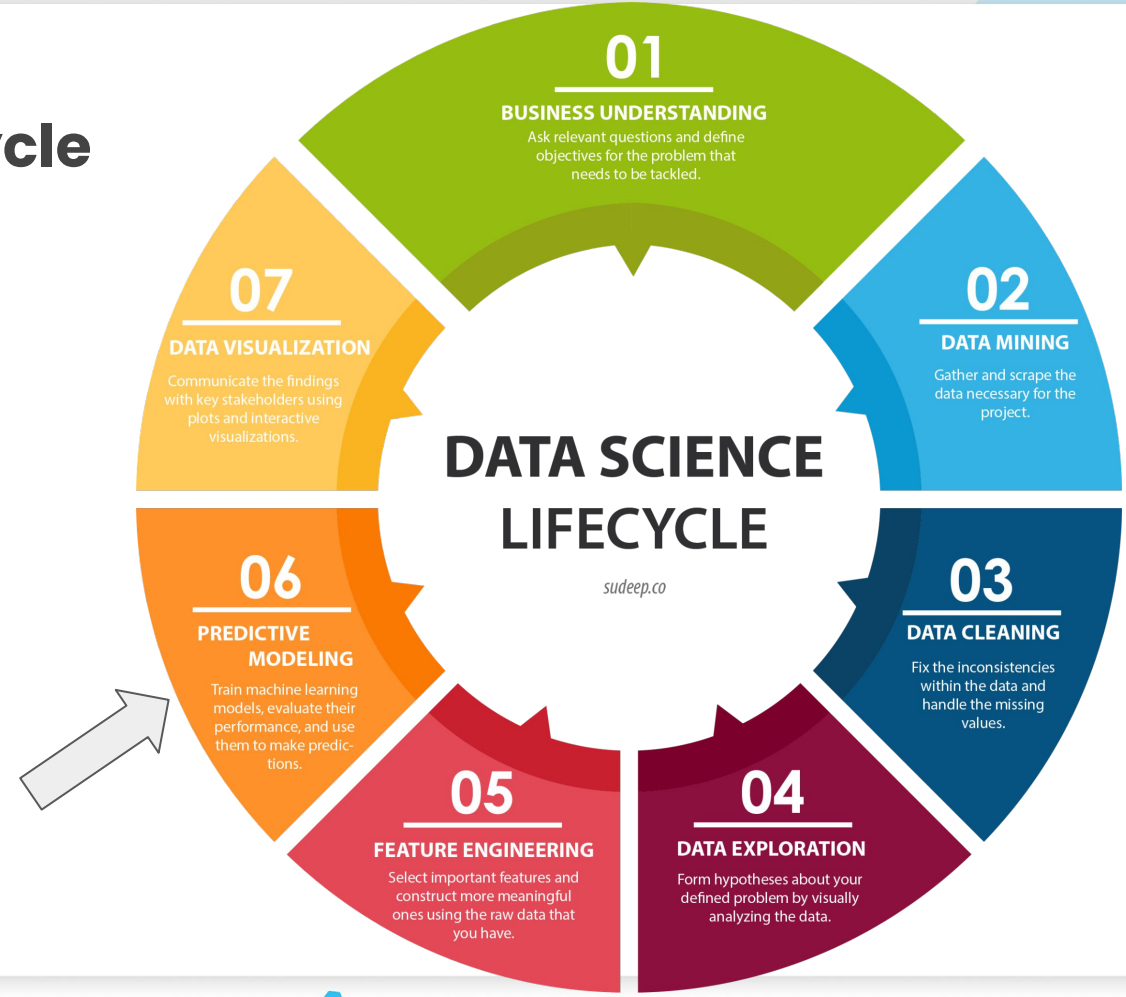
IRON HACK

Supervised Learning

Week 7

Lifecycle


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Supervised Learning

So far we've seen **KNN**, both for **Regression** and **Classification**.

Today we will explore more models, such as:

- Linear Regression
- Logistic Regression
- Decision Trees

Linear Regression

Supervised Learning

Linear Regression

Linear Regression is a fundamental machine learning algorithm used for predictive analysis, when our target is numeric/continuous (Regression).

It models the relationship between a **dependent variable(target)** and **one or more independent variables(features)** by fitting a linear equation to the observed data.

Supervised Learning

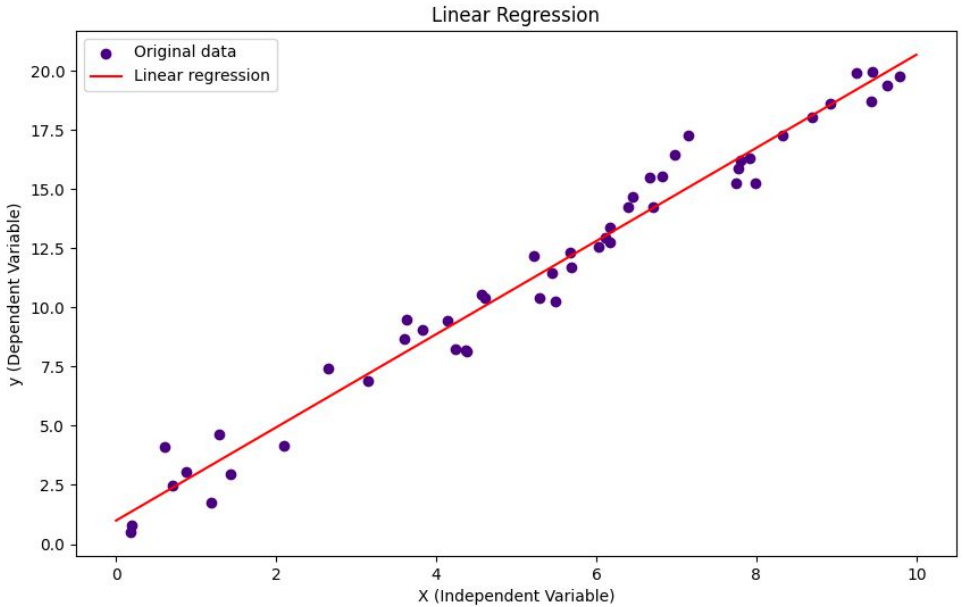
Linear Regression

Linear Regression Equation

$$y = b_0 + b_1 x$$

Where:

- b0 is the y-intercept
- b1 is the slope

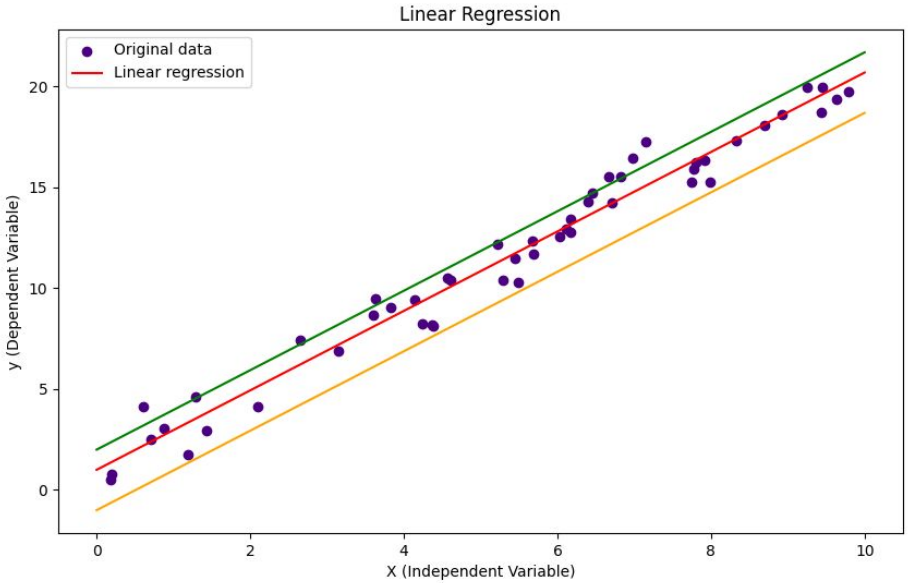


Supervised Learning

Linear Regression

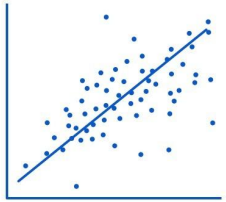
But how does it fit in to our data?

Linear regression finds the best-fitting straight line through the data points by adjusting the line until it minimizes the **overall distance between the line and the points.**

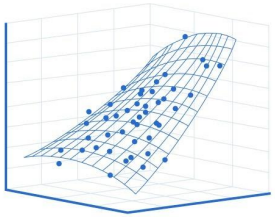


Supervised Learning
Linear Regression

Simple Linear Regression



Multiple Linear Regression



What if we have more features?

Often we use more than a single feature to predict our target.

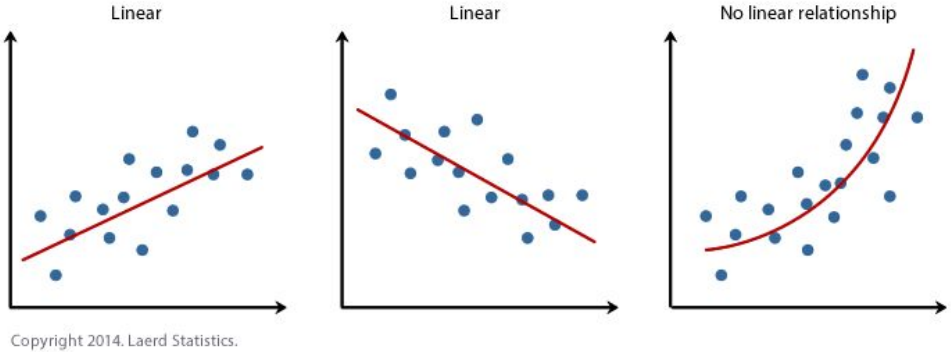
In this case we have:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \epsilon$$

Supervised Learning
Linear Regression

Although Linear Regression is widely employed in various industry solutions, it does have certain limitations.

- **Assumes** that the **relationship** between independent variables and dependent variables is **linear**.
- **Sensitive to outliers.**



Logistic Regression

Supervised Learning

Logistic Regression

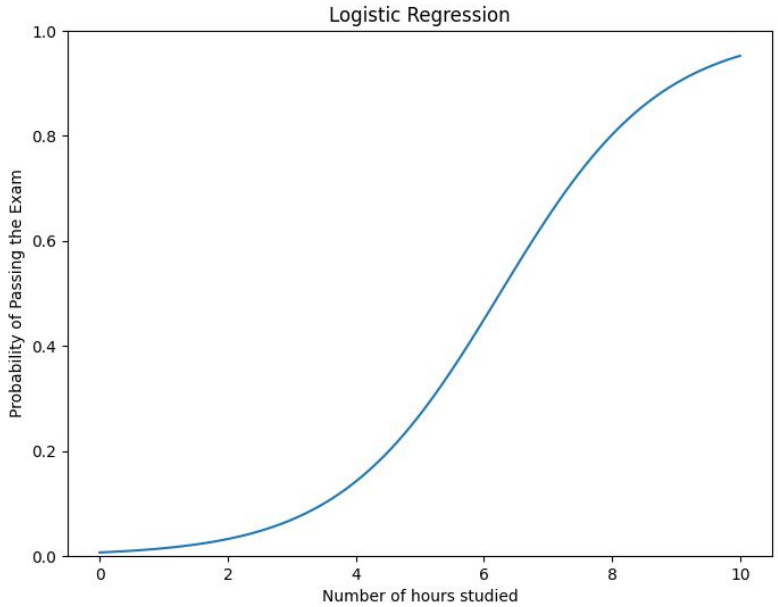
Logistic Regression is also a fundamental machine learning algorithm used for predictive analysis, but in this case we use it for **Classification** problems (usually binary).

It calculates the probability of the **class** of a specific datapoint based on input features.

Supervised Learning
Logistic Regression

Logistic Regression fits an S shaped curve to the data, called Sigmoid.

When stacked, it can generate really powerful models (Deep Learning).

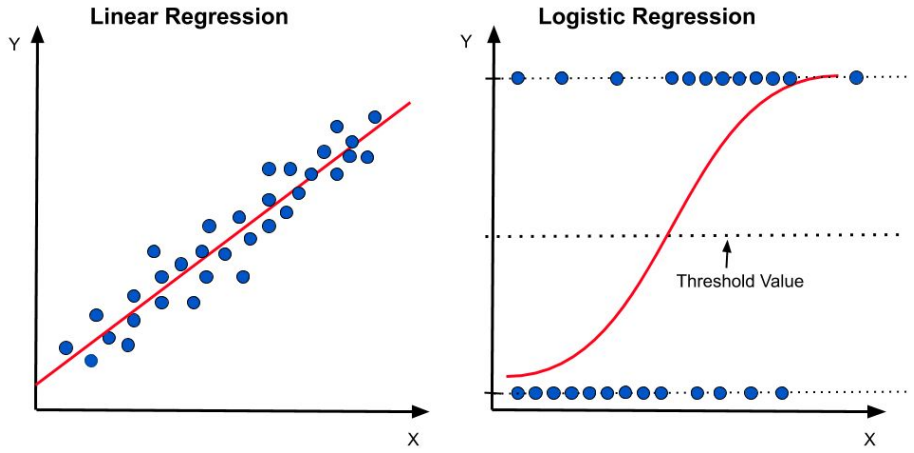


Supervised Learning

Logistic Regression

Logistic Regression Formula

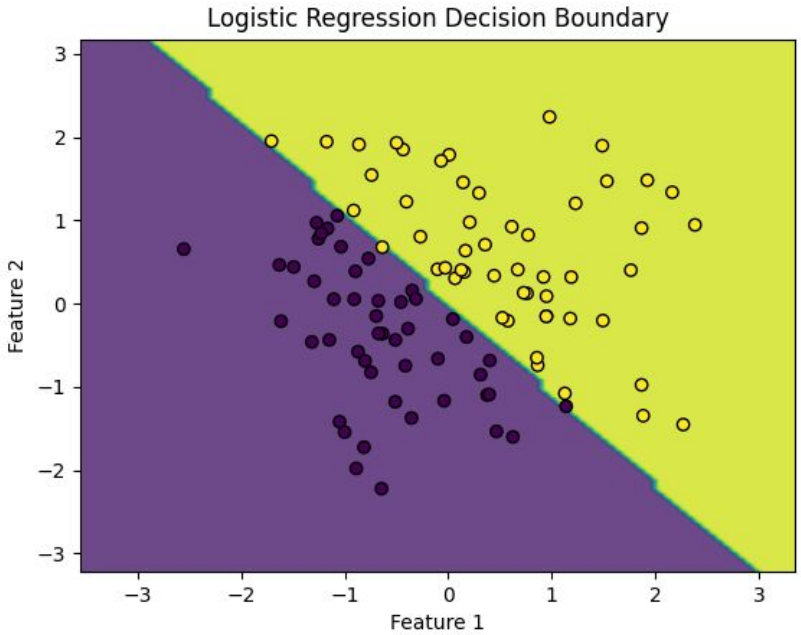
$$y = \frac{1}{1 + e^{(b_1 x + b_0)}}$$



Supervised Learning
Logistic Regression

Logistic Regression Limitations

- Fails in non-linearly separable data.
- Do not work well for multi-class problems



Decision Trees

Supervised Learning

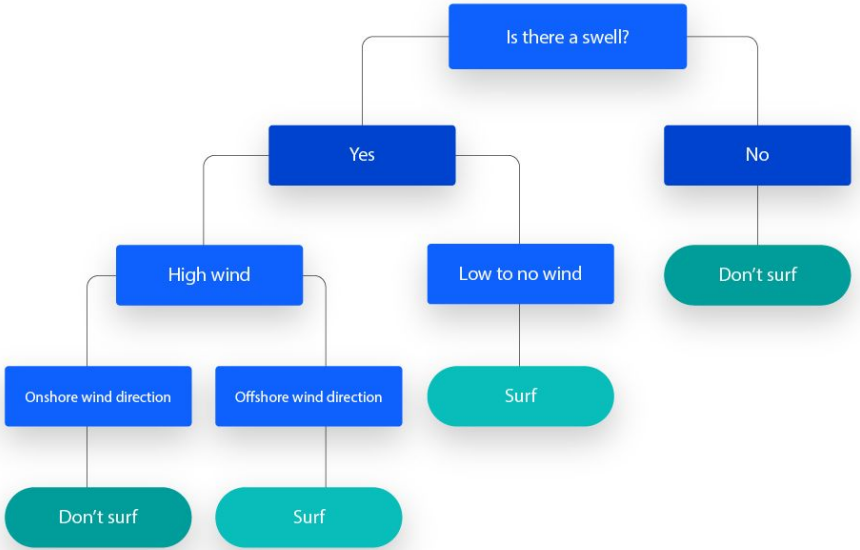
Decision Trees

A **Decision Tree** is a widely-used supervised learning method capable of handling both **classification** and **regression** problems. It dissects a dataset into smaller subsets by employing if-then-else decision rules based on the data's features.

Supervised Learning

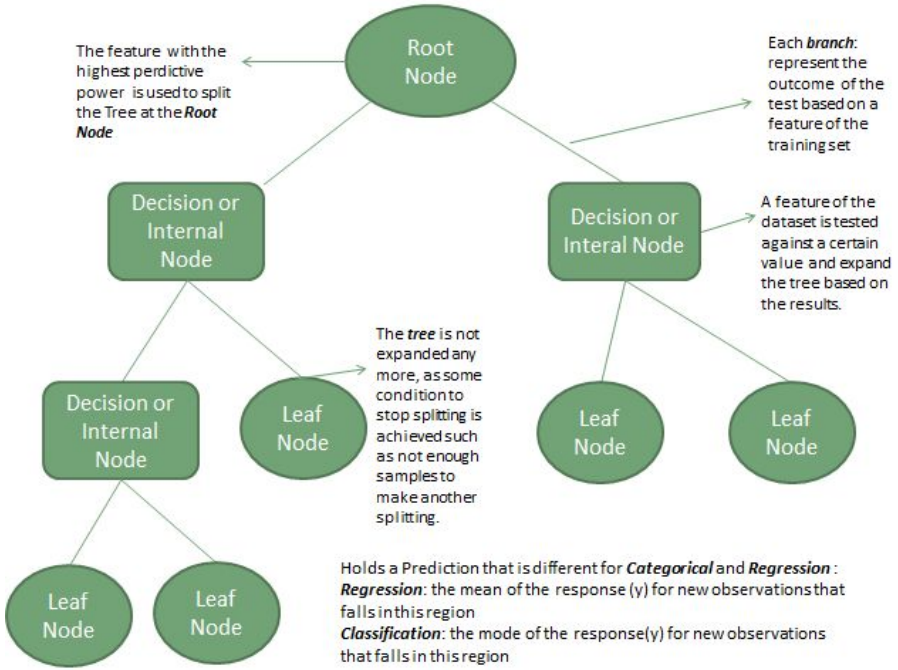
Decision Trees

Each level of the Decision Tree “asks” a question about a specific feature and divides the paths into several nodes.



Supervised Learning

Decision Trees



Supervised Learning
Decision Trees

Decision Tree Limitations

- Affected by noise in the data
- Decision nodes limited to binary outcomes (reducing the complexity that tree can handle)
- Can overfit very quickly