Data Sets in Machine Learning

Structured vs Unstructured Data

- **Unstructured data** is information that has not been structured in a predefined manner. Social media conversations, images, video, and audio etc.

- **Structured data**, on the other hand, is organized in a pre-defined structured format, such as Excel and Google Sheets, where data is added to standardized columns and rows relating to pre-set parameters. The framework of structured data models is designed for easy data entry, search, comparison, and extraction.

- **Semi-structured data**, which is also text-heavy data but loosely organized into categories or “meta tags.” This information can be easily broken into its individual groups, but the data within these groups is itself unstructured. For example: e-mail.

Data set

Area of House (in Sq Yards)	No. of Bed Room	Location	Price (in Lakhs)
200	3	Chandigarh	62.0
300	4	Chandigarh	75.0
400	5	Delhi	120.50
300	4	Delhi	95.25
200	3	Patiala	45.50

Total Features =4

Total Instances =5

Labeled Dataset

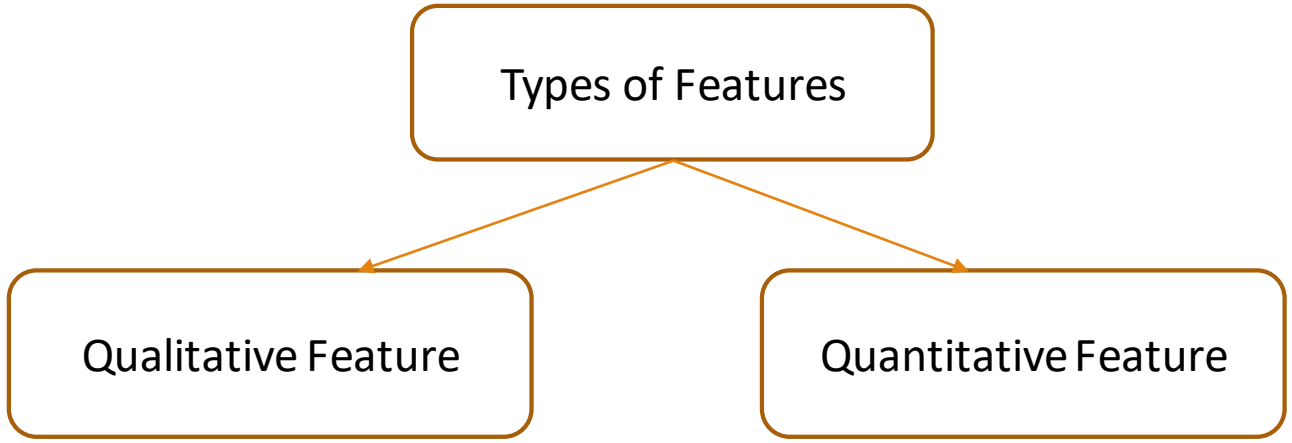
Area of House	No of Bed Rooms	Location	Price (in Lakhs)
200	3	Chandigarh	65.0
300	4	Chandigarh	75.0
400	5	Delhi	120.50
300	4	Delhi	95.25
200	3	Patiala	45.50

Relationship between input features Xi and output feature y

$y=f(X)$

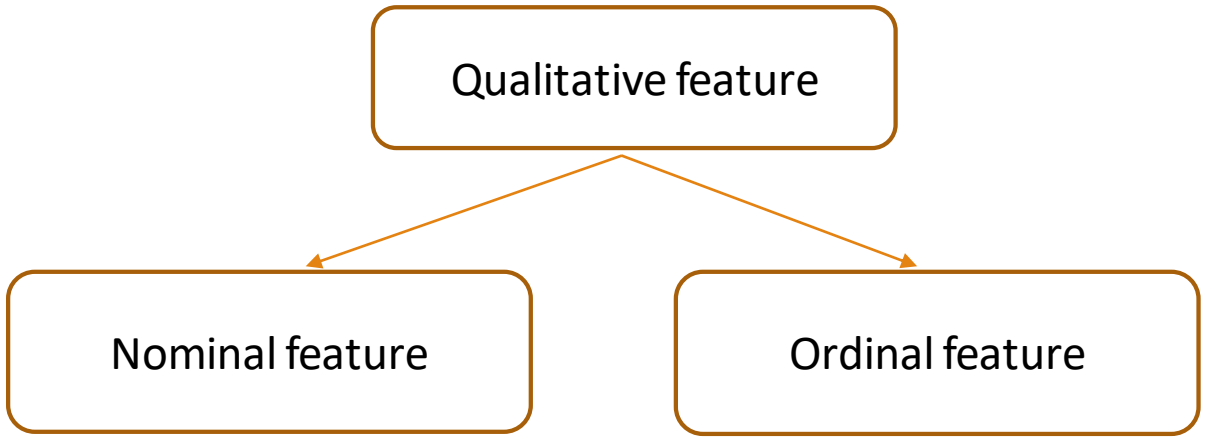
Unlabeled Dataset

Emp Id	Experience (in Years)	Age	Address
101	12	34	Patiala
102	24	47	Delhi
103	16	42	Delhi
104	14	36	Chandigarh
105	16	39	Chandigarh
106	21	42	Chandigarh



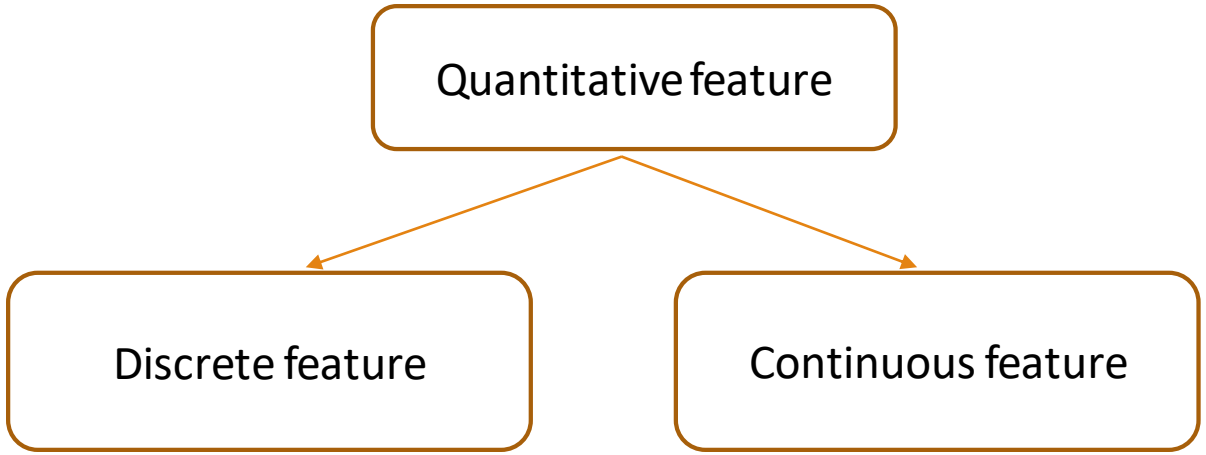
Qualitative or categorical features have non numerical values. E.g. Gender(male or female).
They can’t be measured, but can be grouped.

Quantitative features are numerical in nature. E.g. Salary of a person, Marks obtained by a student.
They are measurable and have some order.



Nominal feature has no numerical values as well as ordering to its categories. For example, gender is a nominal.

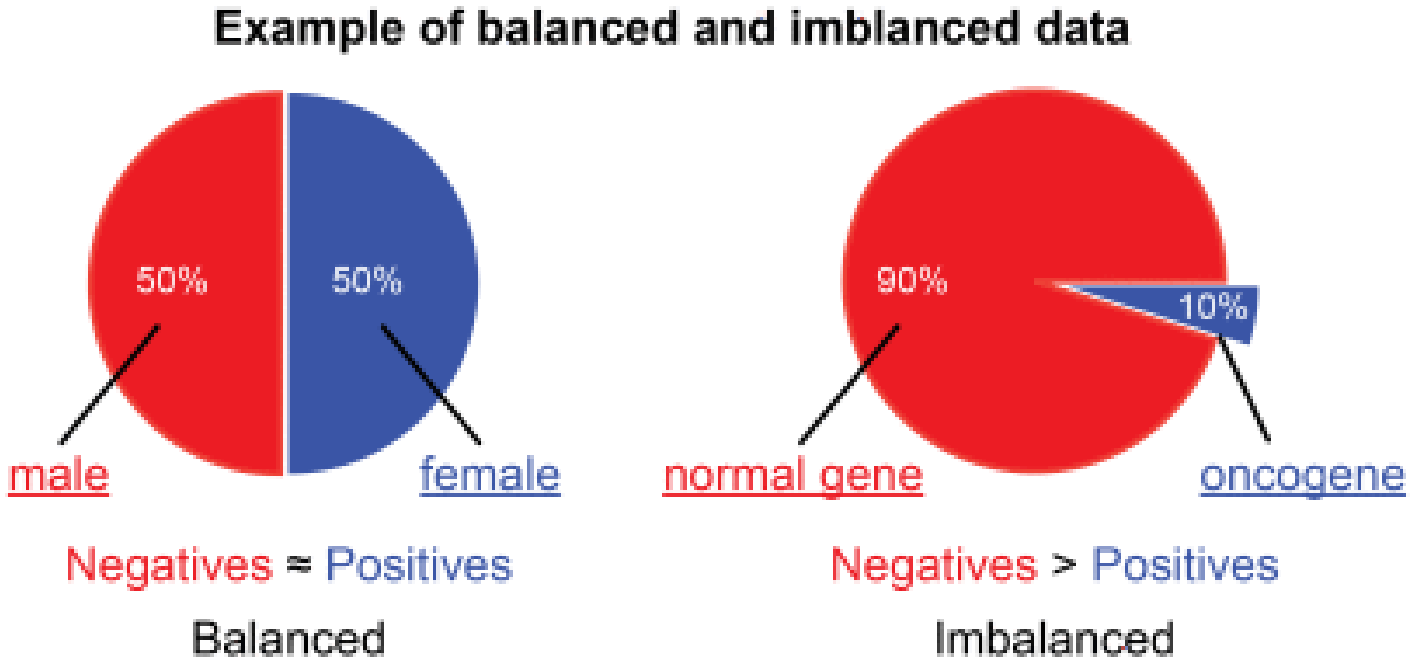
Ordinal feature has a clear ordering. Education level is an ordinal variable as it may be have values- Elementary, Secondary, College/University.



Discrete Feature: It can take finite number of values only. Number of daily admitted patients to a hospital .

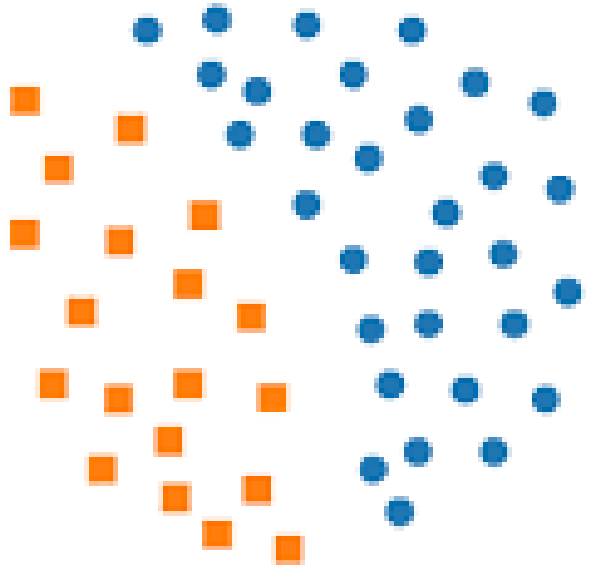
Continuous Feature: It can be measured on a continuum or a scale. For example- Price of a house, weight of a person, blood pressure of a patient *etc*.

Balanced and Imbalanced Dataset



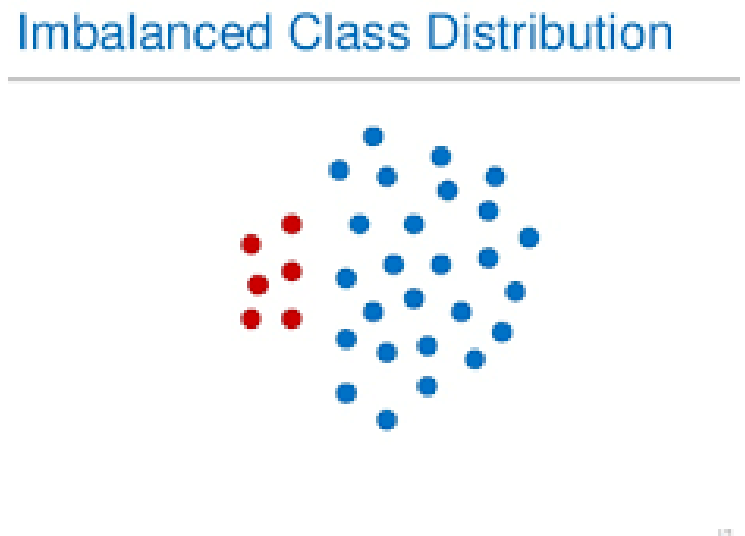
Balanced Dataset

Balanced Dataset: —If in our data set we have positive values which are approximately same as negative values. Then we can say our dataset in balance.



ImBalanced Dataset

ImBalanced Dataset: — If there is the very high different between the positive values and negative values. Then we can say our dataset in Imbalance Dataset.




Thank
You