Can Data Science Give a Credible Indication of Kenya's Economy?

Angwenyi David, Dr. rer. nat.

(The University Of Choice)

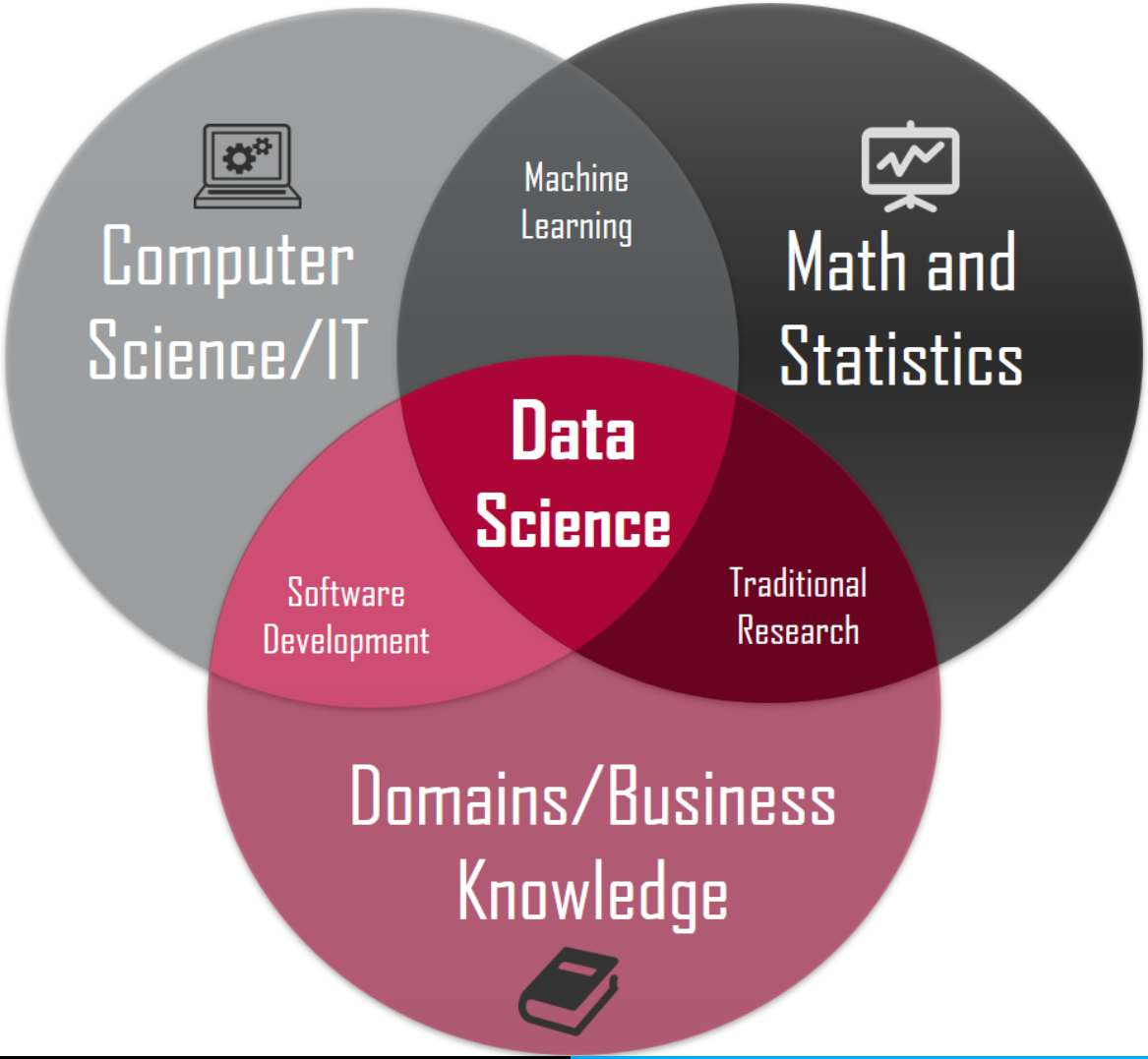
Department of Mathematics

October 28, 2024

Contents

Introduction: What is data science?

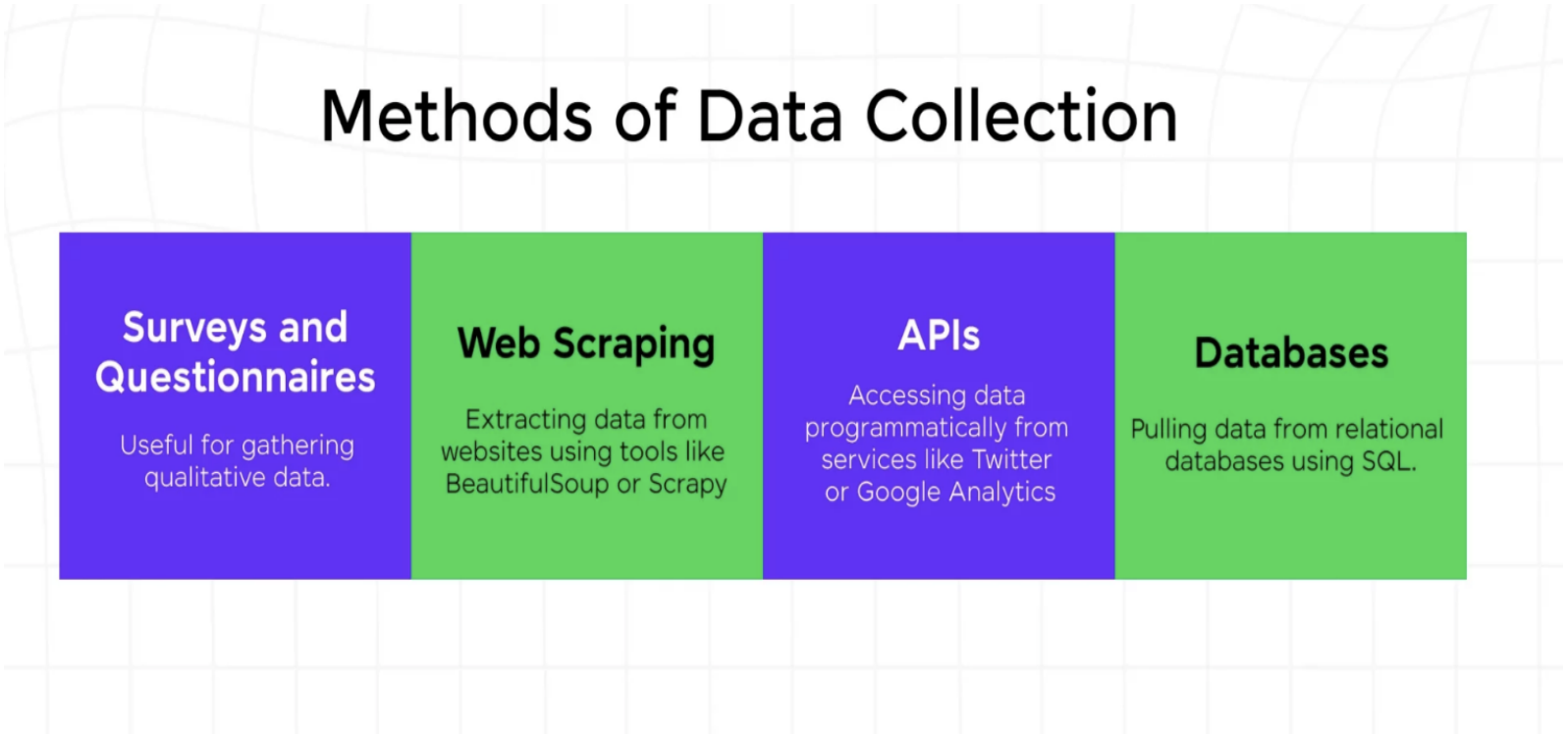
Definition: Data Science is a *multidisciplinary field* that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.



Core components of data science

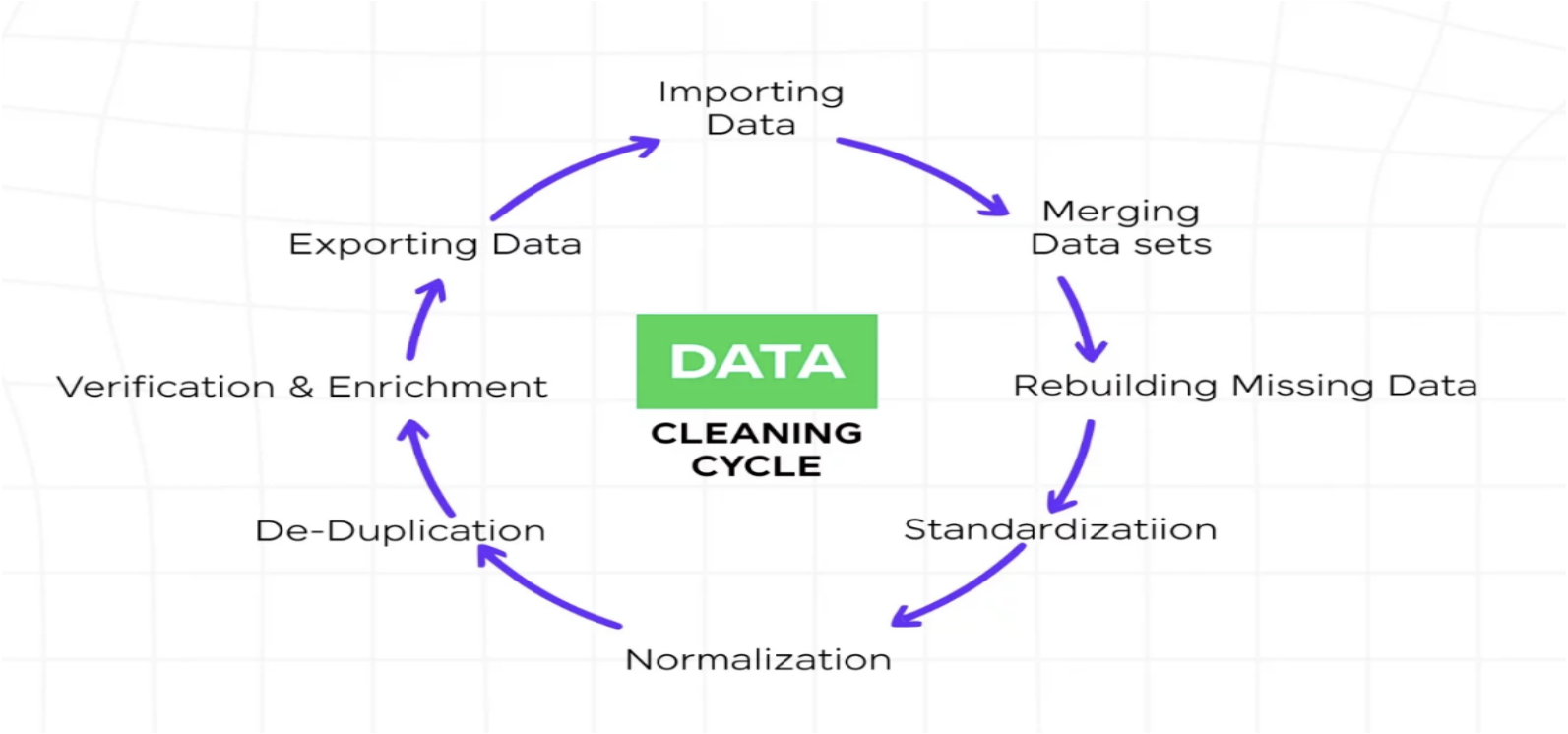
- Data Collection: Gathering data from various sources.
- Data Cleaning: Processing and preparing data for analysis by removing errors and inconsistencies.
- Data Analysis: Using statistical methods to explore data and extract meaningful patterns.
- Data Modeling: Applying machine learning or statistical models to make predictions or decisions based on the data.
- Data Communication: Visualizing and reporting findings to stakeholders to drive decision-making.

Data Collection



In Data Science, data collection is much like gathering your ingredients. You can't proceed without having the raw material, which in this case is the data.

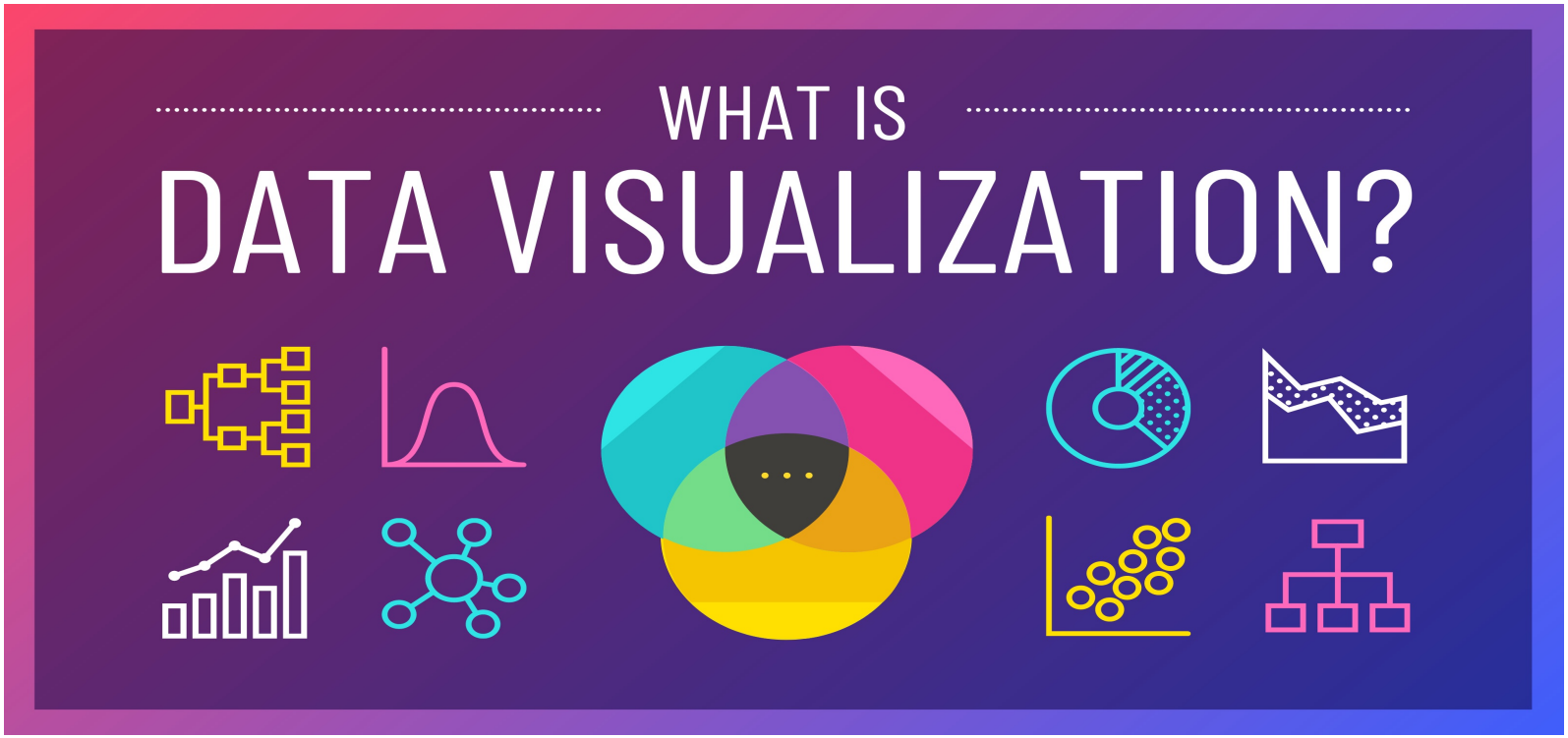
Data Cleaning



Common Data Cleaning Tasks

- Handling Missing Values: Using techniques like imputation or removing incomplete records.
- Removing Duplicates: Ensuring each record is unique to maintain data integrity.
- Correcting Errors: Fixing typos, incorrect entries, or outliers.
- Standardizing Formats: Converting it into a consistent format.

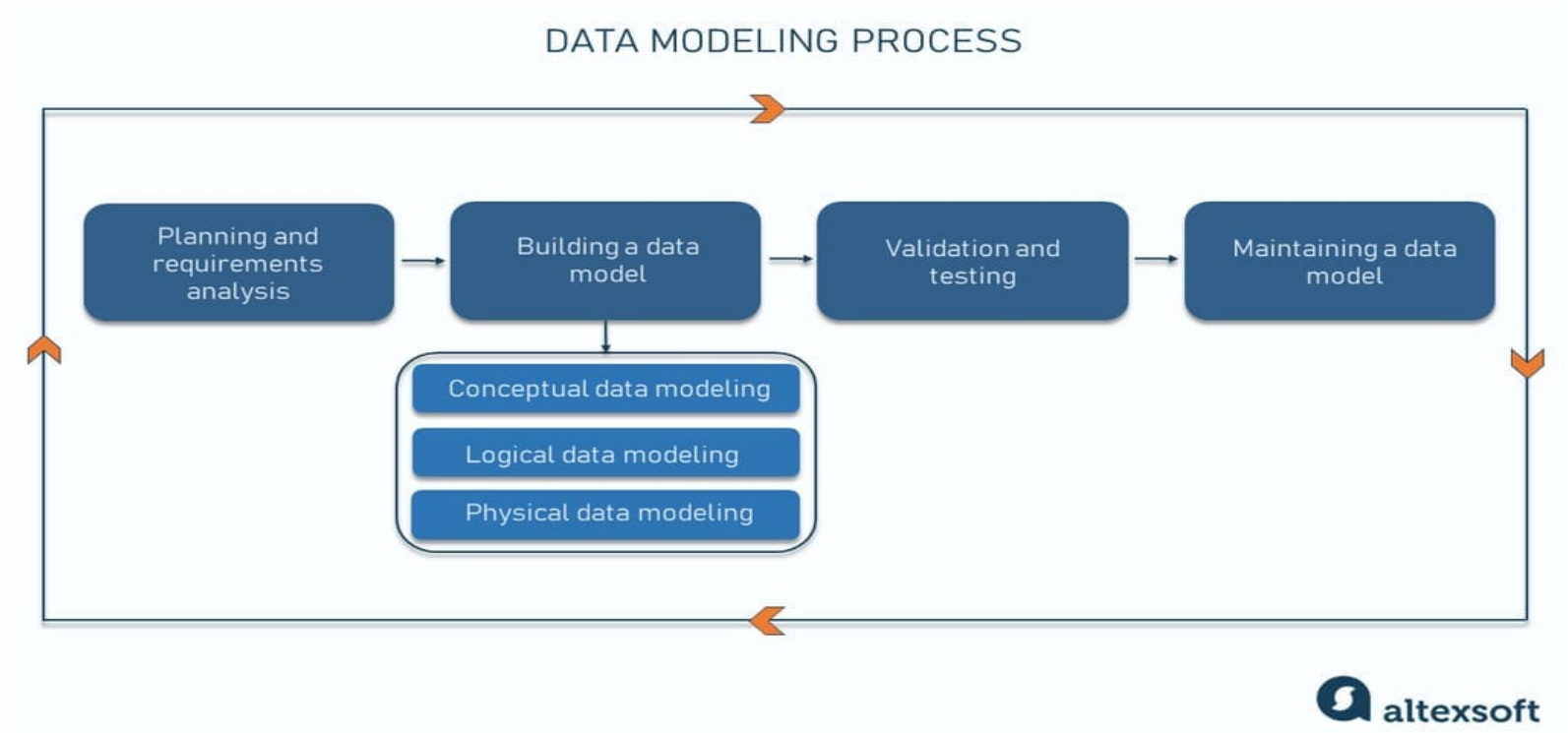
Data Visualization



Tools and Techniques

- Descriptive Statistics: Mean, median, mode, standard deviation, etc.
- Data Visualization Tools: Matplotlib, Seaborn, Tableau, and Power BI.
- Exploratory Data Analysis (EDA): Using plots like histograms, scatter plots, and box plots to understand data distribution

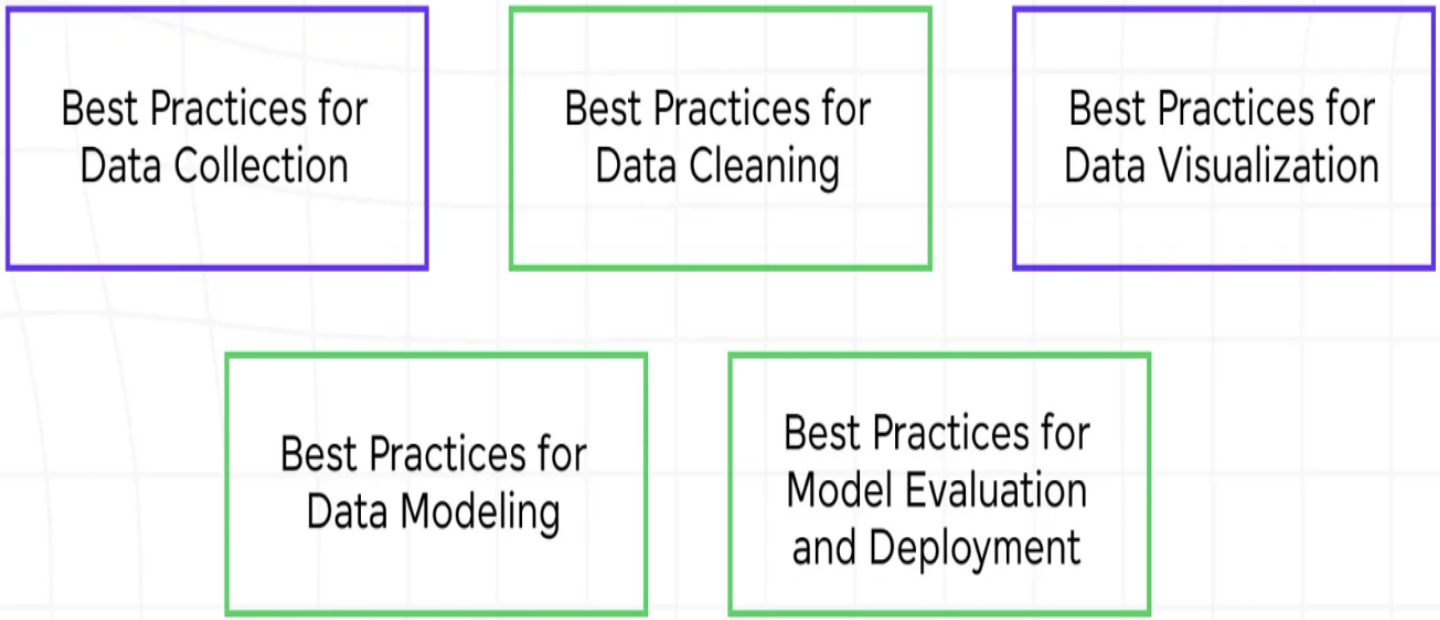
Data Modelling



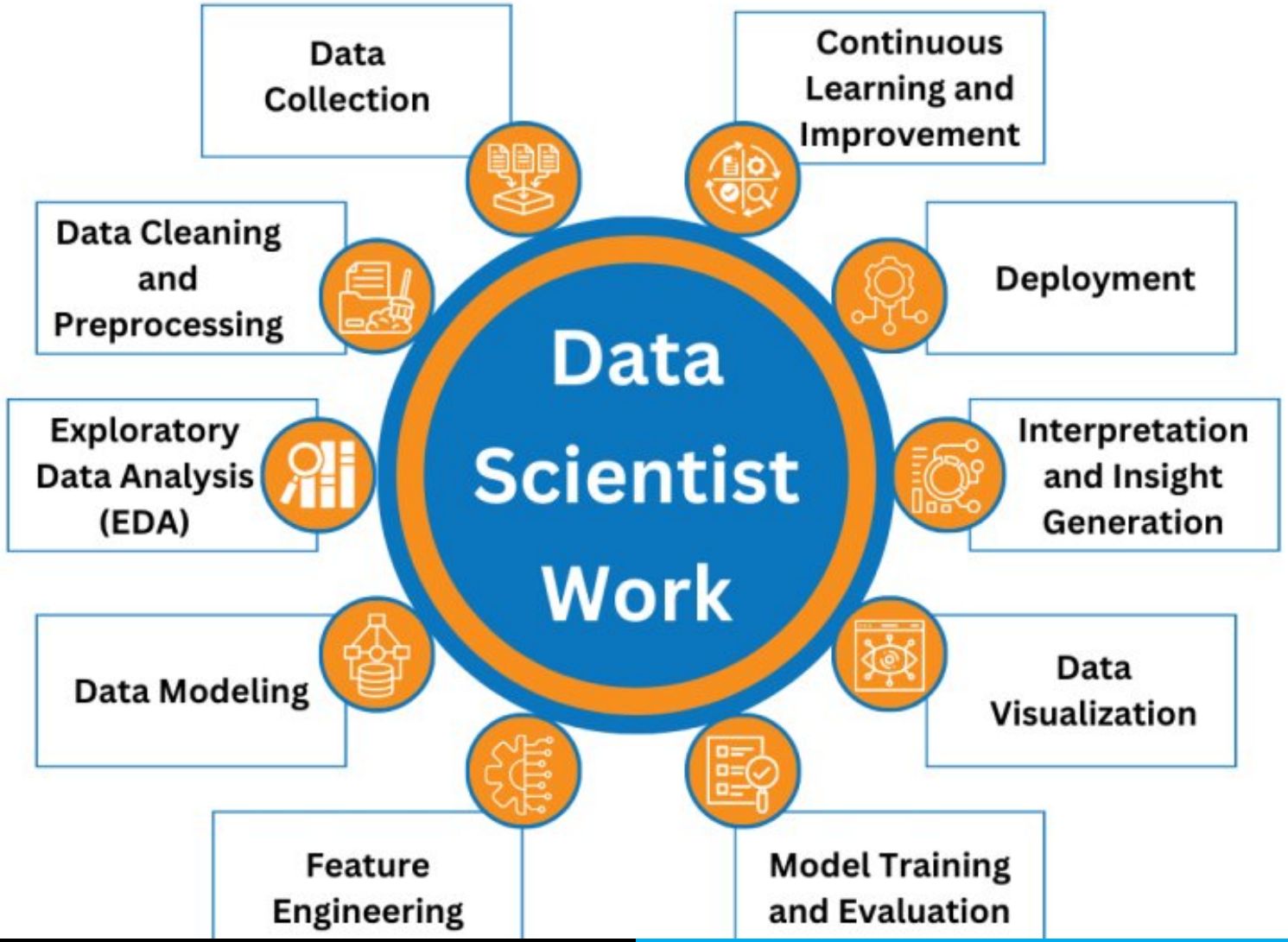
Steps in data modelling

- Defining the Problem
- Choosing the Right Model
- Training the Model
- Evaluating the Model
- Deploying the Model

Best Practices in Data Science



Data Scientist's Work



Skills of a data Scientist

Maths and statistics

- *Machine learning*
- *Statistical Modelling*
- *Exploratory Analysis*
- *Clustering*
- *Regression Analysis*

Programming and Database

- *Computer Science fundamental*
- *Database Management System*
- *Data Visualization*
- *Big Data*

Domain Knowledge and Software skills

- *Keen on working with Data.*
- *Problem Solver.*
- *Strategic proactive and cooperative.*
- *Interested in Hacking.*

Communication and Vizualisation

- *Storytelling Skills.*
- *Convert database insight into the decision.*
- *Collaboration with Sr. Manager*
- *Visual art design.*

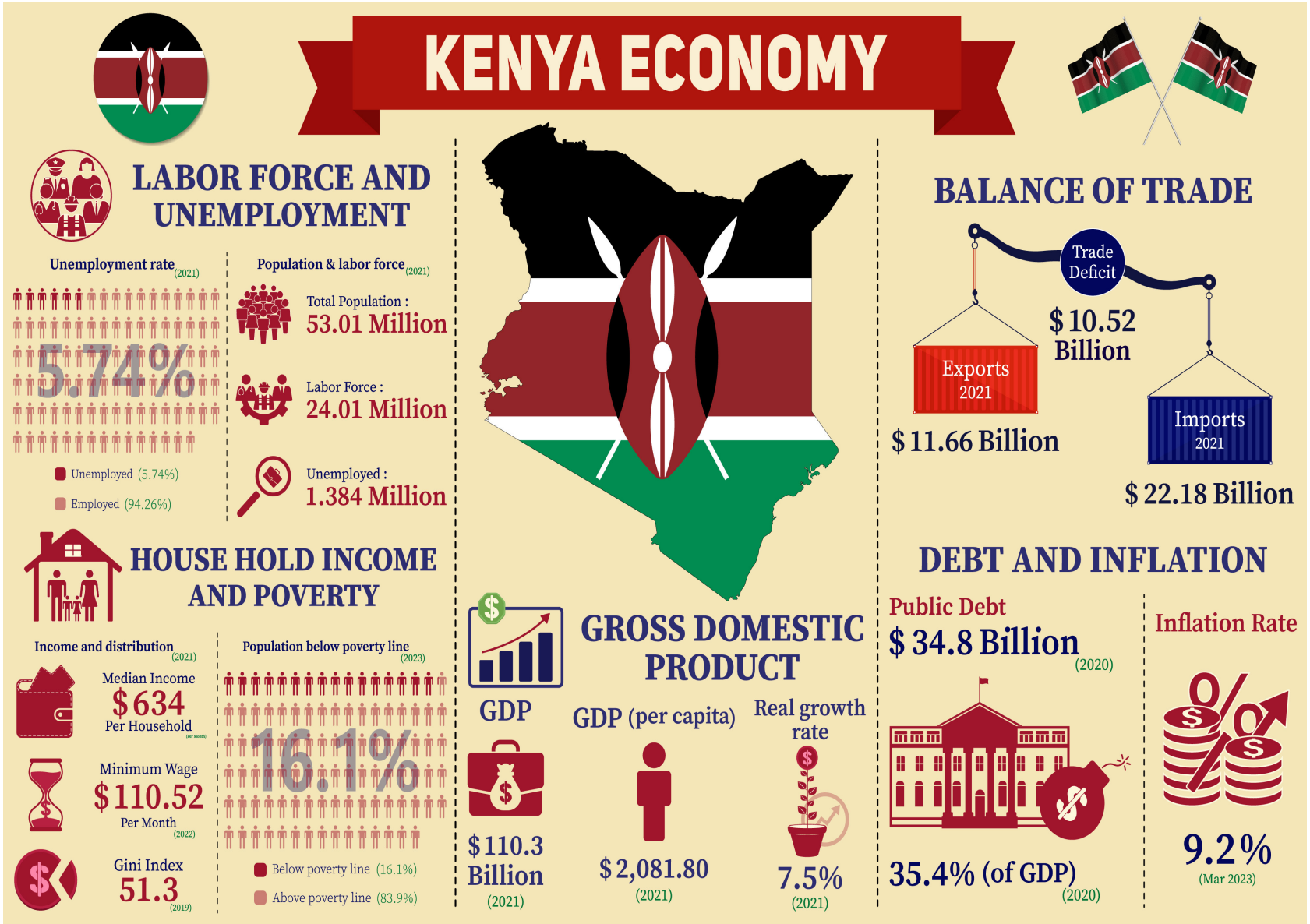
Vista Academy

Application of Data Science

Applications in Various Industries

1. Healthcare: Predictive analytics for patient outcomes, drug discovery.
2. Finance: Fraud detection, risk management, algorithmic trading.
3. Retail: Customer segmentation, inventory optimization, sales forecasting.
4. Marketing: Sentiment analysis, campaign effectiveness, customer behavior analysis.
5. Transportation: Route optimization, demand forecasting, autonomous vehicles.

A Case Study: Kenya's Economy



Markers of an economy:

- Gross Domestic Product (GDP): GDP Growth Rate: The percentage change in GDP from one period to another. A growing GDP indicates economic expansion, while a shrinking GDP may suggest a recession.
- Unemployment Rate: Measures the percentage of the labor force that is unemployed and actively seeking employment. High unemployment rates may indicate economic distress, while low rates suggest a healthy job market.
- Inflation Rate: The rate at which the general level of prices for goods and services is rising, often measured by the Consumer Price Index (CPI) or Producer Price Index (PPI). Moderate inflation is typically seen as a sign of a growing economy, but high inflation can erode purchasing power.
- Interest Rates: Lower interest rates can stimulate borrowing and investment, while higher rates are used to control inflation.

...Markers of an economy:

- Trade Balance: A trade surplus occurs when exports exceed imports, while a trade deficit is the opposite. A persistent trade deficit might indicate problems in competitiveness or an over-reliance on foreign goods.
- Currency Strength: The value of the country's currency relative to others affects trade, investment, and purchasing power. A strong currency generally reflects economic stability, while a weak currency might signal inflation or lack of confidence.
- Poverty and Inequality: TPoverty Rate: The percentage of people living below the poverty line. Gini Coefficient: A measure of income inequality within a country. High inequality can indicate social instability and economic inefficiency.
- Foreign Direct Investment (FDI): The level of investment by foreign entities in domestic businesses and assets. High FDI often reflects confidence in the country's economic prospects and stability.

But where is data?

WORLD BANK GROUP | Data

This page in: English Español Français العربية 中文

Kenya

DataBank Microdata Data Catalog

Kenya

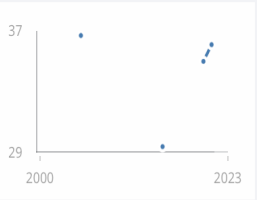
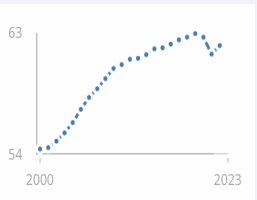
Overview By Theme By SDG Goal

Jump to

Graph, map and compare more than 1,000 time series indicators from the World Development Indicators.

Topic

Social

Indicator	Most recent value	Trend
Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)	36.1 (2021)	
Life expectancy at birth, total (years)	62 (2022)	

Download
CSV XML EXCEL

DataBank
Explore Our DataBank

Country Profile

Importing data to the working platform

jupyter ML_Kenya Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help Trusted

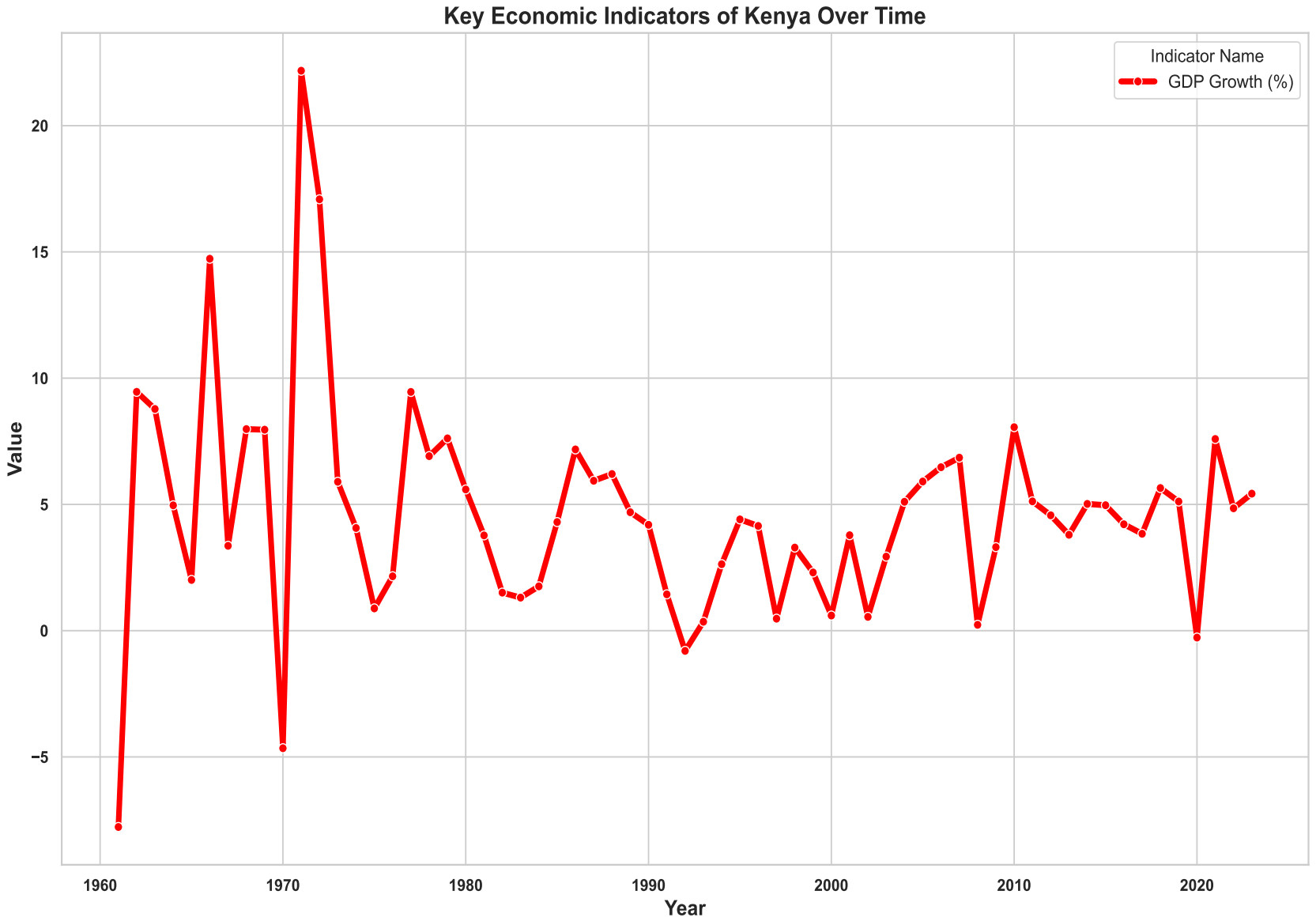
Code JupyterLab Python 3 (ipykernel)

[1]:
```
import wbdata
import numpy as np
import pandas as pd
import datetime
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
import seaborn as sns
```

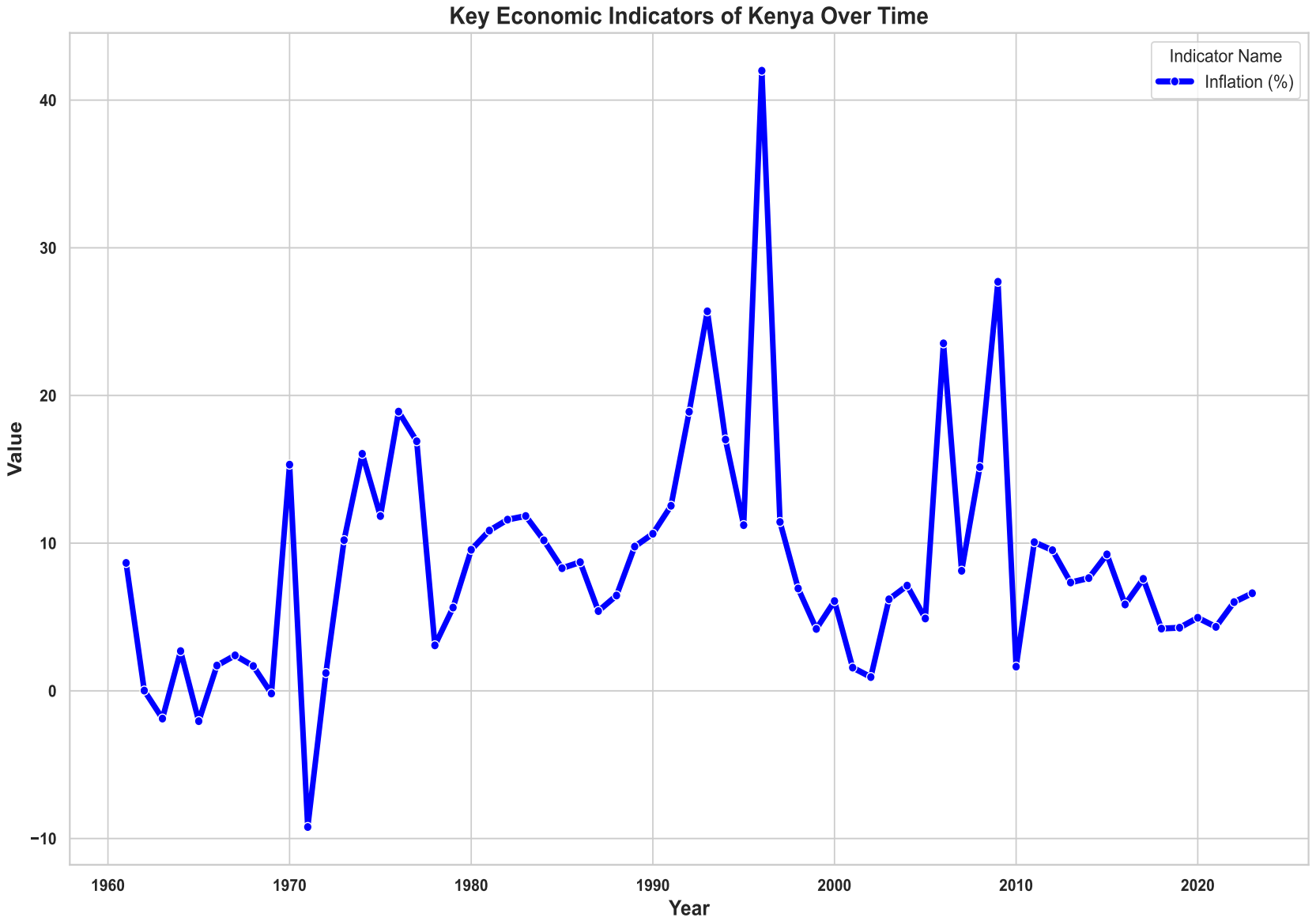
[3]:
```
indicators = {
    'NY.GDP.MKTP.KD.ZG': 'GDP Growth',          # GDP growth (annual %)
    'FP.CPI.TOTL.ZG': 'Inflation',              # Inflation (annual %)
    'SL.UEM.TOTL.ZS': 'Unemployment',           # Unemployment (% of total labor force)
    'FR.INR.LEND': 'Interest Rates',            # Lending interest rate (%)
    'NE.RSB.GNFS.CD': 'Trade Balance',          # Trade balance (current USD)
    'DT.TDS.DPPG.GN.ZS': 'Debt-to-GDP (%)'
}

# Step 2: Get the data for Kenya (country code 'KEN') from 1963 to the present
data = wbdata.get_dataframe(indicators, country='KEN')



# Step 3: Sort the data by date

data = data.sort_index(ascending=True)

# Step 4: Reset index to move 'date' into a column
data.reset_index(inplace=True)

data.loc[(data.index == 63), "Debt-to-GDP (%)"] = 2.451205

# Step 5: Display the DataFrame to ensure data is pulled correctly
print(data)
```

```
    date  GDP Growth  Inflation  Unemployment  Interest Rates  Trade Balance  \
0   1960         NaN   1.243781           NaN             NaN  -2.037001e+07
1   1961   -7.774635   2.457002           NaN             NaN   1.258599e+07
2   1962    9.457359   3.117506           NaN             NaN   1.660398e+07
3   1963    8.778340   0.697674           NaN             NaN   2.844799e+07
4   1964    4.964467  -0.099305           NaN             NaN   4.172000e+07
..   ...         ...        ...           ...             ...            ...
59  2019    5.114159   5.239638         5.014       12.441133  -8.937167e+09
60  2020   -0.272766   5.405162         5.621       11.995785  -8.007858e+09
61  2021    7.590489   6.107936         5.693       12.079998  -1.002792e+10
62  2022    4.846635   7.659863         5.805       12.335841  -1.054698e+10
63  2023    5.425816   7.671396         5.682       13.588502  -9.393300e+09
```

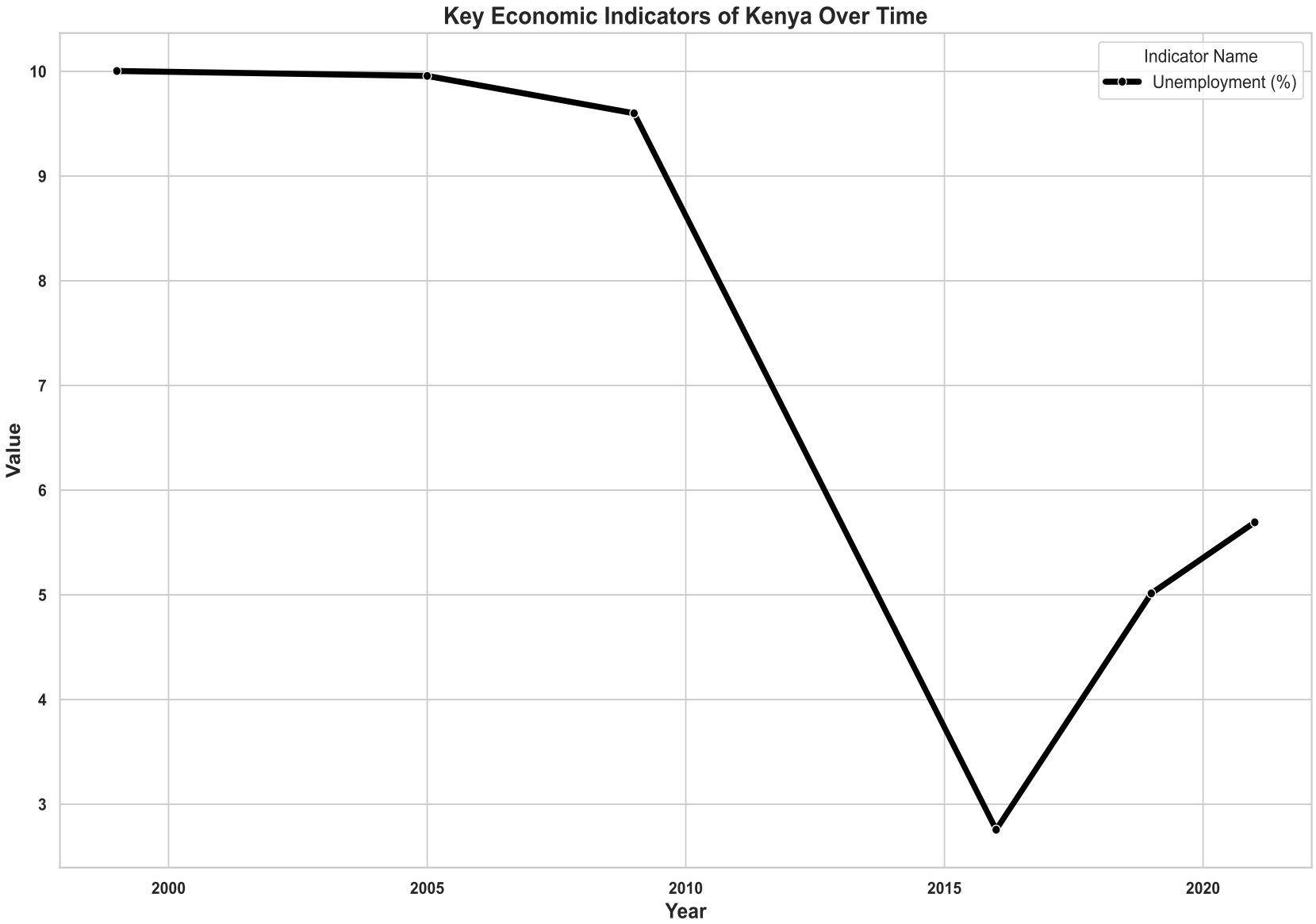
Kenya's GDP over time



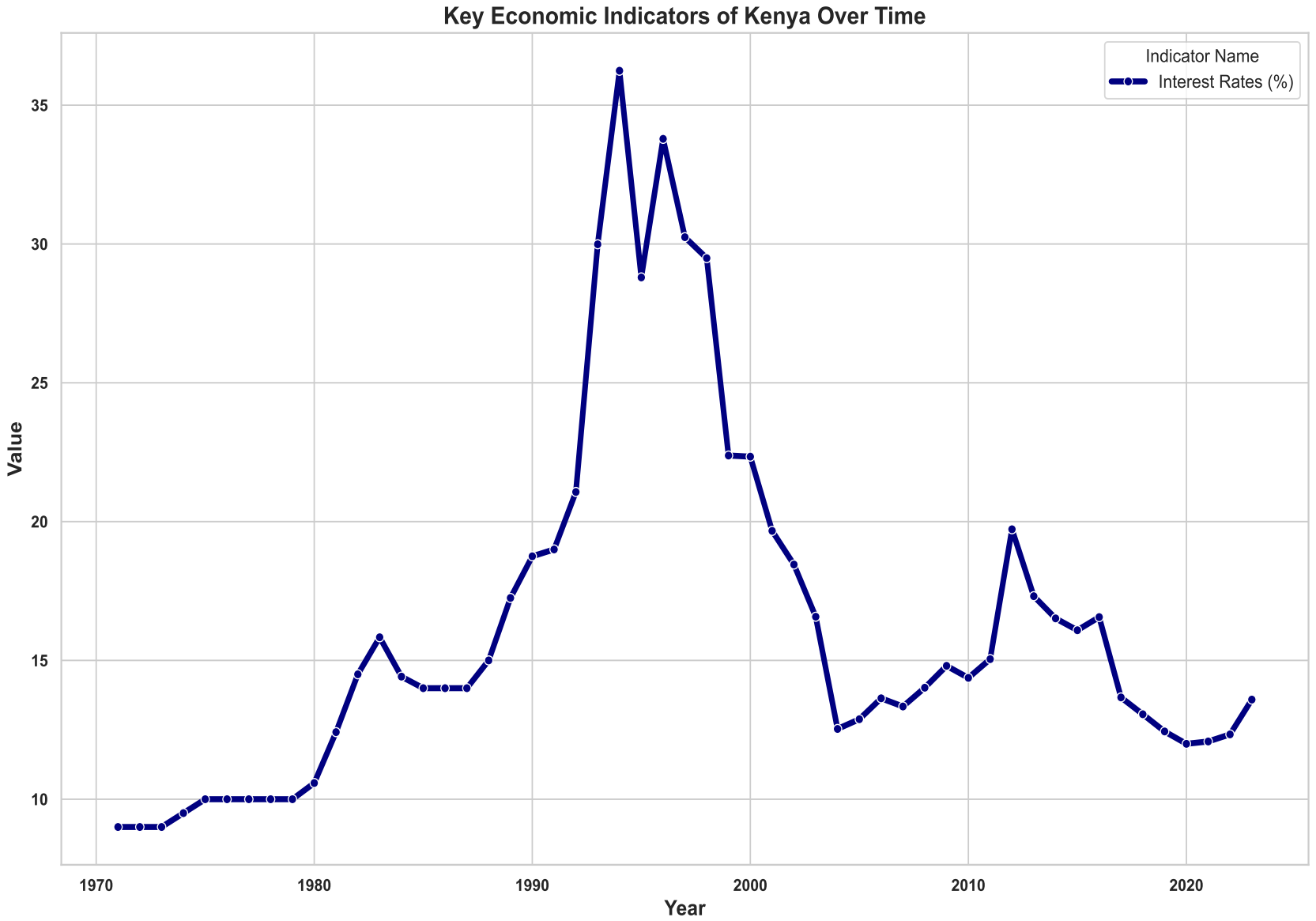
Kenya's Inflation over time



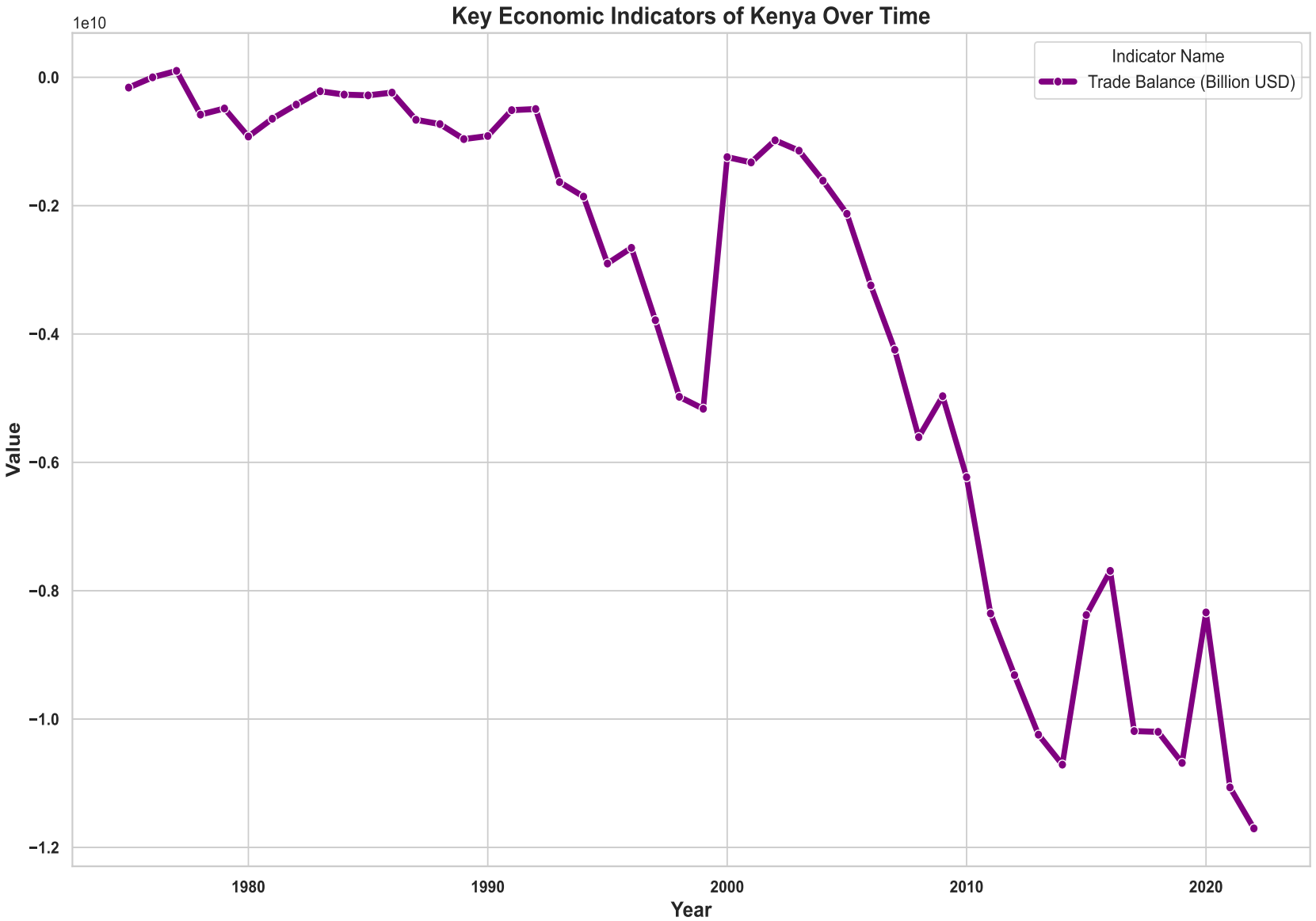
Kenya's Unemployment over time



Kenya's Interest Rates over time

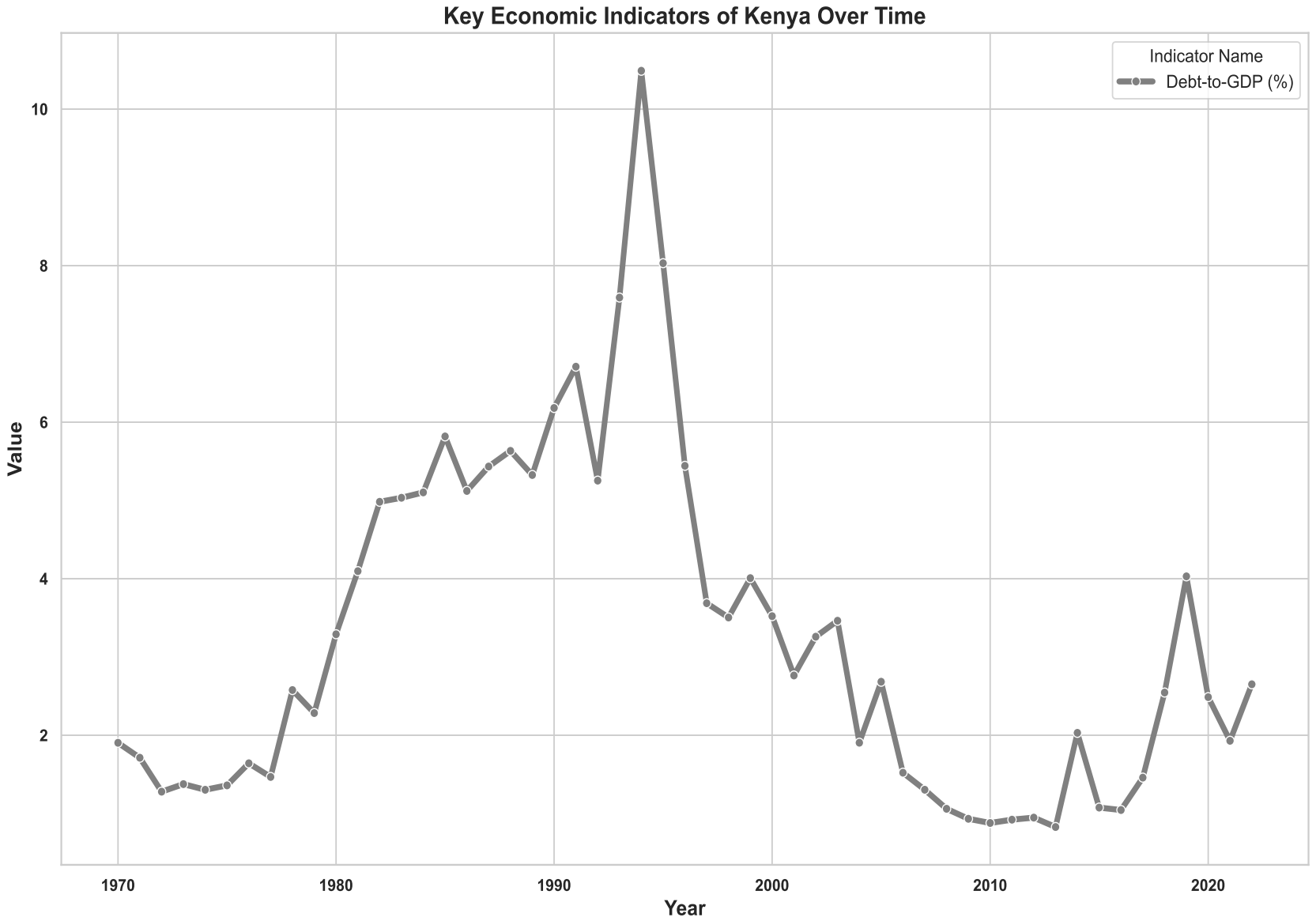


Kenya's Balance of Trade over time



Kenya's Debt-to-GDP ratio over time

Public and publicly guaranteed debt service (% of GNI)



Machine learning model

A **Random Forest Regressor** is a machine learning model that's part of the ensemble learning method, designed to make predictions for continuous outcomes. It's built on decision trees, but instead of using a single tree, it combines many individual decision trees to form a "forest" of predictions. Random forests are widely used for regression tasks because they provide more accurate and stable predictions than single decision trees.

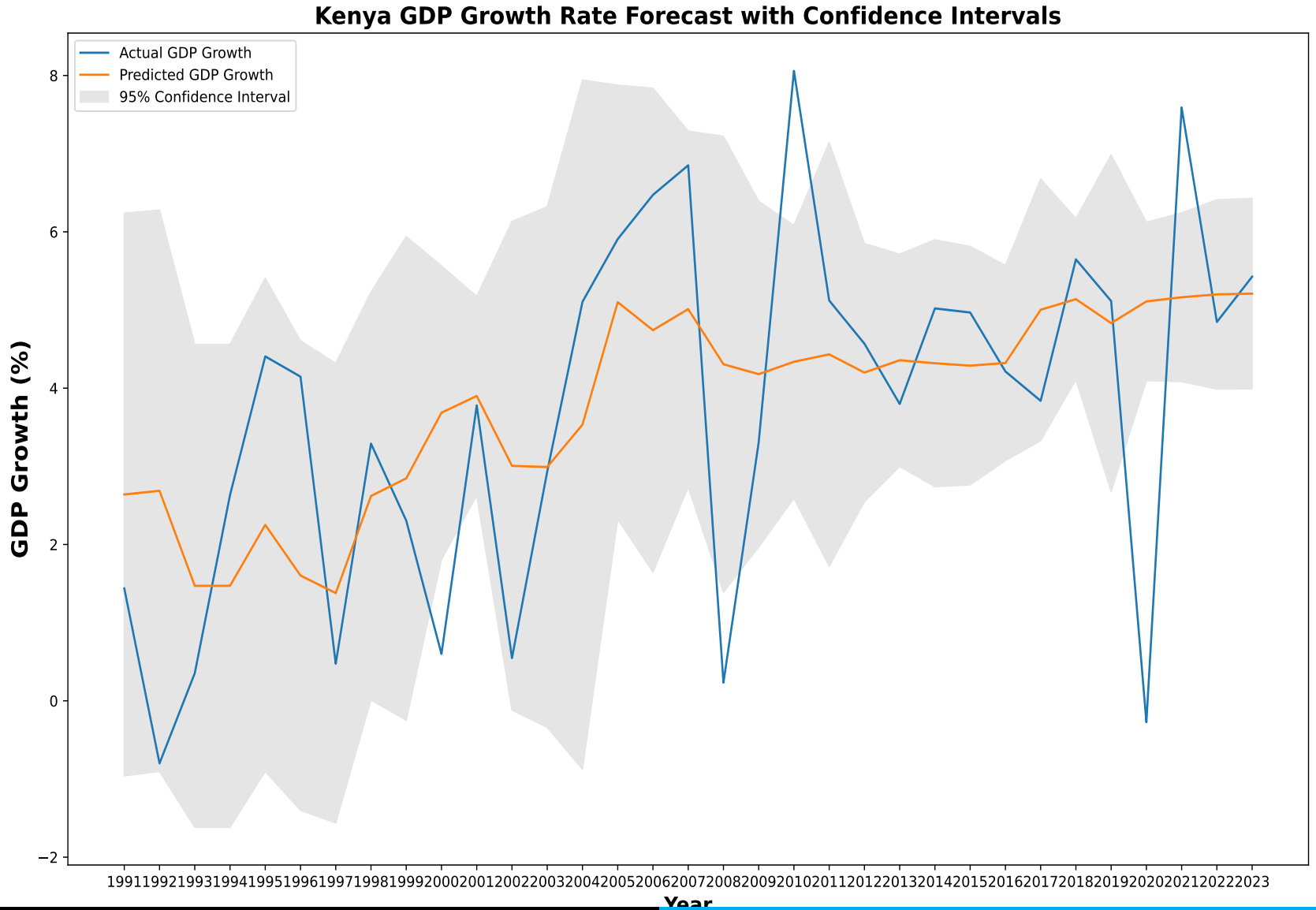
- Handles Non-linearity: By averaging multiple trees, Random Forest can model complex patterns that individual trees might miss.
- Reduced Overfitting: Since each tree in the forest is trained on different data and features, the overall model tends to generalize better than individual trees, reducing overfitting.
- Handles Missing Values: Random Forest can handle missing data relatively well by averaging predictions from different trees, each trained on different parts of the data.

How Random Forest Regressors work

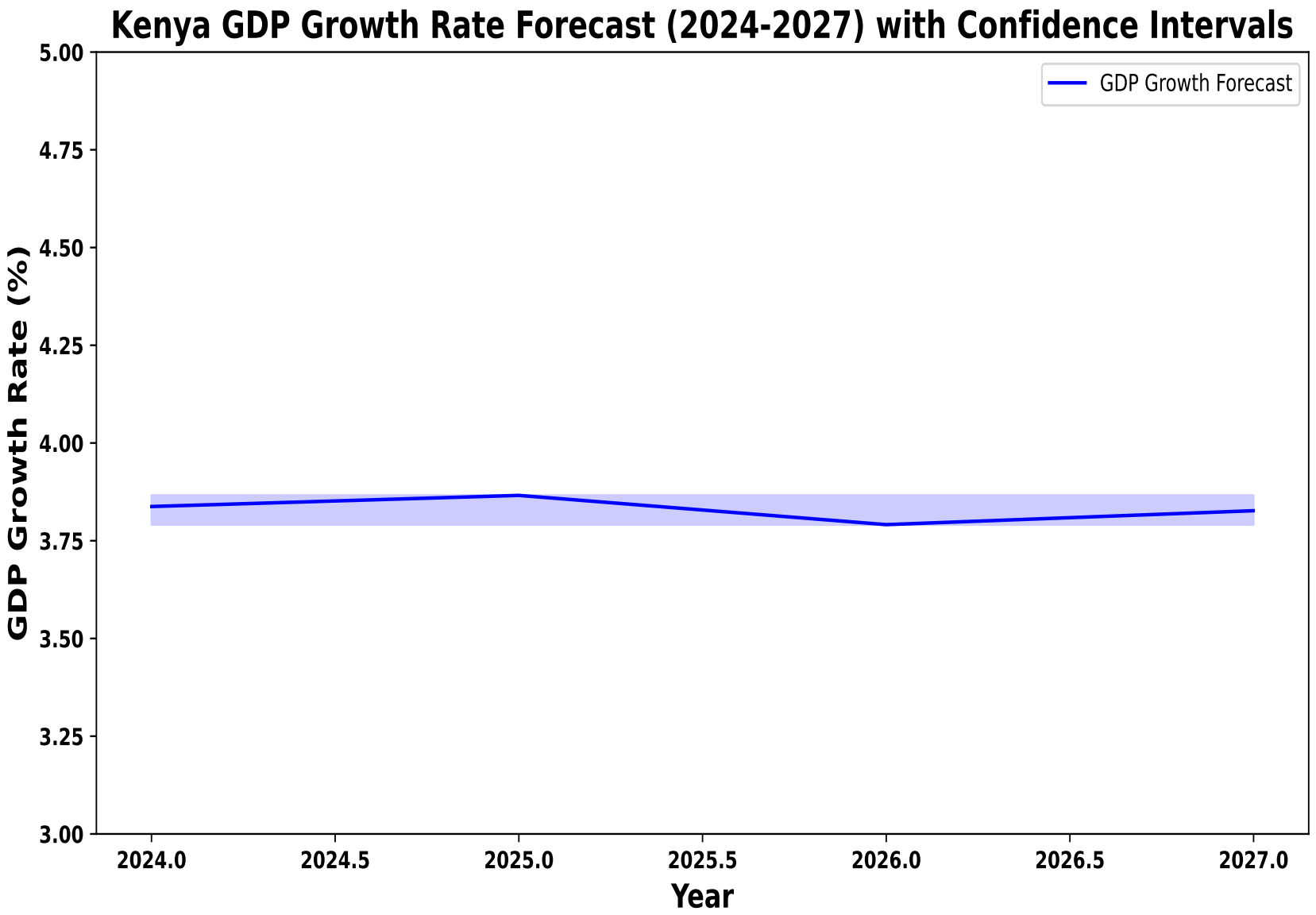
- Creating Decision Trees: Random Forest begins by creating multiple decision trees, each trained on a random subset of the original data. This is called bootstrapping, where data samples are drawn with replacement.
- Random Feature Selection: For each decision tree, only a random subset of features is considered at each split, preventing trees from becoming too similar.
- Ensemble of Predictions: Once all trees are trained, each tree makes a prediction, and the Random Forest takes the average of these predictions to arrive at a final result.

Random Forest Regressor

Predictors: Inflation, Unemployment, Balance of trade, Interest rates, Debt-to-GDP

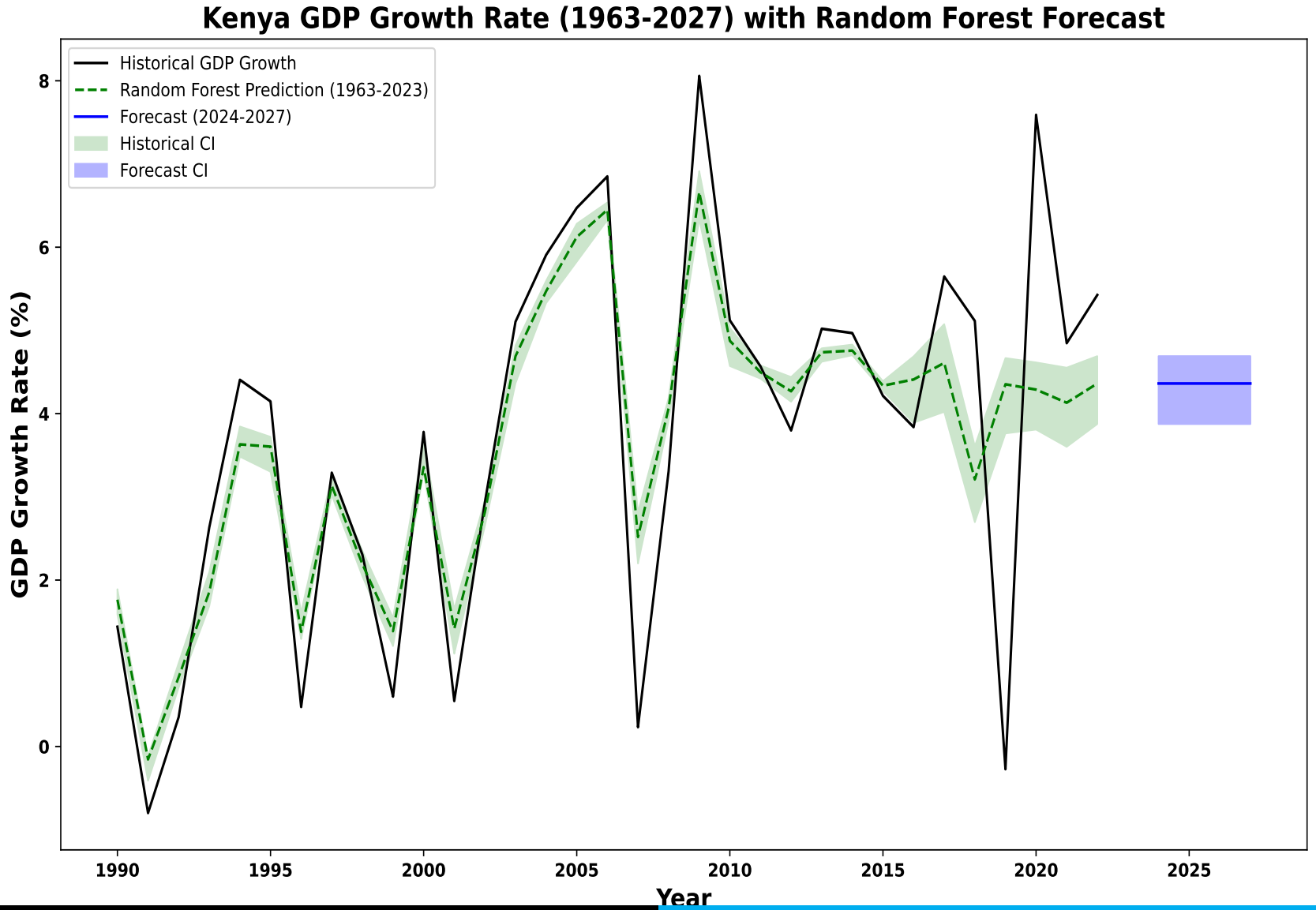


Random Forest Regressor forecast



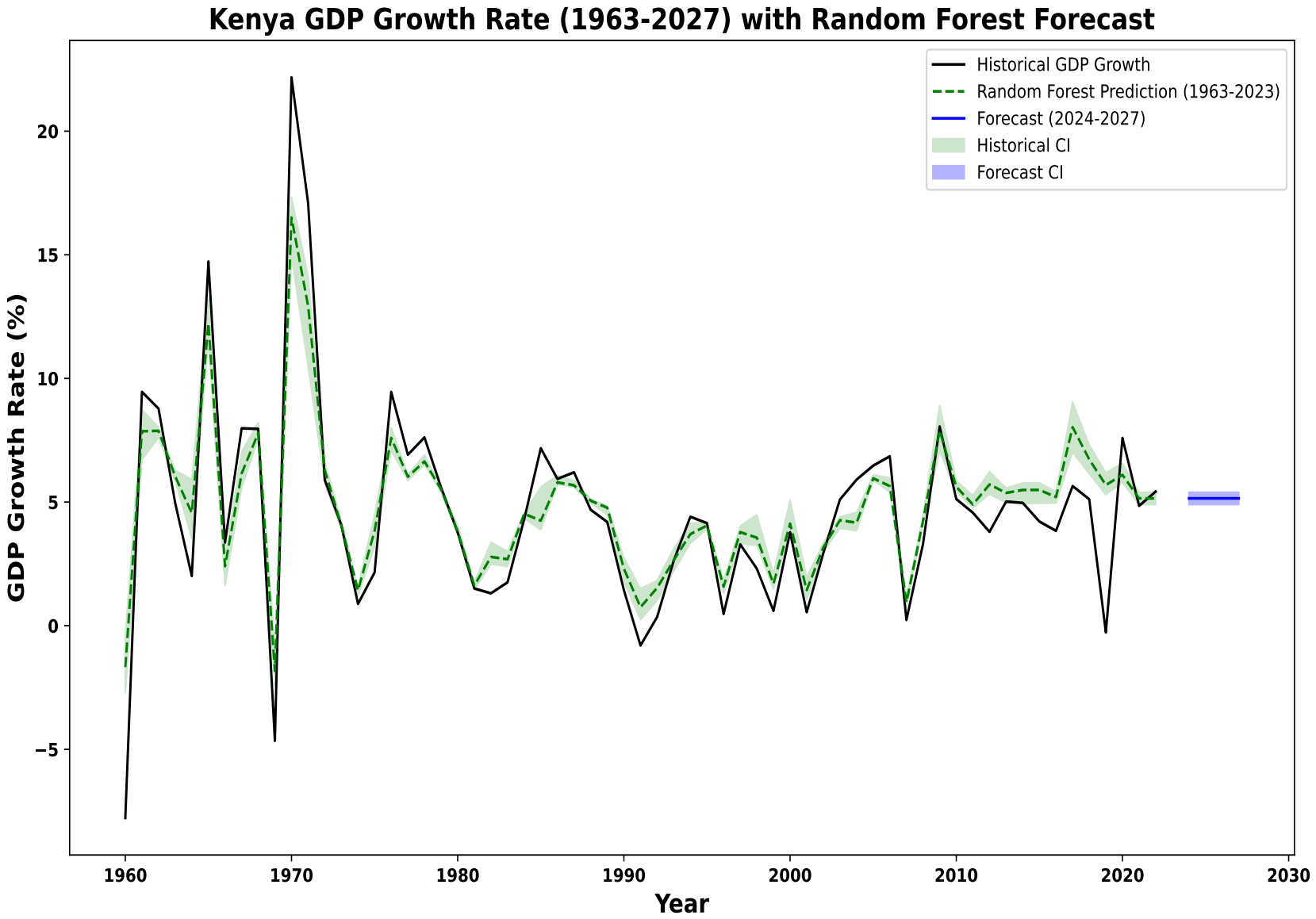
An ensemble of Random Forest Regressors

Predictors: Inflation, Unemployment, Balance of trade, Interest rates, Debt-to-GDP



An ensemble of Random Forest Regressors

Predictors: Inflation & Balance of trade



Deep Learning model

The Long Short-Term Memory (LSTM) model is a type of artificial neural network specifically designed for sequential or time-series data. Developed as a variant of recurrent neural networks (RNNs), LSTMs are particularly good at capturing long-term dependencies in data. This ability makes them valuable in applications where context over time is important, such as:

- Time-series forecasting (e.g., predicting stock prices, economic indicators, or weather patterns)
- Natural language processing (e.g., language translation, speech recognition)
- Sequential decision-making (e.g., in reinforcement learning)

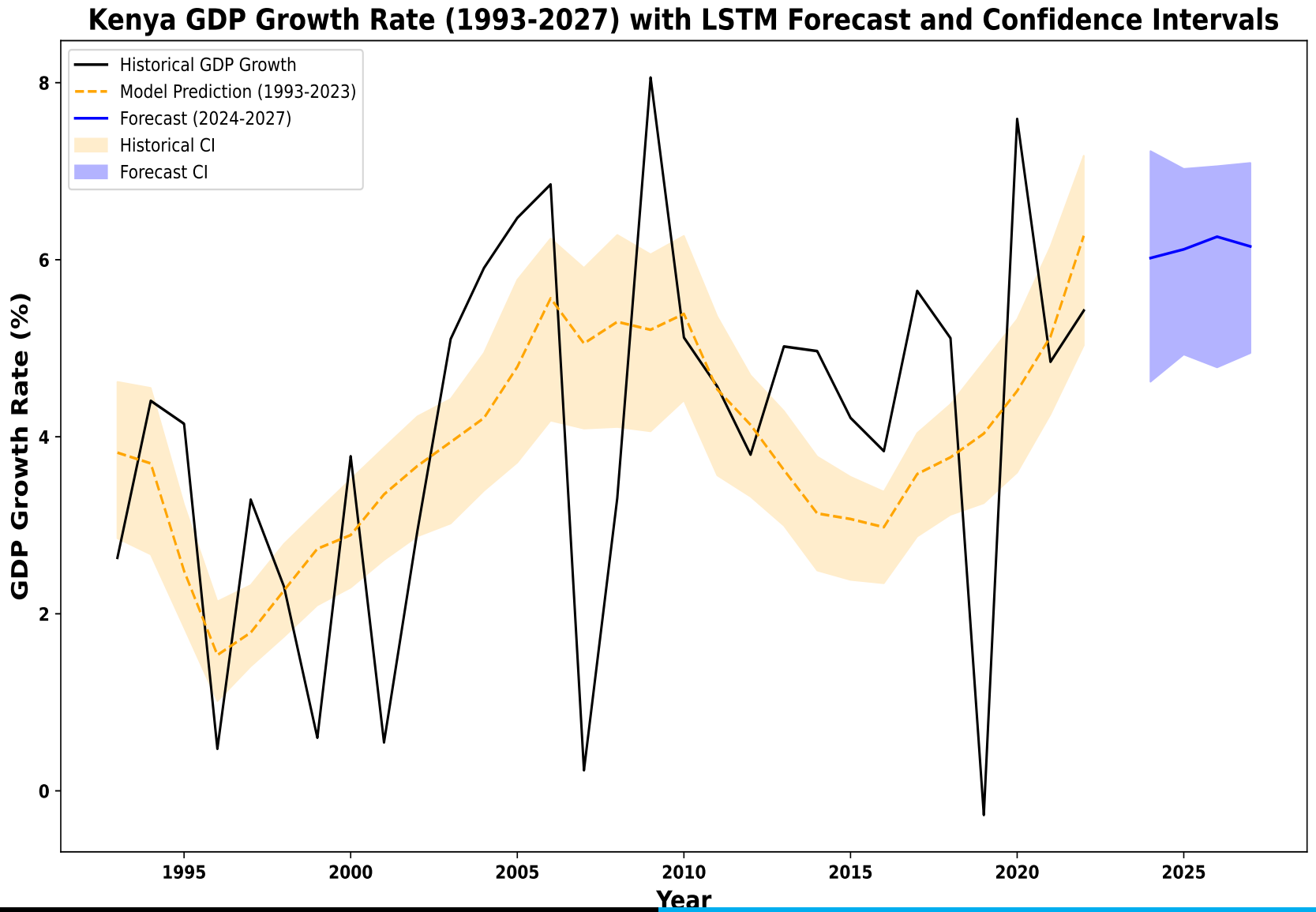
How LSTMs work

Traditional RNNs often struggle with retaining information across many time steps due to the "vanishing gradient problem," where earlier signals in long sequences get diluted as they propagate through the network. LSTMs solve this by using a memory cell with a special structure that regulates the flow of information. This structure is composed of:

- Input Gate: Controls how much new information should enter the cell.
- Forget Gate: Decides which information from the cell should be discarded.
- Output Gate: Determines what information from the cell should be output for the next time step.

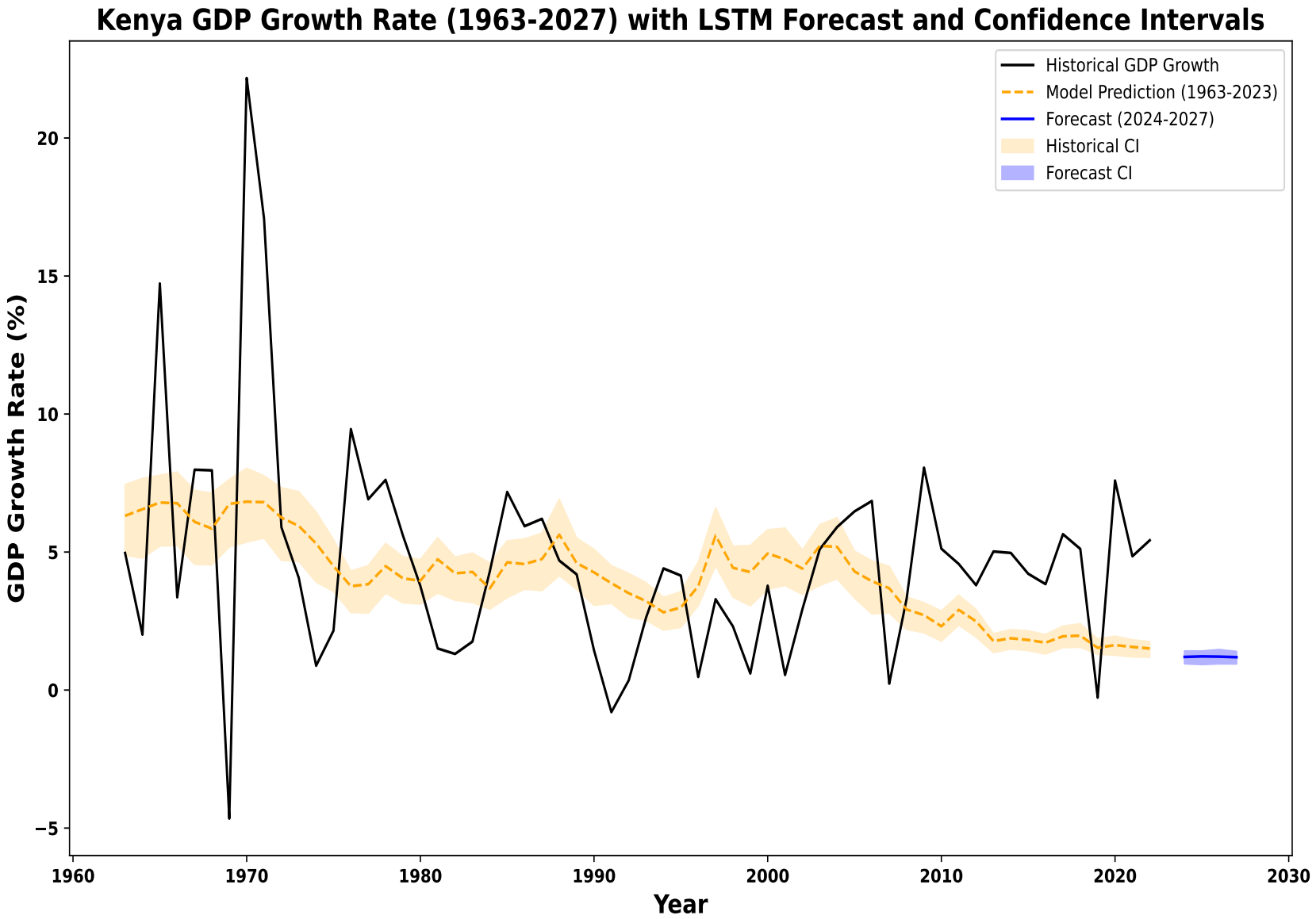
ANN: Long Short-Term Memory (LSTM) model

Predictors: Inflation, Unemployment, Balance of trade, Interest rates, Debt-to-GDP



ANN: Long Short-Term Memory (LSTM) model

Predictors: Inflation, & Balance of trade



Conclusion

What we have seen so far is as follows:

- An overview of data science and what data science can do.
- Application of data science in exploring Kenya's economic indicators.
- Using regression trees to forecast Kenya's economic growth
- Using deep learning model to predict Kenya's GDP
- An ensemble of random forest regressors (with two predictor variables) posted a better prediction than that of LSTM model. LSTM model performs well with more predictor variables than with few.

Acknowledments

1. MMUST for providing a working environment.
2. Directorate of Research in MMUST.
3. Professor Orata for this initiative.
4. Participants in this talk.

The end