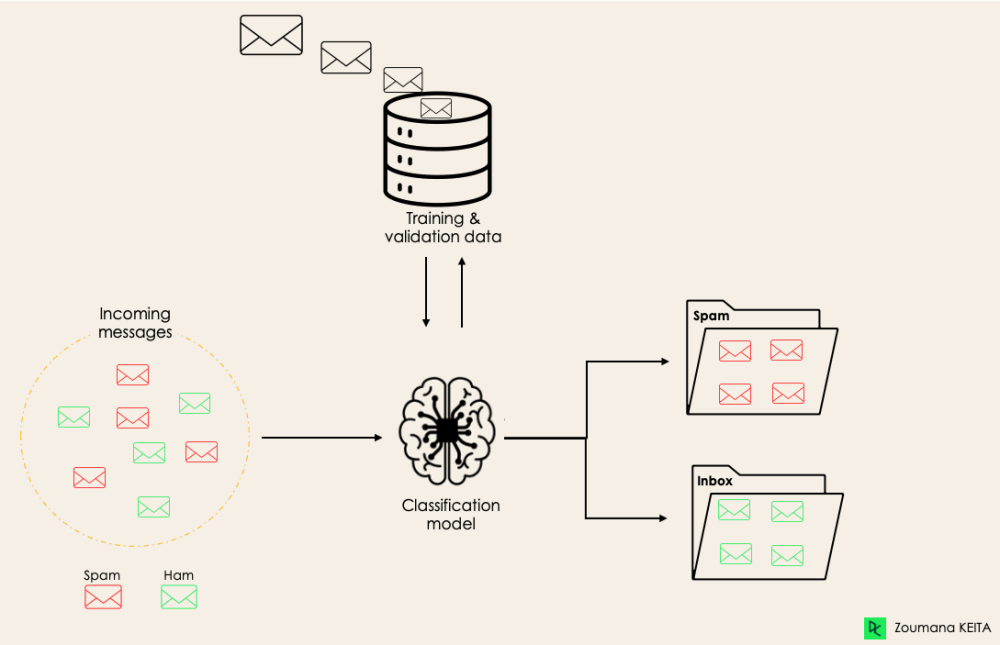
Classification

Classification is method of mechine learning that involves assigning a class label to each instance in a dataset based on its features. The goal of the method is to build a model that accurately predicts the class labels of new instances based on their features.

There are two main types of classification:

- binary : classifying instances into two classes, such as “spam” or “not spam” email
- multi-class: classifying instances into more than two classes.



Common models

- Decision tree
- KNN

The process

Building a classification model typically involves the following steps:

1. Data Collection

In this step, the data relevant to the problem is collected. The data should be representative of the problem and should contain all the necessary attributes and labels needed for classification.

The data can be collected from various sources, such as surveys, questionnaires, websites, and databases.

2. Data Preprocessing

In the second step, the collected data needs to be preprocessed to ensure its quality.

The question is: Why do we have to do this ? Why not putting it to the model directly ?

In reality, there are 2 reasons we can not put it to mechine learning model directly:

- The data usually represented by human's nature language which is not familiar to computer's thinking. For ex: female, male, image, animal, fruit,... Normally, classification algorithms require numerical input. preprocessing will convert the raw data into numerical form
- The collected data might be fault. For ex: missing value, existing outlier, being noise,... These fault result from mistake of collecting data process of human

Preprocessing not only ensure that data are suitable to model but also make model more optimal, accurate, and high performance in predicting the outcome in future

This involves **handling missing values, dealing with outliers, handling missing and duplicated value, and Data Transformation**.

- Handling Missing Values: Missing values in the dataset can be handled by replacing them with the mean, median, or mode of the corresponding feature or by removing the entire record.
- handling missing and duplicated value
- Dealing with Outliers: Outliers in the dataset can be detected using various statistical techniques such as z-score analysis, boxplots, and scatterplots. Outliers can be removed from the dataset or replaced with the mean, median, or mode of the corresponding feature.
- Data Transformation: this involves two steps:
    - scaling or normalizing the data to bring it into a common scale.
    - labeling data by label enconding or one hot encoding.

3. Feature Selection

The third step in building a classification model is feature selection, which means identifying the most relevant attributes in the dataset for classification. This can be done using various techniques, such as **correlation analysis, information gain, and principal component analysis.**

- Correlation Analysis: Correlation analysis involves identifying the correlation between the features in the dataset. Features that are highly correlated with each other can be removed as they do not provide additional information for classification.
- Information Gain: Information gain is a measure of the amount of information that a feature provides for classification. Features with high information gain are selected for classification.
- Principal Component Analysis: Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of the dataset. PCA identifies the most important features in the dataset and removes the redundant ones.

4. Model Selection

The fourth step in building a classification model is model selection. Model selection involves selecting the appropriate algorithm for the problem at hand. There are several algorithms available, such as **decision trees, support vector machines, and neural networks**.

- Decision Trees: Decision trees are a simple yet powerful classification algorithm. They divide the dataset into smaller subsets based on the values of the features and construct a tree-like model that can be used for classification.
- Support Vector Machines: Support Vector Machines (SVMs) are a popular classification algorithm used for both linear and nonlinear classification problems. SVMs are based on the concept of maximum margin, which involves finding the hyperplane that maximizes the distance between the two classes.
- Neural Networks: Neural Networks are a powerful classification algorithm that can learn complex patterns in the data. They are inspired by the structure of the human brain and consist of multiple layers of interconnected nodes.

5. Model Training

The fifth step in building a classification model is model training. Model training involves using the selected classification algorithm to learn the patterns in the data. The data is divided into a training set and a validation set. The model is trained using the training set, and its performance is evaluated on the validation set.

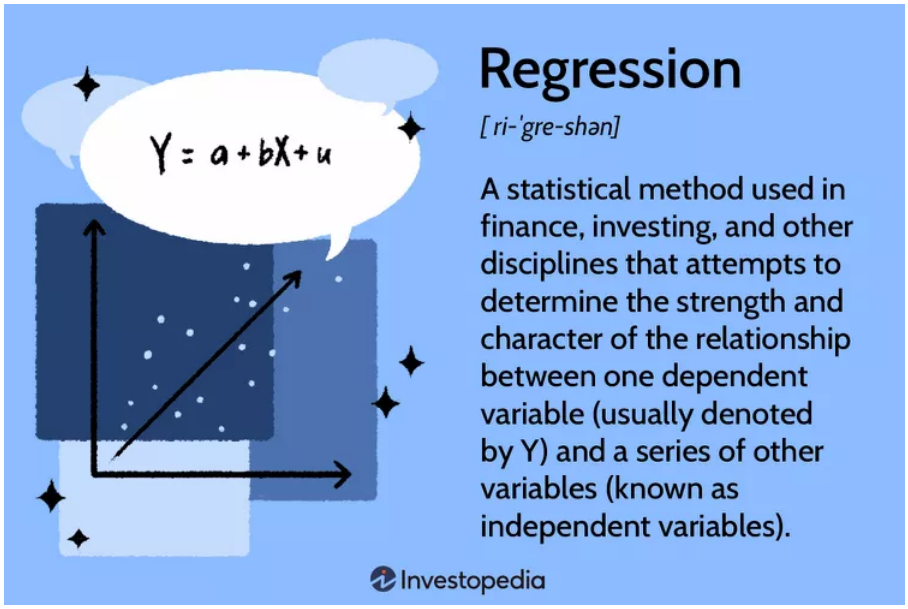
6. Model Evaluation

The sixth step in building a classification model is model evaluation. Model evaluation involves assessing the performance of the trained model on a test set. This is done to ensure that the model generalizes well

Regression

Regression is a machine learning method that uses statistics to predict the dependent variable (Y) based on the value of the independent variable (X). It can be used for cases where we want to predict a continuous quantity. For example, predicting traffic in a retail store, predicting how long users spend on a certain page or the number of pages visited on a certain website, etc.

Regression helps economists and financial analysts in things ranging from asset valuation to making predictions.



Common models

Simple linear regression:

$$Y = a + bX + u$$

Multiple linear regression:

$$Y = a + b_1\, X1 + b_2\, X_2 + b_3\, X_3 + \ldots + b_t\, X_t + u$$

where:

- Y: The dependent variable you are trying to predict
- X: The independent variable you are using to predict or associate with Y
- a: The y-intercept
- b: (beta coefficient) is the slope of the explanatory variables
- u: The regression residual or error term

For example: $Y = 1.0 + (3.2)X1 - 2.0(X2) + 0.21.$

Tree family

- Decision tree
- XGBoost
- Random Forest

The process

Similar to Classification, Regression can be implemented by these following step:

1. Data Collection

2. Data Preprocessing

3. Feature Selection

4. Model Selection

5. Model Training

6. Model Evaluation

Diference between Regression and Classification

- Regression predicts continuous values.
- Classification categorizes data into predefined classes.

Do you have any questions?