Machine Learning

LAB

Lab #1

Introduction

Instructor: Saad Rashad

Course Code: AL3002

Semester Fall 2024

Department of Computer Science,
National University of Computer and Emerging Sciences FAST
Peshawar Campus

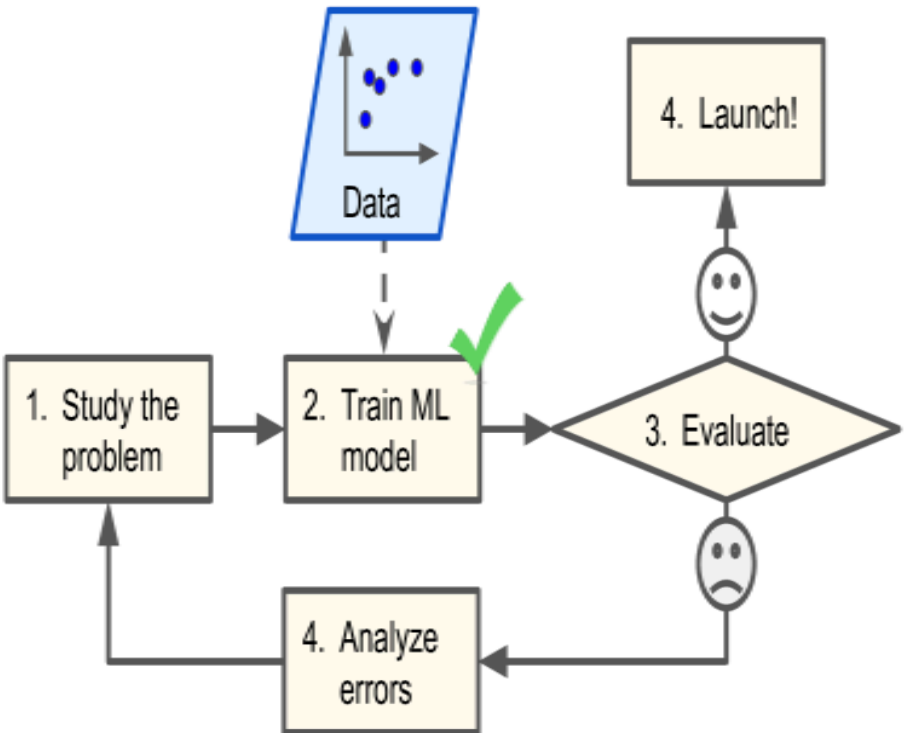
1. **Machine Learning:**

- Science and art of computer programming that can learn from **data**
- Or the process of solving practical problems by gathering **data** and algorithmically building **statistical models or mathematical models** based on that **dataset.**

2. **Types of Machine Learning:**
    a. Supervised Learning
    b. Unsupervised Learning
    c. Reinforcement Learning

3. **Why Use Machine Learning?(Applications):**
    - Spam Filter
    - Fraud Detection
    - Autonomous Vehicles etc.

4. **Machine Learning Flow:**



5. **Main Challenges of Machine Learning:**
    - Insufficient quantity of Data
    - Non-representative training Data
    - Poor-Quality Data
    - Irrelevant Features
    - Overfitting the Training Data
    - Underfitting the Training Data

Mostly the challenges are related to dealing with Data in one way or another

So before going into actual machine learning experience we will deal with the data and how to process it.

6. **Data Preprocessing:**

    Data preprocessing is the process of transforming raw data into an understandable format.

Before diving into the data preprocessing steps, let's get some insights on the data we want to work on.

6.a Data Visualization:

- Each row represents one district. There are 10 attributes. *longitude, latitude, housing_median_age,total_rooms, total_bedrooms, population, households, median_income, median_house_value, and ocean_proximity.*

```
[2] import pandas as pd
```

```
[5] csv_file='/content/drive/MyDrive/Colab Notebooks/housing.csv'
    df = pd.read_csv(csv_file)
```

```
df.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

- The info() method is useful to get a quick description of the data, in particular the total number of rows, each attribute's type, and the number of non-null values.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

- There are 20,640 instances in the dataset, which means that it is fairly small by Machine Learning standards, but it's perfect to get started.

- All attributes are numerical, except the ocean_proximity field.
- Its type is object, so it could hold any kind of Python object.
- But since loaded this data from a CSV file, you know that it must be a text attribute.
- When you looked at the top five rows, you probably noticed that the values in the ocean_proximity column were repetitive, which means that it is probably a categorical attribute.

```
df["ocean_proximity"].value_counts()
```

ocean_proximity	count
<1H OCEAN	9136
INLAND	6551
NEAR OCEAN	2658
NEAR BAY	2290
ISLAND	5

dtype: int64

- Let's look at the other fields. The describe() method shows a summary of the numerical attributes.

```
df.describe()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

The 25%, 50%, and 75% rows show the corresponding percentiles: a percentile indicates the value below which a given percentage of observations in a group of observations fall. **For example**, 25% of the districts have a housing_median_age lower than 18, while 50% are lower than 29 and 75% are lower than 37. These are often called the 25th percentile (or first quartile), the median, and the 75th percentile (or third quartile).

- **Since we have covered data cleaning and transformation in programming for AI Lab**
- **Your task is to do the following with the dataset provided**

Tasks

6.1 Data Cleaning

6.2 Data Transformation