## 

## **Phase-2 Submission – Data Analytics**

**Student Name:** A. Andro Jerslin Jebina  
**Register Number:** 212923205004  
**Institution:** St. Joseph College of Engineering  
**Department:** B. Tech Information Technology  
**Date of Submission:** 07-05-2025  
**GitHub Repository Link:** [Andro-jebina/Naan-Mudhalvan](https://github.com/Andro-jebina/Naan-Mudhalvan)

### **1. Problem Statement**

This project addresses the challenge of evaluating employee performance and productivity using data-driven methods. Many organizations still rely on subjective evaluations, which can lead to unfair assessments and reduced employee motivation. By applying workforce analytics, we aim to analyze patterns in performance-related data such as task completion, attendance, and feedback scores.

This helps companies make better decisions in promotions, training, and resource planning. The analysis supports real-world HR decisions and improves overall efficiency. The project mainly uses **descriptive** analytics (to summarize past data) and **diagnostic** analytics (to find causes behind performance trends).

### **2. Project Objectives**

The goal of this project is to use data to evaluate employee performance and find trends in productivity. We aim to answer questions like:

* What are the patterns in employee performance?
* What factors affect productivity the most?
* Are there signs of burnout or low engagement?

Using data analysis, we will deliver insights, trend summaries, and recommendations to improve performance. After exploring the data, we slightly refined our focus to understand **why** productivity drops and how to fix it.

### **3. Flowchart of the Project Workflow**

**Data Collection**  
 Gather employee-related data (attendance, tasks, feedback, KPIs)

### 

**Data Cleaning  
Handle missing values, remove duplicates, format data correctly**

### 

**Exploratory Data Analysis (EDA)  
 Explore patterns, distributions, and correlations in performance metrics**

**Feature Selection  
 Identify key factors affecting productivity (e.g., working hours, feedback)**

**Insight Extraction  
 Analyze trends, find top/bottom performers, detect anomalies**

**Visualization**

**Use graphs and dashboards to display**

**findings clearly**

****

**Reporting & Recommendations  
 Summarize insights and suggest actions**

**for HR/management**

**4. Data Description**

For this project, we used a dataset named "Employee Performance and Productivity Dataset" sourced from Kaggle.

* **Data Type:** Structured
* **Number of Rows and Columns**: 1,00 rows × 12 columns
* **Dataset Nature:** Static (data does not change in real time)

**Key Fields Relevant to the Problem:**

* Employee\_ID – Unique identifier for each employee
* Department – Department the employee belongs to
* Monthly\_Hours\_Worked – Total hours worked per month
* Performance\_Score – Overall performance rating
* Number\_of\_Projects – Tasks or projects handled
* Last\_Evaluation\_Score – Recent performance evaluation
* Absenteeism\_Days – Number of days absent
* Salary\_Level – Low, Medium, High

### 

### **5. Data Preprocessing**

### To ensure accurate analysis, we performed the following data cleaning and preparation steps:

### Handling Missing Values: Missing values in columns like Last\_Evaluation\_Score and Absenteeism\_Days were filled using mean imputation, as these are numerical and had only a few missing entries.

### Removing Duplicates: Checked for and removed duplicate rows using Employee\_ID to avoid double-counting performance data.

### Formatting and Parsing: Dates (if any) were converted into standard datetime format, and numeric fields were ensured to be of proper data types (int/float).

### Encoding Categorical Variables: Columns like Department and Salary\_Level were encoded using Label Encoding for model compatibility and One-Hot Encoding when needed for visualization clarity.

### Outlier Detection and Treatment: Outliers in Monthly\_Hours\_Worked and Absenteeism\_Days were detected using IQR method. Extreme outliers were either capped or removed to prevent skewing the analysis.

### Transformations: Created new fields like Efficiency\_Score = Performance\_Score / Monthly\_Hours\_Worked to better reflect productivity. These transformations helped in uncovering deeper insights.

### 

* **6. Exploratory Data Analysis (EDA)**
* **Univariate Analysis:**

Used histograms and bar charts to explore variables like Performance\_Score, Monthly\_Hours\_Worked, Department, and Salary\_Level.

*Insight:* Most employees are from Sales and Technical departments with medium salary levels.

* **Bivariate/Multivariate Analysis:**
  + Heatmaps showed a strong correlation between Last\_Evaluation\_Score and Performance\_Score.
  + Box plots revealed that high performers often work more hours and handle more projects.
  + Pair plots helped identify patterns across departments and salary levels.
* **Key Insights:**
  + High evaluation scores and project count are linked to better performance.
  + Employees with low salary levels often show lower engagement.
  + Some overworked employees still perform poorly, suggesting burnout.

### 

**7. Tools and Technologies Used**

* **Programming Language: Python**
* **Notebook/IDE: Google Colab, Jupyter Notebook**
* **Libraries Used:**
  + **pandas, numpy – For data handling and processing**
  + **matplotlib, seaborn, plotly – For data visualization**
* **Optional Tools:**
  + **pandas-profiling – For quick automated EDA reports**

**These tools helped efficiently clean, explore, and visualize the data for performance analysis.**

### **8. Team Members and Contributions**

|  |  |
| --- | --- |
| **Name** | **Contribution** |
| * 1. Andro Jerslin Jebina | Data Cleaning, EDA. |
| J. Janani | Data Collection, Visualization, Insights |
| K.Prithika | Documentation, Flowchart Design, Presentation |