Capstone Project (Data Science)

Why a Capstone Project?

This capstone project brings together all the concepts you have learned into a single, complete data science workflow.
Rather than treating supervised learning, unsupervised learning, and time series as isolated topics, this project requires you to decide which approach is appropriate for a given real-world problem and justify that decision.

In professional data science work, problems are not pre-labeled.
You are expected to analyze the data, understand its structure, and choose a method that fits the problem.

This capstone evaluates your ability to:

- Frame a real-world problem
- Choose an appropriate analytical approach
- Justify modeling decisions using data
- Communicate insights clearly and correctly

Group-Based Project Structure

Students will be assigned into groups for this capstone project.
Each group will work as a single team throughout the duration of the project and submit one unified solution.

This structure reflects real-world data science environments, where projects are completed collaboratively rather than individually.

Why Collaboration Is Required

Data science is a collaborative discipline.
Most errors in analysis arise not from coding mistakes, but from incorrect assumptions or poorly chosen approaches that go unchallenged.

Working in groups encourages:

- Discussion and critical evaluation of ideas
- Shared responsibility for decisions
- Clear explanation of technical choices
- Stronger, more defensible conclusions

  A successful capstone project demonstrates collective understanding, not isolated individual contributions.

General Instructions

Each group must:

1. **Choose ONE track:**
    - Supervised Learning
    - Unsupervised Learning
    - Time Series Analysis

Dataset Sourcing & Justification (Mandatory For All Groups)

1. Dataset Source (Required)

Each group must source their own dataset from a public and credible platform. Manually created or synthetic datasets are NOT allowed. Also, you cannot use the datasets from class

Approved dataset sources:

- Kaggle (https://www.kaggle.com/datasets)
- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
- Government open data portals (data.gov, data.gov.uk, etc.)
- World Bank Open Data (https://data.worldbank.org/)
- IMF & national statistics offices
- Financial datasets (Yahoo Finance, FRED, Quandl)
- Social science and health research repositories
- Environmental and climate data (NOAA, NASA, etc.)

Synthetic or artificially created datasets will result in automatic rejection of the project.

Data Preprocessing (Mandatory For All Tracks)

Data preprocessing is the foundation of any successful machine learning project. Poor preprocessing leads to poor models, regardless of how sophisticated your algorithms are. Every group must thoroughly document and justify all preprocessing steps.

Exploratory Data Analysis (Eda) - Mandatory For All Tracks

Exploratory Data Analysis is the critical bridge between raw data and modeling. A thorough EDA informs your modeling choices, reveals data quality issues, and provides business insights. Every group must conduct a detailed EDA and present findings in a dedicated report section.

Why EDA Is Essential

- Data Understanding: You cannot build a good model if you don't understand the data.
- Problem Validation: Confirms that your chosen approach is feasible and appropriate.Feature Insights: Identifies which features are most informative for your target.
- Data Quality Issues: Discovers missing values, outliers, and inconsistencies before they break your model.
- Assumption Checking: Validates assumptions required by your chosen model.
- Stakeholder Communication: EDA findings drive business decisions and model design choices.

Mandatory EDA Components (All Tracks)

Every group must include the following in their EDA:

- **Dataset Overview**: Rows, columns, data types, memory usage, basic statistics (mean, median, std, min, max for numerical features).
- **Missing Values Analysis**: Visualize missing data patterns using heatmaps or bar charts. Identify which features have missing data and how much.
- **Numerical Features**: For each numerical feature: histogram, distribution shape (skewness, kurtosis), box plot for outliers, descriptive statistics.
- **Categorical Features**: For each categorical feature: value counts, bar chart showing class distribution, percentage breakdown.
- **Univariate Analysis**: Analyze each feature individually: ranges, distributions, concentrations, unusual patterns.
- **Bivariate & Multivariate Analysis**: Examine relationships between features: correlation matrix with heatmap, scatter plots, cross-tabulations.
- **Target Variable Analysis (Supervised Only)**: For regression: histogram of target, check for outliers, skewness. For classification: class distribution, check for imbalance.

Track 1: Supervised Learning

Additional Mandatory Choice

Groups choosing Supervised Learning must further choose ONE:

- Regression
- Classification

A. REGRESSION TRACK

Objective: Predict a continuous numerical value and explain the factors influencing it.

Minimum Requirements

At least 1 baseline model:

- Linear Regression

At least 1 advanced model:

- Decision Tree / Random Forest
- Feature scaling where necessary

Evaluation Metrics

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 (Coefficient of Determination)
- Percentage error

Required Analysis

- Actual vs Predicted plot
- Residual analysis (checking for patterns, normality)
- Feature importance/coefficients interpretation
- Model comparison and justification

B. CLASSIFICATION TRACK

Objective: Predict a category or class label and analyze decision boundaries.

Minimum Requirements

At least 1 baseline model:

- Logistic Regression

At least 1 advanced model:

- Decision Tree classifier / Random Forest classifier
- Class imbalance handling (if present)

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Required Analysis

- Feature importance or coefficient interpretation
- Error analysis (false positives vs false negatives)
- Confusion matrix interpretation
- Class imbalance strategies and their impact

Rules (Strict But Fair)

- Regression must not use accuracy
- Classification must not use RMSE
- Metric misuse = automatic penalty

Track 2: Unsupervised Learning

Focus: K-Means Clustering

Objective: Discover natural groupings within unlabeled data. Identify hidden patterns and segment customers, products, or observations into meaningful clusters without pre-defined labels.

When to Choose Unsupervised Learning

- No labeled target variable exists
- You want to discover hidden patterns
- Customer segmentation or market analysis

Minimum Requirements for K-Means

- Elbow method analysis to determine optimal number of clusters (k)
- Silhouette analysis or other cluster validation method
- K-Means model implementation with optimal values of k
- Cluster interpretation and profiling

Required Analysis

- Elbow curve showing inertia vs number of clusters
- Silhouette plot for chosen k value
- Cluster size distribution
- Feature values comparison across clusters (cluster profiles/centroids)
- Business interpretation: What do these clusters represent in real-world terms?

Track 3: Time Series Analysis

Objective: Analyze and forecast patterns in data that changes over time. Identify trends, seasonality, and cyclical patterns to make predictions about future values.

When to Choose Time Series Analysis

- Data has a temporal component (daily, monthly, yearly observations)
- You need to forecast future values (stock prices, sales, demand, temperature)
- Analyzing trends and seasonality is important
- Comparing performance across time periods

Minimum Requirements

- Time series decomposition (trend, seasonality, residual)
- Stationarity testing (Augmented Dickey-Fuller test)
- Differencing or transformation if data is non-stationary
- At least 1 advanced model: ARIMA or SARIMA (preferred)
- Train-test split with proper time-based splitting (not random)

Evaluation Metrics

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)

Required Analysis

- Time series plot showing original data with trend line
- Decomposition plot (trend, seasonality, residuals)
- ACF (autocorrelation) and PACF (partial autocorrelation) plots - essential for ARIMA selection
- Stationarity test results and interpretation
- Actual vs Predicted plot with forecast horizon highlighted
- Residual analysis: check for white noise properties, autocorrelation
- Model diagnostics

Advanced Models

ARIMA (AutoRegressive Integrated Moving Average):

- ARIMA(p,d,q) where p=AR terms, d=differencing, q=MA terms
- Requires stationarity (check ADF test)
- Use ACF/PACF plots to determine p and q
- SARIMA extends ARIMA to handle seasonality

Deliverables & Deadlines

1. Submission link (Capstone)

2. Submission Format

- **Single Notebook Submission:**
  Each group must submit one notebook containing:
  - Names and email addresses of all group members
  - Description of the chosen track and dataset
  - All code, analysis, visualizations, and results
  - Link to each group member's GitHub repository

GitHub Repository:

Each group member must create a repository named:

`TS_Academy_Capstone_Project`

    - The repository will contain all files related to the project
    - The notebook must include the link to each member's repository

3. Submission Deadline

- **Final Capstone Submission: 16 March 2026**
- **The final submission includes:**
    - **Single notebook containing all code, analysis, visualizations, and results**
    - **Names and emails of all group members**
    - **Links to each group member's GitHub repository**

No submission will be considered complete without all these components.