**UNIT – V**

**Introduction to Data Analytics with R**

**Why R for Data Analytics?**

R is a powerful open-source programming language that is widely used in **data analytics, statistical computing, and machine learning**. It provides a **comprehensive environment** for handling, visualizing, and analyzing large datasets efficiently. Below are some of the key reasons why R is a popular choice for data analytics:

1. **Open-source & Free**
   * R is freely available, making it accessible to researchers, data scientists, and businesses.
   * Large and active community support provides numerous free libraries and resources.
2. **Statistical Computing Capabilities**
   * R is designed for advanced **statistical analysis and data modeling**.
   * Provides inbuilt functions for regression, hypothesis testing, time series analysis, and more.
3. **Rich Ecosystem of Machine Learning Libraries**
   * R supports a variety of machine learning techniques through powerful libraries such as:
     + **caret** – Unified framework for ML models
     + **randomForest** – Random Forest for classification and regression
     + **xgboost** – Gradient boosting algorithm for predictive modeling
4. **Visualization Capabilities**
   * R excels in **data visualization and storytelling**, making it easy to explore and communicate insights.
   * Popular visualization libraries include:
     + **ggplot2** – Advanced data visualization
     + **lattice** – Multi-panel statistical graphics
     + **plotly** – Interactive graphs and dashboards
5. **Integration with Big Data Technologies**
   * R can **handle large datasets** and integrate with Big Data frameworks such as:
     + **Hadoop** – Parallel computing with R using the RHadoop package
     + **Spark** – Distributed ML and big data processing via SparkR
     + **BigR** – Enables R to work with **Big Data** stored in HDFS

**Key Steps in Data Analytics with R**

To perform data analytics in R, a structured workflow is typically followed. Below are the key steps:

**Step 1: Data Collection**

The first step in data analytics is **importing data** from different sources into R. Common data sources include:

* **CSV files** → read.csv("data.csv")
* **Excel files** → readxl::read\_excel("data.xlsx")
* **Databases** (MySQL, PostgreSQL, MongoDB) → DBI and RMySQL
* **Web scraping** (APIs, JSON, XML) → httr, rvest

**Step 2: Data Preprocessing**

Before analysis, raw data needs to be **cleaned and transformed**:

* **Handling Missing Values**
  + Remove missing data → na.omit(dataset)
  + Impute missing values → mean(dataset$column, na.rm = TRUE)
* **Data Transformation**
  + Convert categorical variables → as.factor(dataset$column)
  + Normalize numerical data → scale(dataset$column)

**Step 3: Exploratory Data Analysis (EDA)**

EDA helps in understanding the **distribution, patterns, and relationships** in data.

* **Descriptive Statistics**
  + Summary of data → summary(dataset)
  + Mean, median, standard deviation → mean(), sd(), quantile()
* **Data Visualization**
  + **Univariate Analysis** → Histograms, box plots (ggplot2)
  + **Bivariate Analysis** → Scatter plots, correlation heatmaps

**Step 4: Model Building (Supervised & Unsupervised Learning)**

Depending on the problem type, different machine learning techniques are applied:

* **Supervised Learning (Labeled Data)**
  + Regression: **Linear Regression, Random Forest Regression**
  + Classification: **Logistic Regression, Decision Trees, SVM**
* **Unsupervised Learning (Unlabeled Data)**
  + Clustering: **k-Means, Hierarchical Clustering, DBSCAN**
  + Dimensionality Reduction: **PCA**

**Step 5: Model Evaluation**

After training, models are **evaluated** using various performance metrics:

* **Regression Metrics**
  + **RMSE (Root Mean Squared Error)** → Measures error in prediction
  + **R² (R-Squared)** → Measures model accuracy
* **Classification Metrics**
  + **Accuracy** → (Correct Predictions / Total Predictions)
  + **Precision & Recall** → Performance of classification models
  + **ROC Curve & AUC Score** → pROC package for model evaluation

**Step 6: Deployment & Interpretation of Results**

Once the model is validated, it is **deployed for real-world use**.

* **Deploying as an API** using **Plumber**
* **Deploying on web applications** with **Shiny**
* **Interpreting results** and generating reports using **R Markdown**

**Introduction to Collaborative Filtering**

Collaborative Filtering recommends items by analyzing past interactions between **users and items**.

**How does it work?**

* **User-based filtering:** "People similar to you liked these items."
* **Item-based filtering:** "If you liked this item, you may like similar items."
* **Hybrid Filtering:** Combines **both** user-based and item-based filtering.

**Example Use Cases**

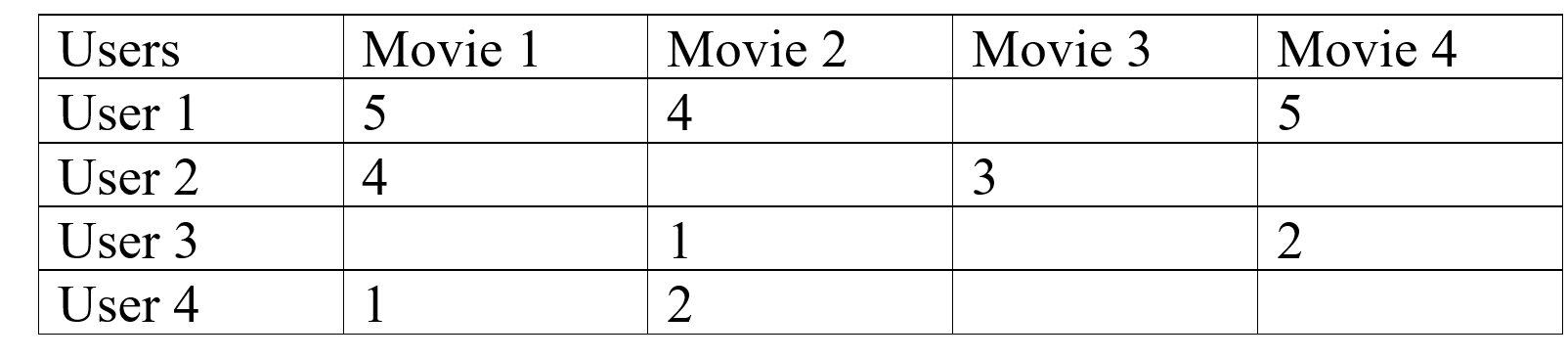
* **E-commerce:** Suggesting products based on past purchases.
* **Streaming Platforms:** Recommending movies based on viewing history.
* **Online Learning:** Suggesting courses based on user activity

**2. Types of Collaborative Filtering**

**2.1. User-Based Collaborative Filtering**

Finds **similar users** and recommends items liked by similar users.

* Example: If **User A** and **User B** have similar movie preferences, then **User A** will get recommendations based on **User B's** likes.



**Mathematical Approach:**

* Measures similarity using **Cosine Similarity** or **Pearson Correlation**.
* similarity=∣*A*∣×∣*B*∣*A*⋅*B*​=∑*i*=1*n*​*Ai*2​​×∑*i*=1*n*​*Bi*2​​∑*i*=1*n*​*Ai*​×*Bi*​​

**2.2. Item-Based Collaborative Filtering**

Finds **similar items** and recommends them to users who liked similar items.

**Example:** If many users who purchased **"iPhone 13"** also bought **"AirPods Pro"**, then a user who buys **"iPhone 13"** will get a recommendation for **"AirPods Pro"** since these items are frequently bought together.

**2.3. Hybrid Filtering**

Combines **User-based and Item-based filtering** for better recommendations.

* Used by Netflix, YouTube, and Amazon.
* New users with no history.

**Social media analytics**

Social media analytics refers to the process of collecting, analyzing, and interpreting data from social media platforms to assess performance, understand audience behavior, and optimize strategies. It helps businesses, marketers, and content creators make informed decisions on how to improve engagement, reach, and overall effectiveness on social platforms.

**Key Metrics to Track:**

* **Engagement:** Likes, comments, shares, retweets, reactions, etc.
* **Reach:** The number of unique users who have seen your posts.
* **Impressions:** The number of times your posts have been viewed, regardless of whether they were clicked or interacted with.
* **Follower Growth:** The increase or decrease in followers over time.
* **Click-Through Rate (CTR):** The percentage of users who click on a link in your post.
* **Conversion Rate:** The percentage of users who take a desired action (e.g., sign up, make a purchase, etc.) after clicking a link.
* **Sentiment Analysis:** Understanding whether the public perception of your brand is positive, neutral, or negative.
* **Hashtag Performance:** How well certain hashtags perform in terms of engagement and reach.

**Tools for Social Media Analytics:**

* **Google Analytics:** Can track traffic from social media platforms to websites.
* **Hootsuite:** Offers analytics for engagement, post performance, and more.
* **Sprout Social:** Helps measure social media campaigns, audience growth, and sentiment.
* **Buffer:** Provides insights on audience interactions, engagement, and post timing.
* **Facebook Insights:** For analyzing Facebook-specific metrics (posts, stories, and ads).
* **Twitter Analytics:** For tracking tweet performance, engagement, and follower demographics.

**Mobile Analytics**

Mobile analytics refers to the process of tracking, measuring, and analyzing the behavior of users on mobile apps or mobile websites. This helps businesses and developers understand how users interact with their mobile apps, identify areas for improvement, and optimize app performance to boost engagement, retention, and revenue.

**Key Metrics to Track in Mobile Analytics:**

* **App Downloads:** The number of times your app has been downloaded from app stores (Google Play, App Store).
* **Active Users (DAU/WAU/MAU):**
  + **DAU (Daily Active Users)**: Number of unique users engaging with your app on a daily basis.
  + **WAU (Weekly Active Users)**: Number of unique users engaging with your app on a weekly basis.
  + **MAU (Monthly Active Users)**: Number of unique users engaging with your app on a monthly basis.
* **Retention Rate:** The percentage of users who return to the app after a specified period (e.g., 1 day, 7 days, or 30 days). This helps measure how sticky your app is.
* **Churn Rate:** The percentage of users who stop using the app after a certain period. A high churn rate is often a sign that there’s a problem with user experience or engagement.
* **Session Length:** The average duration of a user's session in the app.
* **Session Frequency:** How often users return to the app within a given period (daily, weekly, etc.).
* **In-App Events:** Specific user actions like completing a level, making a purchase, or sharing content.
* **Conversion Rate:** Percentage of users who complete a desired action (e.g., sign up, make a purchase).