# Advance AI/ML Training – Telecom Data Analytics

**MODULE 1** - Foundation Recap and Telecom Analytics Basics

# Course Outline

- Module 1: Foundation Recap and Telecom Analytics Basics
  - structured, semi-structured, and unstructured data in telecom with examples
  - Review SQL and Python basics for data collection and cleaning
  - Summarise ETL concepts and role of data pipelines in analytics
  - Visit Telecom predictive Models – Churn, telecom KPIs, and their role in decision making
  - Practical outcomes and use cases
- Module 2: Advanced Data Engineering & Automation
  - Explain advanced ETL pipeline design and automation.
  - Role of data warehouses and data lakes in telecom
  - Introduce scalable data processing methods for large datasets.
  - Explain importance of data governance, privacy, and compliance
  - Practical outcomes and use cases

# Pre-requisites

| S.No | Software | Downlink link | Comments |
|---|---|---|---|
| 1 | MySql DB & Workbench | https://dev.mysql.com/downloads/installer/ | FULL EDITION, REMEMBER USERNAME AND PASSWORD |
| 2 | CSV Editor (Excel/Notepad++) | MS excel or Notepad++ should be available | |
| 3 | Anaconda (incl. Python + Juypter notebook) | https://www.anaconda.com/download | Install – DISTRIBUTION INSTALLER<br><br>Python envirnoment (data & ML libraries will be installed during the classroom) |
| 4 | PowerBI Desktop | https://powerbi.microsoft.com/desktop | Visualization |
| 5 | Web Browser | Edge / Chrome | Web browsing |

# Your Roadmap for this Section

01

## Understanding Telecom Data Types

Explore structured, semi-structured, and unstructured data sources unique to telecommunications

02

## SQL & Python for Data Work

Master essential retrieval and cleaning techniques that analysts use daily

03

## ETL Pipeline Development

Build automated workflows to extract, transform, and load telecom data

04

## Predictive Analytics Applications

Create machine learning models to predict customer churn and revenue

05

## KPIs & Dashboard Design

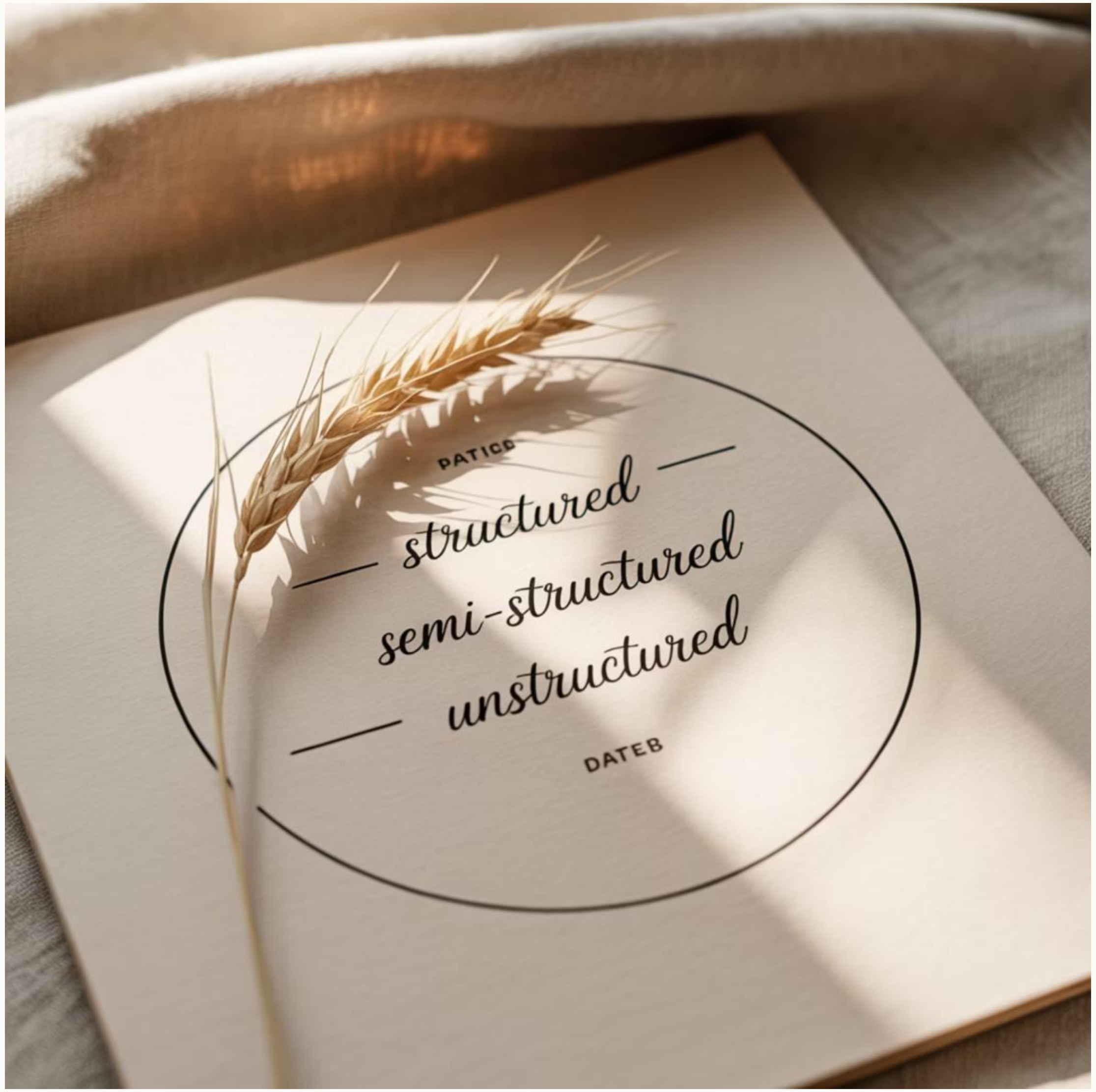
Visualize key metrics that drive telecommunications business decisions

# Session 1: Understanding Telecom Data Types

## What You'll Learn

Telecommunications companies handle massive volumes of diverse data every second. Understanding data types is the foundation of effective analytics. In this session, you'll discover why call detail records differ from customer complaints, and why each requires specialized handling techniques.

We'll explore three fundamental categories that make up the telecom data ecosystem, examining real examples and discussing the tools that analysts use to process each type efficiently.



# Three Critical Data Types in Telecommunications

## Structured Data

Organized in tables with defined schemas

- Call Detail Records (CDRs)
- Billing transaction logs
- Network performance metrics
- Customer subscription databases

**Tools:** SQL databases, Excel, relational database systems

## Semi-Structured Data

Flexible formats with some organizational hierarchy

- JSON API usage logs
- XML network configuration files
- Web server access logs
- IoT sensor data streams

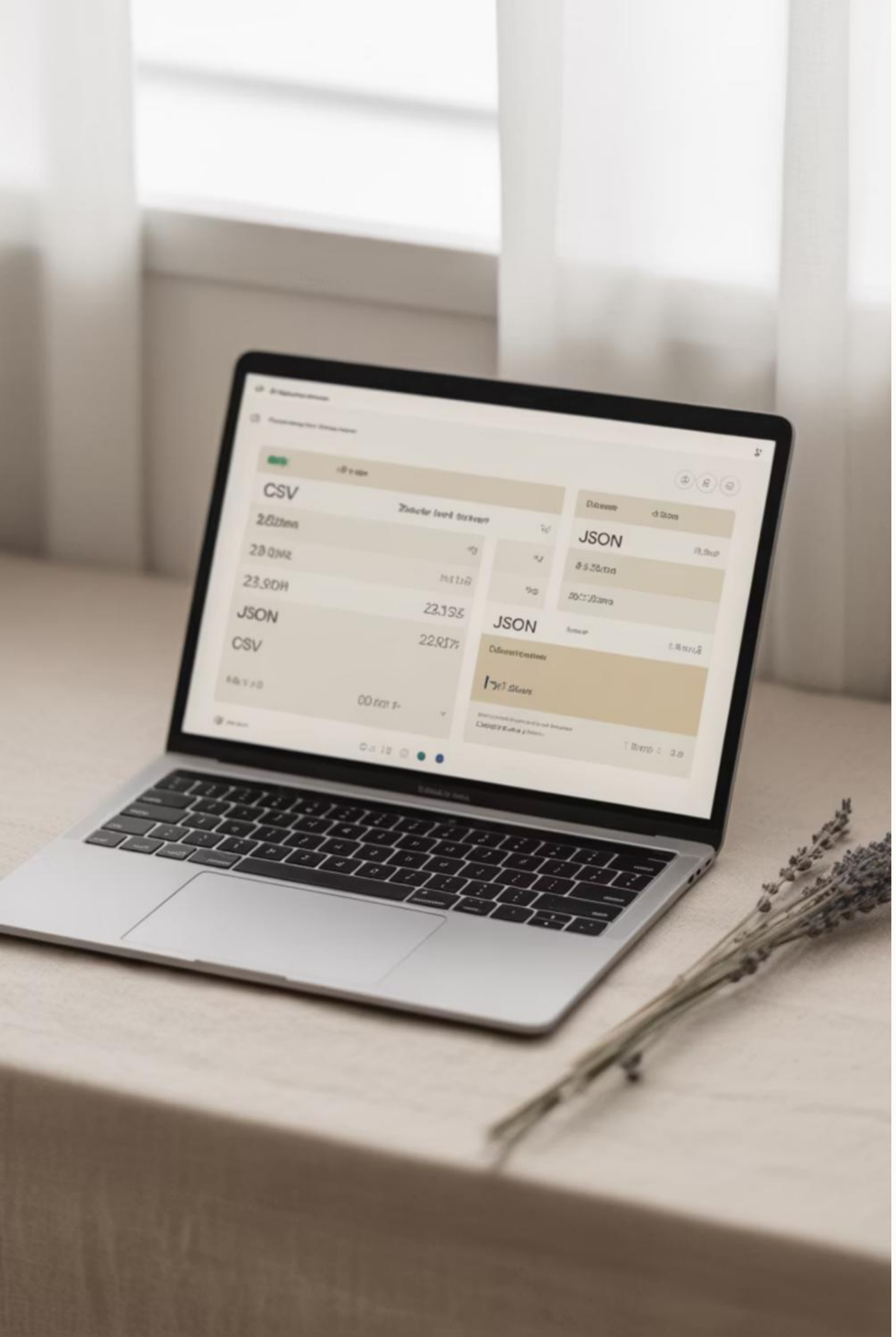
**Tools:** Python libraries, NoSQL databases, JSON parsers

## Unstructured Data

Free-form content requiring advanced processing

- Voice call recordings
- Customer service emails
- Social media feedback
- Network maintenance notes

**Tools:** Natural Language Processing, speech-to-text APIs, text mining




# Hands-On Exercise: Identifying Data Types

Your instructor will share three sample files from a telecom company. Take 15 minutes to examine each one and identify which category it belongs to. This practical recognition skill is crucial because choosing the wrong tool for a data type leads to inefficient analysis and potential errors.

1. **cdr_sample.csv**

   Contains call records with timestamp, caller ID, duration, and data usage in neat columns

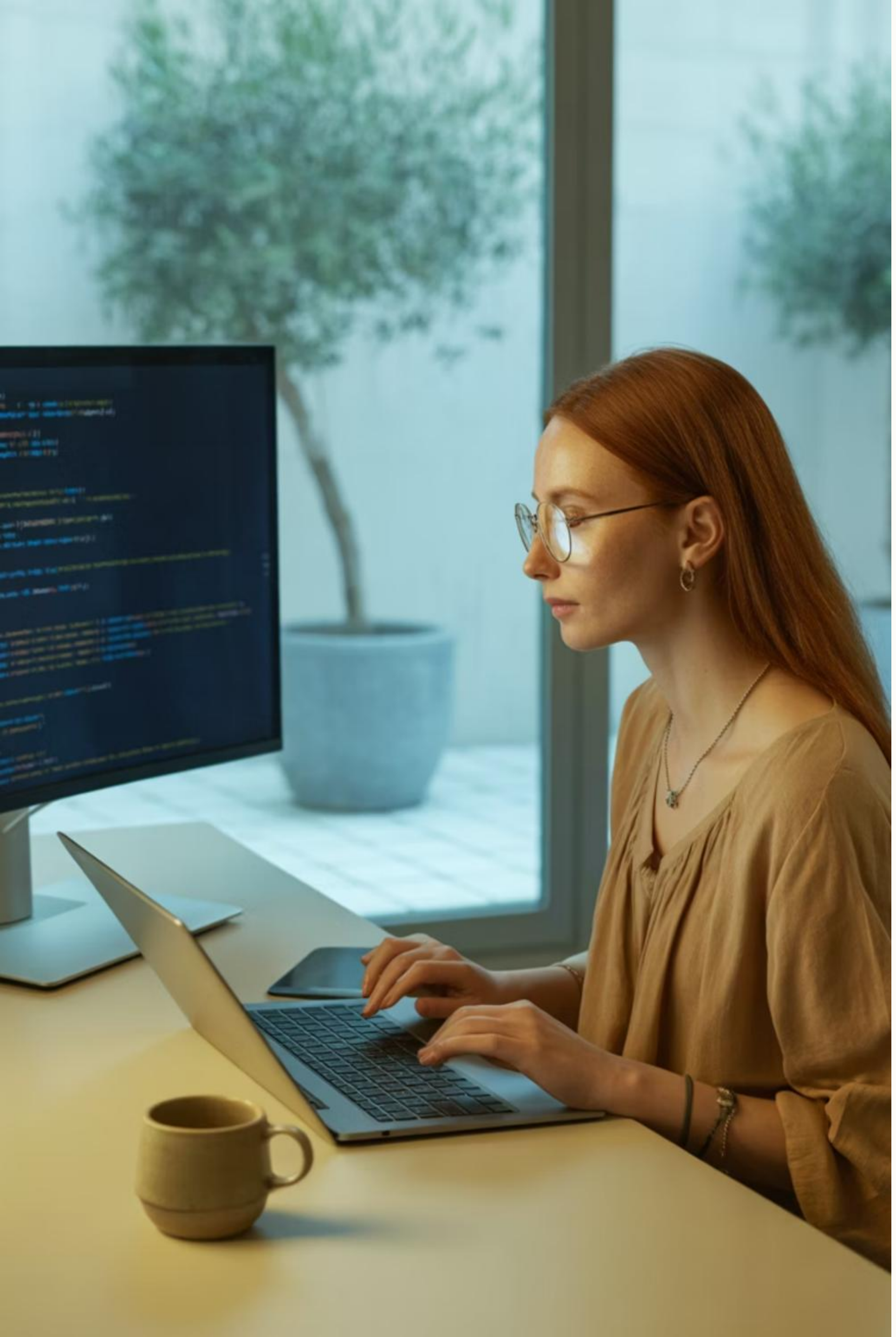
2. **usage_logs.json**

   Nested JSON structure with user sessions, app usage patterns, and device information

3. **complaints.txt**

   Free-form customer complaint descriptions written in natural language

**Discussion point:** Why does each require different processing approaches? What happens if you try to analyze JSON data with SQL alone?



# Session 2: SQL & Python for Data Retrieval and Cleaning

## The Reality of Data Work

Here's a truth that surprises many students: professional data analysts spend 60-70% of their time cleaning and preparing data, not building fancy models. Real-world telecom data arrives messy—with missing values, duplicates, formatting inconsistencies, and human errors.

This session equips you with the essential SQL and Python skills that form the backbone of daily analytics work. You'll learn to query databases efficiently and clean datasets systematically, two skills that employers consistently demand from telecom analysts.

# Essential SQL Operations for Telecom Analytics

1

### SELECT & WHERE

Filter data to find specific records, like customers exceeding usage thresholds

2

### JOIN Operations

Combine customer data with billing records to create complete analytical views

3

### GROUP BY Aggregations

Calculate metrics like average revenue per region or total data consumption

## Practical Exercise 1: High-Usage Customer Query

Write a SQL query to identify customers consuming more than 10 GB of data. This type of analysis helps telecoms identify power users for targeted upgrade campaigns or network capacity planning.

```
SELECT customer_id, data_used_in_GB
FROM usage_data
WHERE data_used_in_GB > 10
ORDER BY data_used_in_GB DESC;
```

Refer - MOP 2 - Exercise 1 – SQLPLUS

**Questions:**

- ➢ Why is SQL still essential even with Python or BI tools available?
- ➢ What are possible errors if data types (numeric vs. string) are mixed incorrectly?
- ➢ How could we automate such queries for daily telecom reports?

# Python Data Cleaning with Pandas

## Why Pandas?

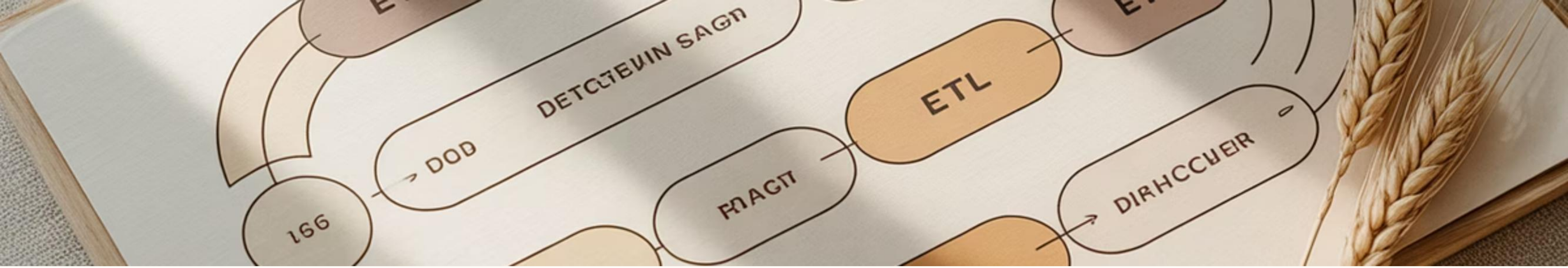
Pandas is the industry-standard Python library for data manipulation. It provides intuitive functions for the most common cleaning tasks: removing null values, eliminating duplicates, and standardizing formats.

In telecom analytics, you'll use Pandas to prepare data extracted from databases before feeding it into visualization tools or machine learning models.

## Exercise 2: Clean a Telecom Usage Dataset

```
import pandas as pd# Load datasetdf =
pd.read_csv("telecom_usage.csv")# Remove rows with missing
valuesdf.dropna(inplace=True)# Eliminate duplicate
recordsdf.drop_duplicates(inplace=True)# Standardize date
formatdf['date'] = pd.to_datetime(df['date'])# Preview
cleaned dataprint(df.head())print(f"Cleaned dataset:
{len(df)} records")
```

Refer - MOP 2 - Exercise 2 – Python




# Session 3: ETL Concepts and Pipeline Building

## What is ETL and Why Does It Matter?

ETL stands for Extract-Transform-Load, the fundamental process of moving data from source systems into analytical databases. In telecommunications, data constantly flows from network equipment, billing systems, customer service platforms, and IoT devices. ETL pipelines automate the collection and preparation of this data for analysis.

Without reliable ETL processes, analysts waste hours manually downloading, cleaning, and combining data. Professional telecoms invest heavily in automated pipelines that deliver clean, integrated data on schedule.

# The Three Stages of ETL

### Extract

Pull data from multiple sources: SQL databases with customer records, CSV files from network monitoring systems, and JSON logs from mobile apps. Telecom data rarely lives in one place, so extraction requires connecting to diverse systems and APIs.

### Transform

Clean, standardize, and combine the extracted data. This includes removing duplicates, handling missing values, converting data types, calculating derived metrics, and joining datasets from different sources into unified analytical tables.

### Load

Store the transformed data in a destination where analysts can access it efficiently. This might be a data warehouse, a CSV file for sharing, or a database optimized for querying and visualization tools.

## Enterprise Tools Overview

Professional telecom companies use tools like Apache Airflow and Apache NiFi to orchestrate complex ETL workflows. While we won't dive deep into these today, understanding that such automation tools exist helps you appreciate the scalability requirements of production analytics.

# Hands-On Mini-Project: Build Your First ETL Pipeline

## Scenario: Customer Complaint Integration

You work for a telecom company that needs to combine customer profile data from a SQL database with complaint records stored in CSV files. Your task is to create a merged dataset that enables analysts to identify which customer segments generate the most complaints.

```
import pandas as pdfrom sqlalchemy import create_engine# EXTRACT: Get customer data from databaseconnection = create_engine('sqlite:///telecom.db')customers = pd.read_sql('SELECT * FROM customers', connection)# EXTRACT: Read complaint CSV filecomplaints = pd.read_csv('complaints.csv')# TRANSFORM: Clean both datasetscustomers.dropna(subset=['customer_id'], inplace=True)complaints.dropna(subset=['customer_id'], inplace=True)# TRANSFORM: Merge on customer_idmerged = customers.merge(complaints, on='customer_id', how='left')# LOAD: Save to new CSV for analysismerged.to_csv('etl_output.csv', index=False)print(f"Pipeline complete! Created dataset with {len(merged)} records.")
```

Refer - MOP 2 - Exercise 3 - ETL pipeline

# SQLAlchemy Engine and PyMySQL

**What is the SQLAlchemy Engine?**

Think of the **Engine** as the **“database gateway”** — it manages:

- the **database connection pool**
- the **dialect** (the database-specific language translator, e.g., MySQL, SQLite, PostgreSQL)
- and executes **SQL commands** sent from your Python code

You typically create the engine **once** in your app and reuse it everywhere you need to talk to the database.

**PyMySQL**: Implements Python’s DB-API for **MySQL only**. Gives you connect(), cursor(), execute().

# Session 4: Predictive Analytics in Telecommunications

## From Descriptive to Predictive

Most analytics begins with describing what happened: how many customers called support last month, what was average revenue per user. Predictive analytics asks a more powerful question: what will happen next?

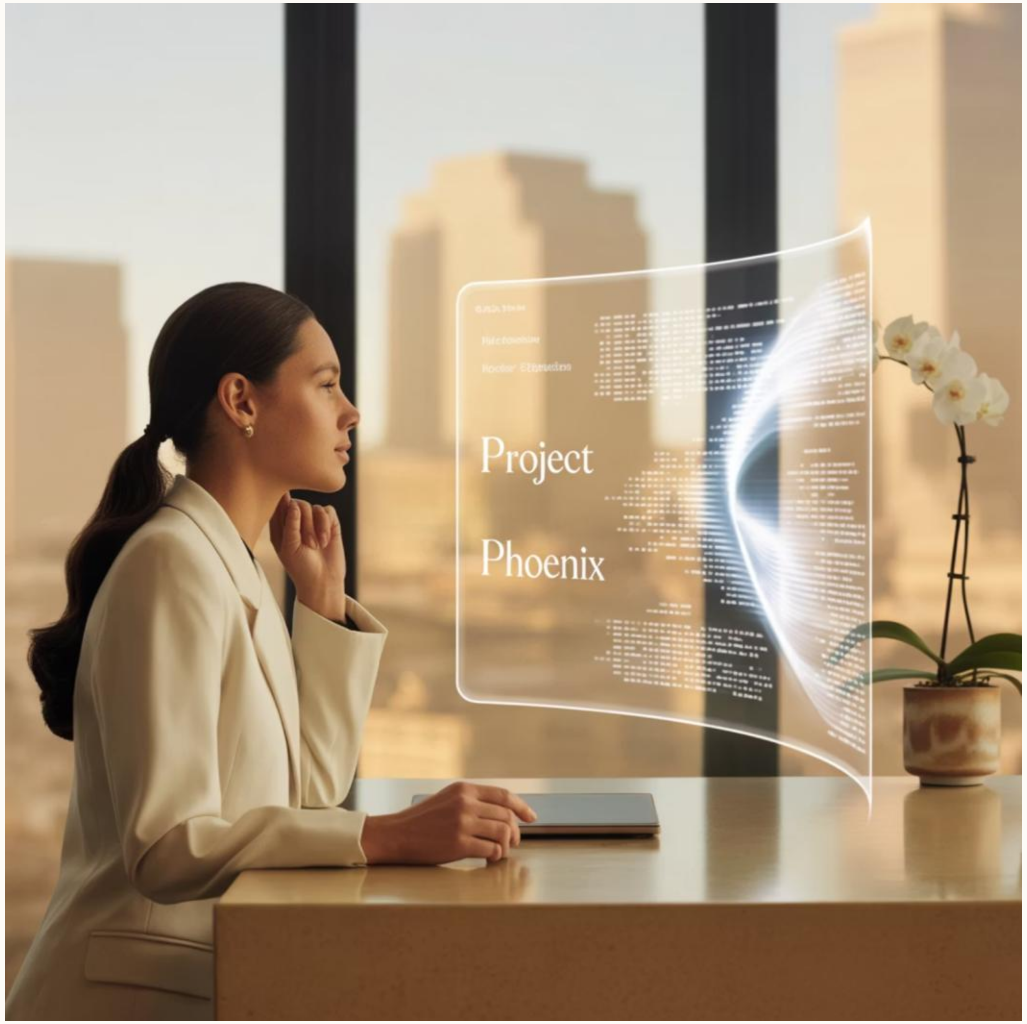
Telecommunications companies use predictive models to anticipate customer behavior, forecast network demand, and optimize operations. The two most common model types are regression (predicting numbers like revenue) and classification (predicting categories like "will churn" or "won't churn").

### Regression Models

Predict continuous numeric outcomes like ARPU (Average Revenue Per User), monthly data consumption, or expected customer lifetime value

### Classification Models

Predict categorical outcomes like whether a customer will churn, which service plan a prospect will choose, or if a network node will fail



# Real-World Telecom Prediction Use Cases

## Churn Prediction

Identify customers likely to cancel their service in the next 30-90 days. By predicting churn early, retention teams can offer targeted incentives or address service issues proactively, significantly reducing customer loss.

## Revenue Forecasting

Predict ARPU (Average Revenue Per User) for upcoming quarters based on usage patterns, service subscriptions, and customer demographics. This helps finance teams set realistic targets and investors understand growth trajectories.

## Network Capacity Planning

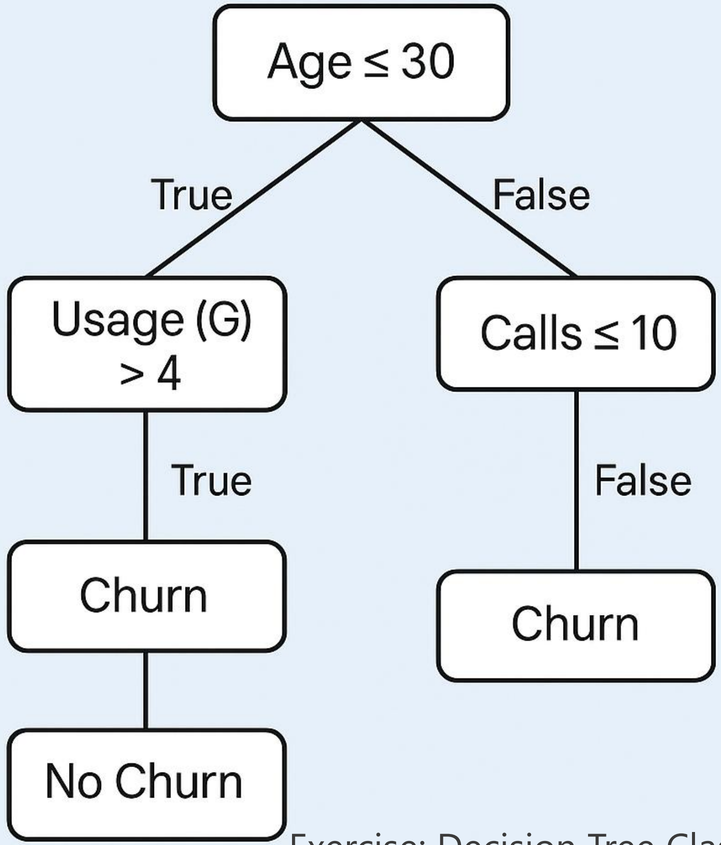
Forecast data usage patterns to determine where network infrastructure investments are needed. Accurate predictions prevent both costly over-building and frustrating service degradation.

## Upsell Targeting

Predict which customers are most likely to upgrade to premium plans based on usage patterns, demographics, and engagement metrics. This enables sales teams to focus efforts on high-probability prospects.

# Building Your First Churn Prediction Model

Customer churn costs telecom companies billions annually. In this exercise, you'll build a simple classification model that predicts which customers might leave based on their age, usage patterns, and complaint history. While real-world models are more sophisticated, this introduces the fundamental workflow.



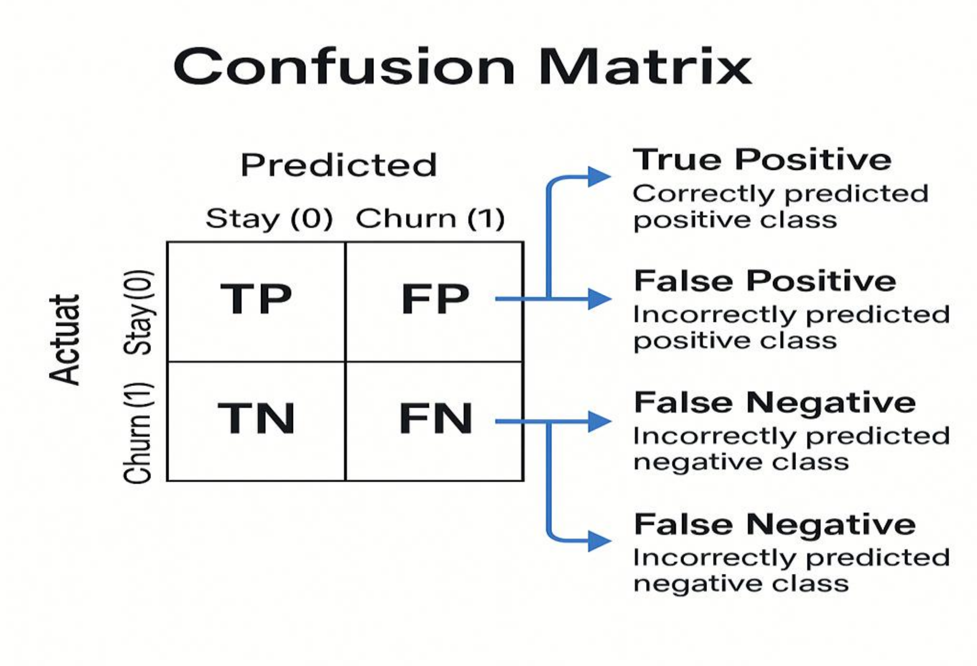

Exercise: Decision Tree Classifier

Refer - MOP 3 - Exercise 4 – Churn Prediction

# Interpreting Your Model Results

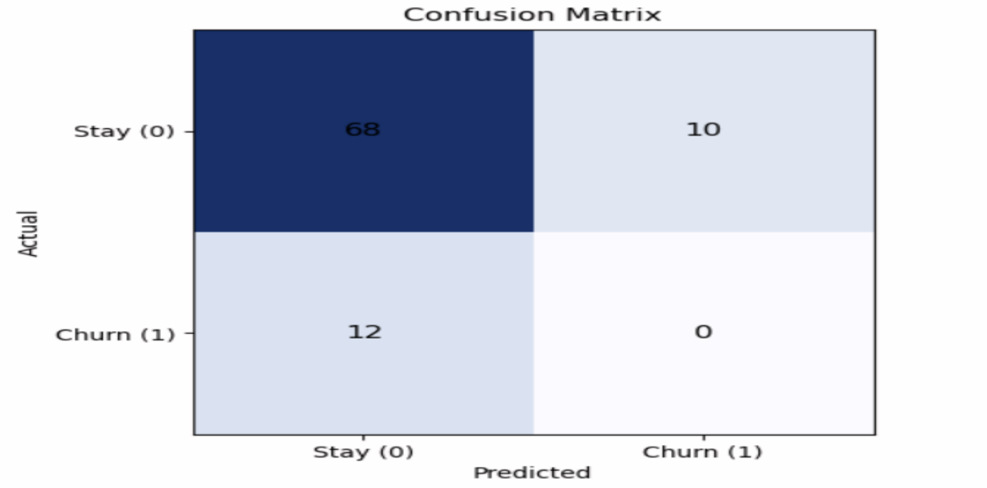
## What Does Accuracy Mean?

If your model achieves 85% accuracy, it correctly predicts churn status for 85 out of 100 customers. However, accuracy alone doesn't tell the whole story. In telecom churn, false negatives (predicting a customer will stay when they actually leave) are often more costly than false positives (predicting churn for loyal customers).




## Business Application

Once deployed, this model would score every customer monthly. Those with high churn probability (say, above 70%) trigger retention workflows: personalized offers, service reviews, or proactive customer support. Companies using churn prediction effectively reduce customer loss by 15-30%, translating to millions in retained revenue.




> "Predictive analytics transformed our retention strategy. Instead of reacting to cancellations, we now prevent them." — Analytics Director, Major US Telecom

# KEY HIGHLIGHTS -

**Using Python**
- Install Pandas Scikit-learn libraries

**EXTRACT - Generating customer dataset**
- With all the details and intentionally missed some values

**TRANSFORM – the dataset**
- Update all the discrepancies
- Add missing values etc.

**LOAD & Training**
- DecisionTreeClassifier is a machine learning algorithm used and trained with the fixed dataset

**PREDICT & EVALUTE**
- Makes predictions using your trained model
- Showcase accuracy of the model
- Display confusion matrix models

**SAVE & REUSE – Model**


# Session 5: Telecom KPIs and Dashboard Design Design

## Why KPIs Matter in Telecommunications

Key Performance Indicators (KPIs) are the vital signs of a telecommunications business. Executives, operations managers, and analysts monitor these metrics constantly to track business health, identify problems early, and make data-driven decisions. Without clear KPIs, companies operate blindly.

In this final session, you'll learn the most important telecom metrics and build an interactive dashboard that brings these numbers to life. Great dashboards don't just show data—they tell stories and enable action.

# Essential Telecom KPIs You Must Know

$45

ARPU: Average Revenue Per Per User

Total revenue divided by active users. Tracks how much value the company extracts from each customer relationship. Growing ARPU indicates successful upselling or premium migrations.

2.3%

Monthly Churn Rate

Percentage of customers who canceled service in a given month. Industry average hovers around 1.5-2.5%. High churn signals competitive pressure or service quality issues.

12

Network Downtime Hours

Total hours of service interruption across the network. Even brief outages frustrate customers and violate service level agreements. World-class telecoms maintain 99.99% uptime.

4.2

CSAT: Customer Satisfaction Satisfaction

Average rating from customer surveys, typically on a 1-5 scale. Strong correlation exists between CSAT scores and customer retention. Scores below 3.5 require immediate attention.

## The KPI Formula Reference

- **ARPU** = Total Monthly Revenue ÷ Number of Active Users
- **Churn Rate** = (Customers Lost in Month ÷ Customers at Month Start) × 100
- **Customer Lifetime Value (CLV)** = ARPU × Average Customer Lifespan in Months

# Hands-On: Build an Executive KPI Dashboard

## Your Mission

Using PowerBI Desktop or Tableau Public (both free), create an interactive dashboard that telecommunications executives could use in Monday morning meetings. You'll work with a provided dataset containing monthly KPIs across different regions.

01

### Connect Your Data

Import the kpi_data.csv file into your visualization tool. Verify that date fields are recognized correctly and numeric fields display proper formatting.

02

### Create Three Core Visualizations

Build: (1) Line chart showing ARPU trend over 12 months, (2) Bar chart comparing churn percentage by region, (3) Column chart displaying downtime hours by month.

03

### Add Interactive Filters

Create slicers for Month and Region so users can dynamically filter the entire dashboard. This interactivity transforms static reports into analytical tools.

04

### Polish Your Design

Apply consistent colors, add descriptive titles, and include a text box explaining what actions should be taken when KPIs exceed thresholds (e.g., "ARPU below $40 requires pricing review").

Use - MOP 3 - Exercise 5 – kpi_data

# Dashboard Design Best Practices

## What Makes Dashboards Effective

- **Prioritize the critical:** Place the most important metric prominently in the top-left, where eyes naturally land first
- **Use color strategically:** Reserve red for problems, green for success. Too many colors create confusion
- **Provide context:** Show trends, not just numbers. A churn rate of 3% is meaningless without knowing if it's increasing or decreasing
- **Enable action:** Every dashboard should answer "so what?" and "what should I do?" Include benchmarks and thresholds
- **Keep it simple:** Executives have 30 seconds. Don't force them to hunt for insights

**Real-world impact:** Well-designed KPI dashboards reduce decision-making time by 40% and improve cross-functional alignment by creating a shared understanding of business performance.

# Congratulations – Completion of Module 1

We have journeyed from understanding basic telecom data types to building predictive models and executive dashboards. You can now distinguish structured data from unstructured formats, write SQL queries and Python scripts for data cleaning, construct ETL pipelines, predict customer churn with machine learning, and visualize KPIs that drive business decisions.

## Your Next Steps

### Practice Daily

Apply these skills to public telecom datasets on Kaggle. Repetition builds mastery.

### Deepen Your Knowledge

Explore advanced topics: time series forecasting, network optimization, big data tools like Spark.

### Build Your Portfolio

Document projects on GitHub. Employers value demonstrated skills over theoretical knowledge.

### Connect with Industry

Join telecom analytics communities, attend webinars, and follow industry trends on LinkedIn.

**Thank you for your engagement and curiosity. The telecommunications industry needs talented analysts who can transform data into strategic advantage. Keep learning, keep building, and keep asking questions!**

# Thank YOU

By ISHAN MARWAH - +91 97178 56777 - Ishan Marwah | LinkedIn