

**ព្រះរាជាណាចព្ររម្ពុជា**

**​​​      ជាតិ សាសនា ព្រះមហាក្សត្រ**

Assignment Introduce to Data Science

              ​​​​​​GROUP: I3-AMS-A

Name of Students ID of Students     Score

1. KE MANUT                    e20221205                          ………

2. LENG VICHEKA            e20221686                          ………

3. LOT SOKLANG              e20221464                          ………

4. KHAY LINA                    e20220518                          ………

5. CHHEANG PHANY       e20220349                          .……..

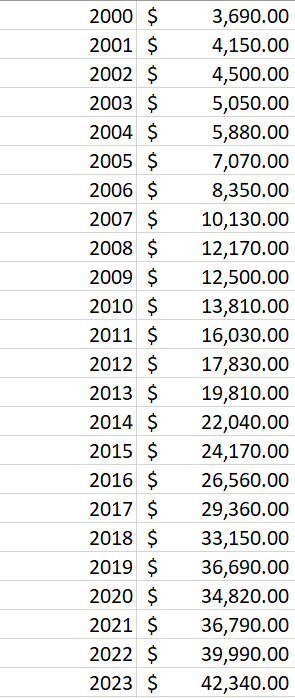
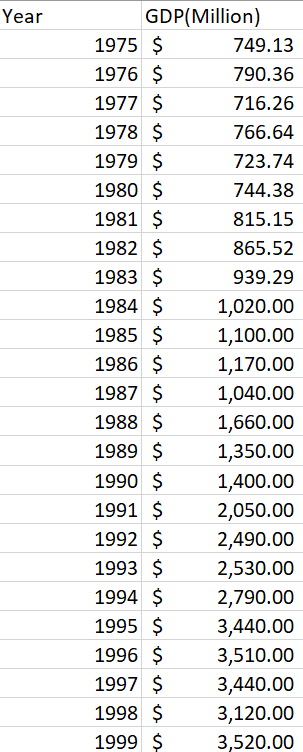
Lecturer:  Dr. PHAUK SOKKHEY

1. **Introduction**

Economic growth is a fundamental indicator of a country's development and stability, and Gross Domestic Product (GDP) serves as a critical metric in assessing the health of an economy. For Cambodia, a rapidly growing nation in Southeast Asia, accurate GDP prediction is essential for effective policymaking, resource allocation, and sustainable development.

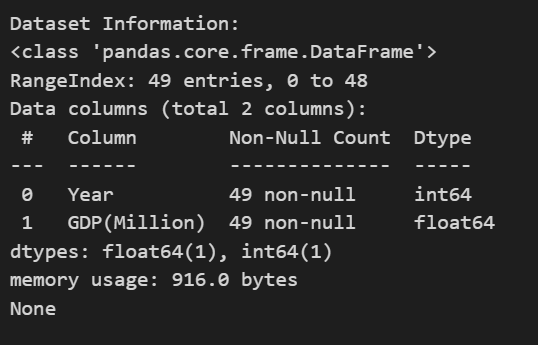
This project focuses on predicting Cambodia's economic performance using GDP as a primary indicator. By leveraging historical data, statistical analysis, and predictive modeling techniques, the study aims to provide insights into economic trends and potential future outcomes. With advancements in data science and machine learning, predictive analytics offers a powerful approach to understanding economic complexities, identifying key growth drivers, and anticipating challenges.

1. **EDA (Exploratory Data Analysis)**

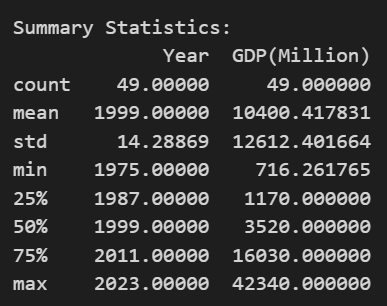
**The dataset:** below is the dataset we took from **the World Bank** which started from 1975 to 2023 which I selected and put into excel and import to Python.

* **Data Information and Summary**

This is the information of our dataset which show that our data contains:

* **two columns**
  + **Year with 49 rows**
  + **GDP(Million) with 49 rows**

This is the summary of our data which shows there are 49 data contains:



**Year**

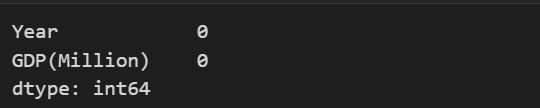
* **Count:49**
* **Mean=1999**
* **Standard Deviation:14.29**
* **25th Percentile:1987**
* **Median(50th Precentile)=1999**
* **75th Percentile=2011**
* **Maximum:2023**

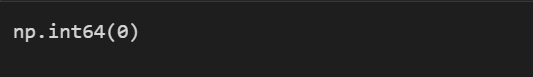
**GDP (Million)**

* **Count:** 49 (total number of GDP values)
* **Mean:** 10,400.42 million (average GDP)
* **Standard Deviation (std):** 12,612.40 million (how much the GDP values vary from the average)
* **Min:** 716.26 million (lowest GDP value)
* **25% Percentile:** 1,170 million (25% of GDP values are below 1,170 million)
* **50% Percentile (Median):** 3,520 million (middle GDP value)
* **75% Percentile:** 16,030 million (75% of GDP values are below 16,030 million)
* **Max:** 42,340 million (highest GDP value which in **2023**)
* Data Cleaning:

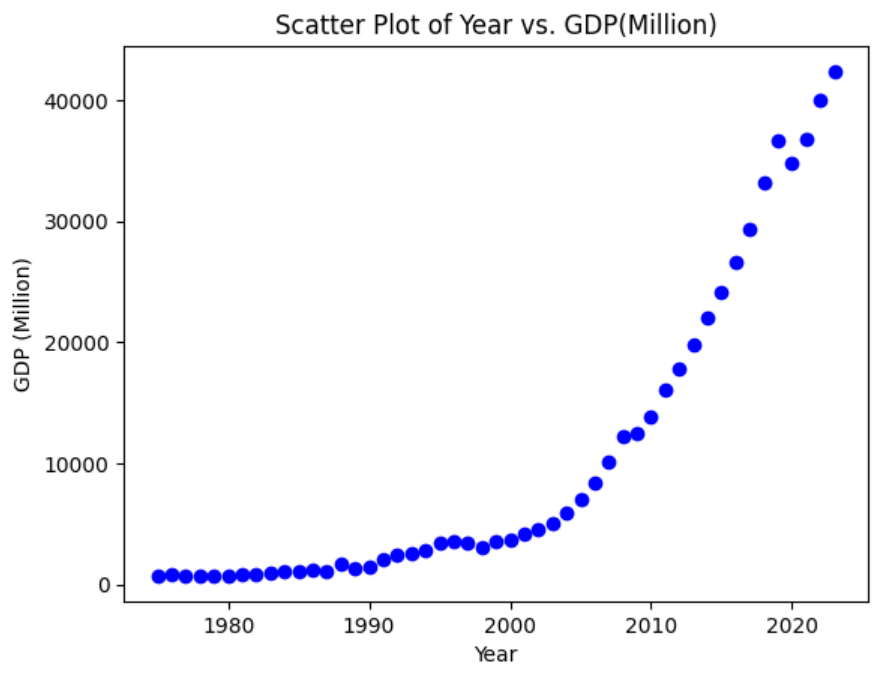
We import data to python and have checked for **duplicates value and missing value in our dataset**. As a result there are no obvious issues such as missing values, duplicates, or inconsistencies in our dataset. So the dataset that we have is clean data. We use method:

* data.isnull().sum() for checking missing data
* data.duplicated().sum() for checking duplication data





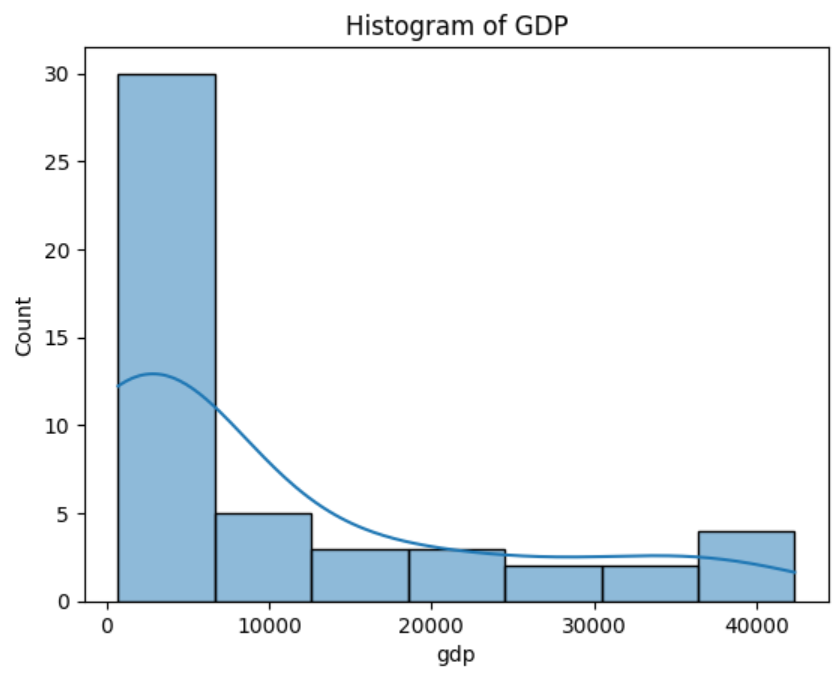
* **Visualizing Numeric Variables**

**Scatter plot of Year vs. GDP(Million)**

The relationship between the year and GDP, the plot shows how GDP ​​(in million ) has changed over time, from 1975 to 2023. As this plot graph above:

* X-axis: represents the year from 1975 to 2023
* Y-axis: represents the GDP values in millions

This plot helps to visually identify the long-term trend of economic growth, as well as the increasing rate of growth in more recent decades. The outliers or clusters of points at the higher end of the scale show the rapid GDP growth in recent years, which may be influenced by factors such as technological advancements, global economic conditions, or government policies. From the plot graph, each data point is represented by a blue dot, indicating the GDP for each specific year. The scatter plot clearly reveals an upward trend, with relatively steady growth in GDP over the years.

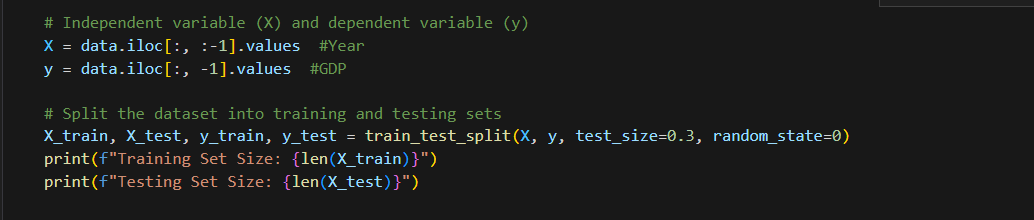
     **Histogram of GDP**

* **Majority of Data:** Concentrated in the lower GDP range (0-10,000 million).
* **Decreasing Count:** As GDP values increase, the count decreases significantly.
* **Skewness:** Indicates that most GDP values are clustered at the lower end, with fewer values at higher GDP ranges.

This histogram visually represents the distribution of GDP values, showing that the majority of GDP values are on the lower end, while higher GDP values are less common.

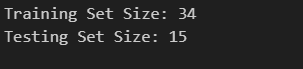
1. **Modeling and Evaluation**

To do data training and data testing test, we split the dataset into training and testing sets using programming language to do:



For the testing size we use 30% and 70% for training data. Generally, we use 30% for data testing and 70% for data training to be balanced and good for data training.

Then we got the value:

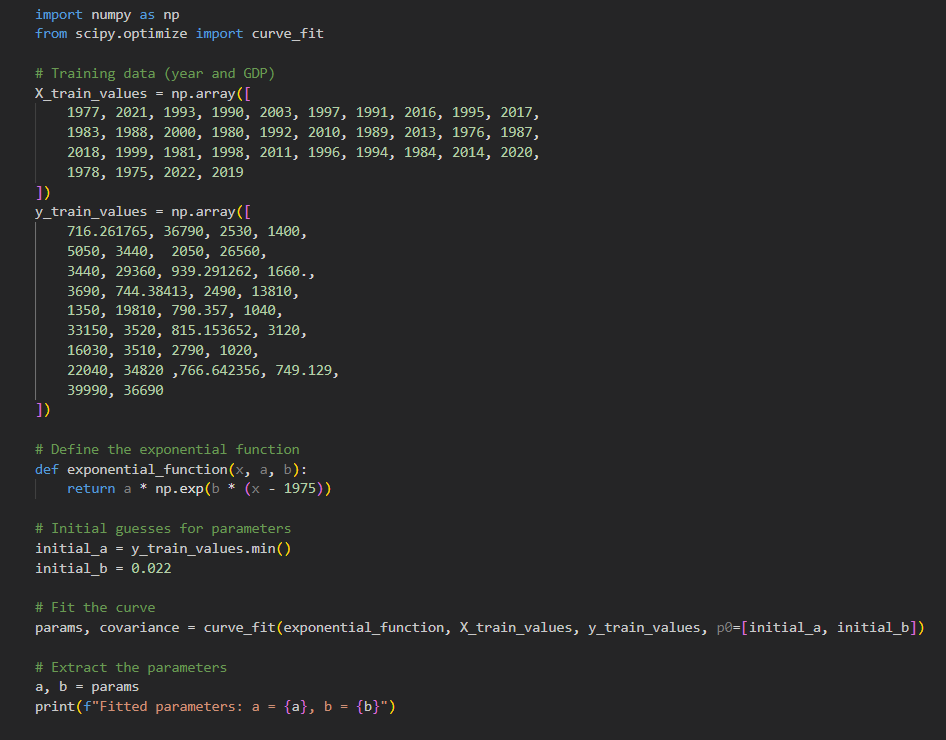


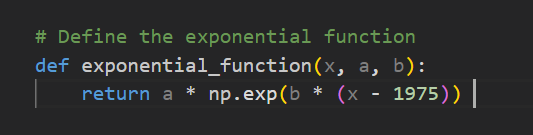
* Training set: for 70% of data, the Training set size=34(34 years) it means that the model will learn from these 34 year-GDP pairs to understand the relationship between time and GDP for the economy in Cambodia.
* Testing set: 30% of data is used to do testing, which 15 year-GDP used to do testing.
* **Machine Learning and Model**

From our graph we assume that it fitted an exponential model to the data and extracted the parameters which the equation form

Where:

* represents GDP
* **Training data**

****

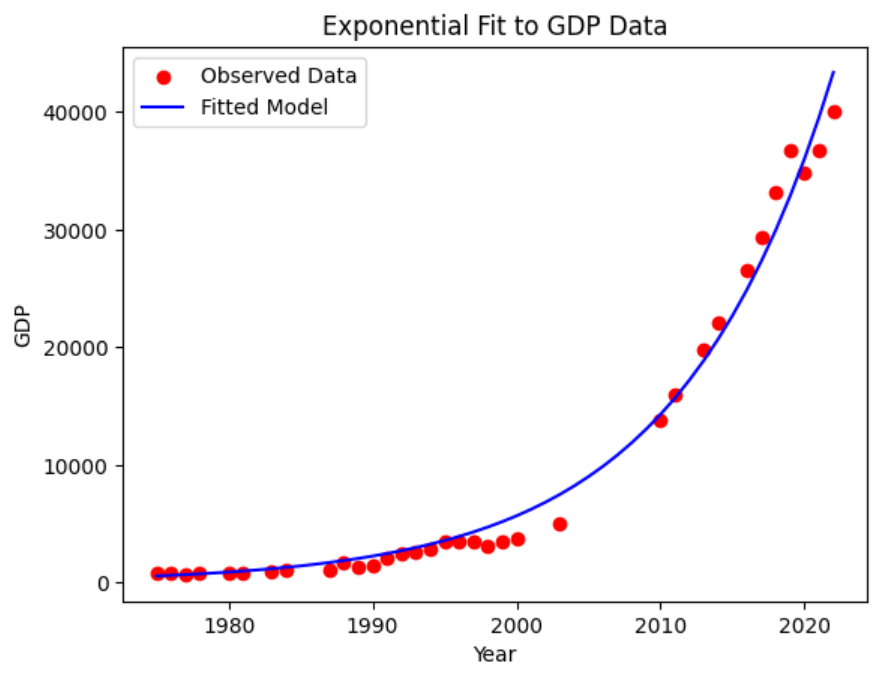
1. Starting from our training data which has 34 training data, we will make model by using the curve\_fit function from the scipy.optimize module is used to fit the exponential function
2. **Initial Parameter Guesses**:
   * initial\_a is set to the minimum GDP value in the training data
   * initial\_b is set to 0.022 as an initial guess for the growth rate.
3. **Curve Fitting**:
   * The curve\_fit function from the scipy.optimize module is used to fit the exponential function to the training data.
   * This function estimates the parameters a and b by minimizing the difference between the predicted and actual GDP values.
4. **Parameter Extraction**:
   * After fitting the curve, the estimated parameters a and b are extracted from the params array.

**Results**: The fitted parameters obtained from the curve fitting process are:



We only round our a and b two to make it look convenience. Thus,

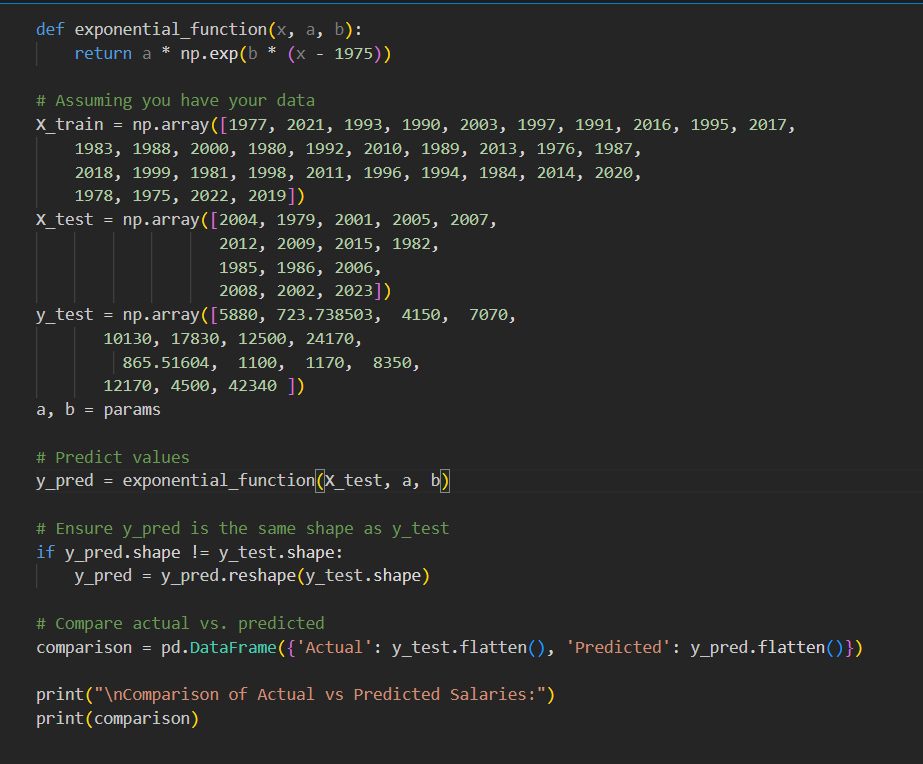
The exponential function, with the estimated parameters, provides a model to predict GDP growth based on historical data. This model can be used to forecast future GDP values and understand the growth trend over the years which we will use in Tested data

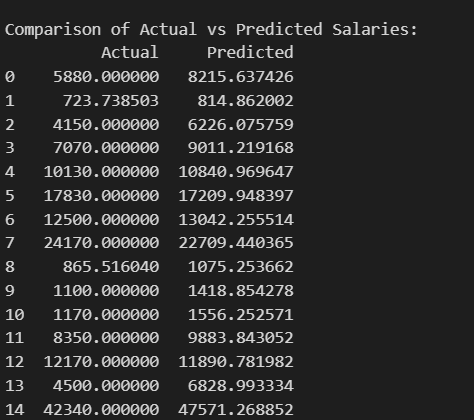
* **The plot of the training model****:** we predict GDP values for the years in the dataset by the above Model and visualize the original and predicted values that shows in the graph.

From the graph, we can say, the scatter points, which represents the real data or observed data, and fitting model, blue curve, is almost fit.

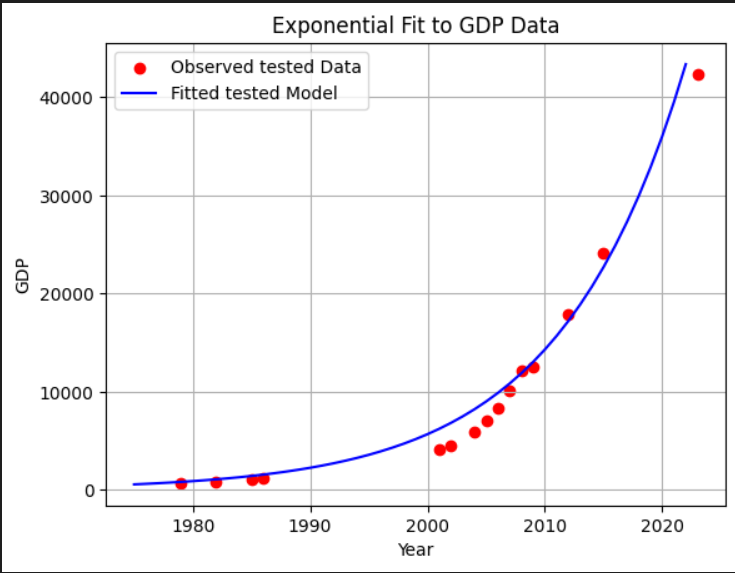
* **Tested Data**

We can use our Training Model to test on Tested data (15 datasets) that our Model never experience before, then we compare.

****

We get:

Visualize Data-Test set



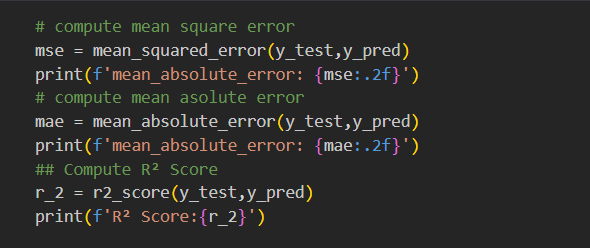
The graph represents the Tested data which shows that Red dots on the graph represent the "Observed tested Data" points, and the blue curve represents the "Fitted tested Model" the graph appears well-constructed and effective for showcasing GDP growth over time. The use of red dots for observed data and a blue curve for the fitted model makes it visually clear and easy to understand. The exponential growth pattern is evident, and the graph effectively highlights the correlation between time and GDP

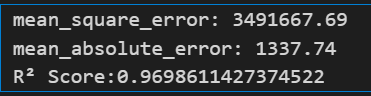
* **Evaluation:**

To make sure that our Model is accurate, we need to calculate

* + **Mean square error**
  + **Mean absolute error**

**We use library from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score**

****

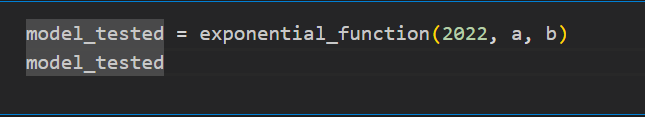


* Mean Squared Error (MSE): This measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. A lower MSE indicates a better fit. In your case, an MSE of 3,491,667.69 suggests that the squared differences between the predicted and actual GDP values are relatively large.
* The **Mean Absolute Error (MAE)** measures how far, on average, the model's predictions are from the actual values, without considering whether the predictions are too high or too low. An MAE of **1,337.74** means the model’s predictions are typically off by about **1,337.74 million units of GDP**. This gives a clear idea of the average error size in simple terms.
* R² Score (Coefficient of Determination): This provides an indication of how well the independent variables explain the variability of the dependent variable. An R² score of 0.9698611427374522 means that approximately 97% of the variance in the GDP data is predictable from the model. This is a very high value, indicating a good fit.

Overall, your R² score suggests that your model fits the data very well, explaining a high proportion of the variance. However, the MSE and MAE values indicate that there are still some significant errors in the predictions, which might be due to the scale of the GDP values.

* To be noted that our Model is not always accurate or closed to the real GDP value because the GDP of each country depends on the situations of the world.
  + Example, in 2022:
    - The actual data was 39990 million dollars
    - The Predicted data was 43371.300764423795 million dollars

Because of covid 19, and the war between Russia and Ukraine.

* Result: If we want to predict GDP for other year, we can use our model like this:

In this case, we want to know the value of GDP in 2022

1. **Conclusion**

This project explores the economic growth of Cambodia through historical data analysis and future predictions using machine learning techniques. When data has been collected, we can identify trends and patterns that have shaped Cambodia’s economic trajectory.

The machine learning model, trained on historical data, provided future projections of Cambodia’s economy, indicating in GDP driven by advancements in the industrial and service sectors. The results emphasize the importance of policy interventions to ensure inclusive and sustainable growth.

Future work can improve prediction accuracy by incorporating more granular datasets and using advanced machine learning models like deep learning. Especially increasing qualitative factors such as policy shifts and geopolitical events could further enhance the model’s reliability.

1. **References**

* **Dataset:** [**GDP (current US$) - Cambodia | Data**](https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=KH)
* **Presentation in slide:** [**Cambodia\_GDP - Presentation**](https://www.canva.com/design/DAGdHULZBQo/pVeTshH4uPJSAsJsRYL3kw/edit)
* **Source of code:**