Research Methodology

Module 3 →

Importance of data, data understanding

Methods of data collection

Data → (?)

- → decision making
- → Innovation

} almost every field.

→ Informed Decision-Making
  - → Data provides insights that guide organizations in making evidence-based decisions.

→ Problem Solving: Patterns in data help identify issues and uncover soln

→ Innovation & Development
  Data fosters innovation
  - → AI
  - → ML
  - → business strategies

→ Performace Evaluation
  it allows the measurment & monitoring of Performace accross various domains

→ Competitive Adv. : organizations that effectively use data gain a strategic edge in Market

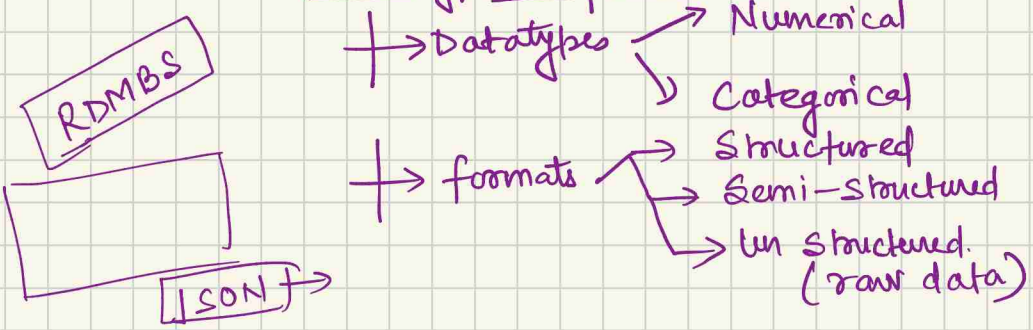
2) Data understanding

it involves comprehending the nature, structure, and meaning of data being analysized

→ Domain knowledge: Understanding the context and purpose of the data

→ Exploratory Data Analysis (EDA):
use
→ statistical & visualization tools to summarize and analyze data distribution, trends, and anomalies

→ Data types & formats:



→ Data Sources: Knowing where data originates, such as surveys sensors, databases its reliability

Methods of Data collection
the purpose and nature of the study

→ Surveys & Questionnaires:
Opinions or feedback from individuals

→ Interviews: Gathering qualitative insights through direct Interaction

→ Observation: Recording a behaviour or events in natural settings

→ Web Scraping: Extracting data from websites using automated tools

→ Sensor and IoT devices: Capturing real-time data from connected devices

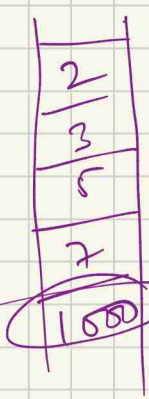
→ Transactional Data: Extracting data from systems like ERP or CRM

→ Public and Open data Sources: Utilizing datasets available from government or organization

4. → Data pre-processing techniques

ensures the data is clean, consistent, and suitable for analysis or modeling

0-100



→ Data cleaning
- → handling missing values
- → Identifying and correcting errors Outliers.
- → Removing duplicate entries.

→ Data Integration
→ combining data from multiple Sources into a unified dataset

→ Resolving conflicts in data formats or units

→ Data Transformation
  → Normalization: Scaling data to fall within a specific range
  → Standardization: Converting data to have a mean of zero and S.D = 1
  → Encoding categorical variables.
    → One-hot encoding
    → Label encoding

→ Data reduction
  → Dimensionality Reduction techques. PCA
  → Feature Selection to retain only the most important variables

→ Data splitting:
  → training
  → validation
  → test sets

1) Surveys & Questionnaires.
→ collecting data from respondents
→ Structured
→ Semi-Structured.

Mobile Application
→ feedback.
Conducting a Survey to understand developer preferences for programming language

→ tools → Google form
→ Typeform (large-scale data collection)

2. Interviews
→ Structured
→ Semi-structured
→ unstructured.

→ Conducting interviews with Software Engineers to understand debugging practices.

→ Talking to it professionals to gather insights into cybersecurity/AI/ML/A challenges in their organizations.

3 Observation
Directly observing subjects or processes without interference.

→ Observing user behavior during usability testing of a website or software
→ Monitoring network traffic in a cybersecurity experiment to detect pattern of malicious activity

4. Experiments

Conducting controlled experiments to establish cause-effect rel^{n}ship.

→ Testing the performace of different machine learning algo. on the same dataset

→ Evaluating the energy consumption of various scheduling algo in cloud computing environments

5. Case Studies

In-depth analysis of a single case or small number of cases.

→ A detailed study of a startup's adoption of Agile methodology.

→ Analyzing the arch. of a successful open-source project (linux or Kubernetes)

6. Focus group.

A moderated discussion with a small grp. to gather opinions or insights

→ Gathering feedback from a focus grp of students on an educational software tool.

→ Discussing potential features of a new version of a programming IDE with developers.

- cursor
- vs-code

7. Online Data collection

Gathering data from online sources like websites.

social media or public datasets

↓ ↓

Reddit Kaggle.

Using Api to collect data from social media.

8. logs & System Monitoring

Collecting data generated by systems during their operation

9. Secondary Data collection

Using existing datasets for research purpose.

10. Simulation

Creating artificial env. to test hypotheses or models.

→ Digital twin.

Semi-manual Data collection

1) web scraping with manual oversight
   → web scraping tools
   → manually check or refine the collected data

2) Data Annotation with preprocessing,
   → Use a tool or script to generate initial annotations.
   → Have humans review & refine the annotation

3) log data collection with manual filtering.
   1) Use tools like Splunk or ELK Stack. to gather system logs.
   2) Manually inspect and remove irrelevant or redundant log entries.

4) Sensor Data collection with Manual contextualization
   → collect the data from IoT devices. (raw data)
   → Manually tag data points with additional context.

5) Data Entry with Automation
   → Use OCR (Optical character Recognition) or Scanning tools. to digitize documents
   → Manual Data entry for correction

Benefits of Semi-Manual Methods.

→ Efficiency
→ flexibility
→ Accuracy : Reduces errors compared to fully automated methods.

→ Gen-AI Role in Data collection

1) Data Augmentation

Gen-AI can create synthetic data to augment small data-sets, making them more diverse and robust for research & training ML models.

→ Generating synthetic images using Generative Adversarial Networks.
→ Creating synthetic text data using GPT models to expand natural language datasets

2. Automated Survey Generation

Gen-AI can create dynamic, well structured surveys or questionnaires based on research objective.

3. Data cleaning & preprocessing.

Gen-AI assists in cleaning raw data by identifying anomalies

4) Simulated Data collection

→ Generate real-word scenarios and generate data for testing or training purposes.

5.) Real-time Assistance During Interviews.
Gen-AI can assist researchers during Interviews by suggesting follow-up questions

6) Social Media Data Collection & Analysis.
Gen-AI collect & analyze data from social media platforms, including hastags comments
→ GROK.

7.) Conversational AI for Surveys.
Gen-AI based chat bots act as survey administrators, interact with users & collecting their responses in real-time.

8) Natural Language Data transformation
Gen AI transform unstructured text data (reviews, feedback).

Benefits of using AI in dat collection
→ Scalability : AI can process large volume data enabling researcher to handle datasets that would otherwise be unmanageable manually.

→ Efficiency Automate repetitive tasks like annotation, cleaning & summarization
→ Cost Effectiveness: Reduce the time & effort required for manual data collection & preprocessing.
→ Customization : Gen AI can adapt to specific research needs, creating tailored data or insights.

Limitation & Ethical Considerations.

1) Bias : AI model might introduce bias if trained on unbalanced datasets.
2) Data privacy : Ensure compilance with regulations like GDPR or India's Digital Personal Data protection act. when using sensitive date
3) Accuracy : AI-generation or annotations should by varified manally to ensure reliability

Text data processing & Generation
- → OpenAI Gpt models.
- → Hugging Face Transformers.
- → TextBlob/VADER.
- → Spacy

Image & video data.
- → DALL-E
- → Deeplabcut
- → GAN
- → Labelling.

Surveys Automation
- → Typeform.
- → Qualtrics

Web scraping.
- → Beautiful soup / scrapy.
- → Gpt models with web scraping.
- → Netlytic

Data Annotation.
- → Label Studio
- → Scale AI
- → Amazon SageMaker.

Chatbots for surveys.
- → Rasa
- → Dialogflow

Data Preprocessing

→ critical step in any technical or research-oriented task.

→ ML, Data science, Big data Analytics.

→ transforming raw data into a clean, organized format effectively utilized by algorithms & model.

Why Data preproc. Important?

→ **Improves Data Quality:**
Raw data → inconsistencies. → missing values
→ noise
→ Irrelevant feature

→ **Enhance Model Performance:**
Clean, normalized, and well-organised data ensures that models learn patterns effectively without biases or errors introduced by poor data qualities.

→ **Facilitates feature Engineering**
Preprocessing sets the stage for creating and selecting feature that are more relevant to the problem.

Steps in Data Preprocessing

1> → raw data from various sources such as databeses, sensors, APIs

2> → **Data cleaning**
→ Handle Missing Data.
→ fill up these missing Data.
imputation technique.
→ mean value.
→ Median value.
→ mode value.

→ remove rows & or columns

→ Remove noise Eliminate anomalies or outliers. using Statistical or machine learning based approaches.

→ Discard network/latency values in a dataset if they are abnormally high compared to the average.

3) Data Integration
Combine data from multiple sources to form a unified dataset.

4) Data Transformation
→ Normalization
Scale data to specific range ([0,1]) to ensure uniformity

Normalize cpu utilization data in %

→ Encoding Convert categorical data into numerical format
→ One hot Encoding

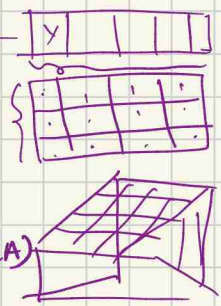
5) Feature Selection & Extraction
→ feature Selection: Identify the most relevant attributes for the problem
→ correlation analysis.
→ feature importance (XAI)

In predicting software reliability, select metrics.
→ number of commits
→ line of code changed

→ Feature Extraction
Create a new feature that better represent the data.

Extract the frequency of API calls from log data as a feature.

6) Data reduction



Reduce data dimensionality to improve computation efficiency and avoid overfitting.

Use Principal Component Analysis (PCA) to reduce the number of I/P variables in an IoT dataset

7) Data Splitting.
→ training
→ validation } sets
→ testing

{ 70-20-10
80:20 }

IDS for Cybersecurity —

raw data log of network activity
- → packet timestamps.
- → Source & Destination IPs.
- → Protocol type. (UDP / TCP)
- → packet size
- → Labels (malicious or normal)

→ Data cleaning. : fill missing packet size values with the median size.

→ Encoding Convert protocol type (into numerical categories.

→ Normalize Data Scale the packet size to a range of [0 1] for uniformity.

→ feature selection ! Retain only features. like Protocol type, packet size, and Source IP

→ Data splitting : training -testing validation Dataset.