Unit III Predictive Analysis Process and R

3.1 Introduction to R: R graphical User Interfaces, Data import and Export, Dirty Data, Data Analysis, Linear regression with R, clustering with R hypothesis testing, Data cleaning and validation tools: MapReduce

📌 Introduction to R

R is a programming language mainly used for **statistical computing**, **data analysis**, and **visualization**. It is widely used by data scientists and statisticians.

💻 R Graphical User Interfaces (GUIs)

These are user-friendly tools to interact with R without writing much code.

Examples:

- **RStudio** – The most popular IDE (Integrated Development Environment) for R.
- **R GUI** – Comes with base R installation; simple and basic.
- **Jupyter Notebooks** – Can also run R with an R kernel.

📁 Data Import and Export in R

R supports **reading and writing** different types of files:

Task	R Function	Example
Import CSV	read.csv()	read.csv("data.csv")
Export CSV	write.csv()	write.csv(data, "output.csv")
Import Excel	readxl::read_excel()	From readxl package
Import from Web/API	read.table(), APIs	Use with URLs or APIs

🧹 Dirty Data

Dirty data means **data with issues** such as:

- Missing values
- Duplicates
- Inconsistent formatting
- Wrong data types

💡 We clean dirty data before analysis to get accurate results.

📊 Data Analysis in R

You can do many types of data analysis in R:

- **Descriptive statistics**: mean, median, mode, standard deviation
- **Visualizations**: using ggplot2 or base R plot
- **Correlation and relationships** between variables

📈 Linear Regression with R

Used to predict a continuous value (e.g., price, score).

Example:

R

CopyEdit

```
model <- lm(y ~ x, data = my_data)

summary(model)
```

- lm() = linear model
- y ~ x means "predict y using x"

🔍 Clustering with R

Used to **group similar data points** (unsupervised learning).

Popular methods:

- **K-Means**: kmeans()
- **Hierarchical clustering**: hclust()

Example:

R

CopyEdit

kmeans_result <- kmeans(my_data, centers = 3)

🧪 Hypothesis Testing in R

Used to **test assumptions** about data (e.g., "does a drug work?").

Common tests:

- **t-test**: t.test()
- **Chi-square test**: chisq.test()
- **ANOVA**: aov()

🧼 Data Cleaning and Validation Tools

R provides packages and tools for cleaning and validating data:

- **dplyr**, **tidyr** – for cleaning and reshaping data
- **validate** – to define rules and check if data meets them
- **janitor** – to clean messy datasets quickly

🗺️ MapReduce (Big Data Concept)

While R is not primarily for big data, you can connect it with **MapReduce** systems (like **Hadoop**) using packages such as:

- **rhdfs** – connect R with Hadoop File System
- **rmr2** – write MapReduce code in R

💡 **MapReduce** helps process **large-scale data** by splitting it into parts (Map) and then combining results (Reduce).

✅ Summary

Topic	**Simple Description**
R GUI	Tools like RStudio to work with R easily
Data Import/Export	Reading/writing CSV, Excel, web data
Dirty Data	Incomplete, wrong, or inconsistent data
Data Analysis	Exploring and summarizing data
Linear Regression	Predicting a value using other variables
Clustering	Grouping similar data points
Hypothesis Testing	Checking assumptions (e.g., does A affect B?)
Data Cleaning Tools	R packages like dplyr, tidyr, validate
MapReduce	Big data processing with R + Hadoop (split → process → combine)

3.2 Data Analytics Lifecycle: Discovery, Data Preparation, Model Planning, Model Building, communicate results, Operationalize, Building a Predictive model.

🔄 Data Analytics Lifecycle

The Data Analytics Lifecycle is a **step-by-step process** that data scientists follow to **solve problems using data**.

✅ 1. Discovery

🔍 **Goal:** Understand the business problem.

- Identify the problem you want to solve.
- Understand the goals and what data might help.
- Ask: *What are we trying to predict or improve?*

📝 Example: A company wants to **predict customer churn** (who will leave their service).

🧹 2. Data Preparation

🧼 **Goal:** Clean and organize the data.

- Gather data from different sources.
- Remove missing or duplicate values.
- Format data for analysis.

📝 Example: Remove customers with missing email IDs, convert dates into standard format.

📊 3. Model Planning

📘 **Goal:** Choose the right approach and tools.

- Decide which techniques to use (e.g., regression, clustering).
- Explore data visually and statistically.
- Choose the evaluation metrics (like accuracy, F1-score, etc.)

📝 Example: Use **logistic regression** to predict whether a customer will churn.

🏗 4. Model Building

🧠 **Goal:** Create the actual predictive model using machine learning.

- Apply algorithms like decision trees, SVM, neural networks, etc.
- Train the model on your data.
- Tune hyperparameters to improve performance.

📝 Example: Build a model that predicts churn with 85% accuracy.

📢 5. Communicate Results

📈 **Goal:** Share insights with stakeholders.

- Visualize results with graphs and dashboards.
- Explain the model's findings in **simple terms**.
- Show how the model helps solve the original business problem.

📝 Example: "The model shows that customers with low usage are more likely to leave."

🚀 6. Operationalize

⚙️ **Goal:** Deploy the model into the real world.

- Integrate the model with business systems.
- Start using it in daily operations.
- Monitor performance regularly.

📝 Example: The model runs daily to flag customers likely to churn.

🔮 7. Building a Predictive Model

✨ **Goal:** Predict future outcomes using past data.

- A predictive model uses patterns in historical data to **forecast what will happen next**.

- Examples: churn prediction, sales forecasting, fraud detection.

📘 **Steps**:

1. Select features (important columns).
2. Train a machine learning model.
3. Evaluate its accuracy.
4. Use it to make predictions on new data.

✅ Summary Table

Stage	**What It Means**
Discovery	Understand the problem and goals
Data Preparation	Clean and organize data for analysis
Model Planning	Choose the right tools and techniques
Model Building	Build the machine learning model
Communicate Results	Share results in a clear and visual way
Operationalize	Deploy the model for real-world use
Predictive Model	Use the model to forecast future events