Regression



Logistic Regression

```
X_train, X_test, y_train, y_test = train_test_split
            (X, y, test_size=0.33, random_state=42)

clf = LogisticRegression()

clf.fit(X_train, y_train)

THRESHOLD = 0.25

preds = np.where(clf.predict_proba(X_test)[:,1] > THRESHOLD, 1,
0)
```