**Net Value Added Prediction Project**

**1. Introduction**

**Problem Statement**

The objective of this project is to predict the **Net Value Added** by industries using input features such as:

* **Invested Capital**
* **Number of Workers**
* **Wages to Workers**
* **Total Emoluments**

By leveraging machine learning techniques, specifically regression, we aim to create a model that can accurately predict the Net Value Added for unseen or new industry groups. The dataset used for this project is derived from the Annual Survey of Industries 2020-21, focusing on Table 2: Principal Characteristics by Major Industry Group.

**Dataset Source**

* Dataset: [Annual Survey of Industries 2020-21](https://microdata.gov.in/nada43/index.php/catalog/186/download/2214)

**Expected Outcome**

The machine learning model will provide accurate predictions of Net Value Added for new industries based on their input features. This will aid decision-making for industrial planning and economic analysis.

**2. Data Description**

**Features in the Dataset**

1. **Invested Capital**: Total capital invested in the industry.
2. **Workers**: Number of workers employed.
3. **Wages to Workers**: Total wages paid to workers.
4. **Total Emoluments**: Total compensation paid to employees, including wages and other benefits.
5. **Net Value Added**: Target variable representing the value contributed by the industry after accounting for input costs.

**Data Preprocessing**

* **Missing Values**: Rows with missing data were dropped to ensure data integrity.
* **Column Renaming and Stripping**: Column names were standardized by removing extra spaces.
* **Feature Selection**: Focused on the necessary columns for analysis and modeling.

**3. Methodology**

**3.1 Tools and Libraries**

* **Python**: Chosen for its versatility and rich ecosystem of libraries.
* Libraries:
  + **pandas**: Used for data manipulation and cleaning. It provides an intuitive DataFrame structure for handling tabular data.
  + **numpy**: Facilitated numerical computations, especially for handling arrays and mathematical operations.
  + **plotly**: Enabled interactive visualizations to better understand relationships in the data and present results.
  + **scikit-learn**: A machine learning library used for model building, evaluation, and prediction.

**3.2 Regression Algorithm**

* **Linear Regression**: This algorithm was selected due to the linear nature of the relationship between input features and the target variable. It models the relationship as a weighted sum of the features.

**3.3 Workflow**

1. **Data Loading**: The dataset was read from an Excel file into a pandas DataFrame for analysis.
2. **Data Preprocessing**: Missing values were handled, and relevant columns were selected to ensure clean input data.
3. **Feature Selection**: Only numeric features relevant to the prediction task were retained to simplify the modeling process.
4. **Train-Test Split**: The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance on unseen data.
5. **Model Training**: A Linear Regression model was trained on the training set to learn the relationship between input features and the target variable.
6. **Model Evaluation**: The model’s accuracy was assessed using performance metrics such as Mean Squared Error (MSE) and R-squared (R²).
7. **Prediction**: The trained model was used to predict the Net Value Added for a new industry group based on input features.

**4. Explanations of Code Components**

**4.1 Data Loading and Preprocessing**

The dataset was read from an Excel file using pandas, a library designed for easy data manipulation. Missing values were dropped to ensure that the analysis was based on complete data. The column names were standardized to avoid errors during data manipulation, and only the necessary columns were retained to focus on relevant features.

**4.2 Train-Test Split and Model Training**

To evaluate the model’s performance, the dataset was split into training and testing subsets. This approach ensures that the model is trained on one portion of the data and evaluated on a different, unseen portion, mimicking real-world scenarios. Linear Regression was used for its simplicity and effectiveness in modeling relationships between continuous variables. The model learns weights for each feature that best fit the training data.

**4.3 Model Evaluation**

Performance metrics were used to assess the model:

* **Mean Squared Error (MSE)**: Measures the average squared differences between actual and predicted values. A lower MSE indicates better performance.
* **R-squared (R²)**: Represents the proportion of variance in the target variable explained by the model. Higher values indicate a better fit.

**4.4 Visualization**

**Scatter Plot: Actual vs Predicted**

This plot was used to visually assess the model’s performance. A strong alignment along a diagonal line indicates that predictions closely match the actual values, demonstrating the model’s accuracy.

**Correlation Heatmap**

A correlation heatmap highlights the relationships between numeric features in the dataset. This visualization helps identify multicollinearity and the strength of associations between input features and the target variable.

**5. Prediction for New Industry**

**Input Data**

The new industry’s data was provided as input to the trained model. This included values for Invested Capital, Workers, Wages to Workers, and Total Emoluments. These features were used to predict the Net Value Added.

**Visualization of Prediction**

The predicted Net Value Added for the new industry was visualized alongside existing industries to compare its performance. Adding the predicted value to the dataset allowed for an intuitive comparison using bar charts, highlighting the position of the new industry relative to others.

**6. Results and Observations**

**Model Performance**

* **Mean Squared Error (MSE)**: The average squared difference between actual and predicted values, indicating the model’s prediction accuracy.
* **R-squared (R²)**: Represents the proportion of variance explained by the model. A higher value indicates better performance.

**Predictions**

The model successfully predicted the Net Value Added for the new industry group based on the provided input features.

**Visualizations**

* Scatter plot: Showed a strong linear relationship between actual and predicted values.
* Correlation heatmap: Highlighted relationships between features and the target variable.

**9. References**

* [Annual Survey of Industries 2020-21](https://microdata.gov.in/nada43/index.php/catalog/186/download/2214)
* Scikit-learn Documentation: https://scikit-learn.org
* Plotly Documentation: https://plotly.com