What is Machine Learning?

“Learning is any process by which a system improves performance from experience.”
– *Herbert Simon*

Definition by Tom Mitchell (1998):

Machine Learning is the study of algorithms that:

- Improve their performance **P**
- At some task **T**
- With experience **E**

A well-defined learning task is represented as (**P, T, E**).

Traditional Programming vs Machine Learning

Traditional Programming

- **Input:** Data + Program
- **Output:** Result

Machine Learning

- **Input:** Data + Output
- **Output:** Program

When Do We Use Machine Learning?

ML is useful when:

- Human expertise does **not exist**
- Humans can’t explain their expertise (e.g., speech recognition)
- Models need to be **customized** (e.g., personalized medicine)
- Huge amounts of data are involved (e.g., genomics)

ML **is not needed** when:

- The task is rule-based (e.g., payroll calculation)

Examples of ML Use Cases

- **Pattern Recognition:**
  - Facial identity/expression
  - Handwriting/speech
  - Medical imaging
- **Pattern Generation:**
  - Images or motion sequences
- **Anomaly Detection:**
  - Credit card fraud
  - Nuclear sensor anomalies
- **Prediction:**
  - Stock prices
  - Currency rates

Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Robotics
- Social networks
- Software debugging
- Space exploration
- [Insert your domain here]

Historical Insight

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.”
–*Arthur Samuel (1959)*

Defining the Learning Task

Improve on task **T**, with respect to performance **P**, based on experience **E**.

Task	**Performance (P)**	**Experience (E)**
Playing checkers	% of games won	Self-play games
Handwriting recognition	% words correctly classified	Human-labeled image dataset

Task	Performance (P)	Experience (E)
Highway driving	Avg. distance before human correction	Recorded sensor/steering data
Email classification	% correctly classified	Labeled email dataset

Types of Learning

1. Supervised Learning
   - Input: Training data + labels
2. Unsupervised Learning
   - Input: Data without labels
3. Semi-supervised Learning
   - Input: Data + few labels
4. Reinforcement Learning
   - Input: Sequence of actions with rewards

Supervised Learning

Regression

- Predict real-valued outputs
- Example: House prices over years

Classification

- Predict categorical outputs
- Example: Cancer classification (Benign = 0, Malignant = 1)

Data Representation in Supervised Learning

- Input x can be multi-dimensional
- Each dimension = feature (e.g., Age, Tumor Size, Cell Shape, etc.)

Unsupervised Learning

- Input: xi, X_2, ..., x_n (no labels)
- Goal: Discover hidden structure
- Example: Customer segmentation using clustering

Reinforcement Learning

- Output: Policy (mapping states to actions)
- Examples:
  - Robot navigation
  - Game playing o
    Balancing a pole

Data Pipeline Overview

Machine Learning is only as good as the data it learns from.

Two Key Steps Before Training:

1. **Data Collection**
2. **Data Filtering (Preprocessing)**

Data Collection

What is it?

- Gathering data relevant to the ML task

Data Types:

- **Structured:** Tables, CSVs (e.g., sales records)
- **Unstructured:** Text, images, videos (e.g., tweets, CCTV)

Sources:

- Web scraping
- IoT sensors
- APIs (e.g., Twitter, Weather)
- Surveys & logs

CSV Files in ML

What is a CSV?

- Plain-text, tabular format
- Each line = data record
- Fields separated by comma, semicolon, or tab

Why CSV is Popular:

Feature Advantage

Simplicity Easy to read/write Portability Supported by all programming languages Compatibility Works with Excel, Python, R, etc. Lightweight No metadata overhead ML-Friendly Compatible with Pandas, Scikit-learn

Example CSV Data:

```
Patient ID,Age,BMI,Glucose,Blood Pressure,Outcome
P001,45~29.5,150,85,1
P002,34,,135,92,0
P003,50,31.2,,88,1
```

Data Storage

- Should be **scalable**, **secure**, and **accessible**

Data Filtering (Preprocessing)

Goals:

- Remove **duplicates**, **missing values**, and **outliers**
- Select relevant **features**
- Normalize/standardize data

Steps:

1. Remove duplicates
2. Handle missing values (drop/fill)
3. Remove outliers (Z-score/IQR)
4. Feature selection
5. Data transformation (e.g., normalization)

Examples of Data Filtering

Example 1: House Prices

- Filled missing `Lotsize` with median
- Removed irrelevant column `ID`

Example 2: Sensor Data

- Applied moving average for smoothing
- Converted temperature to Celsius

Why Data Filtering is Important

- Improves model performance
- Reduces **bias** and **variance**
- Ensures data **consistency**
- Avoids "garbage-in, garbage-out"

Real-World Use Case: Spam Detection

- **Collected Data:** Email content, metadata
- **Storage:** NoSQL database
- **Filtering:**
  - Removed HTML tags
  - Removed stopwords
  - Selected keywords as features

Domain-Wise Use Cases

1. Healthcare - Predicting Diabetes (Supervised Learning)

Before Filtering:

P001	45	29.5	150	85	1
P002	34		135	92	0
P003	50	31.2	NULL	CO CO	1

Filtering:

Filled missing BMI = 30.0 Removed row with missing Glucose

After:

P001	45	29.5	150	85	1
P002	34	30.0	135	92	0

2. Retail - Customer Segmentation (Unsupervised Learning)

Before:

C101	22	15	39
C102	35	95	81
C103		45	66
C104	28	-100	70

Filtering:

- Removed negative income
- Filled missing age with mean = 28.3

After:			
C101	22	15	39
C102	35	95	81
C103	28	45	66

3. Education - Predicting Dropouts (Supervised Learning)

Before:

S001	80	3.1	N
S002	45	2.0	Y
S003	NULL	3.4	N
Filtering:			

- Filled missing attendance with average = 62.5

After:			
S001	80	3.1	N
S002	45	2.0	Y
S003	62.5	3.4	N

4. Finance - Anomaly Detection (Unsupervised Learning)

Before:

T1001	5,000	Mumbai	10:35	AM
T1002	1,20,000	Dubai	3:15	AM
T1003	-3,000	Bangalore	1:00	PM

Filtering:

Removed negative transaction Flagged foreign transaction > ?1,00,000

After:

T1001	5,000	Mumbai	10:35 AM	Normal
T1002	1,20,000	Dubai	3:15 AM	Suspicious

5. Social Media - Sentiment Analysis (Supervised Learning)

Before:

TW001	"I love this phone!"	Positive
TW002	"Worst service ever!!"	Negative
TW003	"Just okay...nothing special"	

Filtering:

- Removed special characters
- Filled missing sentiment

After:

TW001	love this phone	Positive
TW002	worst service ever	Negative
TW003	just okay nothing special	Neutral

Other CSV Examples (Practice Datasets)

1. Gaming & Esports — player stats.csv
   Columns: Player_ID, Game_Title, Hours_Played, Highest_Score, Country
2. Food Delivery — restaurant reviews .csv
   Columns: Order_ID, Restaurant_Name, Rating, Delivery_Time, Review_Text
3. Streaming Platforms - watch_history.csv
   Columns: User_ID, Video_Title, Genre, Duration, Watched_Fully
4. Travel & Tourism — trip_bookings. csv Detect and filter
   unrealistic costs (e.g., ?10,000,000)
5. Mental Wellness Apps — mood tracker.csv Handle missing
   sleep hours and simulate data gaps