

Principles of Data Science and Computing Systems

DATA PREPROCESSING, MODELING, AND REPORTING

A picture containing text, sign

Description automatically generated

SAIF HADDAD

21110214

**Supervised and Unsupervised Learning:**

**Supervised learning:**

The supervised learning method in machine learning makes use of labeled datasets to teach computers to accurately categorize input or predict outcomes. The model evaluates the significance of several features to gradually enhance the model fit to the known result using the labeled data.

(alteryx, 2022)

**Unsupervised learning:**

Algorithms are used to investigate and classify unlabeled datasets in unsupervised learning. Without human oversight, these algorithms may find hidden patterns in data. (alteryx, 2022)

**The difference between Supervised and Unsupervised**

In supervised learning, classification or prediction algorithms are trained using labeled datasets. The 'training' data is introduced, and the model iteratively modifies how it weights various data aspects until it is well suited to the intended result. Models for supervised learning are far more accurate than the alternative method. To confirm that the labels on the information are suitable it needs people's help, they do, however, need people to be involved in the data processing process also.

Unsupervised learning models, in contrast, operate continuously with no human intervention. Using unlabeled data, they discover and arrive at a structure of sorts. Only the output variables need to be validated by humans in this situation. (alteryx, 2022)

|  |  |
| --- | --- |
| **Supervised** | **Unsupervised** |
| The input data is labeled | The input data is Unlabeled |
| Depends on training Dataset | Depends just on input dataset |
| A training dataset is used to classify the data. | classifies supplied data using its properties. |
| Used for prediction | Used for Analysis |
| split into two categories Classification & Regression | split into two categories Clustering and Association |

**Regression Model:**

This statistical technique measures the correlation between a dependent variable and one or more independent variables using algorithms. Regression models allow the user to forecast causes and effects based on a variety of data points. (alteryx, 2022)

**Types of Regressions:**

**Simple Linear Regression:**

If both variables are quantitative, you may use a straight line to estimate the connection between an independent variable and a dependent variable.v

**Multiple linear regression:**

use for predicting a dependent variable's values based on the values of two or more independent variables. (About Sakshi Gupta , 2022)

**Polynomial regression:**

finding a nonlinear link between the dependent and independent variables through modeling. (About Sakshi Gupta , 2022)

**Classification Model:**

is a predictive model that makes use of discrete output variables, such as labels or categories, to find a rough mapping function from input variables. Discrete and real-valued variables can both be used in a classification process, but it still needs the instances to fall into one of at least two groups. (About Sakshi Gupta , 2022)

**Types of Classification:**

**Decision tree classification:**

A classification model is built by constructing a decision tree, where each node represents a test case for an attribute and each branch deriving from the node represents a potential value for that attribute. (About Sakshi Gupta , 2022)

**Random forest classification**

A portion of the main training set's decision trees are randomly chosen as part of this tree-based approach. The ultimate output prediction made by the random forest classification method, which is more accurate than any of the individual trees, is made by combining the outputs from all the various decision trees. (About Sakshi Gupta , 2022)

**K-nearest neighbor**

The K-nearest neighbor method posits that related objects may be found nearby. For estimating the values of fresh data points, it makes advantage of feature similarity. The algorithm assists in putting nearby data items that are comparable together. The algorithm's primary objective is to assess a data point's likelihood of belonging to a certain category. (About Sakshi Gupta , 2022)

**The Difference Between the Classification and the Regression Models:**

is that classification predicts discrete class labels, whereas regression helps forecast a continuous quantity.

|  |  |
| --- | --- |
| **Regression Algorithm** | **Classification Algorithm** |
| The output variable in regression must have a continuous or real value. | The output variable for classification must have a discrete value. |
| The regression algorithm's job is to map the continuous output variable to the input value (x) (y). | The classification algorithm's job is to associate the discrete output variable with the input value (x) (y). |
| Continuous data are employed with regression algorithms. | With discrete data, classification algorithms are applied. |
| We look for the best fit line in regression in order to more precisely forecast the result. | When classifying data, we seek out the decision boundary that may categorize the dataset. |
| May be used to resolve regression issues like weather forecasting and house price prediction, among others. | May be used to identify spam emails, recognize speech, identify cancer cells, and other classification-related issues. |

Discuss and explain the data science life cycle of your developed prediction model (any of them).

-List the data preprocessing techniques utilized for building prediction models.

-Discuss the use of each preprocessing technique utilized for building prediction models.

alteryx (2022) *Supervised vs. unsupervised learning; which is best?*, *Alteryx*. Available at: https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning#:~:text=Supervised%20and%20unsupervised%20learning%20have,tagged%20with%20the%20right%20answer.&text=A%20classification%20problem%20uses%20algorithms%20to%20classify%20data%20into%20particular%20segments. (Accessed: January 22, 2023).

About Sakshi Gupta Sakshi is a Senior Associate Editor at Springboard. She is a technology enthusiast who loves to read and write about emerging tech. She is a content marketer and has experience working in the Indian and US markets. (2022) *Regression vs. classification in machine learning: What's the difference?*, *Springboard Blog*. Available at: https://www.springboard.com/blog/data-science/regression-vs-classification/#:~:text=The%20most%20significant%20difference%20between,types%20of%20machine%20learning%20algorithms. (Accessed: January 22, 2023).

www.javatpoint.com. (n.d.). *Regression vs Classification in Machine Learning - Javatpoint*. [online] Available at: https://www.javatpoint.com/regression-vs-classification-in-machine-learning.

‌ *Regression vs Classification in Machine Learning - Javatpoint*. (n.d.). www.javatpoint.com. <https://www.javatpoint.com/regression-vs-classification-in-machine-learning>