PRIVATE HIGHER SCHOOL OF INFORMATION TECHNOLOGY AND MANAGEMENT OF NABEUL

Machine Learning Mini Project Report

Made By

Abderrazak Bouasker

Predicting Gdp with Machine Learning

Made Within

ITBS

Supervised By

University Supervisor

Mr. Ahmed Ben Taleb

College Year
2023-2024

Acknowledgments

I would like to express my heartfelt gratitude to Mr. Ahmed Ben Taleb, whose unwavering support and guidance have been instrumental in the completion of this project. Mr. Ben Taleb's insightful feedback, dedication to teaching, and encouragement have been invaluable throughout this journey. His expertise and passion for the subject matter have inspired me to strive for excellence. I am truly thankful for the impact he has had on my academic and personal growth.

Contents

List of Figures

General Introduction

In the era of data-driven decision-making, the application of machine learning techniques to economic forecasting has emerged as a powerful tool. This project focuses on leveraging machine learning algorithms to predict the Gross Domestic Product (GDP) for Tunisia, a nation at the crossroads of Africa and the Mediterranean. The GDP is a key indicator of a country's economic health, representing the total value of all goods and services produced over a specific time period.

Tunisia, with its diverse economic sectors, ranging from agriculture and manufacturing to services, presents a complex landscape for economic analysis. Accurate and timely predictions of GDP can provide invaluable insights for policymakers, investors, and businesses, aiding in strategic planning, resource allocation, and risk management.

Chapter 1

Project Presentation

1.1 Introduction

This project revolves around the application of machine learning techniques to predict the Gross Domestic Product (GDP) for Tunisia. The GDP serves as a vital economic indicator, reflecting the overall health and performance of a nation's economy. With Tunisia's diverse economic landscape, encompassing agriculture, manufacturing, and services, accurate GDP predictions are essential for effective policy-making and strategic planning.

By leveraging historical economic data and employing advanced machine learning algorithms, we aim to develop a predictive model that can unveil patterns and trends within the data. The outcomes of this project have the potential to offer valuable insights for policymakers, investors, and businesses, contributing to informed decision-making in the dynamic economic context of Tunisia. This introduction sets the stage for a comprehensive exploration of machine learning applications in economic forecasting for the nation.

1.2 Problem Statement

1.2.1 Introduction

In the realm of project development, the problem statement serves as the cornerstone, defining the central issues or challenges that a project aims to address. It is a concise articulation of the core problem at hand, providing a clear understanding of the gap or deficiency the project seeks to remedy.

1.2.2 Problem

Tunisia have been in a bad situation economically in the last ten years with major shifts in its budget , debt, and raising interest rates there are a lot of uncertainties around the future of the country and its economy .

1.2.3 Solution

The solution that we've found was to create a GDP prediction model using Machine Learning .

1.2.4 Conclusion

After determining the problem statement and solution for it we pass to the specification document .

1.3 Specifications Document

1.3.1 Introduction

In the landscape of project management and development, precision is paramount. The Specification Document stands as a foundational pillar in this pursuit, providing a detailed blueprint that outlines the requirements, functionalities, and parameters essential for the successful execution of a project.

1.3.2 Functional requirements

The Machine Learning model need to be accurate well generalized and not over-fitted or under-fitted .

1.3.3 Non-Functional requirements

- Easy to use user interface .
- Results are simple to interpret and understand .

1.3.4 Conclusion

In this section we determined the functional and the non functional requirements of our project that we will be trying to achieve in it .

1.4 Conclusion

In this chapter we've showcased the problem statement the solution and function and non functional requirements of our project and will be passing to the next chapter for the Implementation .

Chapter 2

Implementation

2.1 Introduction

As we transition from the theoretical groundwork to the tangible manifestation of our project, the Implementation chapter serves as the crucible where ideas crystallize into reality. This section delves into the practical application of the methodologies, strategies, and technologies discussed in preceding chapters. It is here that the blueprint transforms into code, algorithms are executed, and the envisioned solution takes shape. In this chapter we'll see the development environment , the dataset and the exploratory data analysis .

2.2 Development Environment

2.2.1 Introduction

In this section we'll vue the development environment what technologies and softwares we used for the project implementation .

2.2.2 Used Technologies

- Python [The programming language used].
- Anaconda [A Python work environment].
- Jupyter / Visual Studio Code [A code editor].
- Flask [A Python web backend framework].

2.2.3 Conclusion

After getting an idea of our work environment we'll go to the next section of dataset exploration .

Figure 2.1: Jupyter logo

Figure 2.2: Anaconda logo

Figure 2.3: Flask logo

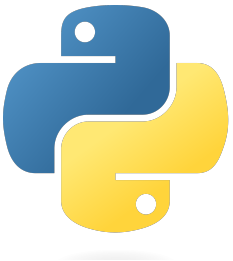

Figure 2.4: Python logo

Figure 2.5: Visual Studio Code logo

Figure 2.6: Used technologies

2.3 Dataset Presentation

2.3.1 Introduction

In this section we'll be exploring the dataset it's sources it's description and the exploratory data analysis .

2.3.2 Dataset Source

The Source for all the data in our Dataset is the Central Bank of Tunisia .

Figure 2.7: Central Bank of Tunisia Logo

2.3.3 Features Description

- GDP :
  - Type : Float
  - Unit : Million Dinar
  - Description : Gross domestic product measures the monetary value of final goods and services that is, those that are bought by the final user, produced in a country in a given period of time . It counts all of the output generated within the borders of a country .
- Total Indebtedness :
  - Type : Float
  - Unit : Million Dinar
  - Description : The total amount of debt the country have be it interior and exterior debt .

- Investment Rate :
  - Type : Float
  - Unit : Percent
  - Description : Percentage of the budget invested .
- Jobs Created :
  - Type : Float
  - Unit : Thousands
  - Description : The number of jobs created in that period of time .
- Trade Deficit :
  - Type : Float
  - Unit : Million Dinar
  - Description : The amount of money spent or gained from trade in that period of time depending on if the trade balance is negative or positive i.e., when the country import more than it exports or export more than it imports .

2.3.4 Exploratory Data Analysis

For the Exploratory Data Analysis we did multiple things such as :

- Boxplot to study the data distribution, the central tendency, spread, skewness of the data and seeing if there is outliers .
- Pairplot is a grid of scatterplots, histograms, and probability density functions, providing a quick overview of the relationships between multiple variables .
- Heatmap is used for the study of correlation between variables showcasing the degree of correlation by representing it with the darkness or lightness of the colors .

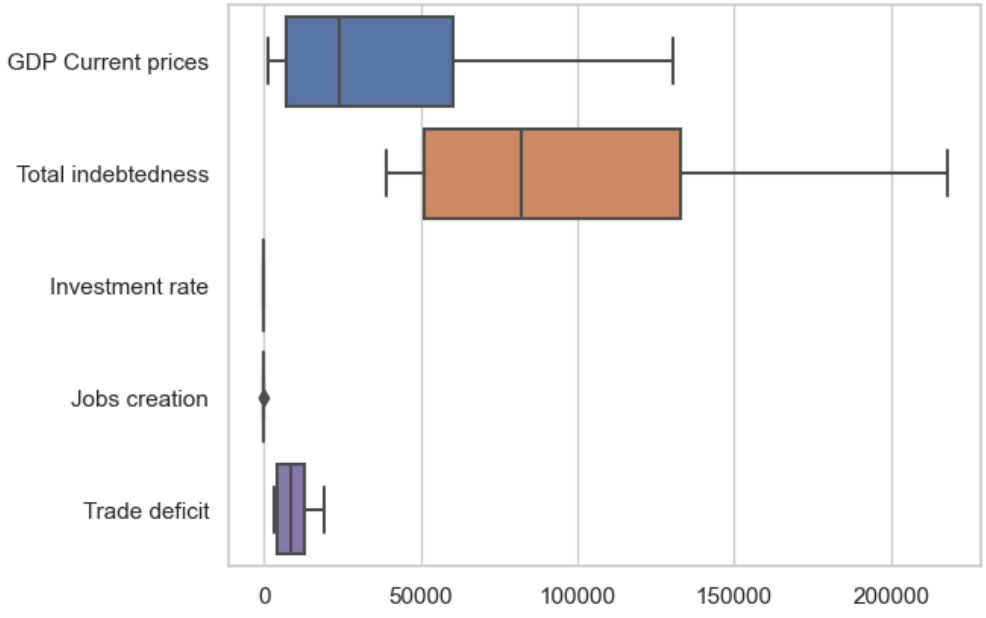

Figure 2.8: BoxPlot Before cleaning and normalization

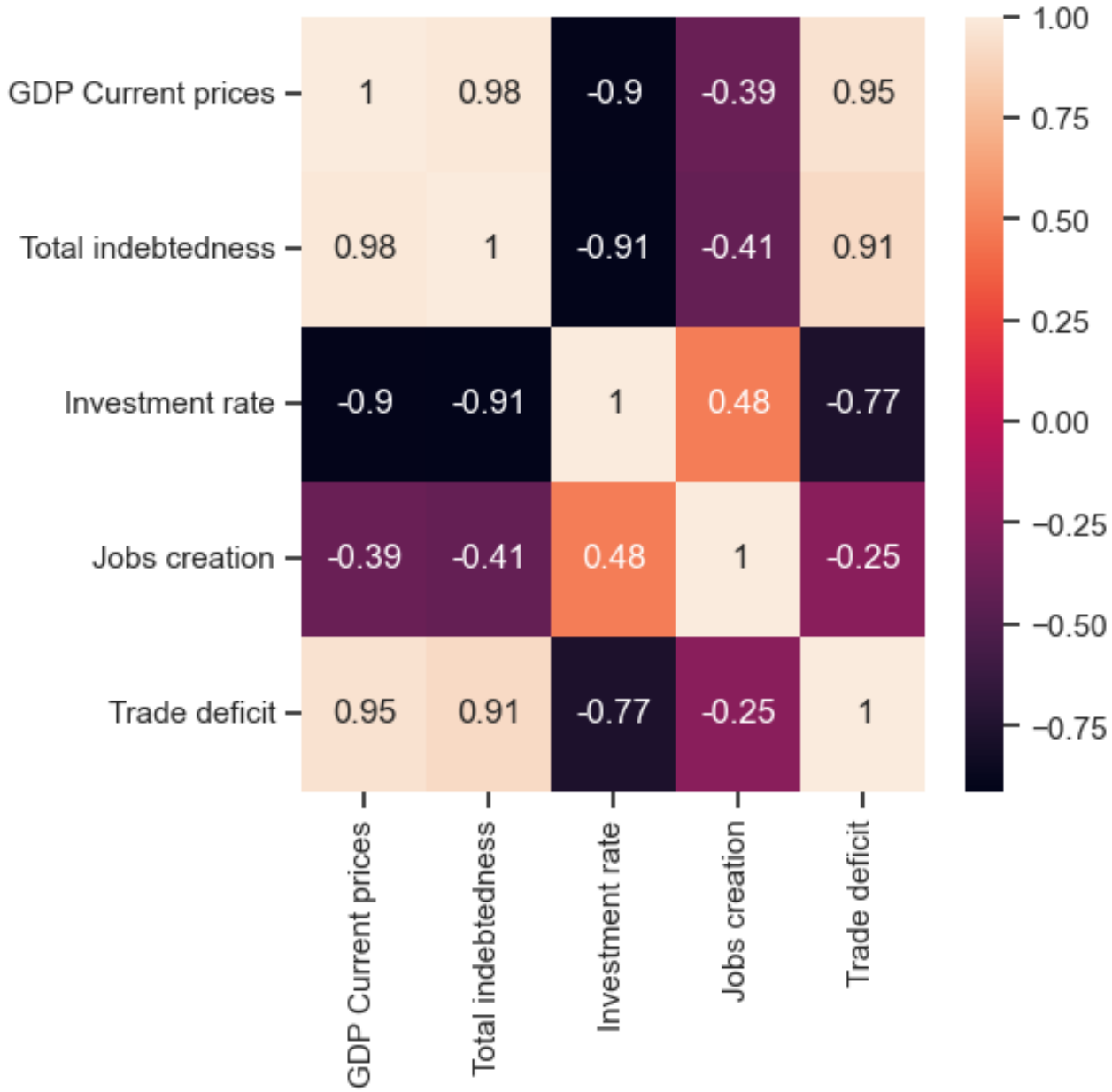

Figure 2.9: HeatMap Before cleaning and normalization

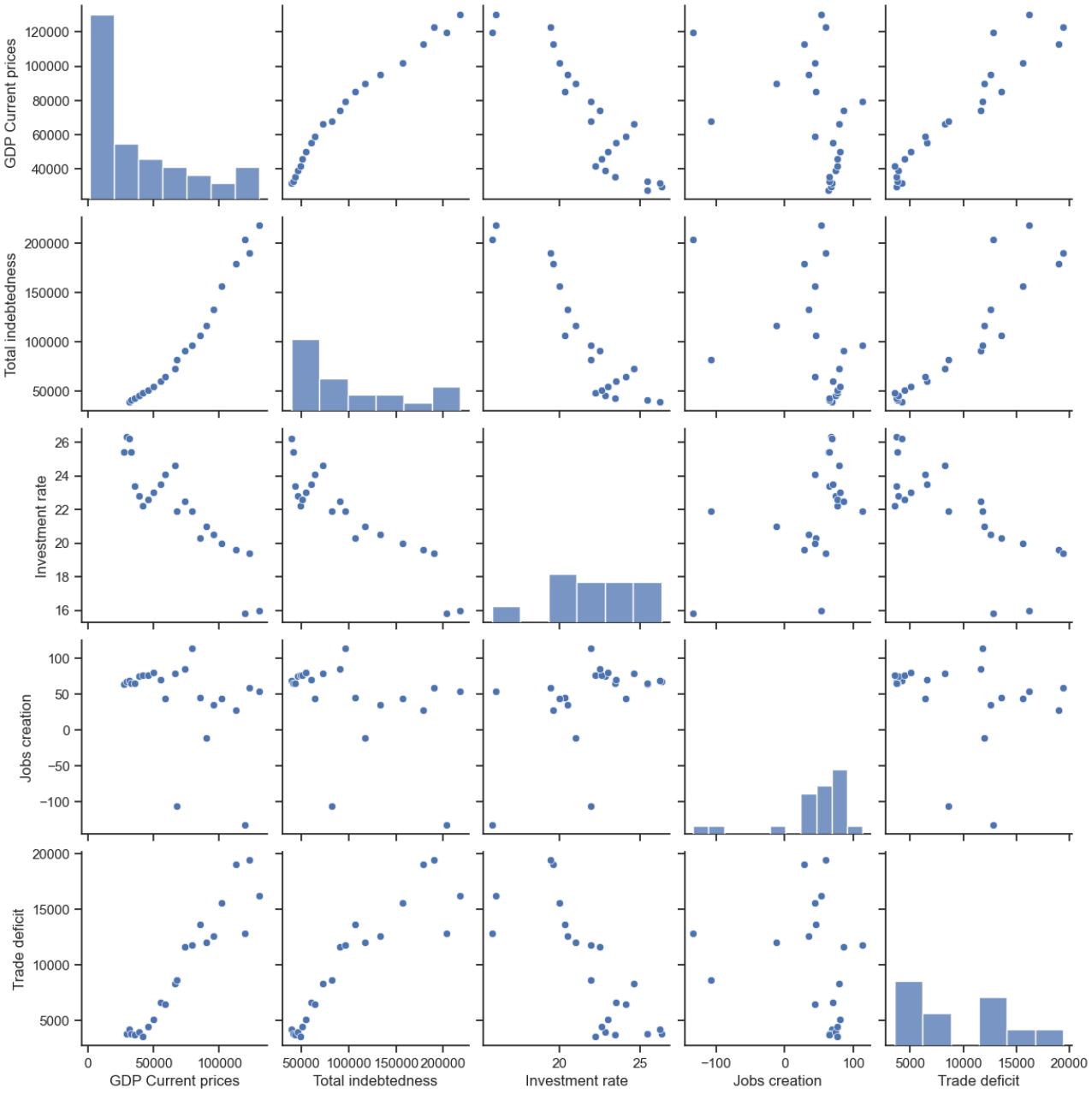

Figure 2.10: PairPlot Before cleaning and normalization

For the data cleaning we did :

- Remove the empty rows of data .
- Normalization of data to make all variables in the same dimension , we used the following methods :
  - Min Max Scaler for : GDP and Jobs Created .
  - Robust Scaler for : Total Indebtedness and Trade Deficit .
  - StandartScaler for : Investment Rate .

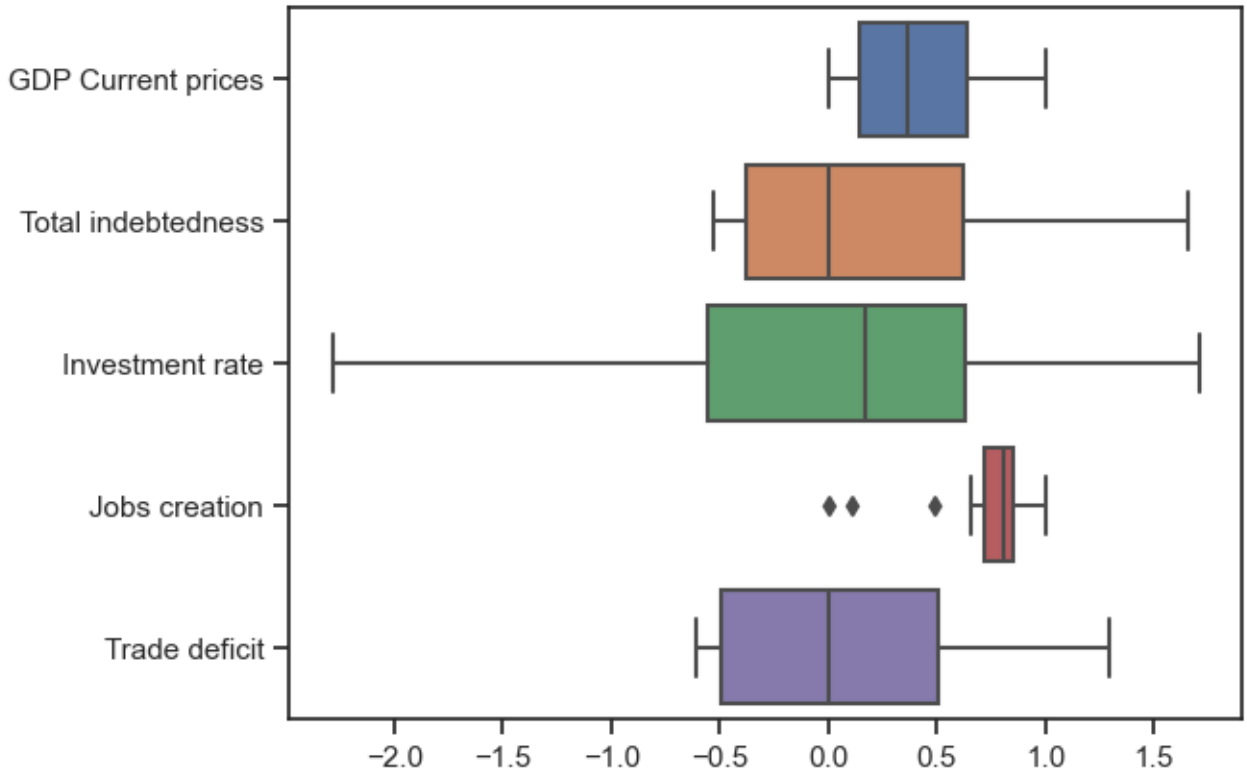

Figure 2.11: BoxPlot After cleaning and normalization

2.3.5 Conclusion

In this section we explored our data went through how we did exploratory data analysis and the different steps we went through to achieve our final dataset .

2.4 Conclusion

In this chapter we've explored our data beginning from the source where we got it from to describing it's variables and finishing with the exploratory data analysis in the next chapter we will be training our machine learning model .

Chapter 3

Model Training

3.1 Introduction

In the journey from data exploration to actionable insights, the chapter on model training marks a pivotal stage where the theoretical foundations and data-driven insights converge to give rise to predictive power. Model training is the transformative process where algorithms learn patterns and relationships within the data, allowing them to make informed predictions or classifications. This chapter unravels the intricacies of this training process, shedding light on the methodologies, considerations, and optimizations that underpin the development of robust and accurate machine learning models.

As we delve into the realm of model training, our focus extends beyond the theoretical constructs explored in earlier chapters. Here, we embark on a hands-on exploration of the data, leveraging machine learning algorithms to uncover hidden patterns and trends. From the selection of appropriate algorithms to the fine-tuning of hyperparameters, this chapter serves as a comprehensive guide to the practical aspects of model development.

The overarching goal is not merely the creation of predictive models but the cultivation of models that generalize well to unseen data, making them valuable tools for decision-making in real-world scenarios. Through a systematic journey, we will explore the nuances of feature engineering, model evaluation, and the iterative process of refining models to achieve optimal performance.

3.2 Model Choice

3.2.1 Introduction

The Choice of the Model is a crucial part of the process as it is the algorithm that will be doing the task the project is aiming to achieve so choosing the correct model is of the atmost importance .

3.2.2 Model Linear Regression

3.2.2.1 Introduction

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The simplest form is simple linear regression, which deals with the relationship between two variables, while multiple linear regression deals with two or more independent variables. The general form of a linear regression equation is expressed as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \epsilon$$

where:

- Y is the dependent variable (the variable we are trying to predict) .
- X1,X2,...,Xn are the independent variables .
- β is the intercept (the value of Y when all independent variables are zero) .
- $\beta 1, \beta 2, .., \beta n$ are the coefficients (representing the change in Y for a one-unit change in the corresponding X) .
- ϵ is the error term (representing the unobserved factors affecting Y) .

3.2.2.2 Advantages

- Simplicity and Interpretability :
  - Linear regression models are simple to understand and interpret. The coefficients provide insights into the strength and direction of relationships between variables .
- Interpolation :
  - Linear regression can be effective for predicting values within the range of the observed data, making it useful for interpolation .
- Efficiency :
  - Linear regression models are computationally efficient, making them suitable for large datasets and quick model training .

3.2.2.3 Disadvantages

- Assumption of Linearity :
  - The model assumes a linear relationship between the dependent and independent variables. If the relationship is not truly linear, the model may not capture the underlying patterns effectively .
- May Not Capture Complex Relationships :
  - For highly complex relationships or interactions among variables, linear regression may not capture the nuances as effectively as more advanced modeling techniques .
- Assumption of Independence :
  - The model assumes that the residuals (errors) are independent. If there is correlation among residuals, it can affect the model's accuracy .
- Assumption of Homoscedasticity :
  - Linear regression assumes that the variance of the errors is constant across all levels of the independent variable. If this assumption is violated (heteroscedasticity), the model's predictions may be less reliable .

3.2.3 Choice of linear regression

- Linear Relationships :
  - There is a linear relationship between the predictor variables (features) and GDP and historical data suggests that changes in certain economic indicators have a consistent linear impact on GDP and linear regression can capture this relationship effectively .
- Simplicity :
  - Linear regression is a simple and computationally efficient model. If the relationship between variables is straightforward and there's no need for a highly complex model, linear regression can provide a good balance between accuracy and simplicity .

3.2.4 Conclusion

In this section we went through our choice of linear regression it's advantages , disadvantages and the reasons why we chose it as our model .

3.3 Training

3.3.1 Introduction

Training a model is the step where we feed our model the data to learn from it and develop to be able to predict on other data it hasn't seen yet .

3.3.2 Data Split

Data splitting is the process of splitting our dataset to two part training and testing in our case because the dataset is small in size we split it in a 90% training 10% testing . We did try other splits like 80%-20% and 70%-30% they all gave close results so we adopted the 90%-10% to give the model the most data possible to train on .

3.3.3 Cross Validation

Cross Validation is splitting the train data into multiple parts to train and test on them . Cross Validation is mainly used to find the best model parameters . we use the train set because we don't want to train the model on the test set which we need as an isolated dataset to test our model generalization on it . In our project because of the small size of the dataset we adopted the "Leave one out cross validation" method which split the set into all possible segments i.e., every row is a segment and leaves only one segment to test on and the rest trains on them . after testing multiple models by changing everytime a different combinaison of features fed to the model and after comparing their error value we concluded that giving the model all the features is the best course of action that will give us the best possible result .

3.3.4 Test

The test part is giving the model after training the test values and letting it do the prediction and them comparing the given predictions with the actual result test values , we can determin the quality

and accuracy of our model by calculating the error using different methods , in our case we used 3 methods :

- MAE or Mean Absolute Error : Measure the average absolute differences between the predicted values and the actual values .
- MSE or Mean Squared Error : Measure the average squared differences between the predicted values and the actual values .
- R2 R squared : Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. R-squared values range from 0 to 1, with higher values indicating a better fit .

If the error values are good then we succeded in training the model and didn't fall to overfitting or underfitting . In the end our model gave us a value of R2 = 0.95 which is a very good value that satisfy the requirement for a good indicator of quality .

3.3.5 Model Saving

We saved the model as a pickle file to use it later directly without the need to retrain the model .

3.3.6 Conclusion

Training is the most important step and after completing it we are basically done and can start using the model as we wish .

3.4 Conclusion

After completing the mentioned steps in this chapter such as choosing a model , training a model and testing it we have achieved our functional requirements and we will be continuing to the next chapter in which we will create a user interface to ease the use of our model .

Chapter 4

Deployment

4.1 Introduction

The deployment of the project is the final step of our work , what is the use for the project if we cannot share it with other . In this chapter we will see how we made the user interface, how it looks and how to use it .

4.2 User Interface

For this project we have a created a simple user interface using html and css and for the backend of it we used flask . In the frontend we created we have two pages the first one contains four fields that the user fills with data that will be used to do the prediction , after pressing the button to predict , the backend will take the submitted values and give it to the model , take the returned prediction value from the model and return it to the user . The prediction value will be added to a dataset file and will be presented to the user which will be directed to our second page in a graph that showcases the GDP in function of the Total indebtedness . The graph is easy to understand and the user can see all the prediction points added to the graph and their values to explore how the different entered values effect the outcome .

4.2.1 Hosting

For the hosting of the site we used a cloud hosting service called ”Render” where we uploaded the application and run it . The link to access the application is : ”https://deploy-flask-bnbx.onrender.com”

Figure 4.1: Render logo

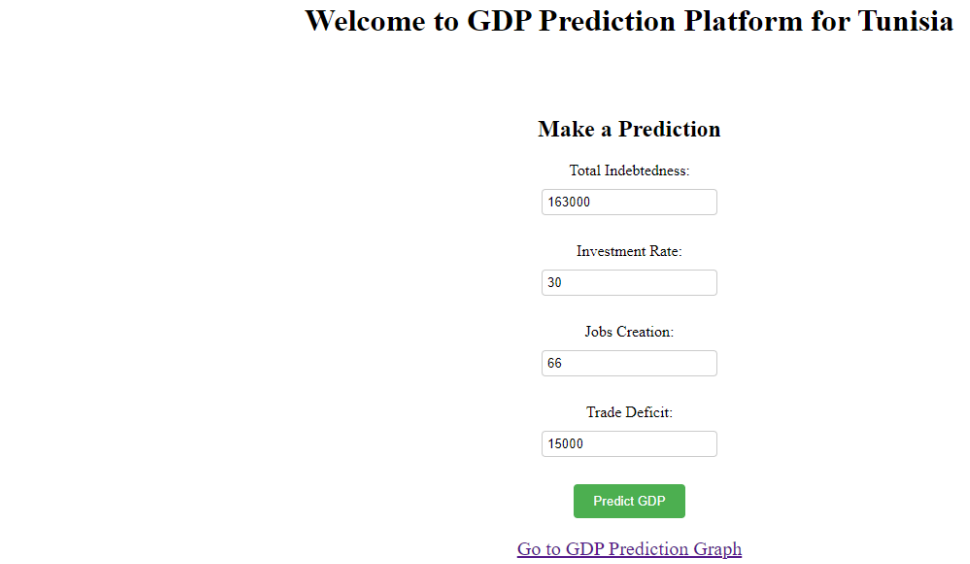

Figure 4.2: Prediction input User Interface

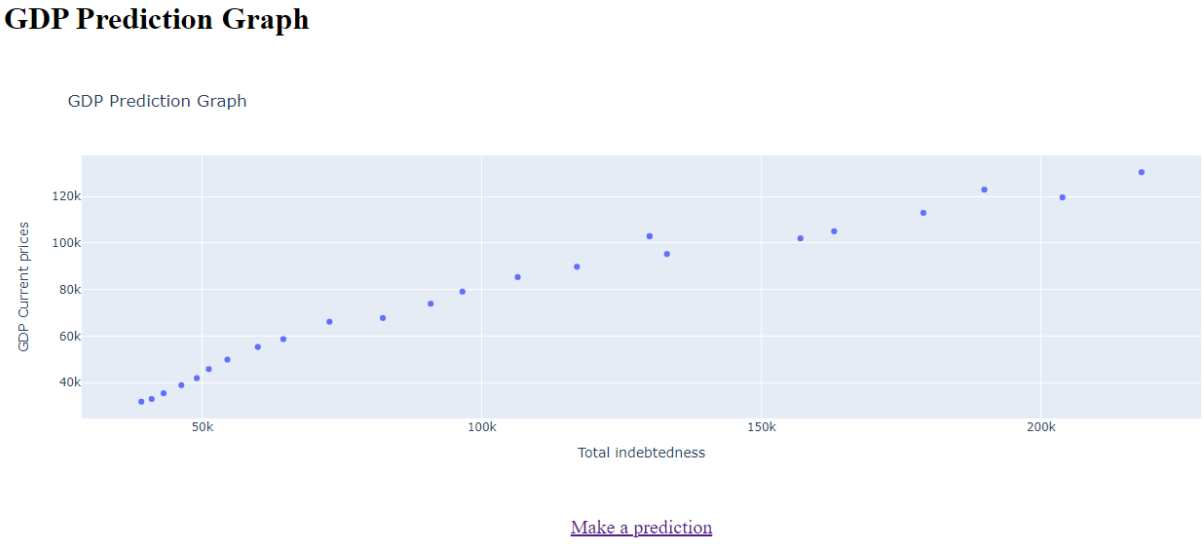

Figure 4.3: Prediction graph User Interface

4.3 Conclusion

In this chapter we talked about our User Interface its details have explained how to use it , going next to the last part of this report with the final conclusion .

Final Conclusion

In drawing the curtains on this comprehensive exploration, it becomes evident that our endeavor has yielded valuable insights and contributions to learning about Machine Learning and the world of Artificial Intelligence. Through meticulous data analysis, model development, and thoughtful interpretation, we have navigated the intricacies of GDP prediction with a commitment to precision and depth.

In concluding our project, our sincere hope is that the outcomes of our endeavors extend beyond the confines of this report and prove beneficial to others in the field of machine learning and artificial intelligence. We aspire for the knowledge generated and methodologies employed to become a valuable resource for fellow students and anyone with an interest in these cutting-edge disciplines. By sharing our findings and insights, we aim to contribute to the collective pool of knowledge, fostering a collaborative environment where ideas can flourish. Additionally, we hold the aspiration that our project not only meets but exceeds the expectations of our esteemed teacher.

In closing, we extend our gratitude to all those who contributed to this undertaking, recognizing the collaborative spirit that propels advancements and success. As we embark on the next steps in this journey, we do so with a renewed sense of purpose and the conviction that our endeavors today pave the way for a more informed and enlightened future.

Thank you for joining us on this intellectual voyage.

Bibliography

[bct,] Central bank of tunisia. https://www.bct.gov.tn/bct/siteprod/index.jsp.

[mlu,] Mlu explain. https://mlu-explain.github.io.

[sci,] Scikit-learn documentation. https://scikit-learn.org/stable/.

[bct, , mlu, , sci,]