Employee Dataset Feature Engineering Project

Group Member Names & Roll Numbers:

Kunal Parihar – 243301018

Mehtab Bano – 243301021

Mohit Acharya – 243301023

Nikhil Bhati – 243301029

Nikhil Daiya – 243301030

Sakshi Kalla – 243301045

Course Code & Name: CCMCA311 – Data Science

Faculty Guide Name: Dr. Seema Loonkar

Semester & Year: MCA III Semester - 2025

Abstract

This project focuses on improving the quality of an employee dataset by performing various feature engineering steps. Real datasets often contain issues such as missing values, errors, inconsistent formats, and outliers, which must be fixed before any machine learning work. This project applies simple and clear data cleaning methods including filling missing values, correcting wrong entries, removing duplicates, and handling outliers. After cleaning, different transformation techniques such as encoding, scaling, log transformation, and salary discretization are applied to prepare the dataset for further analysis. PCA is also used to reduce the number of features while keeping most of the important information. The dataset used is a synthetic CSV file containing more than 1000 employee records with details like age, salary, education, experience, and job role. The final output is a clean and organized dataset suitable for machine learning. This project helps in understanding how feature engineering improves data quality and makes it ready for analysis.

Introduction

Feature engineering is one of the most important stages in any data science project. Raw datasets often contain mistakes and missing information that need to be fixed before the data can be used. In this project, we work on an employee dataset that contains attributes such as age, salary, experience, education, and job role. The goal is to clean and transform the data so that it becomes accurate, consistent, and ready for machine learning.

The main objectives of this project are:

- - Clean and correct errors in the dataset
- - Transform categorical and numerical features
- - Remove unnecessary and repeated columns
- - Apply PCA for dimensionality reduction

Dataset Description

The dataset used in this project is a CSV file created specifically for learning and practice. It has over 1000 employee records with features such as Employee ID, Name, Gender, Age, Experience Years, Salary, Bonus, Overtime Hours, Department, Job Role, City, State, Country, and Attrition. The dataset includes both numerical and categorical values, making it suitable for testing multiple feature engineering techniques. Some values are purposely kept incorrect or missing so that proper cleaning methods can be applied.

Methodology

The project follows these major steps:

1. Data Cleaning:

- - Handling missing values using mean, median, and mode
- - Fixing unrealistic values like wrong age or experience
- - Converting INR salary values to USD
- - Removing duplicate rows
- - Treating outliers using the IQR method

2. Data Integration

    - - Standardize column names
    - - Resolving salary unit conflicts

3. Data Transformation:

- - Label Encoding and One-Hot Encoding
- - Scaling using Min-Max and Standard Scaler
- - Log transforming skewed features
- - Creating salary categories using discretization

4. Data Reduction:

- - Removing unnecessary columns like Name and Joining Date
- - Applying PCA to reduce dimensionality

Tools Used: Python, Pandas, NumPy, Scikit-learn

Results & Analysis

After applying all the preprocessing steps, the dataset became clean, consistent, and ready for machine learning. Missing values were filled, inconsistent salary units were corrected, and noisy entries were fixed. Outliers were handled properly using the IQR method. Categorical data was encoded into numerical form, and numerical data was scaled to improve consistency. PCA helped reduce multiple features while keeping most of the important information. This processed dataset is suitable for building models like employee attrition prediction, salary prediction, or performance analysis.

Conclusion & Future Scope

This project shows how important data cleaning and feature engineering are in any data science work. By applying simple techniques, we were able to fix missing values, correct

errors, handle outliers, apply encoding, scaling, and PCA. The final dataset is clean and ready for machine learning. In the future, this project can be extended by building predictive models, performing deeper analysis, or adding more advanced feature selection and visualization techniques.

References

-Lab knowledge

- ChatGPT

- Online learning resources

Appendix

GitHub Repository Link: https://github.com/nikhildaiya/feature-engineering-project