Capstone Project1

MANUFACTURING EQUIPMENT OUTPUT PREDICTION

WITH LINEAR REGRESSION

Team Members

T120100312: Sharon A Dobbin
T120100316: Sherin Nayana B
T120100317: Prithvi Prabhu Pani V
T120100318: P Shreya
T120100324: Tejushree R
T120100330: Varshitha Sunnampalli

Table of Contents

Problem Statement

- The main problem focused on this project is related to optimizing manufacturing efficiency.
- Manufacturing industries often face inefficiencies due to variations in process parameters like temperature, pressure, and cycle time.
- Goal: Build a predictive model that analyzes machine and process data to estimate Efficiency Score and improve manufacturing outcomes. This prediction will help the industries to optimize their production settings, reduce downtime, and maintain consistent product quality.

Overview of the Project

This project falls under the domain of Smart Manufacturing or Industrial AI.
The idea is to create a complete end-to-end ML pipeline — that predicts machine efficiency and visualizes performance metrics.

The project uses a synthetic dataset with 1000 records and 19 features that represent real manufacturing conditions.

The tech stack includes:

- ❑ Python for data analysis and modeling
- ❑ Scikit-Learn for ML algorithms
- ❑ Matplotlib/Seaborn for data visualization
- ❑ The final outcome is a model that can predict the machine efficiency.

Data generation & Loading

- File: manufacturing_dataset_1000_samples.csv
- Rows: 1000, Columns: 19
- Key columns: Injection_Temperature, Injection_Pressure, Cycle_Time, Cooling_Time, Material_Viscosity, Ambient_Temperature, Machine_Age, Operator_Experience, Maintenance_Hours, Shift, Machine_Type, Material_Grade, Efficiency_Score, Machine_Utilization, Parts_Per_Hour, Total_Cycle_Time, Timestamp
- Missing values: Material_Viscosity, Ambient_Temperature, Operator_Experience — 20 missing values each (980 non-null).
- The dataset was loaded using the Pandas read_csv() function

```
#Sample code
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('manufacturing_dataset_1000_samples.csv')
df.info()
```

Data Exploration & Understanding

Definition:
Data exploration and understanding is the process of **looking at your data carefully** to know what it contains and how it is structured before doing any analysis. Think of it like **getting to know your ingredients before cooking**.

Why it is important:

- You can **see if the data has missing or wrong values**.
- You can **know what type of data each column has** (numbers, text, dates).
- You can **understand the patterns, ranges, and distributions** of your data.

What we usually do in this step:

1. Check dataset size – How many rows (records) and columns (features) are there?
2. Check data types – Are they numbers, text, or dates?
3. Get basic statistics – For numbers: average, min, max. For text: most common values, unique values.

Goal:

- To **understand the data completely** so we can clean it, process it, and use it for analysis or modeling without mistakes.

Exploratory Data Analysis

Goal:

Discover patterns and relationships in manufacturing data

- Understand how machines, settings, and operations affect production.

What we do in EDA for this project:

1. Histograms for all variables

1. Show how each variable (like machine speed, temperature, etc.) is distributed.
2. Helps to **see trends, ranges, and unusual values**.

2. Correlation matrix

1. Shows **how variables are related** to each other.
2. Helps identify which factors **influence production most**.

3. Scatter plots between key parameters and output

1. Visualize **how changing a machine setting affects production**.
2. Helps find patterns or trends.

4. Analyze machine settings vs production

1. Check **which settings lead to better output**.
2. Helps understand **performance of each machine**.

5. Identify optimal operating ranges

1. Determine the **best values for settings** to maximize production.
2. Helps improve efficiency and reduce errors.

Data Preprocessing & Feature Engineering

Step: Data Preprocessing and Feature Engineering
Goal:
Prepare data so that the model can learn efficiently and give accurate results.
What we do in this step:
1. Handle outliers using IQR method
1. Outliers are values that are **very different from others** (too high or too low).
2. The **IQR (Interquartile Range) method** helps to detect and remove these unusual values.
3. This prevents the model from being **misled by extreme values**.
2. Create derived features
1. Sometimes, combining existing columns creates **new features that are more meaningful**.
2. Example: Efficiency = Output / Machine Runtime
3. These **new features can improve model performance**.
3. Scale features if necessary
1. Some models work better when all features are **on the same scale**.

Model Building & Training

AIM : To Develop a **Linear Regression model** to predict *Parts_Per_Hour* using production parameters, after proper data cleaning and preparation.

Workflow :

Feature & Target Selection:
Extracted Parts_Per_Hour as target; all others as predictors.
Train-Test Split (80/20):
Ensured reproducibility using random_state=42.
Data Preprocessing:
- Handled missing numeric values → Mean imputation
- Handled missing categorical values → Most frequent value

Model Selection:
Linear Regression chosen for interpretability & simplicity.
Training:
Fitted model using training data (X_train, y_train).
Prediction:
Generated predictions for test data (y_pred).

Model Evaluation & Performance Analysis

R^2: The proportion of variance in the target variable explained by the model.
MSE: The average of the squared differences between predicted and actual values.
RMSE: The square root of MSE, representing the average error in the original units.
MAE: The average of the absolute differences between predicted and actual values.

Visual Evaluation

Residual Plot

X-axis: Predicted values
Y-axis: Residuals (y_true - y_pred)
Goal: Random scatter around 0 → model assumptions hold

Predicted vs Actual Plot

X-axis: Actual values
Y-axis: Predicted values
Goal: Points close to diagonal line = high accuracy

Manufacturing Insights and Feature Interpretation

- **Feature Impact Analysis Results :** This shows the **impact strength and direction** of every feature in your Linear Regression model.
- **Optimal Operating Range Identification:**
  - These ranges represent **safe, efficient, and stable operating conditions.**
  - Used **Interquartile Range (IQR)** to determine stable performance ranges for key manufacturing variables.
  - IQR = Middle 50% of data (between 25th and 75th percentile)
- **Top Parameters Driving Production Output:**
  - Identified the top 3 parameters most strongly influencing output using model coefficients.
  - These parameters form the **primary focus** for process optimization.
- **Production Optimization Recommendations:**
  - Focused on the **top 3 most impactful parameters** from the regression model.
  - Extracted **optimal operating ranges** (based on IQR analysis).
  - These ranges define **efficient and stable working zones** for process tuning.

Conclusion

This project successfully demonstrates how data analysis and machine learning can be applied to understand and improve manufacturing efficiency.

It provides valuable insights into how process and environmental parameters influence productivity, supporting data-driven decisions in industrial operations.

TNS INDIA FOUNDATION
Partners in Economic Transformation

Thank you