**Capstone Project 1 Details**

**1. Project Title:**  
Data Science Capstone Project: Manufacturing Equipment Output Prediction with Linear Regression

**2. Category:**  
Supervised Learning (Regression)

**3. Problem Statement:**

You are working as a data analyst for a manufacturing company that operates injection molding machines to produce plastic components. The company wants to optimize production efficiency by predicting the hourly output (number of parts produced per hour) based on various machine operating parameters. Your task is to build a linear regression model that can predict machine output using factors like temperature, pressure, cycle time, and material properties. This will help the production team optimize machine settings, plan production schedules, and identify when machines are underperforming.

**4. Objective:**

* **Build a linear regression model** to predict hourly machine output using operating parameters like temperature, pressure, cycle time, and material properties.
* **Use predictions** to optimize machine settings, improve production efficiency, plan schedules, and detect underperforming machines.

**5. Dataset Information:**

* **Dataset Name/Source:** manufacturing\_dataset\_1000\_samples
* **Size & Features:**
  + No of records: 1000
  + Target variable: Parts\_Per\_Hour
  + Features: 17 Features + 1 Class Label (Parts\_Per\_Hour)
* Features are Injection Temperature, Injection Pressure, Cycle Time, Cooling Time, Material Viscosity, Ambient Temperature, Machine Age, Operator Experience, Maintenance Hours, Shift, Machine Type, Material Grade, Day\_of\_Week, Temperature\_Pressure\_Ratio, Total\_Cycle\_Time, Efficiency Score, Machine Utilization
* **Pre-processing Required:**

**1. Remove/Process Unnecessary Columns**

* **Timestamp** is not directly useful for regression unless you extract features like hour, day, or month.
* Drop it or convert to useful time-based features.

**2. Handle Missing Values**

* Columns with missing values:
  + Material\_Viscosity (20 missing)
  + Ambient\_Temperature (20 missing)
  + Operator\_Experience (20 missing)
  + Options:
  + **Numerical:** Fill with mean/median.
  + **Categorical:** Fill with mode.

**3. Encode Categorical Variables**

Categorical columns:

* Shift (Day/Night/Evening)
* Machine\_Type (Type\_A, Type\_B)
* Material\_Grade (Economy, Standard, Premium)
* Day\_of\_Week (Monday–Sunday)  
  → Use **One-Hot Encoding** or **Label Encoding**.

**4. Feature Scaling**

* Since linear regression is sensitive to feature scales, apply **StandardScaler** or **MinMaxScaler** to numerical features:
  + Example: Injection\_Temperature, Injection\_Pressure, Cycle\_Time, etc.

**5. Define Target Variable**

* **Target:** Parts\_Per\_Hour (hourly output)
* **Features:** All other columns after preprocessing.

**6. Train-Test Split**

* Split dataset into training and test sets (e.g., 80%-20%).

**7. Tools & Technologies:**

Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, SciPy, FastAPI (for deployment), Docker (for containerization)

**8. Evaluation Criteria:**

* **Primary Metrics:** Recall (Sensitivity), Precision, F1-score, RMSE, MSE
* **Secondary Metrics:** Accuracy
* Model Evaluation report and recommendations
* Quality of EDA and feature analysis
* Clarity and professionalism of presentation and documentation

**8. Expected Deliverables:**

* **Code Files:** Python scripts/notebooks implementing EDA, model training and evaluation
* **Documentation/Report:** Detailed project report with insights, methodology, and results
* **Presentation Slides:** No of slides 15-20

**9. Timeline:**

Day 1: Data exploration, EDA, preprocessing

Day 2: Baseline model development & evaluation

Day 3: Model Evaluation, final evaluation, Deployment (FastAPI + Docker),

Day 4: final report & presentation