FDP Pre-Read Material

Data Science Pipeline — Hands-on with Google Colab

Speaker: Kapil Panwar

Organization: Gyannetra Pvt Ltd, Pune

Date: 30 May 2026

Venue: JIET, Jodhpur

1. Introduction

Welcome to the **Data Science Pipeline** Faculty Development Program session! This pre-read document will prepare you for a completely hands-on, no-installation-needed workshop.

> **Good News:** You do not need to install Python, Docker, or any software on your laptop. Everything runs in your browser via **Google Colab**.

What is a Data Science Pipeline?

A Data Science Pipeline is the end-to-end process of taking raw data and transforming it into actionable insights. The standard pipeline has these stages:

1. **Data Collection** – Gathering data from CSV, databases, APIs, etc.
2. **Data Cleaning & Preprocessing** – Handling missing values, fixing data types
3. **Feature Engineering** – Creating new features, encoding categorical variables
4. **Exploratory Data Analysis (EDA)** – Visualizing patterns and relationships
5. **Model Training** – Applying machine learning algorithms
6. **Evaluation** – Measuring how well the model performs
7. **Interpretation** – Understanding what the model tells us

> **Session Tip:** Don't worry if some terms feel new. The session is designed for beginners. Just bring curiosity and a working internet connection.

2. Google Colab — Your Browser-Based Lab

Google Colaboratory (Colab) is a free, cloud-based Jupyter notebook environment that runs entirely in your browser. No installation. No setup.

Why Colab for this FDP?

- ✅ **Zero setup** – Works on any laptop (i3, i5, any OS)
- ✅ **Free GPU/CPU** – Google servers do the heavy lifting
- ✅ **Pre-written code** – Just run each cell with Shift+Enter
- ✅ **Auto-saves to your Google Drive**
- ✅ **Works offline** after initial load

How to Open a Colab Notebook

1. Open your browser (Chrome / Firefox recommended)
2. Go to `colab.research.google.com`
3. Click **File → Upload Notebook**
4. Select the `.ipynb` file we share during the session

> **Shortcut:** We will also share a direct Colab link — just click and it opens automatically. No upload needed.

Colab Cell Types

Cell Type	Use	Example
Code Cell	Runs Python code	`print("Hello")`
Text Cell	Markdown documentation	Explanations, headings

3. Categorical Data & One-Hot Encoding

What is Categorical Data?

Categorical data represents groups or categories rather than numbers. Examples: Department (Science, Commerce), Gender (Male, Female), or **Favorite Color** (Red, Blue, Green).

Machine learning models understand only numbers. So we must convert text categories to numbers. This is called **encoding**.

The Problem with Simple Numbering

If we encode Red=1, Blue=2, Green=3, the model might think Green > Blue > Red (i.e., order/ranking). But colors have no inherent order. This is why we use one-hot encoding.

How One-Hot Encoding Works

One-hot encoding creates a separate column for each category, with 1 (present) or 0 (absent).

Original	Red	Blue	Green	Yellow
Red	1	0	0	0
Blue	0	1	0	0
Green	0	0	1	0
Yellow	0	0	0	1

In Python with pandas, it's just one line:

```
import pandas as pd
encoded = pd.get_dummies(df['Favorite_Color'], prefix='Color')
```

> **Visual Memory Aid:** Think of one-hot like a light switch panel — each color has its own switch, and exactly one switch is ON at a time.

4. Session Dataset — Student Stress Prediction

During the session, we will use a dataset about **college student stress levels**. This dataset is specially designed to be relatable for faculty members.

Feature	Type	Description	Example
Department	Categorical	Student's department	Science, Arts, Commerce, Engineering
Attendance Level	Categorical	Attendance category	Regular, Irregular, Borderline
Favorite Color	Categorical	Self-reported favorite color 🌈	Red, Blue, Green, Yellow, Purple, Orange
Weekly Study Hours	Numerical	Hours studied per week	2 – 25 hours
Assignment Submission %	Numerical	Assignment submission rate	40% – 100%
Backlogs Count	Numerical	Number of pending backlogs	0 – 5
Stress Level	Binary (Target)	0 = Low Stress, 1 = High Stress	0 or 1

Session Outline (Hands-On)

1. **Open Colab & Load Data** — 5 min
2. **Explore the Data (EDA)** — 10 min
3. **One-Hot Encode Categorical Columns** — 10 min
4. **Train a Machine Learning Model** — 10 min
5. **Evaluate & Interpret Results** — 10 min
6. **Q&A / Open Discussion** — 10 min

Total: ~55 minutes

> **No prior coding experience needed.** All code is pre-written. You just click "Run" and watch the magic happen.

What You Will Learn

- ✅ How to load and view data in Python
- ✅ How to handle categorical data (one-hot encoding with colors!)
- ✅ How to train a simple ML model
- ✅ How to check if the model is accurate
- ✅ How these skills apply to your own courses & research

5. Quick Reference — Python & Colab Cheatsheet

Common Colab Shortcuts

Action	Shortcut
Run current cell	`Shift + Enter`
Add a code cell	Click **+ Code** button
Add a text cell	Click **+ Text** button
Delete a cell	Click cell → Trash icon (top-right)
Toggle cell output	Double-click output area

Key Python Commands We'll Use

Command	What it does
`df.head()`	Show first 5 rows of data
`df.info()`	Show data types and missing values
`df.describe()`	Show statistical summary
`df['col'].value_counts()`	Count unique values in a column
`pd.get_dummies()`	One-hot encode categorical columns
`train_test_split()`	Split data into train/test
`RandomForestClassifier()`	Create a Random Forest model
`accuracy_score()`	Calculate model accuracy

Before the Session, Please:

- ✅ Ensure you have a working internet connection
- ✅ Have a Google account (Gmail) logged in your browser

- ✅ Open Chrome / Firefox browser (latest version)
- ✅ Keep this pre-read handy for reference

> **See you at the session!** If you have questions beforehand, please reach out via the FDP coordination team. We're here to help.

See You at the Session!

Data Science Pipeline • 30 May 2026 • JIET Jodhpur

Kapil Panwar

Gyannetra Pvt Ltd, Pune, Maharashtra

Gyannetra — India's Premier AI Consultancy