# GDP PREDICTION USING

# LINEAR REGRESSION

## ABSTRACT:

The goal of this project is to create and assess a linear regression model to forecast the dependent variable "GDP ($ per capita)" using the dataset "GDP_Country.csv." The first two columns of the dataset, which contained data about "country" and "region," were eliminated as part of the preprocessing process. The major goals of this project were to develop linear regression with gradient descent optimization from scratch, assess the model's performance using independent testing instances, and conduct various metrics-based analyses of the outcomes. A 70-30 split of the dataset was used to enable reliable model evaluation, with 70% of the samples going to training and 30% going to testing. To learn more about the properties of the processed dataset, descriptive statistics were also created. Metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R2), and Adjusted R-squared were used to assess the model's performance. Research offers insights into the accuracy and applicability of the linear regression model for this dataset, and it gives a thorough investigation of its capacity to estimate GDP per capita. The outcomes revealed the model's predictive capability and its potential for use in economic forecasting.

**KEYWORDS:** Gradient descent optimization, descriptive statistics

# INTRODUCTION:

## A) PROBLEM DEFINITION:

Understanding the dynamics of a country's financial health requires careful consideration of economic forecasting and analysis. The Gross Domestic Product (GDP) per capita, which gauges a nation's economic success by dividing its total economic production by its population, is one of the most important metrics in this area. For investors, analysts, and politicians, precise GDP per capita forecasts can provide insightful information. In order to achieve this goal, we set out on a data-driven journey to develop a linear regression model specifically designed to forecast the dependent variable, "GDP ($ per capita)". Our comprehensive methodology includes data preprocessing, model building, evaluation, and insightful analysis, all essential elements in the data science and machine learning pipeline.

**Aim of this project:**

To implement the linear regression with gradient descent (GD) optimization method from scratch

## B) BACKGROUND STUDY:

Predicting "GDP ($ per capita)" is a crucial task in economic analysis since it provides information on the health and development of a nation's economy. It is an essential tool for evaluating economic development, living standards, and the effectiveness of economic policy. In this project, we use linear regression with gradient descent optimization to create a predictive model for GDP per capita. In order to study the links between economic factors and GDP per capita, this strategy makes use of both conventional econometric approaches and cutting-edge machine learning techniques. Comprehensive data preprocessing, model construction, meticulous evaluation, and the interpretation of important performance measures are all part of the project. We randomly divide the dataset into training (70%) and testing (30%) subsets to assure the model's dependability. We contribute to the field of economic forecasting by launching this project and offering a useful and data-driven strategy to anticipate GDP per capita, which is significant for economic growth.

### RELATED WORKS:

1. **WORLD'S GDP PREDICTION USING MACHINE LEARNING by Dwarakanath G V, Shivakumara T, Shraddha.**

   The goal of this research is to anticipate global GDP using machine learning and to determine annual growth. This research-based project also determines the significance of the features that have an impact on GDP computation. In this project, they created a predictor that makes it simple to comprehend the GDP of our country or another one.

2. **GDP FORECASTING: MACHINE LEARNING, LINEAR OR AUTOREGRESSION? by Giovanni Maccarrone, Giacomo Morelli, Sara Spadaccini.**

   In order to forecast the real U.S. GDP, this study evaluates the accuracy of various models. They discover that the machine learning K-Nearest Neighbor (KNN) model captures the self-predictive potential of the U.S. GDP and outperforms conventional time series analysis using quarterly data from 1976 to 2020. In order to improve forecasting accuracy, they investigate the incorporation of predictors including the yield curve, its latent features, and a collection of macroeconomic indicators. Only when considering long forecast horizons do the predictions prove to be more accurate.

3. **GROSS DOMESTIC PRODUCT PREDICTION USING MACHINE LEARNING by Vaishnavi Padmawar, Pradnya Pawar, Akshit Karande.**

   This study's objective is to forecast GDP using linear regression and random forest for a specific time period. Applying applied mathematics and mathematical models to forecast future economic events is necessary for GDP prediction. They used linear regression and Random Forest in macroeconomic data forecasting. Based on optimization process, the machine learning algorithm "Random Forest" utilized during this study worked well.

## C) OBJECTIVES AND CONTRIBUTION:

1. **Data preprocessing:**

   Data preprocessing is used to convert data into clean data sets which is used for analysis. It consists of several steps like data cleansing, data reduction, data enrichment organization and transformation. So, it will be easy for machine learning models to read data and learn from data.

2. **Descriptive statistics:**

   Descriptive statistics is used to describe data and visualize data. Understanding all the features of dataset is required to build model effectively and to evaluate the model. It is also used to know the mean, median and mode of all the features of dataset.

3. **Model Training:**

   Model Training includes the creation of machine learning models. Many prebuilt models can be used but we are creating a linear regression model from scratch. We will train the model using train data which is done by splitting the data into train and test datasets.

4. **Model Evaluation:**

Model Evaluation is done by calculating different performance evaluation metrics like Mean Square Error (MSE), absolute error, Root means square error (RMSE), R2 values and Adjusted R2 values. This can be done using test dataset.

## METHODOLOGY:

- Data Encoding and preprocessing: In data preprocessing we removed two columns from the dataset named “country” and “region”. We also removed null values from the data set. Normalizing the numerical data, encoding the categorical variables was also done on the dataset.

- Descriptive statistics: In this step we will describe all the features of the dataset. We will get count, mean, standard deviation, min and max values of all the features of the dataset to understand the data effectively.

- Visualization of data: Some of the dataset features are visualized using statistical view given by google colab.

- Splitting of data: In this step we will divide the entire dataset into two groups, training set and testing set. Training set is used to train machine learning model and learn from it. Testing data is used to evaluate model performance.

- Model creation: In this project we will create a linear regression model using gradient descent (GD) optimization method from scratch. This is used to find GDP values of different countries.

- Model training: Model training is done by using training dataset. The model will get the relationship between the features and output, and it will give us the predicted values.

- Model Evaluation: Model evaluation is done using testing dataset. Using this dataset, we will calculate different performance evaluating features like MSE, Absolute error, RMSE, R2 value and Adjusted R2 values.

- Analysis of results: The performance evaluating features give us whether the model we created is effective or not.

# MODEL DESCRIPTION:

- **StandardScaler:** A preprocessing method used frequently in machine learning and data analysis is called the StandardScaler. It is a phase in the preprocessing or data transformation process that datasets go through before machine learning models are trained. The StandardScaler job is to standardize or normalize a dataset's features (variables). Standard scaler is used to convert feature values to zero mean and unit standard deviation values.

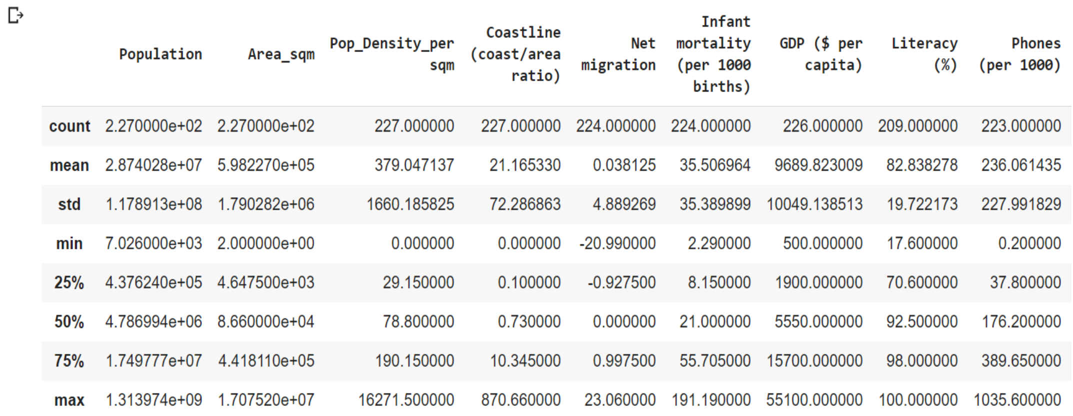
- **Linear regression model using gradient descent optimization method from scratch:** A linear regression model using gradient descent (GD) optimization method from scratch is used. By fitting a linear equation to the observed data, the statistical technique of linear regression is used to model the connection between a dependent variable and one or more independent variables. In the fields of machine learning and statistics for predictive modeling and data analysis, it is one of the most straightforward and extensively used techniques. In its most basic form, linear regression uses a straight-line equation to represent the relationship between two variables, commonly identified as "X" (the independent variable) and "Y" (the dependent variable).

- **GD regression model from sklearn:** A linear regression from sklearn is also used for analysis of data. We will compare the results of linear regression model using gradient descent optimization method from scratch with this GD regression model from sklearn. Sklearn gives a wide range of models to work on datasets. As our dataset is related to linear regression, we are importing that from it.

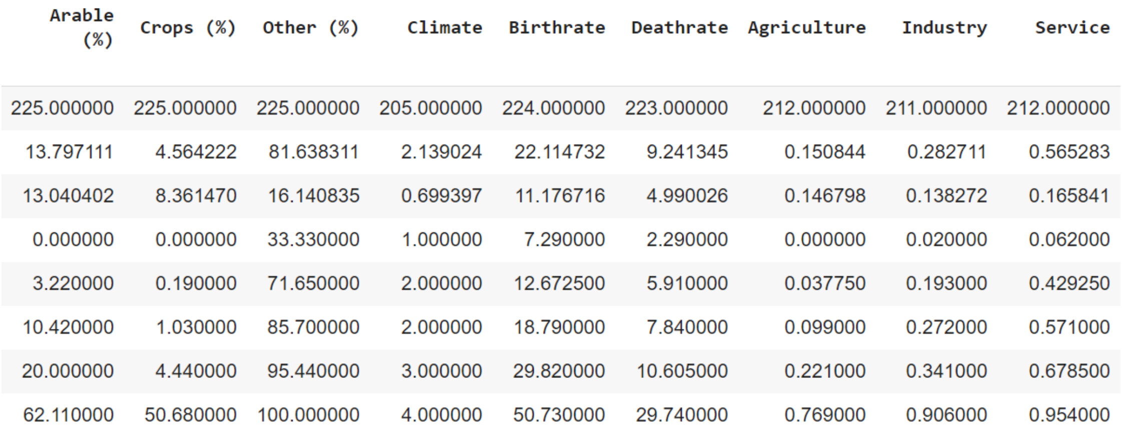
# EXPERIMENT AND RESULTS:

## A) DATABASE:

Database GDP_Country has 227 observations and 20 variables. We removed two variables named “country” and “region” in the data preprocessing step for better analysis of dataset. Here is description of different variables in the dataset.



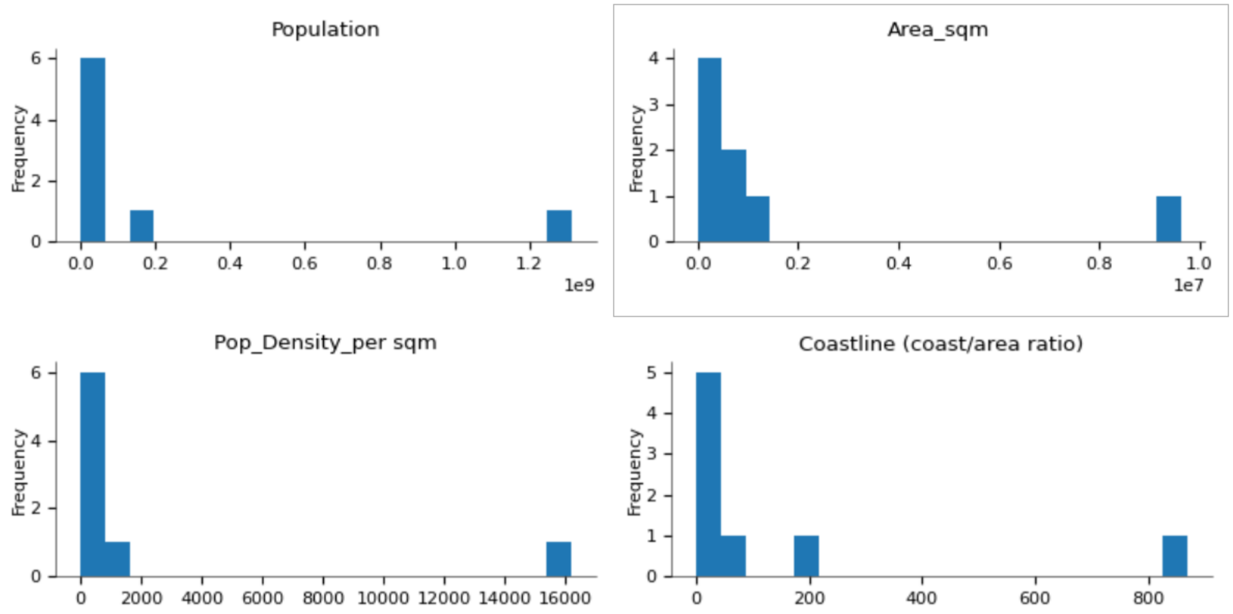
| | Population | Area_sqm | Pop_Density_per sqm | Coastline (coast/area ratio) | Net migration | Infant mortality (per 1000 births) | GDP ($ per capita) | Literacy (%) | Phones (per 1000) |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 2.270000e+02 | 2.270000e+02 | 227.000000 | 227.000000 | 224.000000 | 224.000000 | 226.000000 | 209.000000 | 223.000000 |
| **mean** | 2.874028e+07 | 5.982270e+05 | 379.047137 | 21.165330 | 0.038125 | 35.506964 | 9689.823009 | 82.838278 | 236.061435 |
| **std** | 1.178913e+08 | 1.790282e+06 | 1660.185825 | 72.286863 | 4.889269 | 35.389899 | 10049.138513 | 19.722173 | 227.991829 |
| **min** | 7.026000e+03 | 2.000000e+00 | 0.000000 | 0.000000 | -20.990000 | 2.290000 | 500.000000 | 17.600000 | 0.200000 |
| **25%** | 4.376240e+05 | 4.647500e+03 | 29.150000 | 0.100000 | -0.927500 | 8.150000 | 1900.000000 | 70.600000 | 37.800000 |
| **50%** | 4.786994e+06 | 8.660000e+04 | 78.800000 | 0.730000 | 0.000000 | 21.000000 | 5550.000000 | 92.500000 | 176.200000 |
| **75%** | 1.749777e+07 | 4.418110e+05 | 190.150000 | 10.345000 | 0.997500 | 55.705000 | 15700.000000 | 98.000000 | 389.650000 |
| **max** | 1.313974e+09 | 1.707520e+07 | 16271.500000 | 870.660000 | 23.060000 | 191.190000 | 55100.000000 | 100.000000 | 1035.600000 |



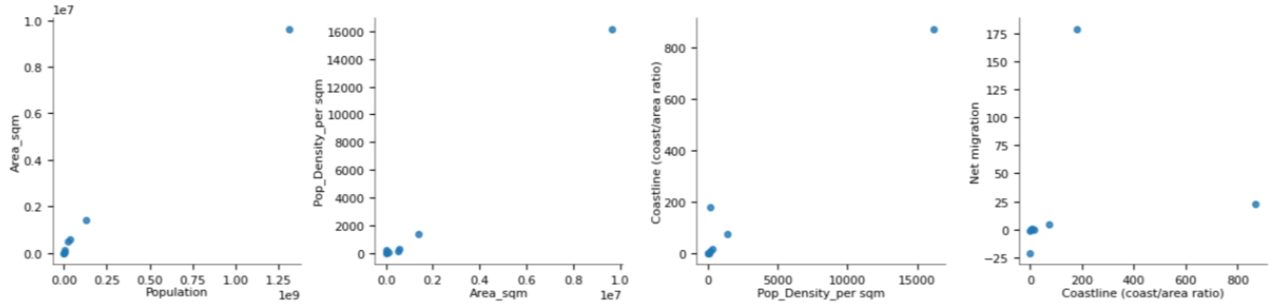
| Arable (%) | Crops (%) | Other (%) | Climate | Birthrate | Deathrate | Agriculture | Industry | Service |
|---|---|---|---|---|---|---|---|---|
| 225.000000 | 225.000000 | 225.000000 | 205.000000 | 224.000000 | 223.000000 | 212.000000 | 211.000000 | 212.000000 |
| 13.797111 | 4.564222 | 81.638311 | 2.139024 | 22.114732 | 9.241345 | 0.150844 | 0.282711 | 0.565283 |
| 13.040402 | 8.361470 | 16.140835 | 0.699397 | 11.176716 | 4.990026 | 0.146798 | 0.138272 | 0.165841 |
| 0.000000 | 0.000000 | 33.330000 | 1.000000 | 7.290000 | 2.290000 | 0.000000 | 0.020000 | 0.062000 |
| 3.220000 | 0.190000 | 71.650000 | 2.000000 | 12.672500 | 5.910000 | 0.037750 | 0.193000 | 0.429250 |
| 10.420000 | 1.030000 | 85.700000 | 2.000000 | 18.790000 | 7.840000 | 0.099000 | 0.272000 | 0.571000 |
| 20.000000 | 4.440000 | 95.440000 | 3.000000 | 29.820000 | 10.605000 | 0.221000 | 0.341000 | 0.678500 |
| 62.110000 | 50.680000 | 100.000000 | 4.000000 | 50.730000 | 29.740000 | 0.769000 | 0.906000 | 0.954000 |

**Table: Descriptive Statistics of variables**

**Distributions**

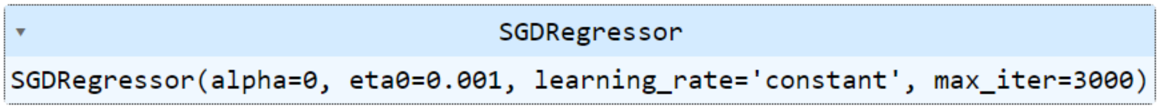


2-d distributions



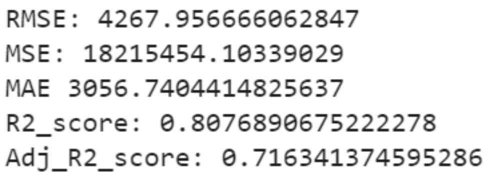
## B) TRAINING AND TESTING LOGS:

### ❖ Model 1: GD Regression model using sklearn.

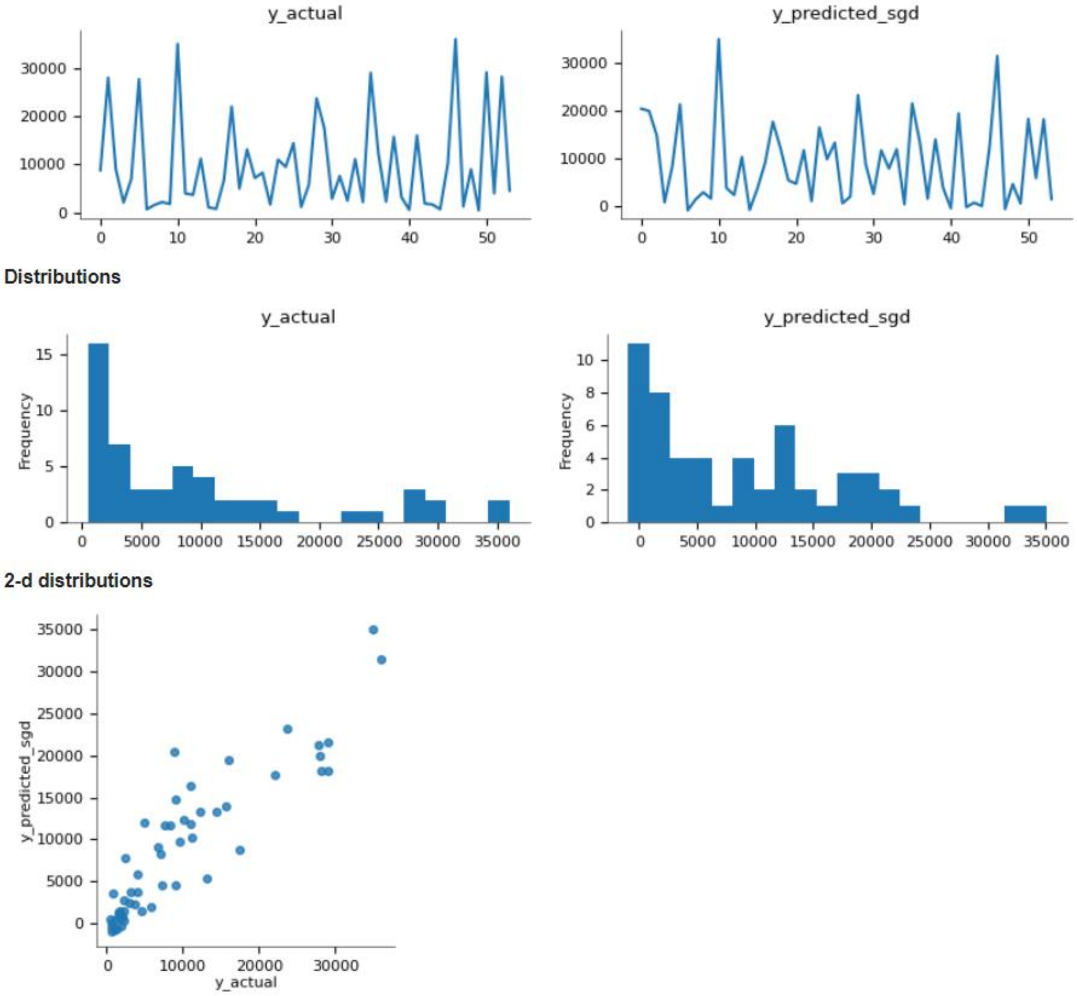


```
SGDRegressor
SGDRegressor(alpha=0, eta0=0.001, learning_rate='constant', max_iter=3000)
```

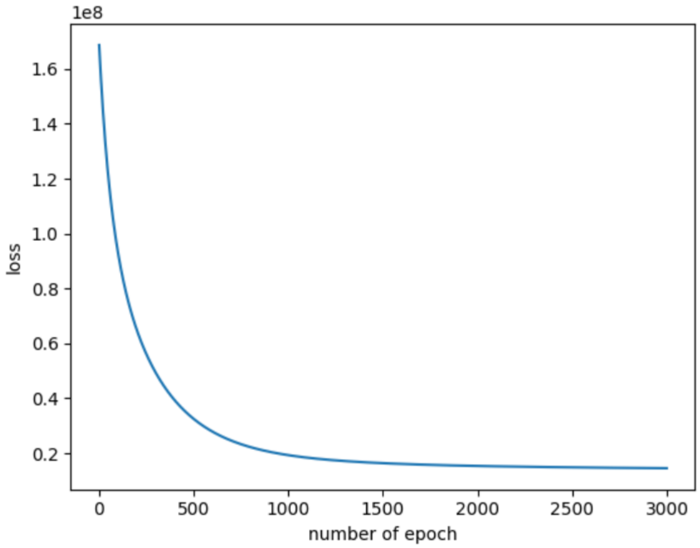
Performance metrics:



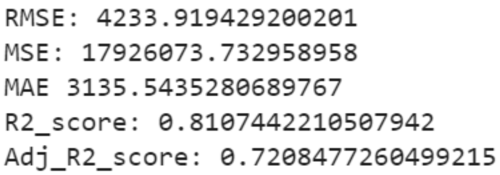
```
RMSE: 4267.956666062847
MSE: 18215454.10339029
MAE 3056.7404414825637
R2_score: 0.8076890675222278
Adj_R2_score: 0.716341374595286
```



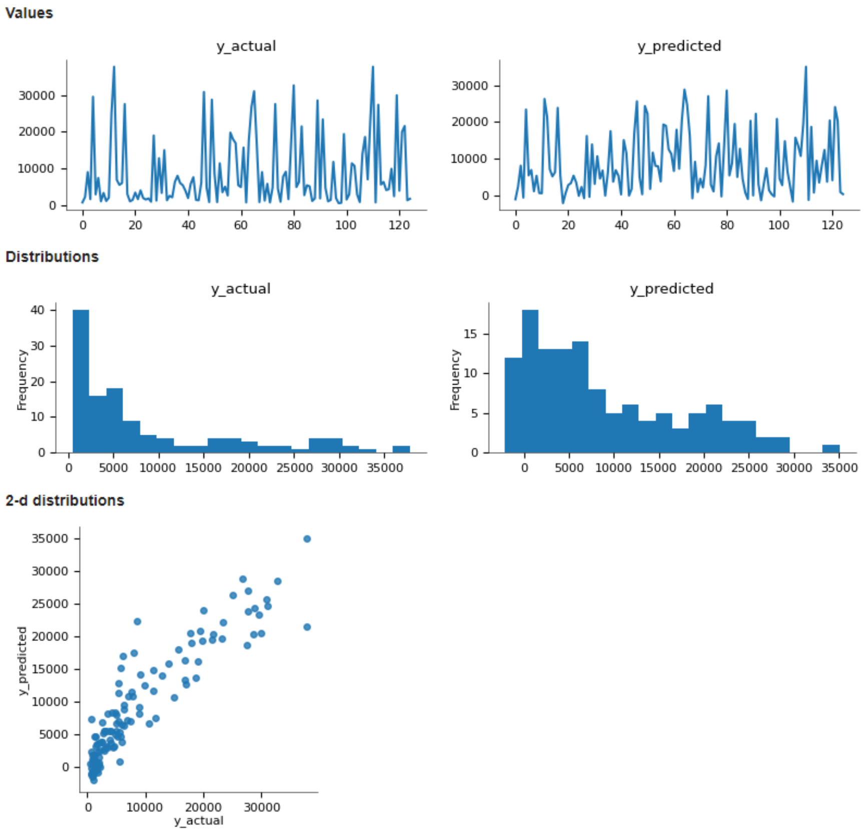
- **Model 2: Linear regression model using gradient descent optimization method from scratch.**



Performance metrics:



```
RMSE: 4233.919429200201
MSE: 17926073.732958958
MAE 3135.5435280689767
R2_score: 0.8107442210507942
Adj_R2_score: 0.7208477260499215
```



## C) DISCUSSION AND COMPARISON:

To predict GDP_Country we implemented two models on the dataset. Both the models worked well on the dataset. R2 value is a performance metrics we can use in comparing two models. R2 values give us the information about the proportion of variance in the dependent variable that can be explained by the independent variable. The higher the R-squared value, the better the model fits your data. The R2 value of GD regression model is 0.8076. The R2 value of Linear regression model using gradient descent optimization method from scratch is 0.8107. So, we can say that Linear regression model using gradient descent optimization method from scratch is better than GD regression model from Sklearn.

## CONCLUSION:

The prediction of GDP of a country is done by using machine learning techniques. The GDP_Country database was used for it. Many preprocessing steps like normalizing the data using StandardScaler were used for it. We split data into two groups like train and test. We fitted two ML models named GD regression model from Sklearn and linear regression model by using gradient decent optimization method from scratch and evaluated them and we calculated performance metrics. Both the models worked well on dataset but linear regression model by using gradient decent optimization method from was better than GD regression model.

# REFERENCES:

- Tanvi Gharte, Himani Patil, Soniya Gawade. *GDP Prediction and Forecasting using Machine Learning* [Online] https://www.irjet.net/archives/V9/i4/IRJET-V9I4362.pdf

- Giovanni Maccarrone, Giacomo Morelli, and Sara Spadaccini. GDP Forecasting: *Machine Learning, Linear or Autoregression?* [Online] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8554645/

- Dwarakanath G V, Shivakumara T, Shraddha. *World's GDP Prediction Using Machine Learning* [Online] https://journalppw.com/index.php/jpsp/article/download/2973/1929/3392

- S.C. Agu a, F.U. Onu b, U.K. Ezemagu c, D. Oden. *Predicting gross domestic product to macroeconomic indicators* [Online] https://www.sciencedirect.com/science/article/pii/S2667305322000229

- Ronan Flannery, *A Machine Learning Approach to Predicting Gross Domestic Product* [Online] https://norma.ncirl.ie/4441/1/ronanflannery.pdf

- RefCodes: https://memphis.instructure.com/courses/98483/files/folder/Ref.%20codes?preview=7771518

- Nikhil Vyas1, Jay Patel, Darshit Vala, Devansh Patel, Rohit Patel, *Machine Learning Based Generic GDP Analysis And Prediction System* [Online], http://ijasret.com/VolumeArticles/FullTextPDF/804_