# HOW TO GET DATASETS FOR MACHINE LEARNING

# What is Dataset?

A dataset is a collection of records usually presented in tabular form.

| Name | Age | City | Department |
|---|---|---|---|
| Sam | 25 | Jodhpur | Digital Marketing |
| Sharon | 27 | Jaipur | Developer |
| Jack | 32 | Pune | SEO |
| Mark | 42 | Mumbai | Content Writer |
| Diana | 50 | Hyderabad | Trainer |

# Types of Datasets

Numeric Data
(Quantitative)

Categorical Data
(Qualitative)

Ordinal Data

# NUMERIC DATA (QUANTITATIVE)

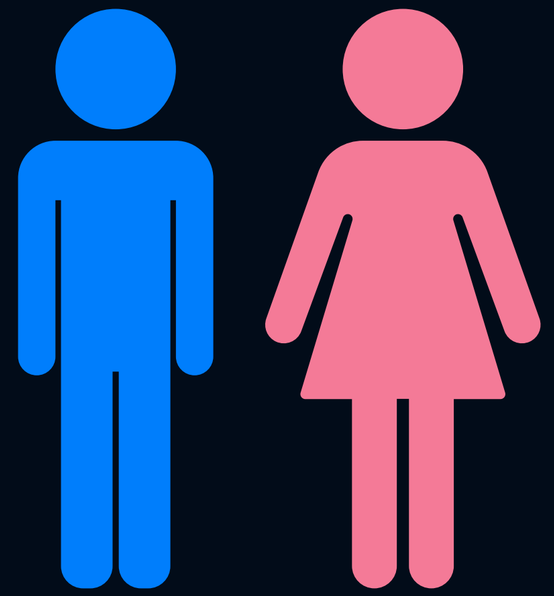
Numeric data, also known as quantitative data, is data that you typically present in number form, and it doesn't include any language or descriptive form.

It's always measurable, and we can add it together.

Example: Age, Blood Pressure, Temperature etc

# CATEGORICAL DATA (QUALITATIVE)

Categorical data is a type of qualitative data that is described using words instead of numbers. It can be grouped into categories, rather than being measured numerically.



Example: Gender etc.

# ORDINAL DATA

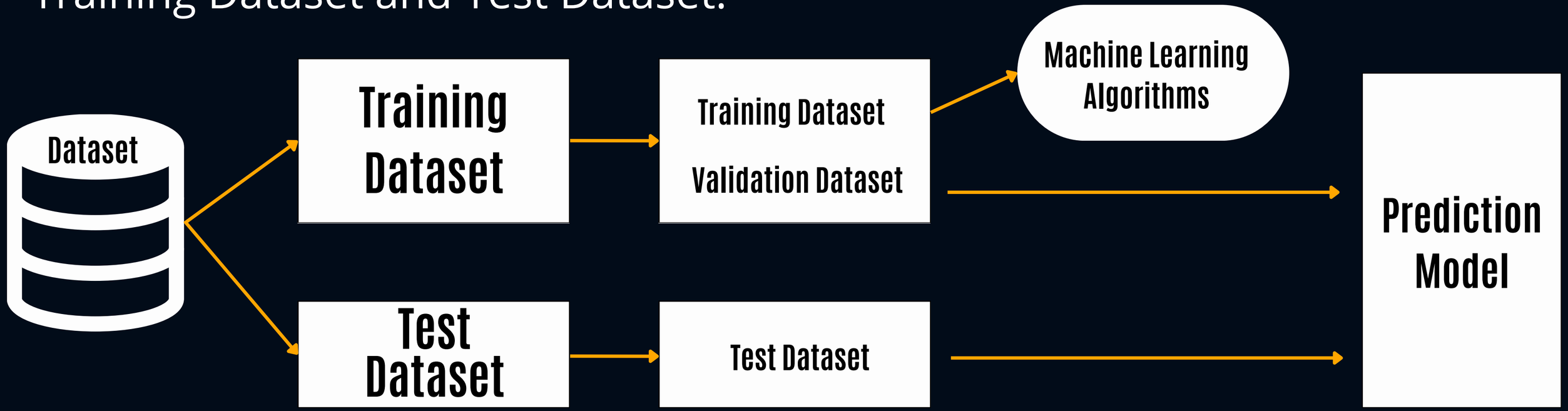
Ordinal data is a type of qualitative data that organizes variables into ordered categories.

The categories are ranked based on a hierarchical scale, like high to low.

Example: Food Rating.

# Need Of Dataset

In the development phase of ML projects, datasets are classified as Training Dataset and Test Dataset.



Home - Canva | Presentation By PROF. NILADRI | WhatsApp | (189) What is Dataset & Types | Dogs vs. Cats | Kaggle

kaggle.com/c/dogs-vs-cats/data

kaggle

Search

Sign In | Register

Create

Home

Competitions

Datasets

Models

Code

KAGGLE · PLAYGROUND PREDICTION COMPETITION · 10 YEARS AGO

Late Submission

# Dogs vs. Cats

Create an algorithm to distinguish dogs from cats

Overview | Data | Code | Models | Discussion | Leaderboard | Rules

www.kaggle.com

nd cats. Train your algorithm on these files and predict the labels for test1.zip (1 =

anually label your submissions. We work hard to fair and fun contests, and ask for

Learn at your own pace

Learn from the best prof

Meet expert teacher

Share knowledge

**Files**
3 files

**Size**
853.96 MB

**Type**
zip, csv

**License**
Subject to Competition Rules

Kaggle uses cookies from Google to deliver and enhance the quality of its services and to analyze traffic. Learn more. Ok, Got it.

Type here to search | Dogs vs. Cats | Kag... | 19°C Haze | ENG | 19:43 16-01-2024

Home - Canva | Presentation By PROF. NILADRI | WhatsApp | (189) What is Dataset & Types | Dataset Search

datasetsearch.research.google.com/search?src=0&query=cat%20vs%20dog&docid=L2cvMTF0YzUyd3Rfeg%3D%3D

Google

cat vs dog

Last updated | Download format | Usage rights | Topic | Provider | Free | Saved data sets

100+ data sets found

H

Cat_and_Dog

huggingface.co

kaggle.com

Updated Oct 2, 2023

+3more

Updated Sep 21, 2019

Cat_and_Dog

Cat and Dog

Bingsu/Cat_and_Dog

Explore at: huggingface.co | Kaggle | kaggle.com

5 scholarly articles cite this dataset (View in Google Scholar)

s/cc0-1.0/

on Hugging Face and contributed by the HF Datasets community

# Google Dataset

Learn at your own pace

Learn from the best prof

Meet expert teacher

Share knowledge

Type here to search

Dataset Search - G...

19°C Haze

ENG

19:50

16-01-2024

# Government Dataset

**INDIA**

https://data.gov.in

**USA**

https://data.gov/

**EUROPEAN UNION**

https://www.oppendatani.gov.uk/

Northern Ireland

https://www.data.europa.eu/euodp/data/dataset

# What is Data Pre-processing?

Data Pre-processing is a process of converting the raw data in suitable form.

The steps of Data processing are following:

1. Getting Dataset
2. Importing Libraries
3. Importing Datasets
4. Finding missing values
5. Encoding Categorical Data
6. Splitting Dataset into Training Dataset and Test Dataset
7. Feature Scaling

# Features and Labels in Dataset

In ML, Feature means property of Training Data.

In ML, Label means the output we get from our model after training.

| 1 | Person | Height(in Feet) | Weight (in kg) | Foot Size(in inches) |
|---|---|---|---|---|
| 2 | Male | 6 | 81 | 12 |
| 3 | Male | 5.92 | 86 | 11 |
| 4 | Male | 5.58 | 90 | 12 |
| 5 | Female | 5.92 | 77 | 10 |
| 6 | Female | 5 | 45 | 6 |
| 7 | Female | 5.5 | 68 | 8 |
| 8 | Female | 5.52 | 58 | 7 |
| 9 | Female | 5.75 | 68 | 9 |

# Features and Labels in Dataset

Case 1: The prediction (Y) Calories here is a label.

Calories is the column that we want to predict using various features like
X1: Gender
X2: Height
X3: Weight

| 1 | Person | Age | Height(in Feet) | Weight (in kg) | Duration | Heart_Rate | Body_Temp | Calories |
|---|---|---|---|---|---|---|---|---|
| 2 | Male | 68 | 6 | 94 | 29 | 105 | 40.8 | 231 |
| 3 | Male | 70 | 5.92 | 79 | 5 | 88 | 38.7 | 26 |
| 4 | Female | 20 | 5.44 | 60 | 14 | 94 | 40.3 | 66 |

# Features and Labels in Dataset

Case 2: The prediction (Y) Heart_Rate here is a label.

Heart_Rate is the column that we want to predict using various features like
X1: Gender
X2: Weight

| 1 | Person | Age | Height(in Feet) | Weight (in kg) | Duration | Heart_Rate | Body_Temp | Calories |
|---|---|---|---|---|---|---|---|---|
| 2 | Male | 68 | 6 | 94 | 29 | 105 | 40.8 | 231 |
| 3 | Male | 70 | 5.92 | 79 | 5 | 88 | 38.7 | 26 |
| 4 | Female | 20 | 5.44 | 60 | 14 | 94 | 40.3 | 66 |

# Identify Features and Lables

# Regression

Regression is a statistical method that helps us to understand and predict the relationship between variables.

Describe how one variable (dependent variable) changes as another variable (independent variable) changes.

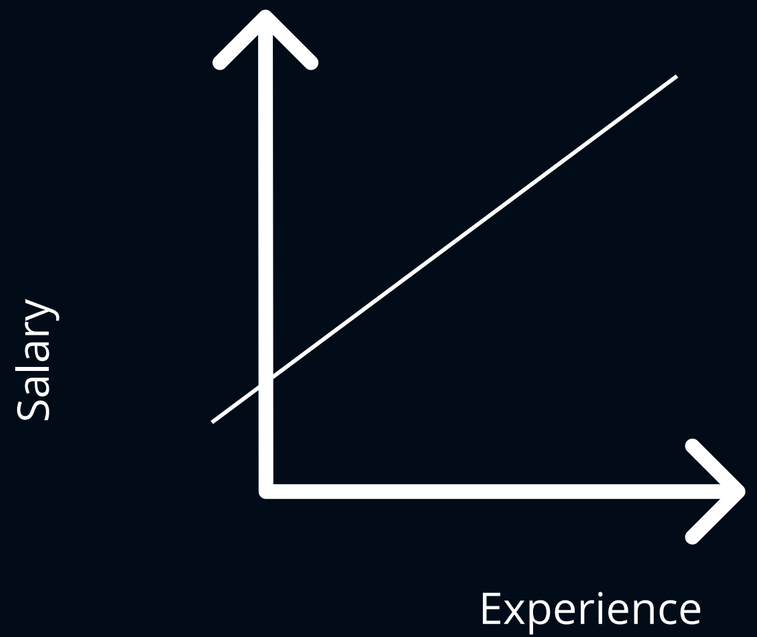
Dependent variable: Trying to predict or explain (Y)

Independent variable: That are used to predict or explain the changes in the dependent variable.

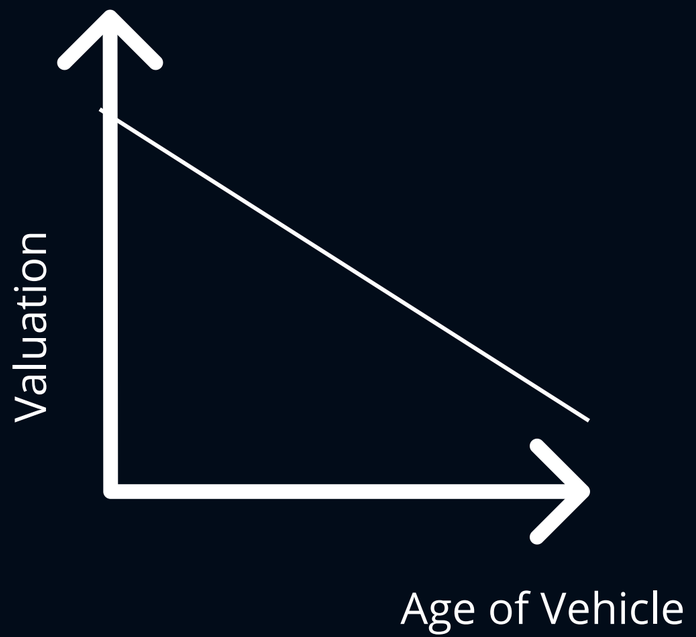
Example: Predict the salary on basis of experience.

# Regression-Positive/ Negetive

Positive Regression




Negative Regression



# Types of Regression

Linear Regression

Multi Linear Regression

Polynomial Regression

# Linear Regression

Equation of Linear Regression : $Y= mX + b$

Here, $Y$ represents dependent variable.
Here $X$ represents independent variable.
Here $m$ represents the slope of the line ( How much $Y$ changes for a unit change in $X$)
Here $b$ represents the intercept ( the value of $Y$ when $X=0$)

Problem: Predict the prize of Pizza

Phase 1: Data Collection
Phase 2: Calculations
Phase 3: Prediction
Phase 4: Visualization

# Linear Regression

| Diameter (X) | Price (Y) | Mean (X) | Mean (Y) | Deviations (X) | Deviations (Y) | Product of Deviations | Sum of Product of Deviations | Square of Deviations for X |
|---|---|---|---|---|---|---|---|---|
| 8 | 10 | | | | | | | |
| 10 | 13 | | | | | | | |
| 12 | 16 | | | | | | | |

m= (sum of product of deviations)/(sum of square of deviations for X)

b= Mean of Y - (m * Mean of X)