Variables and Data Types for Modeling and Analysis

Importance of Data in Modeling and Analysis

- Data is the foundation for predictive modeling, decision-making, and AI-driven insights

What Are Variables in Data Science?

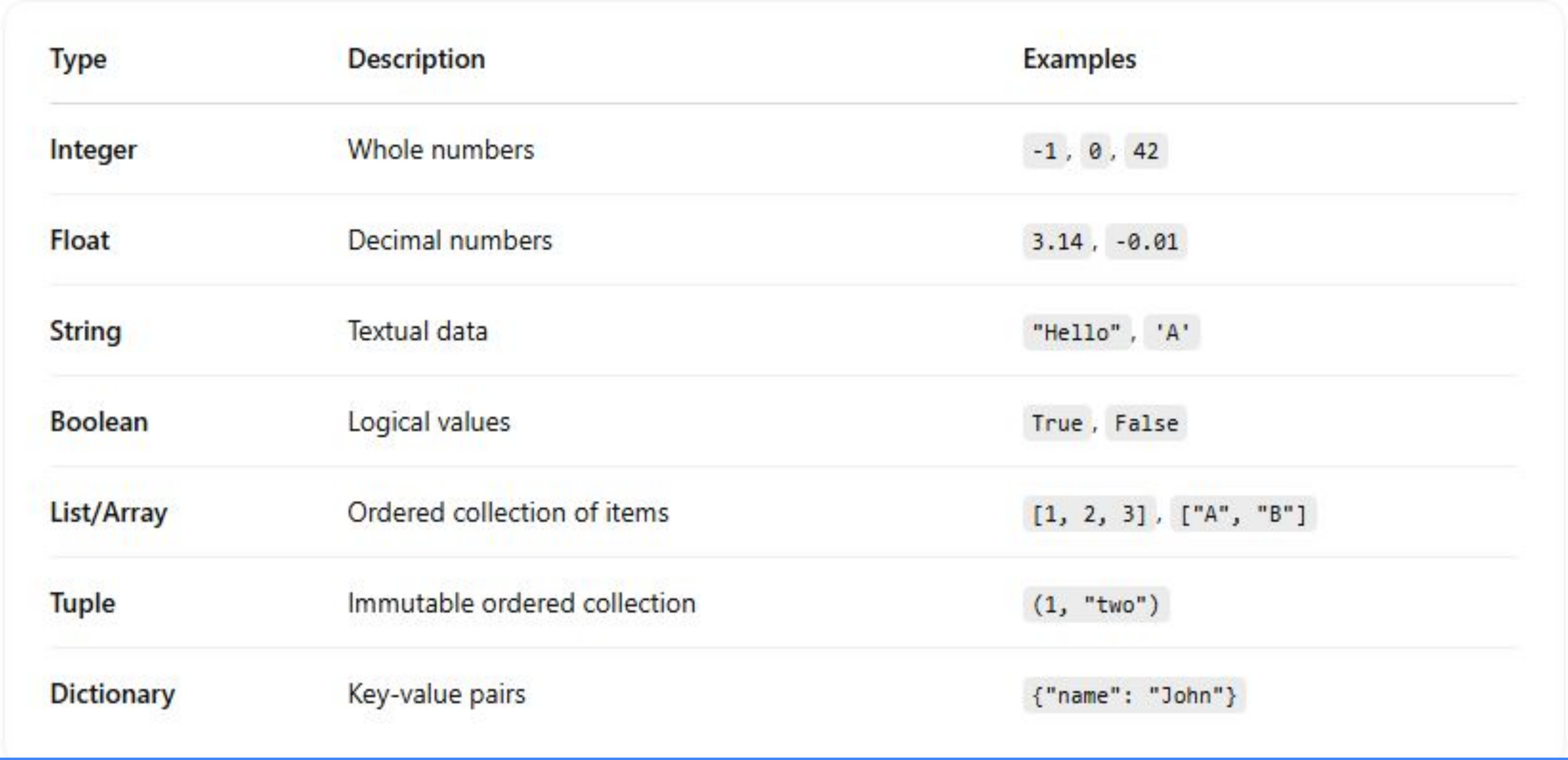
Quantitative (Numerical) Variables

- Discrete variables: Countable values (e.g., number of students)
- Continuous variables: Measurable quantities with infinite possible values (e.g., height, weight)

2. Qualitative (Categorical) Variables

- Nominal variables: No natural order (e.g., gender, city name)
- Ordinal variables: Have a meaningful order (e.g., education level: high school < college < PhD)

Data Types in Programming



Type	Description	Examples
Integer	Whole numbers	`-1`, `0`, `42`
Float	Decimal numbers	`3.14`, `-0.01`
String	Textual data	`"Hello"`, `'A'`
Boolean	Logical values	`True`, `False`
List/Array	Ordered collection of items	`[1, 2, 3]`, `["A", "B"]`
Tuple	Immutable ordered collection	`(1, "two")`
Dictionary	Key-value pairs	`{"name": "John"}`

Dependent vs Independent Variables

- **Independent Variable**

  -Used for prediction

  -Also called as Predictor, Input, Feature

  Ex….

  Exercise Time, Diet
- **Dependent Variable**

  -Used to get results

  -Also called as Response, Output, Target

  -Weight Loss

Ex. Housing Prices Dataset – Key Variables

Key Variables

- Location

  Where the house is (e.g., city, neighborhood) – affects how expensive the area is.
- Area

  Size of the house in square feet – bigger homes usually cost more.
- Bedrooms

  Number of bedrooms – more rooms can mean higher price.
- Price

  The cost of the house – this is what we usually want to predict

What Is Missing Data?

- Sometimes, information is not available in a dataset. This is called missing data

What Are Outliers?

- The point which is very far from actual value is called as Outliers

EDA

- Definition of EDA:-Understanding the data before making any decisions
- Importance in Data Science:-Detecting Missing value,Finds Outliers,Improves Data Quality for Better Models,Saves Time and Prevents Costly Mistakes

Limitations of Quantitative Data Exploration

Skewed Data

- Quantitative data may not always be normally distributed, leading to inaccurate conclusions

Outliers

- Outliers are values that are significantly different from the rest of the data. They can distort statistical results

Data Gaps

- Missing values in quantitative data can lead to incomplete analysis

Limitations of Qualitative Data Exploration

Hard to Measure

- Feelings and opinions are difficult to quantify

Time-Consuming

- Analyzing text, audio, or video takes much more time than numbers

Requires Skilled Interpretation

- Analysts must be trained to accurately results

Structured Data

Def:-Structured data is organized data stored in rows and columns format

Ex..

SQL, Excel

Unstructured Data

- Def:- Unstructured data is messy data that doesn’t fit in tables

Ex…

- Images
- Video
- Audio

What is a Feature/Variable?

Features (or Attributes):

- Examples: Age, income, product , location, Temperature, sales

Variables:

A variable is a specific feature

Type

Dependent (Target) Variable: The outcome you're trying to predict or explain (e.g., sales)

Independent (Predictor) Variables: Features that are used to predict the dependent variable (e.g., age, temperature)

Why Dimension Reduction is Important

- Simplifies the Model:
  -Eliminates irrelevant or redundant features, focusing only on the most important ones.
- Improves Computational Efficiency:

  -Less memory consumption, leading to faster processing, especially with large datasets.

- Reduces Overfitting:

  -Helps generalize the model better to unseen data, improving its performance.

- Enhances Visualization:

  -Makes it easier to visualize data, especially when reduced to 2 or 3 dimensions (e.g., using PCA)

- Improves Accuracy:

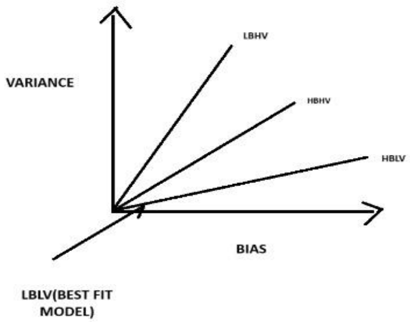
  -By removing noise and irrelevant features, dimension reduction can increase model accuracy by focusing on the most informative aspects of the data

Challenges with High-Dimensional Data

- Computational cost
- difficulty in visualization
- noise

Overfitting vs Underfitting

- LBLV – Low Bias, Low Variance (ideal model)
- LBHV – Low Bias, High Variance (overfitting)
- HBLV – High Bias, Low Variance (underfitting)
- HBHV – High Bias, High Variance (poor model)



Impact on Model Performance

- Reduced dimensions lead to faster and often more accurate models

Lasso Regularization

- Lasso (Least Absolute Shrinkage and Selection Operator) is a form of regularization used to prevent overfitting

Ridge Regularization

- Ridge regularization is another form of regularization used to prevent overfitting

Elastic Net Regularization

- ElasticNet is a regularization technique that combines both **L1 (Lasso)** and **L2 (Ridge)** regularization methods. It aims to improve the performance of models

Introduction to Correlation

- **Definition:-** Correlation means **how two things move Together**

Positive Correlation

- You're observing ice cream sales and temperature.
- On hotter days, ice cream sales go up.
- On colder days, sales go down.
- That’s a positive correlation

Negative Correlation

Negative correlation means that when one thing increases, the other decreases.

- They move in opposite directions.

Ex

- Sleep vs Stress
- Screen Time vs Eye Health

Zero Correlations

Definition:- Zero correlation means there is **no** relationship between two things

Ex.

- Pizza toppings vs Exam marks
- Hat color vs Monthly salary

Pearson Correlation Coefficient

- The Pearson Correlation Coefficient is a number that tells you how strongly two things are related
- If people who study more usually get higher scores, that’s a positive relationship
- If people who study more get lower scores, that’s a negative relationship

- If studying more doesn’t affect scores at all, then there’s no relationship

Principal Component Analysis (PCA)

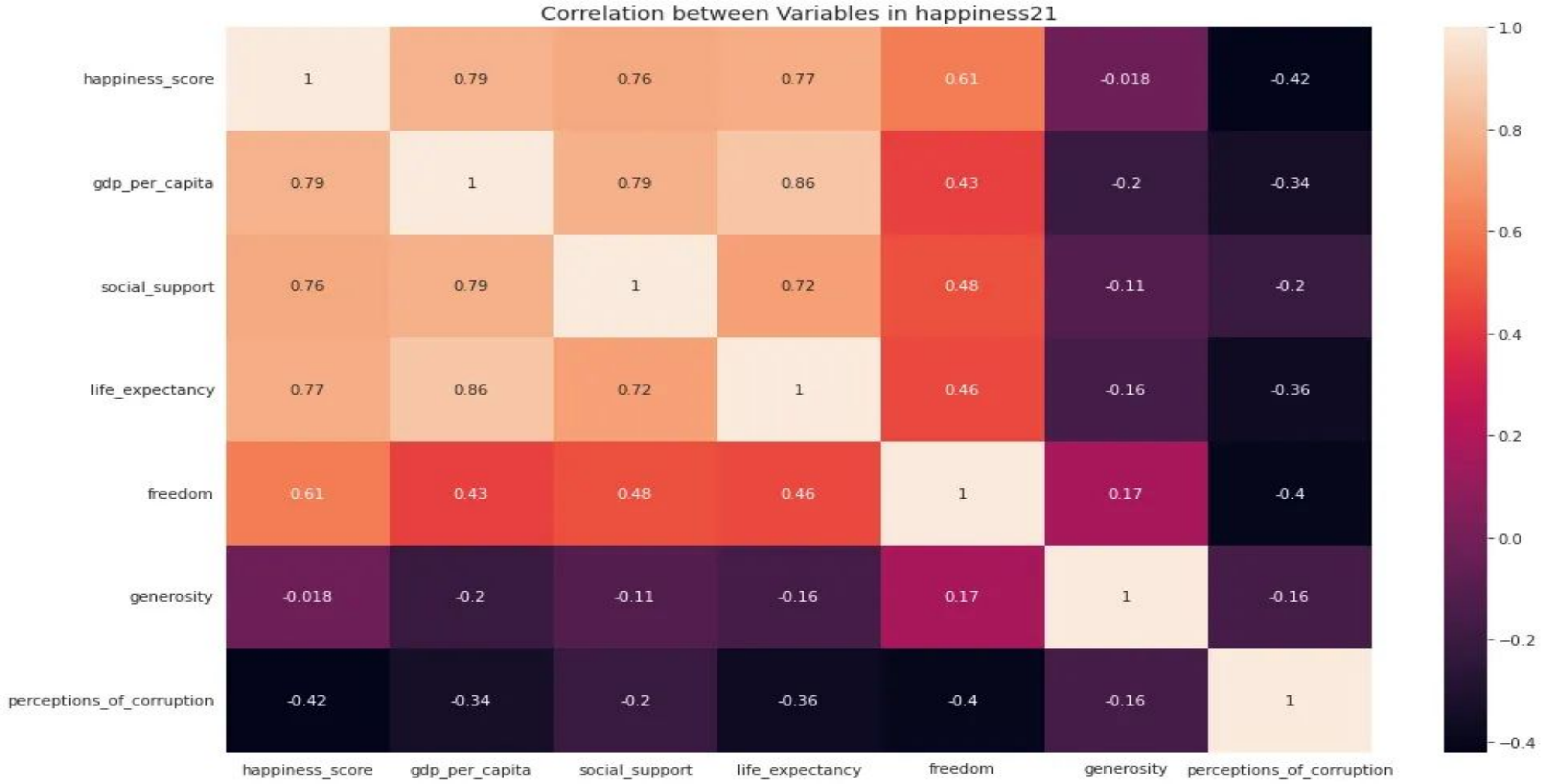
- Principal Component Analysis (PCA) is a technique used to simplify big and complicated data without losing too much important information

  Ex…
  Color
  Size
  Weight

Tools and Libraries

- Scikit-learn
- Pandas
- NumPy
- Matplotlib
- seaborn

Heatmaps & Correlation Matrices



Prescriptive vs Predictive vs Descriptive

- Descriptive Analytics

-Descriptive analytics focuses on summarizing past data to understand what has happened in the business
Example
Monthly sales reports

- Predictive analytics

-Predictive analytics uses historical data and machine learning models to forecast future outcomes.
Example:
Predicting customer churn

- Prescriptive Analytics

-Prescriptive analytics recommends actions to optimize outcomes based on predictive models
Example:
Recommending best routes for delivery to minimize cost

Prescriptive analytics in healthcare sector

- A hospital has lots of patients with diabetes. Some of them often end up in the emergency room because their blood sugar levels go out of control.

What the Hospital Does:

  - It uses a computer system that studies past patient data — like age, diet, sugar levels, and medication history
  - The system predicts which patients might have a sugar
  - But it doesn’t stop there — it goes a step further and recommends what to do:
    i. Sends the patient a reminder to take medicine
    ii. Suggests the right food to eat that day
    iii. Alerts the doctor to schedule a quick check-up
    iv. Recommends changes in insulin dosage

Recommendation Systems

- Product based recommendation
- Content based recommendation

Ex…

Amazon, flipkart, Netflix, Zee5

What is Hypothesis Testing?

Def:-Hypothesis testing is like trying to prove something is true or false using evidence

- Evidence is called as P-Value

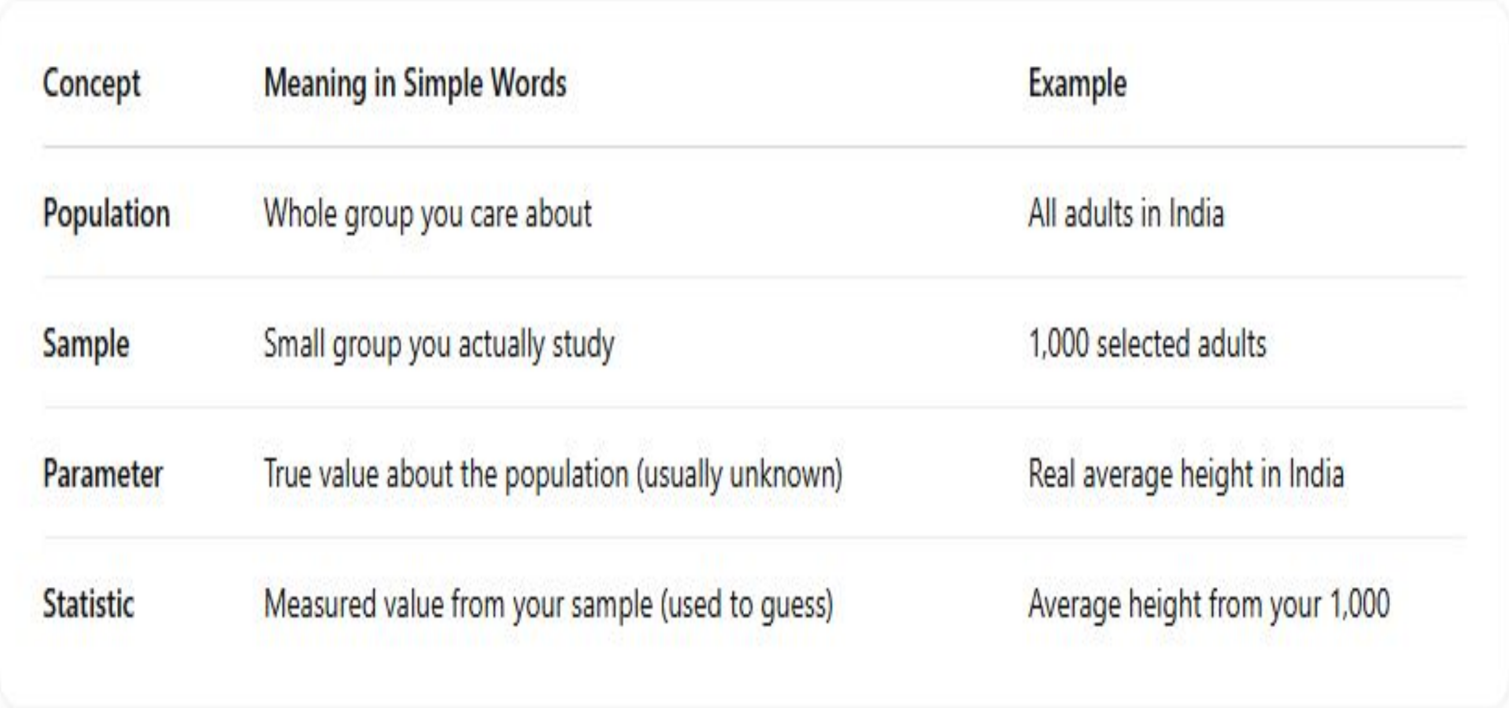
Types with Ex..

- Null Hypothesis (H_0): This new coffee doesn’t make people more awake than regular coffee.
- Alternative Hypothesis (H_1): This new coffee does make people more awake than regular coffee.

Importance of Hypothesis Testing

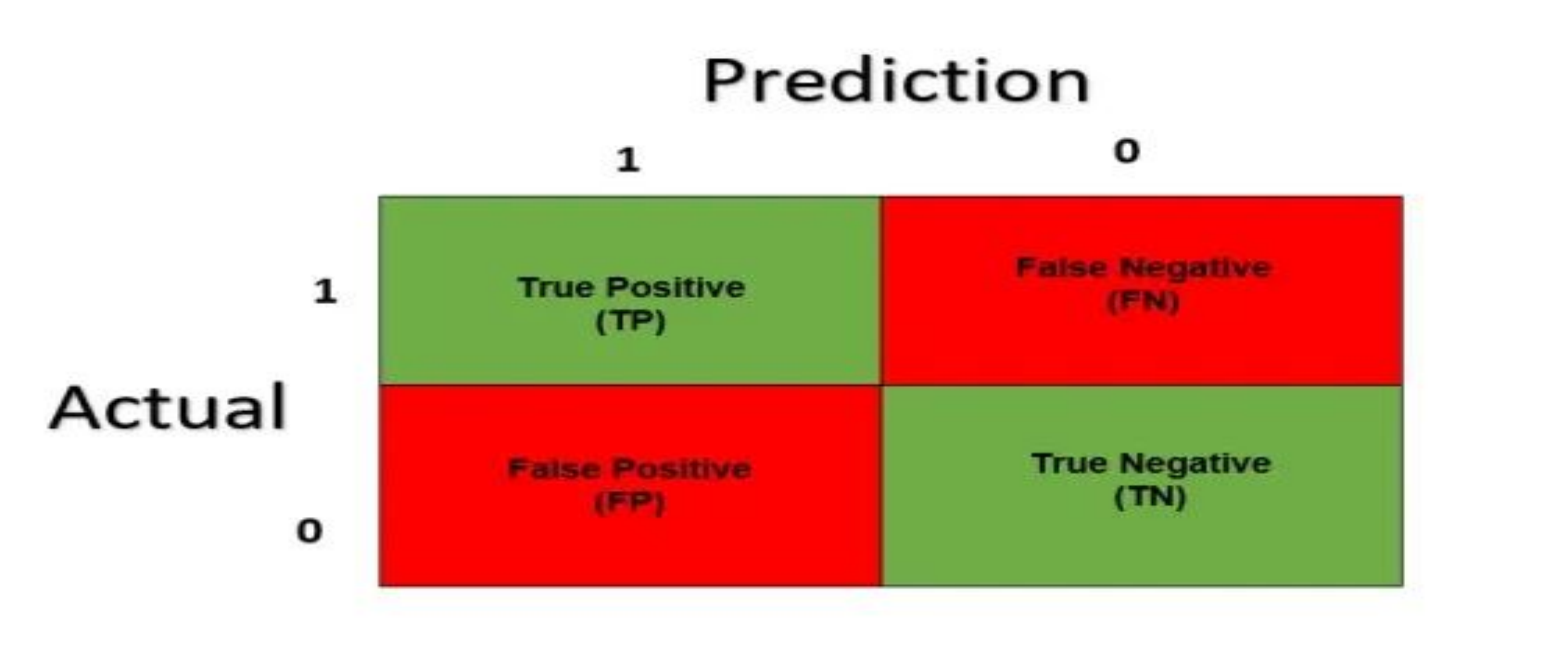
- Hypothesis testing is important because it helps us make decisions using data instead of just guessing or people opinions

Key Terms



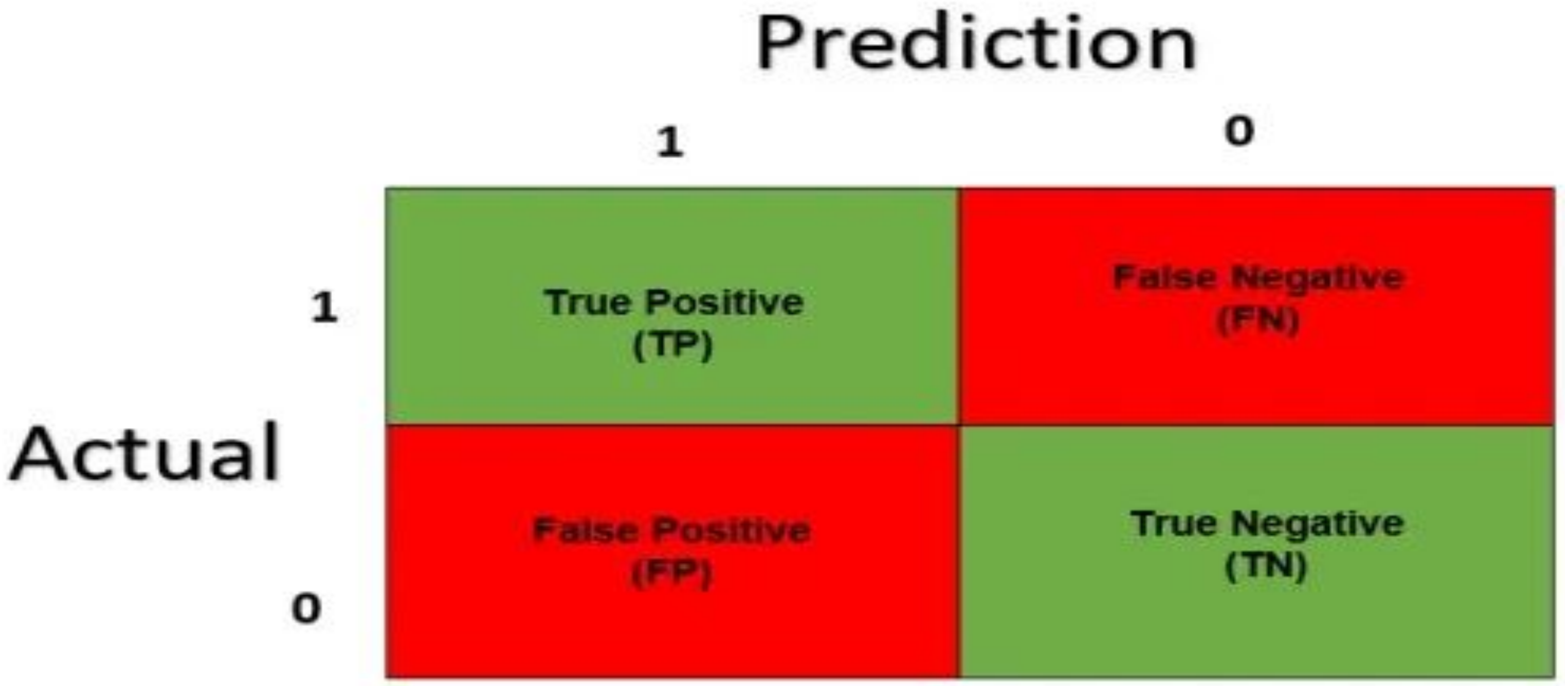
Concept	Meaning in Simple Words	Example
Population	Whole group you care about	All adults in India
Sample	Small group you actually study	1,000 selected adults
Parameter	True value about the population (usually unknown)	Real average height in India
Statistic	Measured value from your sample (used to guess)	Average height from your 1,000

Type I Error



		Prediction	
		1	0
Actual	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

Type II Error



		Prediction	
		1	0
Actual	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)