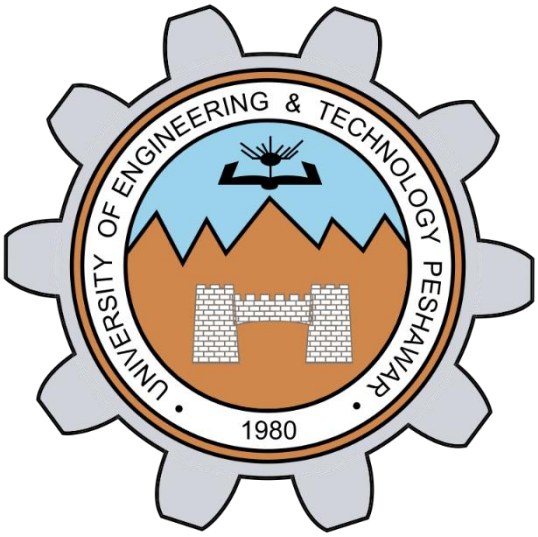
Data Science

(Mini Project) UNIVERSITY OF ENGINEERING AND TECHNOLOGY, PESHAWAR PAKISTAN

Main Campus

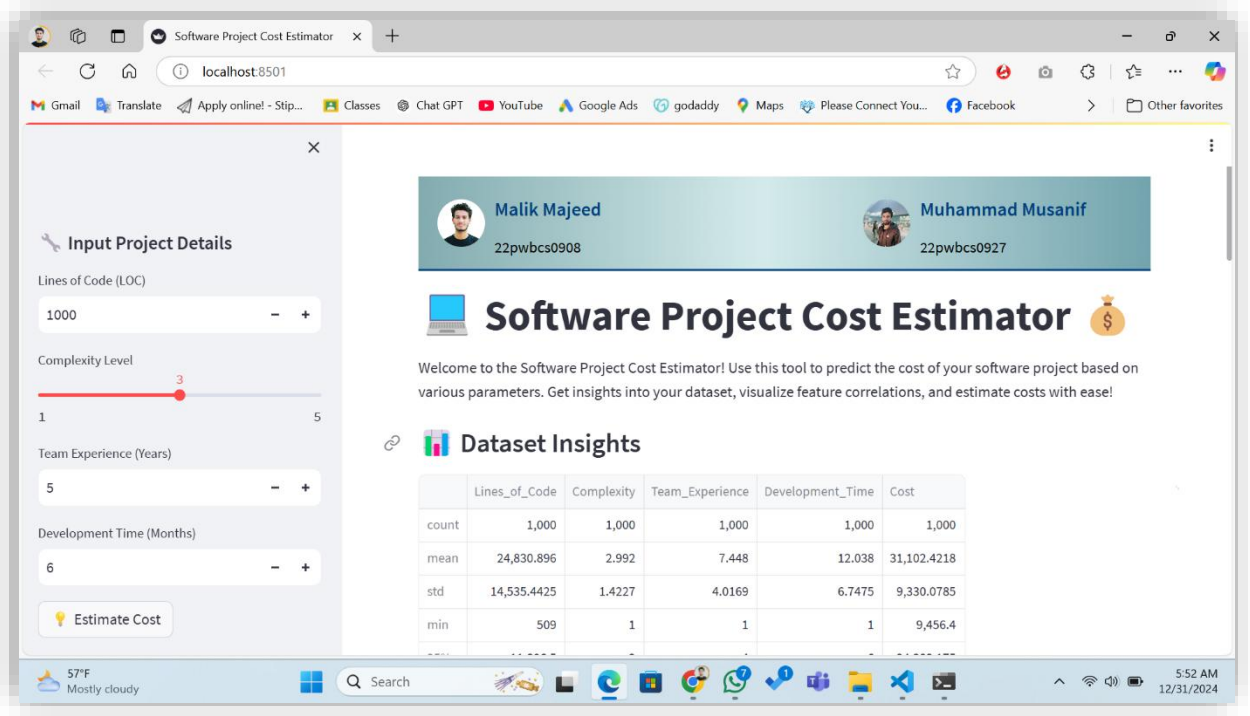


Subject: **Data Science**

"Software's Projects Cost Estimator"

Submitted By:	Malik Majeed (22PWBCS0908)
	Muhammad Musanif (22PWBCS0927)
Semester:	BS CS 5th
Section:	B
Submission Date:	January 15, 2025
Submitted to:	**Prof. Wajeeha Khalil**

DEPARTMENT OF COMPUTER SCIENCE & IT



Overview

The Software Project Cost Estimator is a web-based application designed to provide accurate cost estimations for software projects. The estimator uses Data Science learning algorithms and regression techniques to predict the cost based on project details such as the number of lines of code, complexity, team experience, and development time.

Problem Statement

Estimating the cost of a software project is a critical aspect of project planning and resource allocation. Manual estimation methods are often time-consuming, prone to errors, and lack accuracy, especially for complex projects. The objective of this application is to automate the estimation process using data-driven techniques, ensuring reliability and precision while saving time and effort.

Target Audience

- Software project managers and team leads
- Freelancers and independent software developers
- Organizations planning software development projects
- Students and researchers in software engineering and project management

---

Key Features

1. **User-Friendly Interface**: A simple, interactive web interface built with Streamlit.
2. **Real-Time Predictions**: Provides instant cost estimates based on user inputs.
3. **Data Visualization**: Displays insights about the dataset, including correlations and statistical summaries.
4. **Dynamic Filtering**: Allows users to filter dataset attributes for customized analysis.
5. **Model Insights**: Shows the performance of the underlying regression model.

Input Parameters

The application takes the following inputs:

1. **Lines of Code (LOC)**:
    - Description: Number of lines of code in the project.
    - Input Method: Slider (Range: 500 - 50,000, Step: 500)
2. **Complexity**:
    - Description: The complexity level of the project.
    - Input Method: Dropdown (Values: 1 - 5, where 1 is the lowest and 5 is the highest complexity)
3. **Team Experience**:
    - Description: Average years of experience of the development team.
    - Input Method: Slider (Range: 1 - 15, Step: 1)
4. **Development Time**:
    - Description: Estimated time required to develop the project in months.
    - Input Method: Slider (Range: 1 - 24, Step: 1)

Libraries and Technologies Used

1. **Python**:
    - Core programming language used for the application.
2. **Streamlit**:
    - Used for building the user interface and deploying the application.
3. **Pandas**:
    - For data manipulation and analysis.
4. **NumPy**:
    - For numerical computations.
5. **Seaborn and Matplotlib**:
    - For data visualization.
6. **Scikit-learn**:
    - For implementing the linear regression model.

Approach

1. **Data Preprocessing**:
    - Loaded a dataset containing software project details and corresponding costs.
    - Checked for missing values and constant columns, which were removed if found.
2. **Model Training**:
    - Used a `LinearRegression` model from Scikit-learn.
    - Splitted the dataset into training and testing sets using an 80-20 split.
    - Trained the model on features such as "Lines_of_Code," "Complexity," "Team_Experience," and "Development_Time."
3. **Model Evaluation**:
    - Evaluated the model using metrics like Mean Squared Error (MSE) and R2 Score.
4. **UI Integration**:
    - Developed an interactive Streamlit interface to take inputs and display predictions and insights dynamically.
5. **Visualization**:
    - Provided dataset insights using statistical summaries and heatmaps to show feature correlations.

Implementation Steps

1. **Dataset Loading**:
    - Used a CSV file, "Software_Project_Cost_Estimator_Dataset.csv."
2. **User Input Handling**:
    - Captured user inputs through sliders and dropdowns in Streamlit's sidebar.
3. **Prediction**:
    - Created a `predict` function to calculate project costs based on the trained model.
4. **Dynamic Dataset Filtering**:
    - Added filtering options to allow users to analyze specific subsets of the dataset.
5. **Insights and Visualizations**:
    - Displayed statistical summaries and correlation heatmaps using Seaborn and Matplotlib.

Outputs

1. **Cost Estimation**:
    - Provides an estimated cost in USD based on user inputs.
2. **Dataset Insights**:
    - Count, mean, standard deviation, min, max, and quartile values for each dataset attribute.
3. **Correlations**:
    - Visual representation of feature correlations to aid understanding of relationships between attributes.

Future Enhancements

1. **Advanced Models**:
    - Incorporate more complex machine learning models like Random Forests or Gradient Boosting for better accuracy.
2. **Expanded Features**:
    - Include additional features like project domain, risk factors, or team size.
3. **Export Options**:
    - Allow users to download cost estimations and dataset insights as reports.
4. **Deployment**:
    - Host the application on a cloud platform like AWS, GCP, or Streamlit Cloud for wider accessibility.

Conclusion

The Software Project Cost Estimator simplifies the process of estimating project costs with its user-friendly interface and data-driven predictions. This tool can save time, reduce errors, and improve planning for software development projects.