DATA | MODULE 3

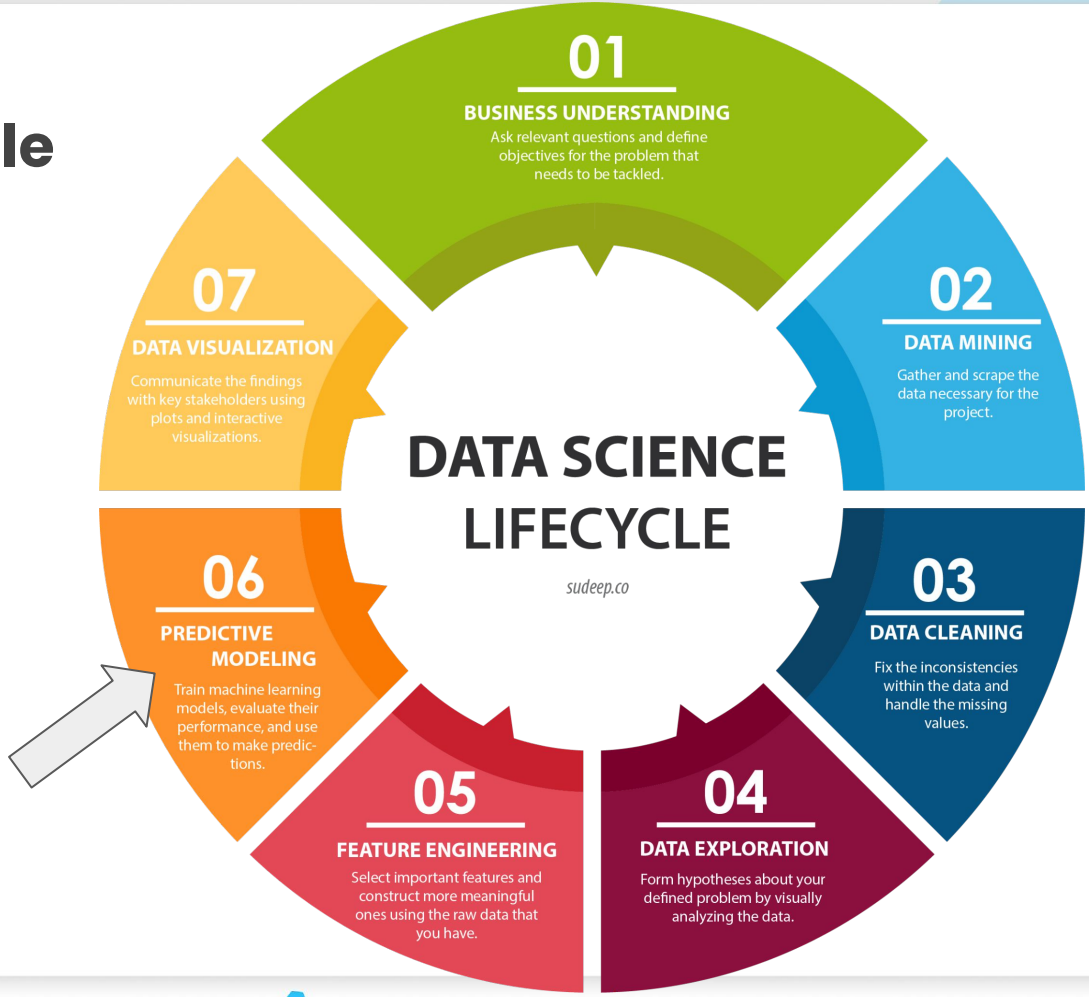
IRON HACK

Intro to Machine Learning

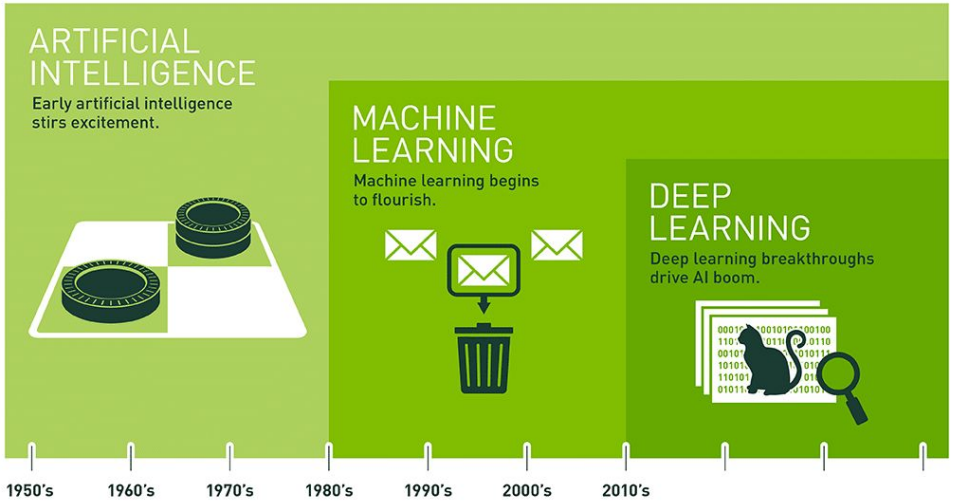
Week 7

Table of Contents

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

Machine Learning
What is it?



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Machine Learning
What is it?

AI it’s been around for a while now, and includes all solutions that mimic human behavior, and in some way replacing human decision making by AI models.

Machine Learning is a subset of Artificial Intelligence, using historical data to predict future behaviors.

Deep Learning is a subset of Machine Learning that also learns with historical data, focusing specifically on algorithms inspired by the structure and function of the brain's neural networks.

Machine Learning
Some applications

Fraud detection

Financial institutions use **machine learning** to spot fraud, like credit card fraud, by analysing historical data in order to find patterns and spotting suspicious behavior.

Machine Learning
Some applications

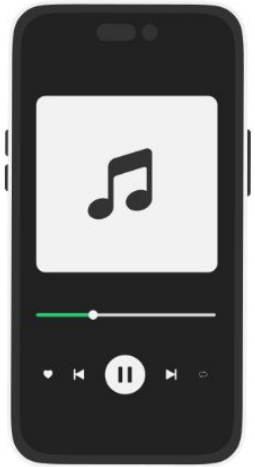
Healthcare Diagnosis

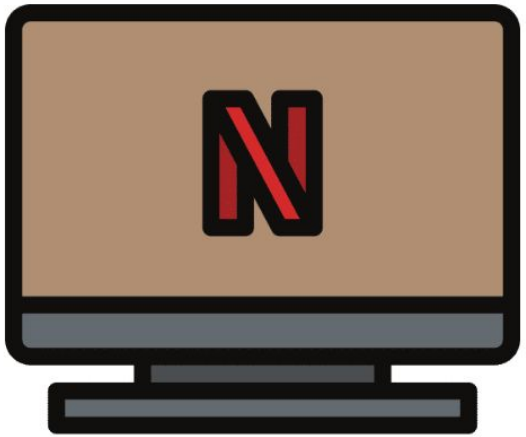
Machine learning models are used to analyze medical data including imaging scans, genetic information, and patient records to assist healthcare professionals in diagnosing diseases

Machine Learning
Some applications

Recommendation Systems

Platforms like Netflix, Amazon, and Spotify use **ML** algorithms to analyze user preferences and behavior, and then recommend personalized content to enhance user experience.





Machine Learning
Some applications

Autonomous Vehicles

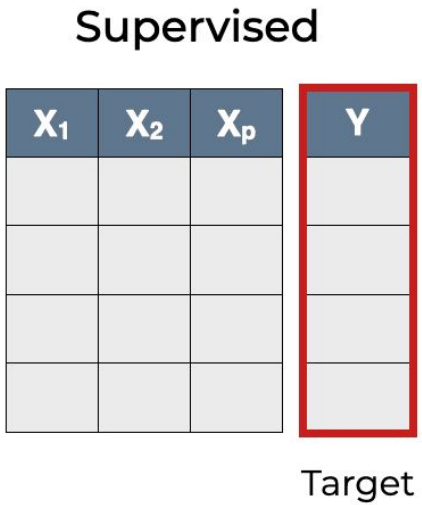
Self-driving cars employ a variety of **AI** and **ML** techniques including computer vision, sensor fusion, and deep learning to perceive their environment, make decisions, and navigate safely without human intervention.

Types of Machine Learning

Machine Learning
Different types of ML

Supervised Learning

- You have data with inputs and corresponding **outputs**. The algorithm learns from this labeled data to make predictions on new, unseen data.



Machine Learning
Different types of ML

Unsupervised Learning

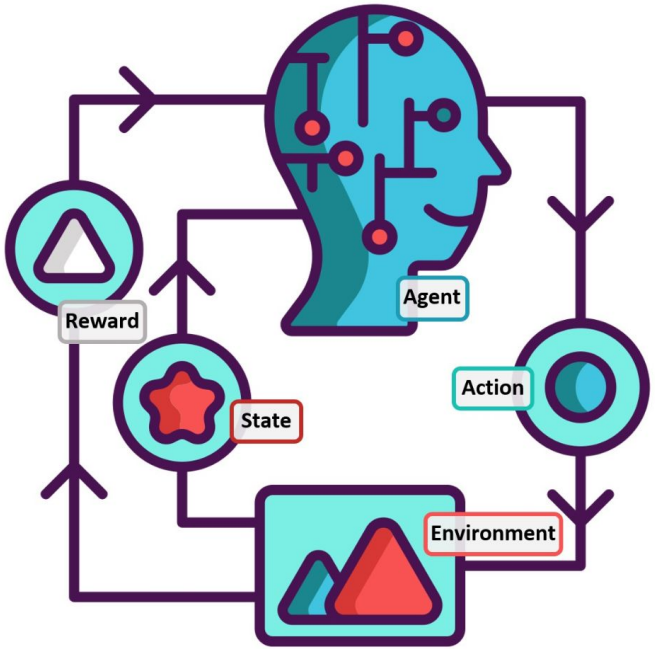
- You have data **without** predefined labels or categories. The algorithm tries to find patterns or structures in the data on its own.

Unsupervised

X_1	X_2	X_p	Y

No Target

Machine Learning
Different types of ML



Reinforcement Learning

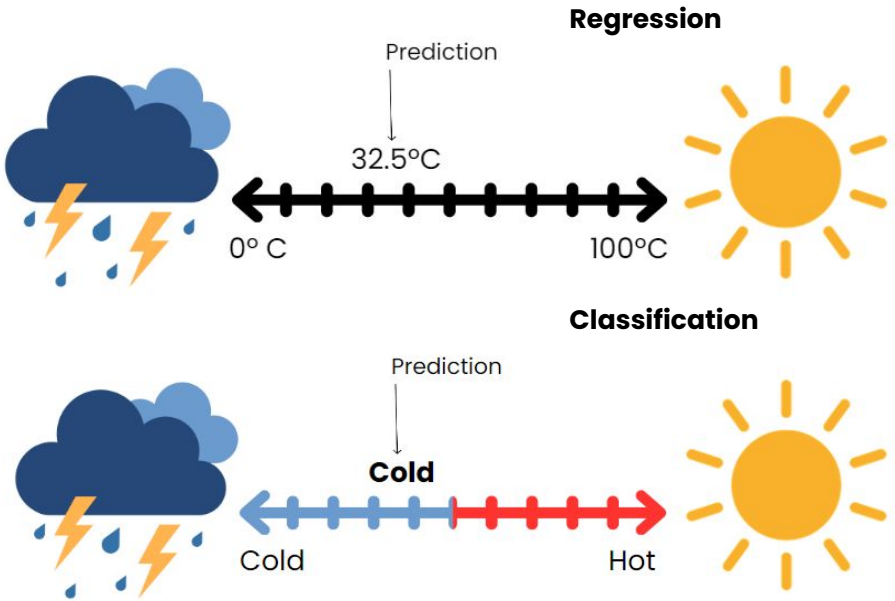
- An algorithm learns to make decisions by trial and error, receiving feedback in the form of rewards or penalties as it interacts with an environment.

Supervised Learning

Machine Learning
Supervised Learning

Supervised Learning

- In Supervised Learning, depending on the type of variable we are trying to predict, a particular problem can either be a **Regression** or **Classification**.



Machine Learning
Supervised Learning

Classification

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.2	4.1	1.5	0.1	Iris-Setosa
1	7.7	3.0	6.1	2.3	Iris-Virginica
2	5.1	3.8	1.5	0.3	Iris-Setosa
3	6.5	3.0	5.8	2.2	Iris-Virginica
4	6.1	3.0	4.6	1.4	Iris-Versicolour
5	6.9	3.1	4.9	1.5	Iris-Versicolour
6	4.6	3.6	1.0	0.2	Iris-Setosa
7	5.2	3.5	1.5	0.2	Iris-Setosa

In a classification problem our target variable is **Categorical.** We have a **finite** number of possible outcomes.

Machine Learning
Supervised Learning

Regression

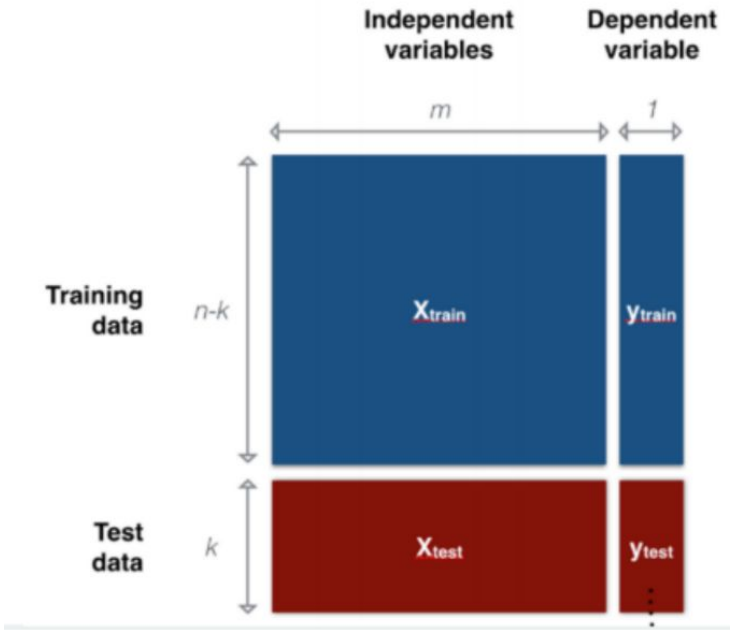
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	5.4103	33.0	5.275748	0.960133	683.0	2.269103	32.81	-117.24	3.32400
1	2.8788	38.0	5.862197	1.057728	1378.0	2.566108	36.35	-119.65	0.85500
2	3.1324	19.0	29.248649	5.632432	441.0	2.383784	34.24	-116.86	1.32000
3	2.6903	34.0	5.729216	1.059382	1558.0	3.700713	35.37	-118.93	0.62800
4	4.3512	24.0	4.976101	1.021384	3196.0	4.020126	33.76	-117.95	1.91400
5	5.5248	22.0	6.141361	0.994764	1217.0	3.185864	32.68	-116.98	1.56300
6	4.1026	25.0	3.549683	0.993658	1314.0	2.778013	33.65	-117.92	2.11500
7	2.0547	35.0	4.750000	1.000000	1183.0	4.349265	32.69	-117.10	0.98000

In a regression problem our target variable is **Continuous/Numerical.** We have a **infinite** number of possible outcomes.

Machine Learning
Supervised Learning

Train Test Split

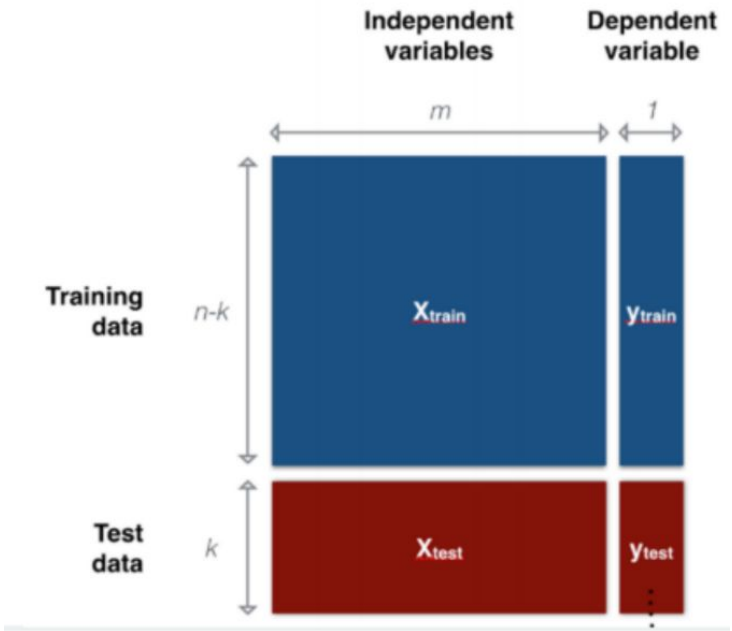
- After identifying the type of variable we're trying to predict, we must separate our data into **Training** and **Testing**.
- Usually we reserve between 20-30% for Test set.



Machine Learning
Supervised Learning

Train Test Split

- We split our data into **Training** and **Test** sets to evaluate our model fairly. During training, the model only accesses the Training set. Once trained, we use the Test set to predict labels for which we already know the target.
- Then, we compare our predictions with the true labels of the data points in the Test set.

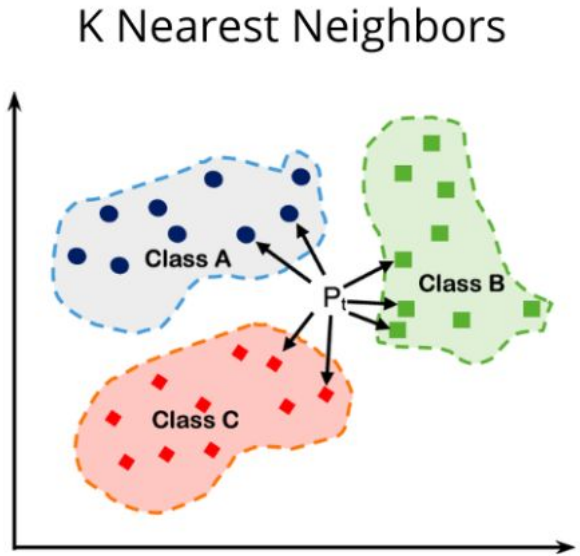


K Nearest Neighbors

Machine Learning
KNN

- **K Nearest Neighbors** or **KNN**, is a ML algorithm that classifies an data point by assigning it the most common class label among its nearest neighbors in the training set.

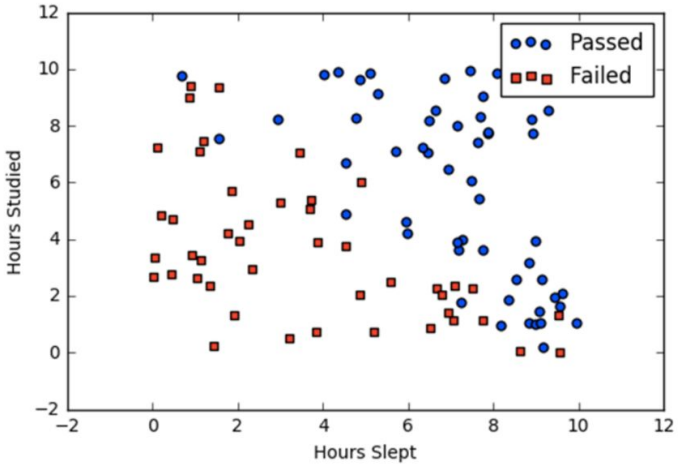
- Can be used both **classification** or **regression.**



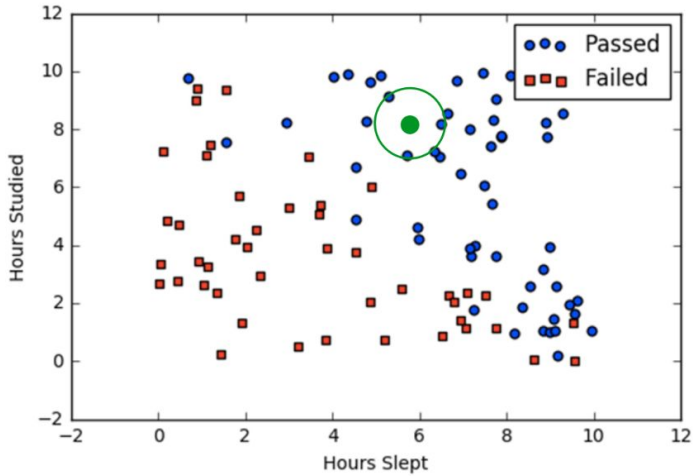
Machine Learning
KNN

- Suppose you have two independent variables(**features**):
  - **Hours Slept** and **Hours studied**

- We want to **predict** if a student will either **pass** or **fail**.



Machine Learning
KNN



- KNN will label a new data point based on the class of the **k** nearest points.
- For instance, when k=3, **KNN** evaluates the class labels of the 3 closest data points and assigns a class to the new data point through a majority voting.
- For **Regression**, we calculate the **average** instead majority voting.

Evaluation Metrics

Machine Learning
Evaluations Metrics

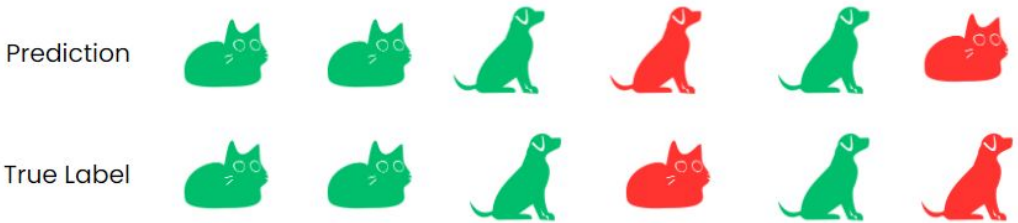
- After training our model, we need to evaluate its performance.

- Metrics compute in order to evaluate the performance, will be different whether we are dealing with a **Classification** or a **Regression** problem.

Machine Learning
Evaluations Metrics
Classification

- In **Classification** problems, **Accuracy** is often to compute in order to assess model performance.

$$accuracy = \frac{Correct\ Predictions}{All\ Predictions}$$

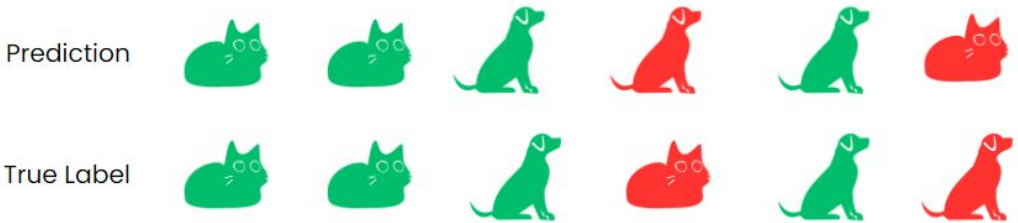


Machine Learning
Evaluations Metrics
Classification

In this case we have correctly predict 4 out a total of 6 predictions

$$accuracy = \frac{Correct\ Predictions}{All\ Predictions}$$

$$accuracy = \frac{4}{6} = 66\%$$



Machine Learning
Evaluations Metrics
Regression

- In **Regression** problems, **R-Squared** is often to compute in order to assess model performance.
- **R-squared (R2)** measures how effectively the independent variables in a regression model predict or explain changes in the dependent variable.
- It varies between 0 and 1, where values close to 1 indicate a better model (though occasionally, we may obtain negative values for R2).

Machine Learning
Evaluations Metrics

Regression

- **R-Squared Formula**