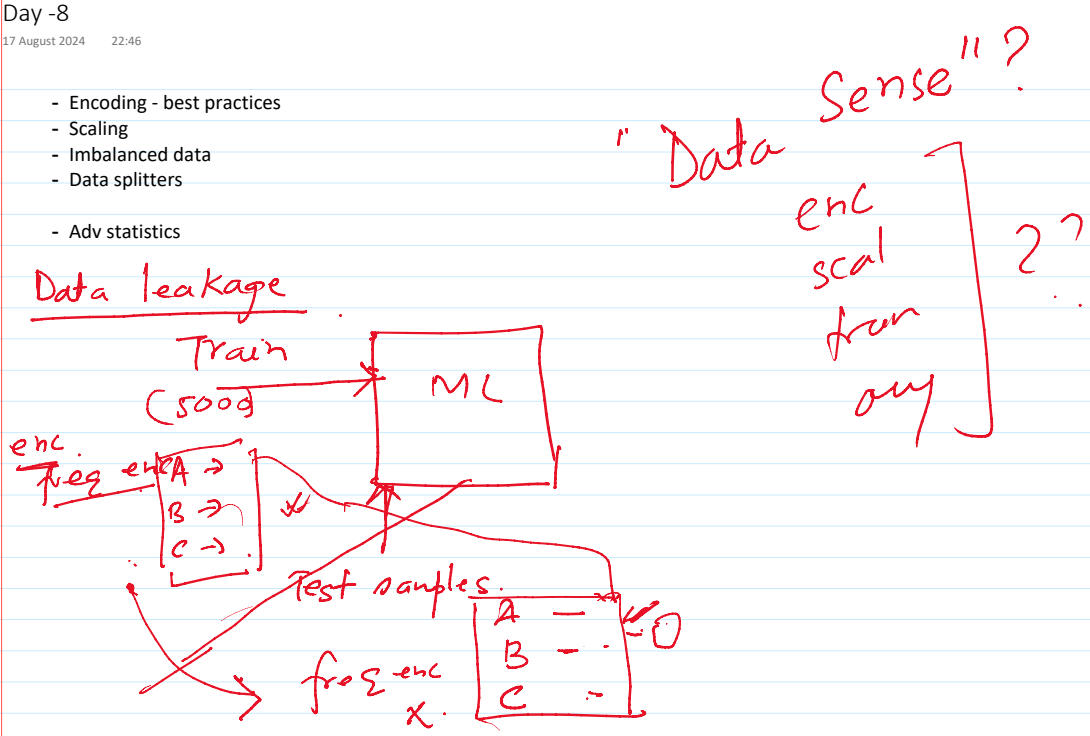
Day -8

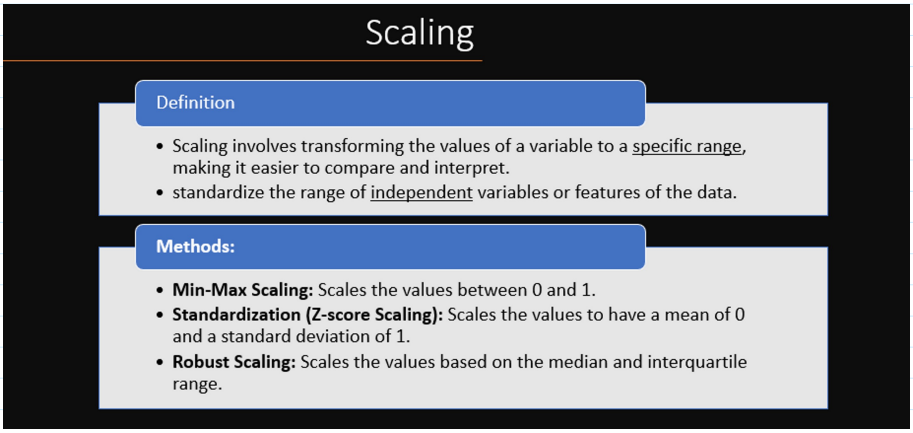
17 August 2024 22:46

- Encoding - best practices
- Scaling
- Imbalanced data
- Data splitters

- Adv statistics



Next topic : Scaling



Min-Max Scaling

- Min-Max Scaling, also known as Min-Max Normalization, is a technique used to transform the values of a numerical variable to a specific range, typically between 0 and 1.

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- X is the original value of the variable.
- $X_{\min}$ is the minimum value of the variable in the dataset.
- $X_{\max}$ is the maximum value of the variable in the dataset.

Robust scaler

Formula

$$X_{\text{robust-scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)}$$

where:

- X is the original value of the variable.
- median(X) is the median of the variable.
- QR(X) is the interquartile range of the variable.

Next topic : Imbalanced datasets

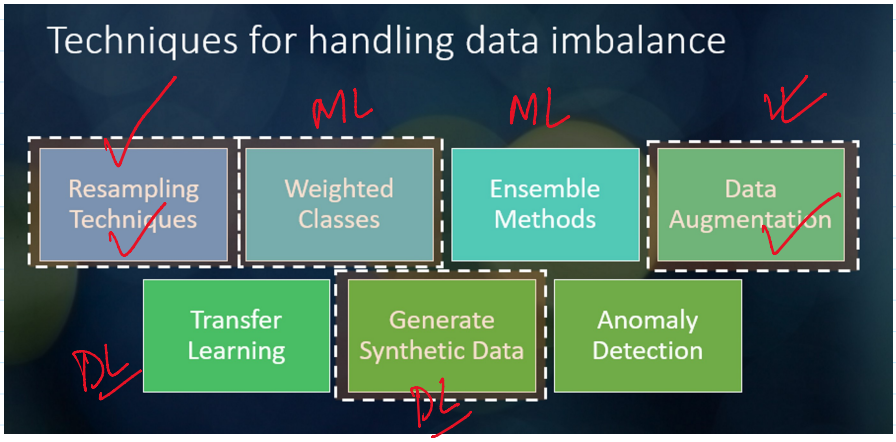
Define data imbalance

1

Data imbalance occurs when the distribution of classes in a classification dataset is uneven, with one or more classes having significantly fewer samples than others.

2

Handling data imbalance is important because it can lead to biased models that perform well on the majority class but poorly on the minority class.



SMOTE ??

6 mins! (self-walkthru).

why image aug??

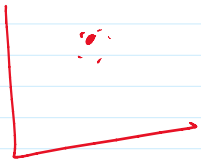
Step 1 - Identify Minority Class Instances

This step involves determining which class in the dataset has fewer instances.

In the context of SMOTE, this is typically the class you want to oversample, referred to as the minority class.

Step 2 and 3

Select a Minority Instance:	A single instance is randomly chosen from the minority class.
Find k Nearest Neighbors:	The k-nearest neighbors of the selected instance within the minority class are identified. The choice of k is a parameter set by the user, defining the number of neighbors to consider.



Step 4

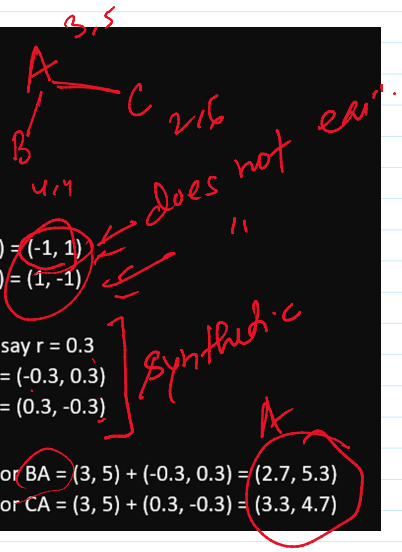
Generate Synthetic Instances:

- For each neighbor, a synthetic instance is created by calculating the vector between the selected instance and the neighbor.
- This vector is scaled by a random value between 0 and 1.
- The scaled vector is added to the selected instance to generate a new synthetic instance.

Illustrating step 4

- Suppose we have a minority class instance A with two features:
  - Instance A: (x_A, y_A) = (3, 5)
- Let's say we choose k=2, and the two nearest neighbors of instance A in the minority class are B and C:
  - Neighbor B: (x_B, y_B) = (4, 4)
  - Neighbor C: (x_C, y_C) = (2, 6)
- Now, we'll generate synthetic instances for instance A based on its neighbors.

Generate synthetic data



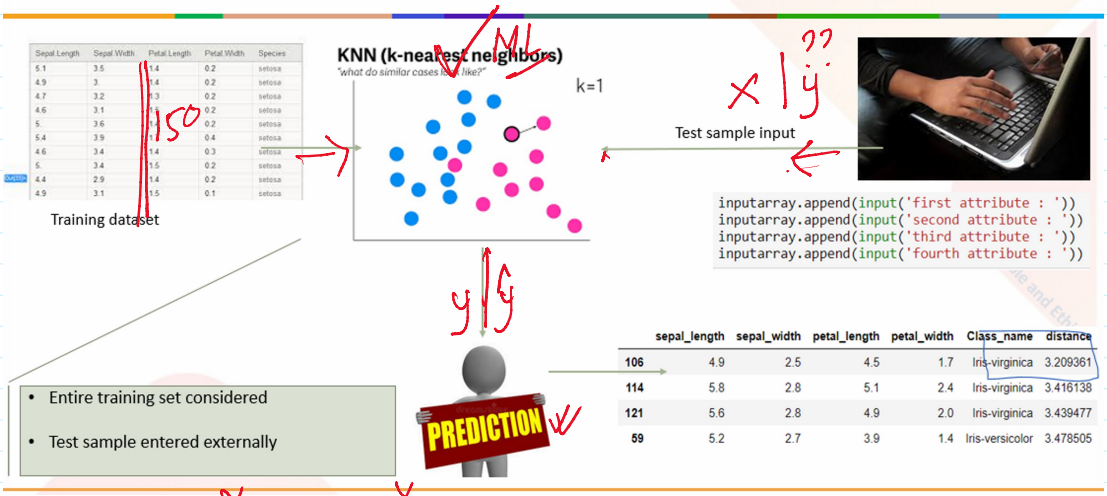
1. **Calculate Vectors:**
   1. Vector BA = (x_A - x_B, y_A - y_B) = (3 - 4, 5 - 4) = (-1, 1)
   2. Vector CA = (x_A - x_C, y_A - y_C) = (3 - 2, 5 - 6) = (1, -1)
2. **Scale Vectors:**
   1. Choose a random value between 0 and 1, let's say r = 0.3
   2. Scaled Vector BA = r * Vector BA = 0.3 * (-1, 1) = (-0.3, 0.3)
   3. Scaled Vector CA = r * Vector CA = 0.3 * (1, -1) = (0.3, -0.3)
3. **Generate Synthetic Instances:**
   1. Synthetic Instance 1 = Instance A + Scaled Vector BA = (3, 5) + (-0.3, 0.3) = (2.7, 5.3)
   2. Synthetic Instance 2 = Instance A + Scaled Vector CA = (3, 5) + (0.3, -0.3) = (3.3, 4.7)

Data splitters

Various types of datasets

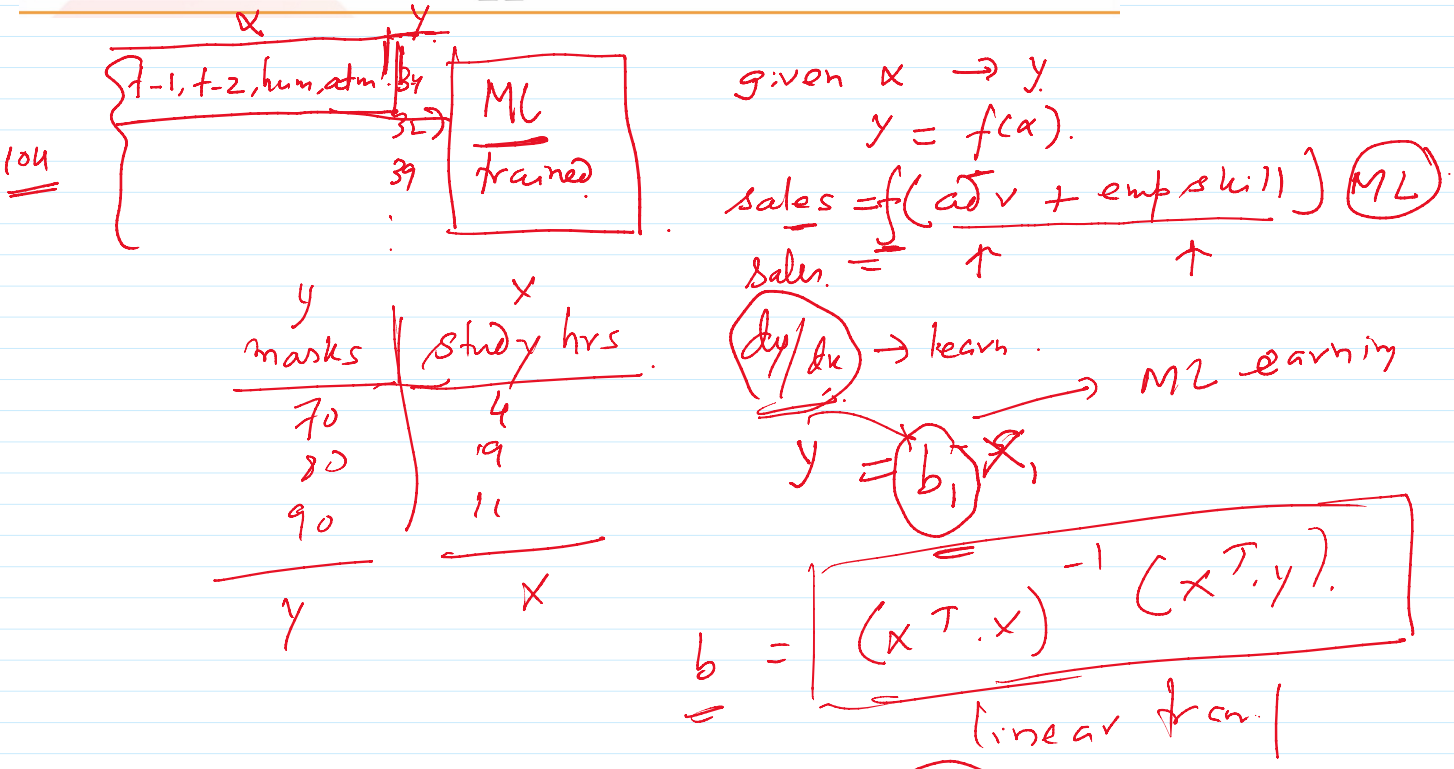
1. Training Set:	2. Validation Set:	3. Testing Set (or Test Set):
• **Purpose:** Used to train the machine learning model. • **Size:** Typically the largest subset (60-80% of the data). • **Usage:** The model learns patterns, relationships, and features from this set.	• **Purpose:** Used for hyperparameter tuning and model selection during training. • **Size:** Smaller than the training set (usually 10-20% of the data). • **Usage:** Helps prevent overfitting by fine-tuning model parameters without contaminating the test set.	• **Purpose:** Reserved for evaluating the model's performance on unseen data. • **Size:** Independent subset not used during training or validation (10-20% of the data). • **Usage:** Provides an unbiased assessment of how well the model generalizes to new, unseen instances.

EVALUATION – 1 (RANDOM INPUT TEST SAMPLE)

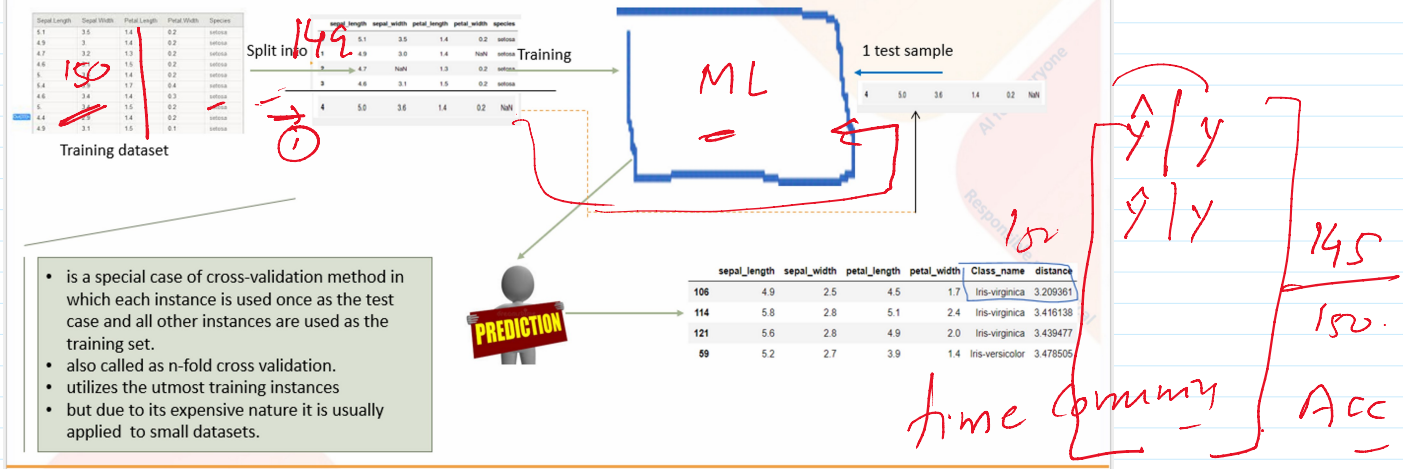


	sepal_length	sepal_width	petal_length	petal_width	Class_name	distance
106	4.9	2.5	4.5	1.7	Iris-virginica	3.209361
114	5.8	2.8	5.1	2.4	Iris-virginica	3.416138
121	5.6	2.8	4.9	2.0	Iris-virginica	3.439477
59	5.2	2.7	3.9	1.4	Iris-versicolor	3.478505

- Entire training set considered
- Test sample entered externally



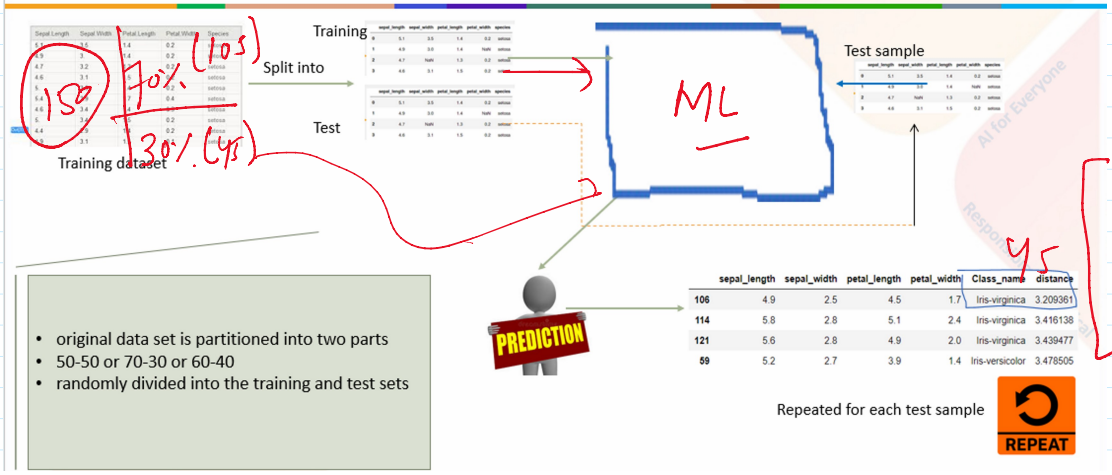
EVALUATION – 2 (LEAVE-ONE-OUT CROSS-VALIDATION - LOOCV)



	sepal_length	sepal_width	petal_length	petal_width	Class_name	distance
106	4.9	2.5	4.5	1.7	Iris-virginica	3.209361
114	5.8	2.8	5.1	2.4	Iris-virginica	3.416138
121	5.6	2.8	4.9	2.0	Iris-virginica	3.439477
59	5.2	2.7	3.9	1.4	Iris-versicolor	3.478505

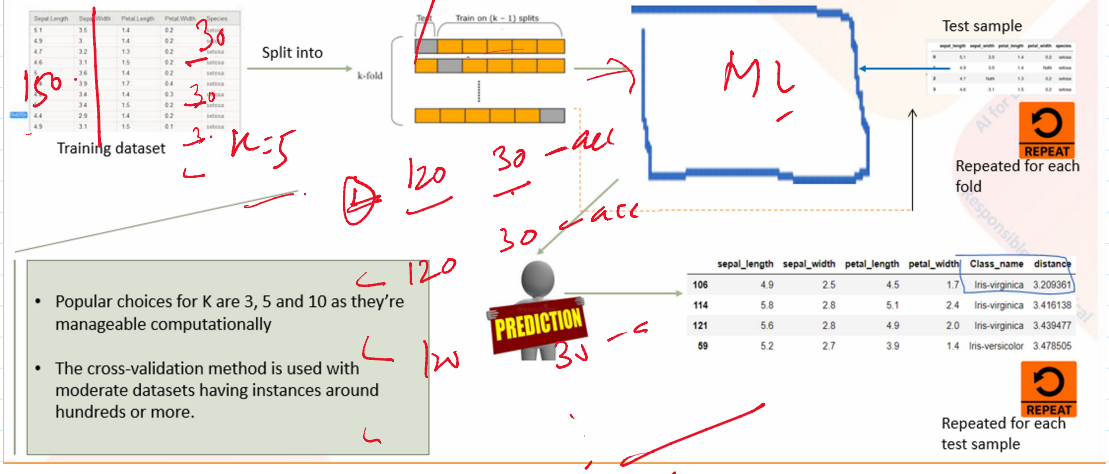
- is a special case of cross-validation method in which each instance is used once as the test case and all other instances are used as the training set.
- also called as n-fold cross validation.
- utilizes the utmost training instances
- but due to its expensive nature it is usually applied to small datasets.

EVALUATION – 3 (HOLDOUT METHOD)



- original data set is partitioned into two parts
- 50-50 or 70-30 or 60-40
- randomly divided into the training and test sets

EVALUATION – 4 (K-FOLD CROSS-VALIDATION METHOD)



- Popular choices for K are 3, 5 and 10 as they're manageable computationally
- The cross-validation method is used with moderate datasets having instances around hundreds or more.

LOOCV → very sparingly; cluster of machines.
holdout method → very often; dev cycle.
K-fold → at end of major milestone.
↳ (variations).

Next topic : Inferential statistics

Define interferential statistics

is a branch of statistics that involves making inferences or predictions about a population based on a sample of data.

allows business leaders and managers to draw conclusions and make decisions about a larger group based on a smaller subset of data.

Why learn inferential statistics?

Generalization from Samples:

Example: A pharmaceutical company tests a new drug on a sample of 1,000 patients to determine its effectiveness. Using inferential statistics, they can generalize the findings to the entire population of potential patients, estimating how effective the drug would be if given to millions of people.

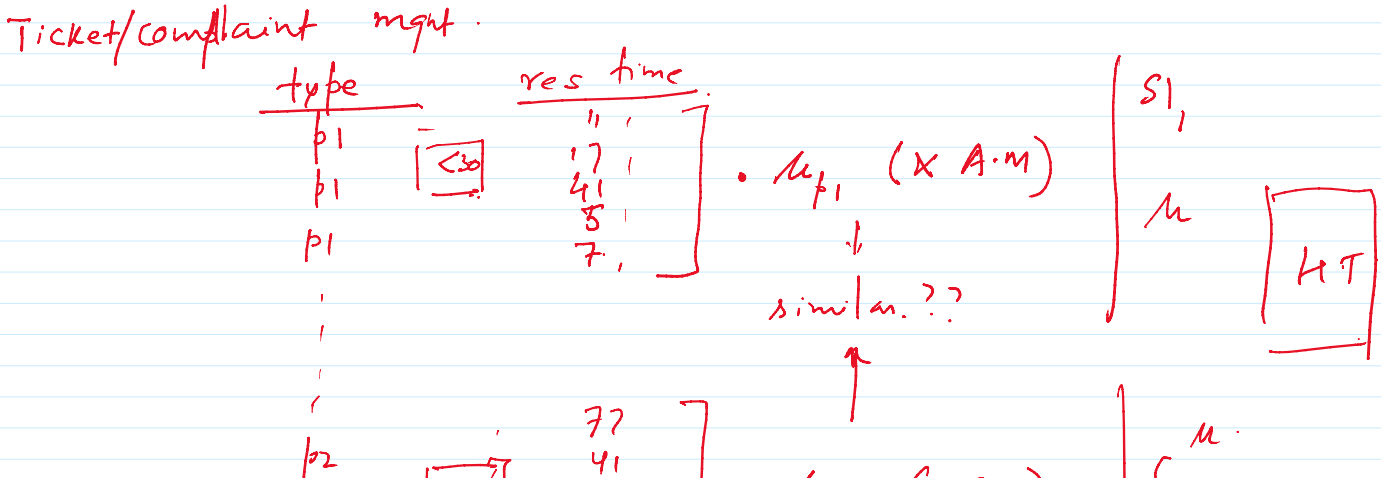
•**Understanding Uncertainty**:

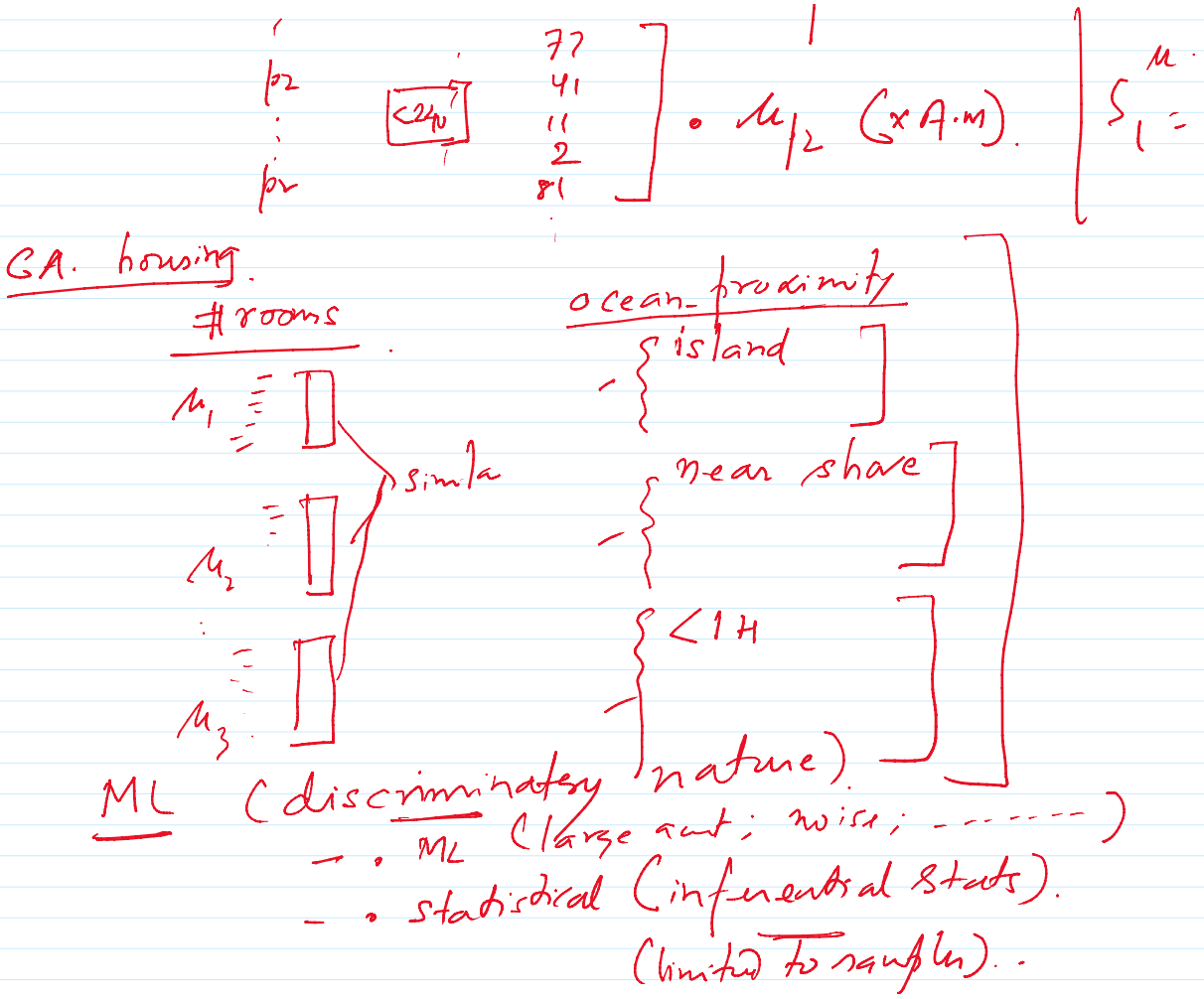
Example: A polling organization surveys 1,200 voters to predict the outcome of an election. They use inferential statistics to calculate a confidence interval, saying that Candidate A is likely to receive 48% to 52% of the vote with 95% confidence. This helps them understand the potential range of the true voter percentage.

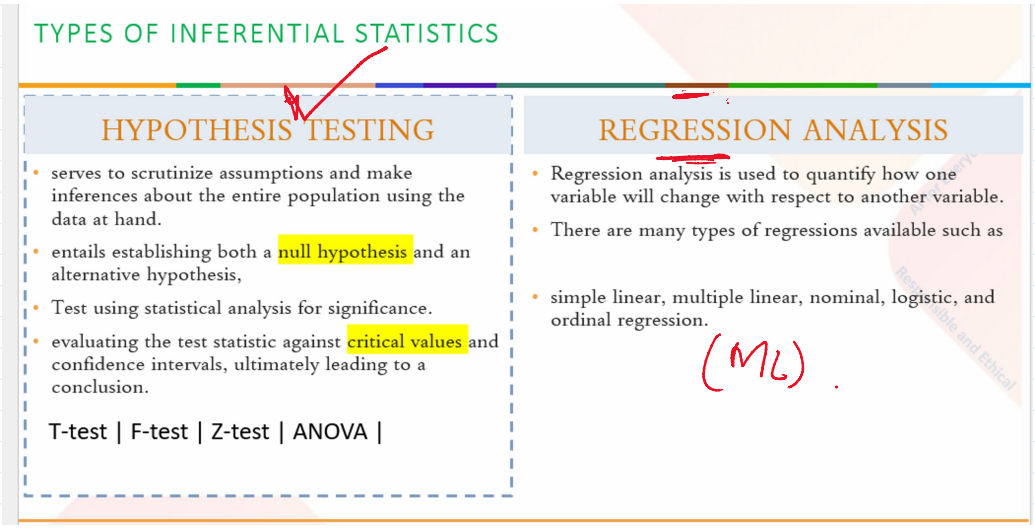
Hypothesis testing

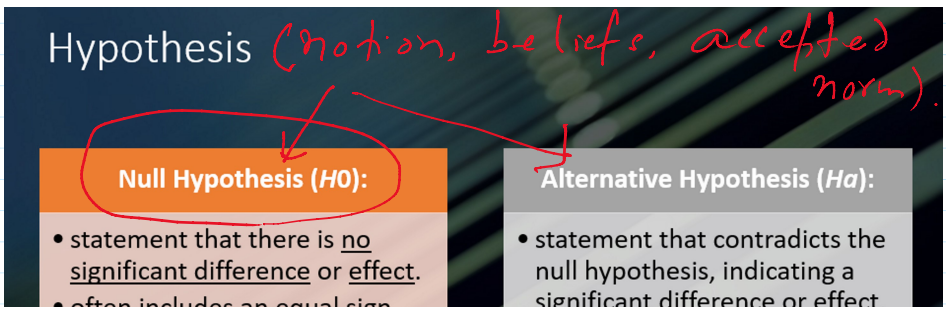
is a statistical method used to make inferences about a population based on a sample of data.

involves formulating a hypothesis about a population parameter, collecting data, and using statistical techniques to determine whether the data provide enough evidence to reject or not reject the null hypothesis.









$\mu_{p_1} = 20$ min.
null hyp: H_0.

Hypothesis (notion, beliefs, accepted norm).

Null Hypothesis (*H0*):

- statement that there is no significant difference or effect.
- often includes an equal sign (=).
- Example: *H0*: μ=50 (population mean is equal to 50).

Alternative Hypothesis (*Ha*):

- statement that contradicts the null hypothesis, indicating a significant difference or effect.
- Example: *H1*: μ *not equal* 50 (population mean is not equal to 50).

$\mu_{p_1} = 20$ min.

null hyp: H_0

$\mu_{p_1} \neq 20$

> 20

< 20 $\Big\}$ H_a

Examples – Market research

- Suppose we have a sample of 100 consumers who participated in the market research study.
- categorize their age groups into three categories: Young (18-30), Middle-aged (31-50), and Older (51 and above).
- For screen size preference, we'll consider two categories: Small and Large.
- **TASK** : the relationship between screen size preference and age group

Consumer ID	Age Group	Screen Size Preference
1	Young	Large
2	Middle-aged	Small
3	Older	Large
4	Young	Large
5	Middle-aged	Small
6	Middle-aged	Large
7	Older	Small
8	Young	Large
9	Young	Large
10	Older	Small
...	...	...
100	Middle-aged	Large

Example - Employee Productivity

- A business leader wants to determine if there is a significant difference in productivity between two teams within the organization.
- use a hypothesis test to compare the average productivity scores of Team A and Team B based on specific metrics (e.g., sales volume, project completion time).
- test would help the leader infer whether the observed difference in productivity between the two teams is statistically significant.

```
Team A Sales Volume: [100, 120, 110, 90,  105, 115,  95, 105, 115, 100]

Team B Sales Volume: [110, 105, 115, 100, 125, 115, 120, 110, 115, 105]
```

```
The customer satisfaction scores before training are represented by the list
[80, 75, 85, 70, 75, 78, 82, 79, 80, 77].

The customer satisfaction scores after training are represented by the list
[85, 82, 88, 75, 80, 84, 87, 86, 85, 82].
```

Example - Training Effectiveness

- A business leader invests in a training program for customer service representatives
- To assess the effectiveness of the training program, the leader could conduct a hypothesis test by comparing the customer satisfaction scores before and after the training.
- test would help infer whether there is a significant improvement in customer satisfaction as a result of the training program

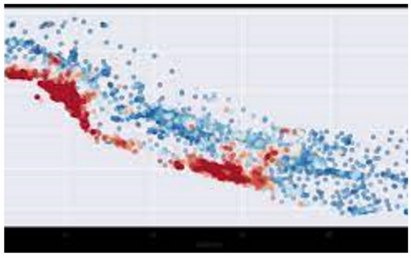
Example: Medical Treatment

- **Null Hypothesis (*H*0):** The new drug has no effect on patient recovery.
- **Alternative Hypothesis (*H*1):** The new drug improves patient recovery.
- **Interpretation:** Evaluating whether a medical intervention has a significant impact on patient outcomes.

Example: Educational Intervention

- **Null Hypothesis (*H*0):** The teaching method has no impact on student performance.
- **Alternative Hypothesis (*H*1):** The teaching method improves student performance.
- **Interpretation:** Investigating the effectiveness of a particular teaching approach on student learning.

Example



- **MedInc (Median Income)**: Median income of households within a district.
- **HouseAge (Housing Age)**: Median age of houses within a district.
- **AveRooms (Average Rooms)**: Average number of rooms per household within a district.
- **AveBedrms (Average Bedrooms)**: Average number of bedrooms per household within a district.
- **Population**: Total population of the district.
- **AveOccup (Average Occupancy)**: Average household occupancy within a district.
- **Latitude**: Latitude coordinate of the district's location.
- **Longitude**: Longitude coordinate of the district's location.
- **MedHouseVal (Median House Value)**: target/y Median house value for households within a district.

Example – test

01

Hypothesis Test -

Age of Houses and House Value:

02

Null Hypothesis (H0):

There is no significant correlation between the age of houses and median house values in California districts.

03

Alternative Hypothesis (H1): Older houses have significantly lower median values compared to newer houses.

Example

- Credit Scoring in Finance:
  - **Hypothesis**: A new credit scoring model is more accurate in predicting loan default.
  - **Null Hypothesis** (H0): The new credit scoring model is not more accurate.
  - **Alternative Hypothesis** (H1): The new credit scoring model is more accurate.
  - **Test**: Collect data on loans, apply both the old and new credit scoring models, and compare their accuracy in predicting loan defaults.

Process

Objective:

Evaluate the importance of each feature independently in relation to the target variable.

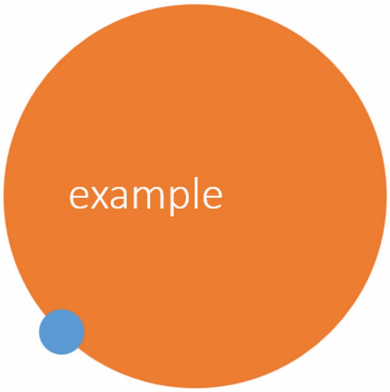
Statistical Measures:

commonly involves statistical tests or scoring methods to quantify the significance or relevance of each feature. Examples include t-tests, ANOVA F-tests, mutual information, and chi-square tests

Limitations

Does Not Consider Interactions or Relationships Between Features:

- **Explanation:** Univariate methods evaluate features in isolation, neglecting potential synergies or dependencies between features that may contribute jointly to predictive power.
- **Mitigation:** Consider employing multivariate feature selection methods that capture relationships between features.



- Let's say we want to test whether the average height of a certain population is different from 65 inches.

hypothesis

Null Hypothesis ($H0$)	Alternative Hypothesis ($H1$):
• μ=65 (population mean height is equal to 65 inches).	• μ not = 65 (population mean height is not equal to 65 inches).

$\neq 65$ (1)
> 65 (2)
< 65 (3)

Null vs alternate

Null Hypothesis:

Think of the null hypothesis as a statement that nothing special is happening. It's like saying "there's no change" or "there's no effect." It's a way of being cautious and assuming that any observed differences are just due to random chance.

Alternative Hypothesis:

The alternative hypothesis is like the opposite of the null hypothesis. It's where you suggest that something interesting or different is happening. It's like saying "there is a change" or "there is an effect."

gender — pol. party

M, F ⟷ A, B, C, A

H_0 = No relation

H_a = gender impact pol.

↳ M → A, F → B

Simply stating

In layman words

Null Hypothesis (H0): "Nothing unusual is going on."

Alternative Hypothesis (H1): "Something interesting or different is happening."

if a new medicine helps people sleep better

Null Hypothesis (H0): "Taking this medicine doesn't make any difference in sleep quality."

Alternative Hypothesis (H1): "Taking this medicine does make a difference in sleep quality."

H_0: $\mu = 5.6$ (Given)

DS: $\mu > 5.6$

Null Hypothesis (H0) - examples

E-commerce Conversion Rates:: "The recent website design change did not lead to a significant change in the conversion rate."

Ad Click-Through Rates: "The new ad campaign did not result in a higher click-through rate compared to the old campaign."

Customer Churn Analysis: "The introduction of a new loyalty program did not result in a significant reduction in customer churn."

Product Recommendation Engine: "The personalized product recommendations do not lead to a higher average purchase value compared to non-personalized recommendations."

1. Significance level
2. Test statistic
3. p-value
4. Critical value

] return related.

(4) Critical value

#	Topic	Number of Use cases/ Examples discussed	Explained Y/N, comments if any
1	Scaling	Min max scaler Std scaler Robust scaler Partial scaling Best practices	
2	Data imb	-SMOTE (std data) -Image aug (PIL package)	
3	Data splitters	-LOOCV, Leave-Pout -Hold out -K-fold	
4.	Inferential stats	-What they are, -Examples HT	