Introduction to Data Science - Practical Sessions

SESSION 1 - Exploratory Data Analysis (EDA) + Data Cleaning

Challenge A - Student Lifestyle & Academic Success

Students receive a dataset containing: study hours, sleep, attendance rate, extracurricular activities, stress level, screen time, and GPA.

Tasks:

- Inspect missing values, outliers, and extreme/unrealistic behaviors.
- Produce full descriptive statistics: mean, median, std, skewness, kurtosis.
- Create at least **6 visualizations**: histograms, boxplots, pairplots, correlation heatmap, distribution curves.
- Identify **5 actionable insights** that may explain GPA differences.
- Perform feature filtering: drop irrelevant features, merge sparse categories.

Challenge B - E-Commerce Transactions Dataset

Dataset contains: product category, quantity, price, timestamp, payment method, customer type.

Tasks:

- Detect anomalies: duplicated rows, negative quantities, timestamps that don't make sense.
- Perform **temporal EDA**: sales per hour/day/month.
- Extract the top 10 best-selling products and categories.

- Create a customer segmentation overview using behavioral features.
- Write a **5-point insight brief** for the marketing team.

Challenge C - Healthcare Patient Monitoring Dataset

Dataset includes: heart rate, glucose level, steps, calories, sleep hours, BMI.

Tasks:

- Clean missing vital signs and unrealistic measurements.
- Visualize distributions of main health indicators (at least 5 plots).
- Investigate correlations between lifestyle variables and health metrics.
- Form hypotheses for future predictive modeling.
- Output a cleaned dataset ready for next sessions.

SESSION 2 - Unsupervised Learning: PCA + Clustering

Challenge A - Nutrition & Food Consumption Patterns

Dataset: foods with nutritional values (protein, fat, carbs, fiber, sodium, calories).

Tasks:

- Normalize all numeric nutrient columns.
- Apply PCA and explain contribution of PC1–PC3.
- Cluster foods into dietary groups (e.g., high-protein, high-carb, balanced).

- Interpret clusters and describe real-world implications.
- Plot PCA biplot and PCA-based cluster scatter.

Challenge B - Smartphone Sensors Dataset

Dataset: accelerometer and gyroscope readings for different human activities.

Tasks:

- Normalize sensor readings.
- Apply PCA to retain 95% of variance.
- Cluster activities: walking, sitting, standing, running.
- Compare **K-means vs Agglomerative Clustering** using silhouette scores.
- Discuss errors and possible improvements.

Challenge C - Social Media Engagement Dataset

Dataset: posts with likes, shares, comments, view duration, retention rate.

Tasks:

- Apply PCA on engagement metrics.
- Cluster posts into performance groups.
- Identify which metrics drive separation.
- Visualize clusters in PCA component space.
- Propose recommendations for content optimization.

SESSION 3 - Supervised Learning: Regression

Challenge A - Predicting Music Popularity

Dataset: tempo, loudness, danceability, energy, acousticness, release year, popularity score.

Tasks:

- Clean and preprocess features.
- Normalize selected features.
- Fit a linear regression model and evaluate using RMSE/MAE.
- Fit a second model using interaction terms or polynomial features.
- Interpret coefficients: what makes a song popular?

Challenge B - Predicting Airline Flight Delays

Dataset: flight number, origin, destination, carrier, weather factors, scheduled time, actual delay.

Tasks:

- Perform EDA: busiest airports, delay distributions, weather impact.
- Encode categorical data: airport, carrier, month.
- Train a regression model to predict delay time.
- Evaluate the model under multiple train–test splits.
- Provide operational recommendations to reduce delays.

Challenge C - Predicting Vehicle CO_2 Emissions

Dataset includes: engine size, cylinders, fuel type, weight, horsepower, CO_2 output.

Tasks:

- Normalize and select relevant features.
- Train linear regression and compare with polynomial regression (degree 2 or 3).
- Plot residuals and check model assumptions.
- Identify top factors affecting CO_2 emissions.

SESSION 4 - Supervised Learning: Classification

Challenge A - Workplace Attrition Prediction

Dataset: employee demographics, satisfaction, overtime, salary, years at company, attrition (Yes/No).

Tasks:

- Preprocess: missing values, scaling, encoding.
- Train logistic regression + a decision tree classifier.
- Compare performance using accuracy, F1, recall.
- Interpret the decision tree structure.
- Suggest HR actions based on findings.

Challenge B - Fake News Detection

Dataset: news headlines + labels (real / fake).

Tasks:

- Clean text: punctuation removal, lemmatization, stopword removal.
- Transform text using **TF-IDF vectorization**.
- Train logistic regression classifier.
- Train decision tree classifier as comparison.
- Identify and interpret most important words/features.

Challenge C - Early Disease Prediction

Dataset: patient age, glucose level, BMI, blood pressure, family history, disease label (binary).

Tasks:

- Clean and preprocess dataset.
- Normalize features where needed.
- Train logistic regression and decision tree.
- Extract feature importance.
- Interpret results in simple medical terms.
- Add a short note on ethical and communication considerations.

FINAL MINI-PROJECT

Group Project - Full Mini Data Science Pipeline (2-3 students)

Students choose a dataset from: music, sensors, e-commerce, health, or environment.

They must:

- Perform **advanced EDA** (≥12 visualizations + correlation study).
- Apply PCA and interpret top components.
- Apply clustering and justify number of clusters.
- Apply one supervised model (regression or classification).
- Submit a **2-page written summary**.
- Deliver a **5-minute presentation** of results.