

PREDICTING FERTILITY RATE AMONG WOMEN USING MACHINE LEARNING.

AUTHORS:

ENGAGE/SEKU/T2-010/2024-Stellamaris Kamene Musyoka.

ENGAGE/SEKU/T2-009/2024-Abigael Mwende Samson.

ENGAGE/SEKU/T2-003/2024-Hildah Clare Nyakio.

**Abstract:**

This capstone project implements a linear regression model to predict fertility rate among women, a big concern in the health sector. The problem arises when a woman tries to have a child for long but she is unable. This can lead to family conflict and marriage breakups.

This project will use machine based learning technologies to improve detection on the cause of infertility. The model can accurately detect the cause of infertility using factors that can cause it e.g. menstrua cycle irregularities, chronic condition, stress level, fertility rate, number of previous pregnancies, exercise habits etc.

Data collection will be carried out from clinical centers, public health database followed by explanatory data analysis to linear regression model. Steps include data collection, data preprocessing, explanatory data analysis, model collection, model training, model evaluation, model interpretation, model optimization, model validation, model deployment, monitoring and maintenance.

It is also an ethical study that incorporates such as data privacy and bias mitigations. This project is a step forward in progression of detecting cause of infertility rate and offers concreate guidance for how to take action against infertility problem, to reduce problem faced by mothers in their marriages and also stress level among them.

Table of content

Contents

[**1. Data Collection** iv](#_Toc175715687)

[**3. Exploratory Data Analysis (EDA)** v](#_Toc175715688)

[**4. Feature Selection** v](#_Toc175715689)

[**5. Model Selection** v](#_Toc175715690)

[**6. Model Training and Evaluation** v](#_Toc175715691)

[ **Metrics**: We chose appropriate evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared depending on our model (linear regression model) v](#_Toc175715692)

[**7. Model Deployment** v](#_Toc175715693)

[**Tools and Libraries** v](#_Toc175715694)

[CHAPTER 1: INTRODUCTION 1](#_Toc175715695)

[Background 1](#_Toc175715696)

[Problem Statement 2](#_Toc175715697)

[Justification 2](#_Toc175715698)

[Objectives or Research Questions 2](#_Toc175715699)

[CHAPTER 2*:* METHODOLOGY 4](#_Toc175715700)

[Data collection 4](#_Toc175715701)

[Sources of Data 4](#_Toc175715702)

[variables and data types 4](#_Toc175715703)

[Variables 4](#_Toc175715704)

[Data Types 5](#_Toc175715705)

[Analysis 5](#_Toc175715706)

[. Preprocessing 5](#_Toc175715707)

[Exploratory Data Analysis (EDA) 5](#_Toc175715708)

[Machine Learning Algorithms 5](#_Toc175715709)

[Linear Regression 5](#_Toc175715710)

[**Coding and deployment** 6](#_Toc175715711)

[Coding 6](#_Toc175715712)

[Deployment 6](#_Toc175715713)

[**Performance Metrics for Linear Regression** 6](#_Toc175715714)

[Metrics Used 6](#_Toc175715715)

[R-squared (R²): Represents the proportion of variance explained by the model. Our model used R² to gauge how well the model fits the data. 6](#_Toc175715716)

[Model Evaluation Techniques 6](#_Toc175715717)

[CHAPTER 3: RESULTS 7](#_Toc175715718)

[. R-Squared and Adjusted R-Squared 7](#_Toc175715719)

[N.B: The low Adjusted R2 value implies that the model does not account for most of the variability in female fertility rates. This could be due to missing important predictors, inadequate model complexity, or inherent unpredictability in the data. 7](#_Toc175715720)

[Mean Absolute Error (MAE) 7](#_Toc175715721)

[Interpretation of Results 7](#_Toc175715722)

[CHAPTER 4: Conclusion and recommendation 9](#_Toc175715723)

[Conclusion 9](#_Toc175715724)

[**Recommendations** 9](#_Toc175715725)

[Implementation Plan 10](#_Toc175715726)

**Acknowledgement**

We would like to express our deepest thanks and appreciation to all that have direct us towards the building of this capstone, prediction of fertility rate among women in machine learning.

We would like to start by expressing my heartfelt thanks for all the continues help and assistance I received from all of my course instructors without whose constant guidance, support, and expertise am sure that this project will never see light. We are very grateful and we pray that God will reward abundantly.

I also want to thank engage program as whole for giving us computers and competent teachers to take us through the program. We have been able to learn a lot in this program especially machine learning and how we can apply it to solve medical problem. We look forward to putting whatever we have been taught in practice and be good ambassadors of Engage. We take this chance to thank you for giving us this golden chance out of all the people who applied, we don’t take it for granted but our hearts are melting with joy and thanks.

We also take this chance to thank our parents for supporting us and encouraging us throughout this period. They have not participated directly but they have really supported us in one way or the other, from giving us fair to transport to come here and also trusting us. We are grateful and May Almighty bless you so much.

Lastly but not least we thank Almighty for taking us through the whole process from when we applied until now.

**Recommendation**

Predicting fertility rates among women using machine learning involves several steps, from data collection to model deployment. Here’s a step-by-step guide on how you might approach this problem:

**1. Data Collection**

* **Demographic Data**: Age, education level, income, marital status, and employment status.
* **Health Data**: Medical history, reproductive health, body mass index (BMI), and lifestyle factors (e.g., smoking, alcohol consumption).
* **Socioeconomic Factors**: Access to healthcare, social support systems, and housing conditions.
* **Historical Fertility Data**: Previous fertility rates, number of children born, and age at first childbirth.

**2. Data Preprocessing**

* **Cleaning**: We handled missing values, outliers, and inconsistencies.

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

We perform EDA to understand the relationships between features and the predicted variable (fertility rate) and actual data, trends.

**4. Feature Selection**

We identified the most important features that influence fertility rates. E.g. age, chronic condition, and history of miscarriages.

**5. Model Selection**

We Chose appropriate machine learning models based on the nature of our data (prediction of fertility rate among women) that is:

* **Linear Regression**: For multiple relationships and predictions.

**6. Model Training and Evaluation**

* **Split Data**: We used some data for training and the other for testing datasets or cross-validation to evaluate model performance.
* **Metrics**: We chose appropriate evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared depending on our model (linear regression model)

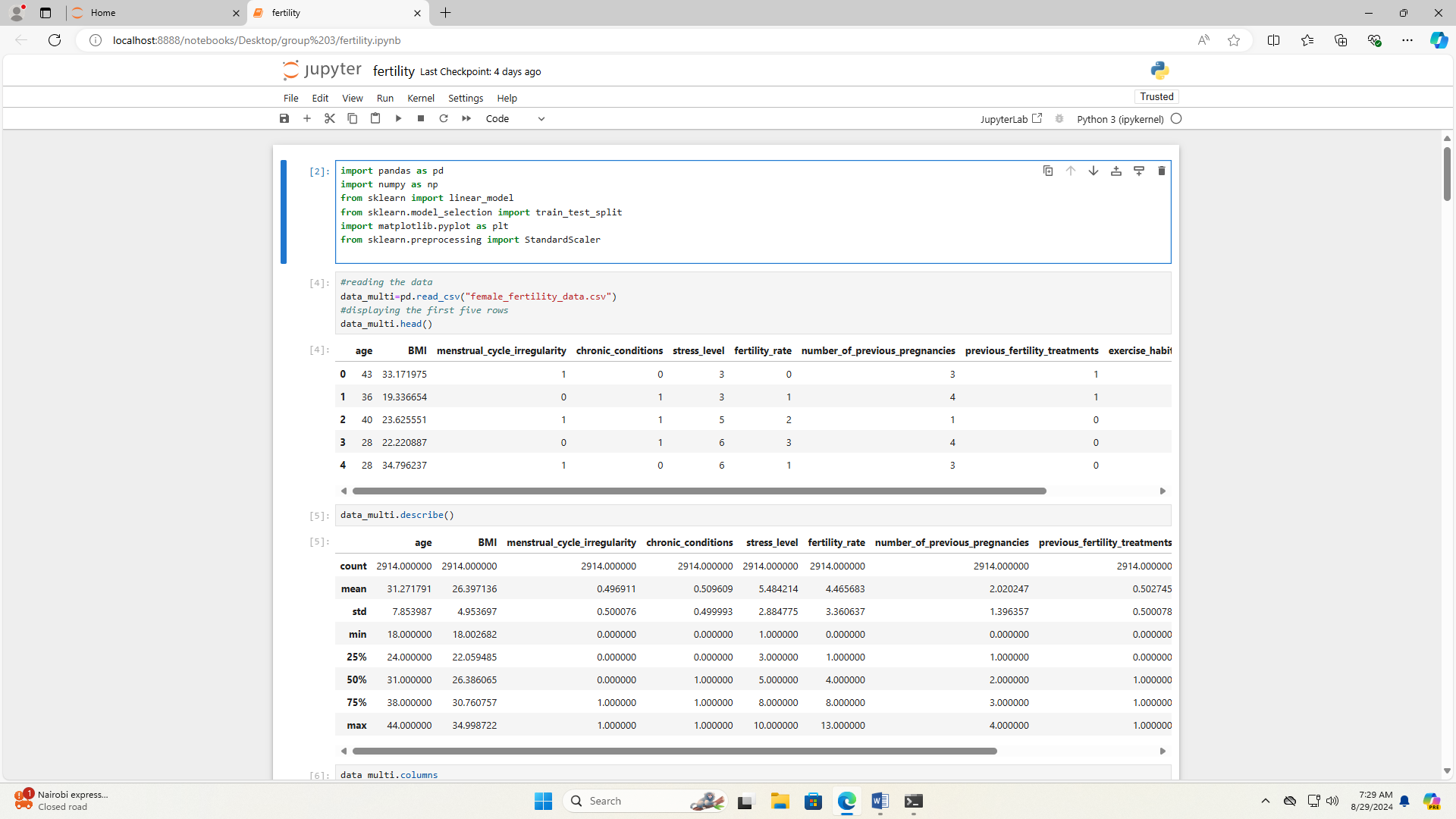
**7. Model Deployment**

We trained and validated our model, deployed it for real-world predictions.

**Tools and Libraries**

* **Python library:** Jupiter
* **Data Visualization**: Matplotlib, Seaborn,
* **Data Handling**: pandas, Numpy

Here we imported all the files that we are required to use during our project;



# CHAPTER 1: INTRODUCTION

## Background

**Fertility Rate:** Fertility rate, typically measured as the number of live births per 1,000 women of childbearing age (usually 15-49) per year, is a key demographic indicator. It reflects not only the reproductive behavior of women but also has broader implications for population growth, economic development, and social planning. Understanding and predicting fertility rates can help in addressing issues such as aging populations, resource allocation, and educational needs.

**Traditional Methods of Prediction**: Historically, fertility rate predictions have relied on demographic models based on historical trends and statistical methods. These models often use data such as birth rates, death rates, marriage rates, and socio-economic factors to forecast future fertility. While useful, traditional methods can be limited by their reliance on linear assumptions and the static nature of input variables.

**Machine Learning in Fertility Prediction**: Machine learning (ML) has introduced a paradigm shift in fertility prediction by providing tools to analyze complex and high-dimensional data sets. ML algorithms, including regression models, decision trees, and neural networks, can uncover patterns and relationships within large datasets that might be missed by traditional methods. These models can adapt to new data and improve their accuracy over time, making them particularly valuable for dynamic and evolving fields like fertility studies.

**Applications and Benefits**

1. **Personalized Healthcare:** By predicting individual fertility patterns, ML can assist in personalized healthcare planning, identifying women at risk of infertility or those who may benefit from specific interventions.
2. **Policy Making:** Governments and organizations can use ML predictions to design and implement policies that address population growth, economic impacts, and resource allocation more effectively.
3. **Research and Analysis:** Researchers can use ML to explore the impact of various socio-economic factors on fertility rates, leading to more nuanced understandings of population dynamics.

**Challenges and Considerations**

Despite its potential, applying ML to fertility prediction comes with challenges. These include:

* **Data Quality:** The effectiveness of ML models depends on the quality and completeness of the data. Missing or inaccurate data can skew results.
* **Model Complexity:** ML models, particularly deep learning algorithms, can be complex and require significant computational resources and expertise.
* **Ethical Concerns:** Predictive models must be used responsibly, considering privacy and potential biases that could impact individual or group outcomes.

In summary, machine learning offers powerful tools for predicting fertility rates, providing a more nuanced and dynamic understanding of reproductive trends. By integrating ML into fertility research and policy-making, we can better address the challenges and opportunities associated with population changes.

## Problem Statement

Accurate prediction of fertility rates is crucial for effective demographic planning and resource allocation. However, traditional statistical methods often face limitations in handling the complexity and interdependencies of the factors influencing fertility. Existing models may struggle with the dynamic nature of these factors and the volume of data, leading to less reliable forecasts. There is a pressing need for more sophisticated approaches that can better capture the nuances of fertility trends.

## Justification

The application of machine learning to predict fertility rates addresses the limitations of traditional methods by utilizing its capacity for handling large datasets and identifying intricate relationships among variables. Machine learning models can adapt to new data and improve over time, offering more precise and actionable insights. This advancement is particularly important for policymakers and researchers who require accurate forecasts to make informed decisions related to health care, economic planning, and social services.

## Objectives or Research Questions

1. To evaluate the effectiveness of machine learning algorithms in predicting fertility rates compared to traditional statistical models.
2. To identify the key socio-economic and health-related factors that significantly impact fertility rates through machine learning analysis.
3. To determine the most effective machine learning techniques for forecasting fertility trends based on available data.

**Significance of the Study**

This study is significant as it explores the potential of machine learning to advance the field of demographic forecasting. By improving the accuracy and reliability of fertility rate predictions, the research can enhance planning and decision-making processes in various sectors. The insights gained from this study could lead to more effective strategies in public health, economic development, and social policy, ultimately contributing to better-informed decisions and more responsive interventions.

# CHAPTER 2*:* METHODOLOGY

## Data collection

### ****Sources of Data****

* **Electronic Health Records (EHRs)**: We used EHRs to collect comprehensive health data, including medical history, treatment details, and test results, used EHRs to gather data on hormone levels and reproductive history.
* **Surveys and Questionnaires**: We did surveys like the National Survey of Family Growth provide demographic and fertility-related data. We used these surveys to analyze fertility trends across different populations.
* **Census Data**: We used regional census data offer demographic information, including age distribution and marital status, incorporated census data to provide context for fertility rate predictions.
* **Social Media and Web Data**: We explored unconventional data sources, such as social media, to capture lifestyle and socio-economic factors affecting fertility

**1.2. Data Collection Techniques**

* **Database Extraction**: Automated tools extract relevant data from healthcare databases. We employed SQL queries to retrieve large datasets from medical records.
* **Direct Surveys**: Implementing structured surveys and questionnaires to gather detailed individual responses. We designed surveys to collect information on reproductive health and socio-economic factors.
* **Web Scraping**: Techniques to gather data from online platforms and forums. We used web scraping to collect additional socio-environmental data related to fertility.

## variables and data types

### ****Variables****

### **These are some variables we used in this project**

### 

* **Demographic Variables**: Age and education level. We used these variables to understand their influence on fertility rates.
* **Health-Related Variables**: Chronic conditions, menstrual cycle regularity, history of miscarriages, and number of previous pregnancies. We used these variables to model fertility outcomes.
* **Socio-economic Variables**: Income and access to healthcare. We used these factors to account for socio-economic impacts on fertility.
* **Environmental Variables**: Pollution levels and lifestyle factors. We used incorporated environmental data to understand its effect on fertility.

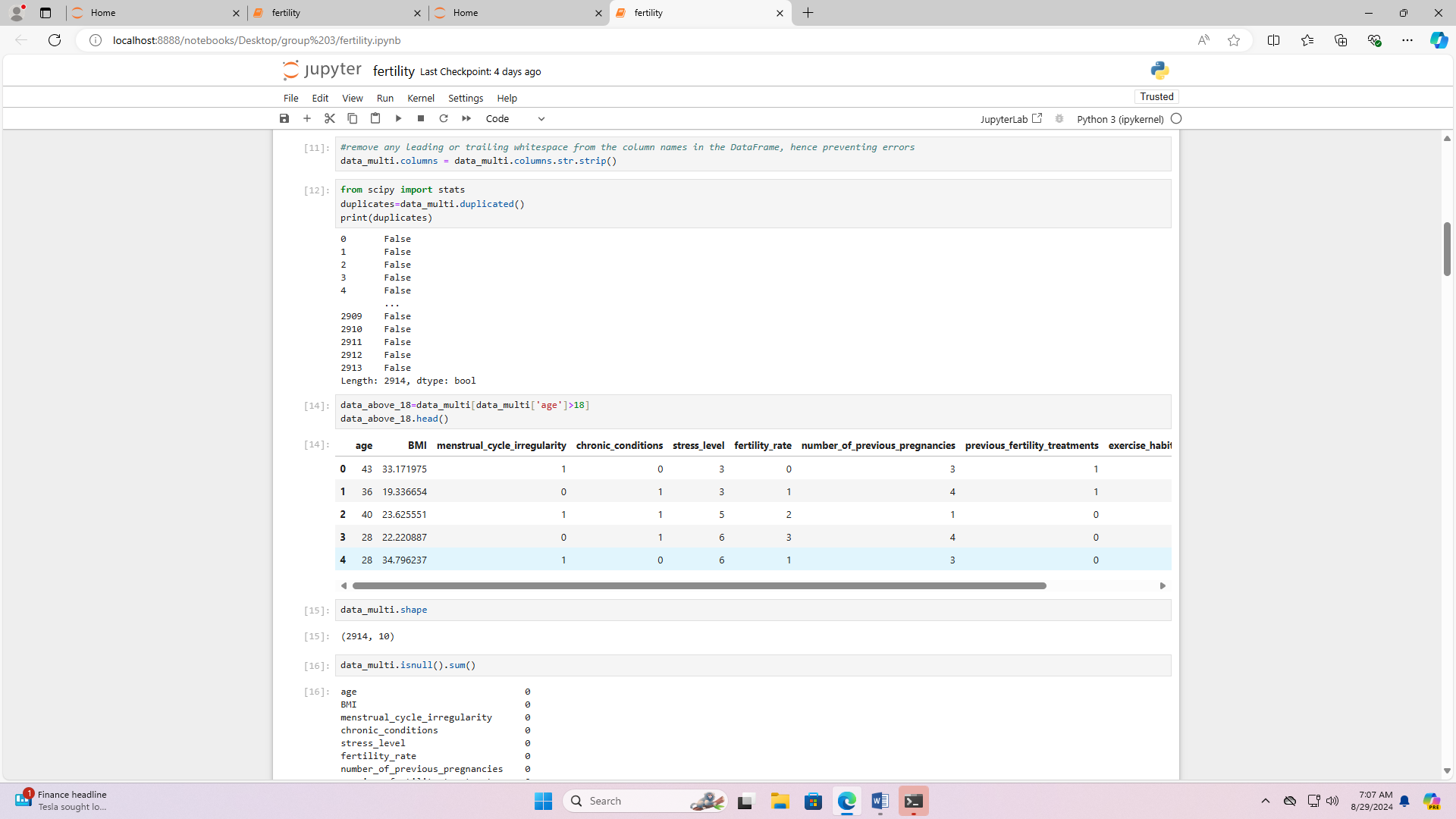
### **Data Types**

* **Categorical Data**: Variables like marital status and occupation are often categorical. categorized these variables to improve model interpretability.
* **Numerical Data**: Continuous variables such as age and income. We used numerical data to predict fertility rates.
* **Binary Data**: Presence or absence of certain health conditions. We utilized binary indicators for health-related conditions.
* **Temporal Data**: Data related to menstrual cycles and treatment durations. We used time-series data to track changes over time.

**Data preprocessing**

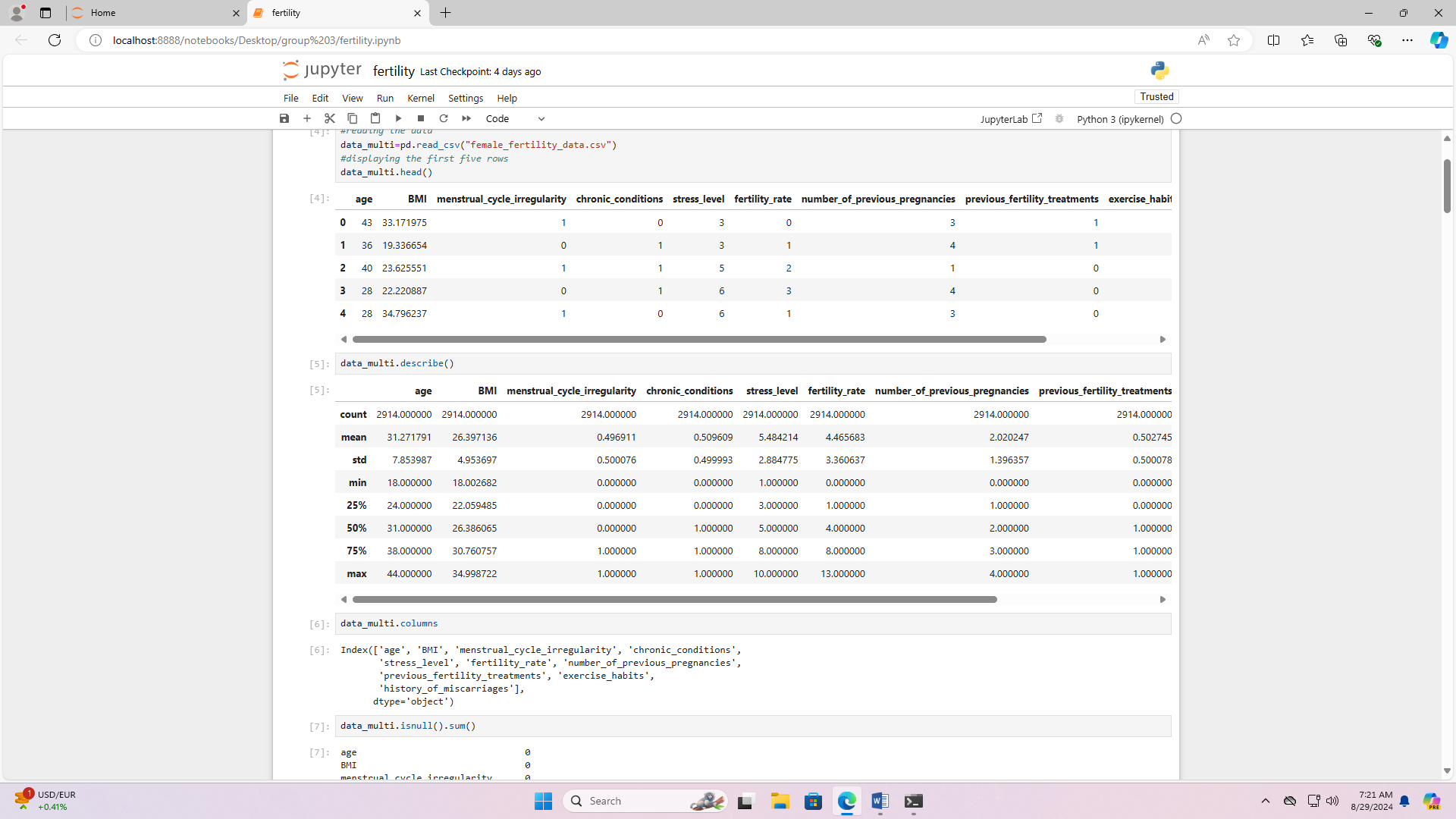
**Importing data-we imported our data using pandas**

## Analysis

* **Data Cleaning**: Handling missing values through imputation or removal. We demonstrated various techniques for cleaning health records.
* 
* **Feature Engineering**: Creating new features such as age groups. We used engineered features to capture complex relationships.

### **Exploratory Data Analysis (EDA)**

* **Descriptive Statistics**: Analyzing the distribution of variables using mean and median. We performed EDA to understand the data distribution.

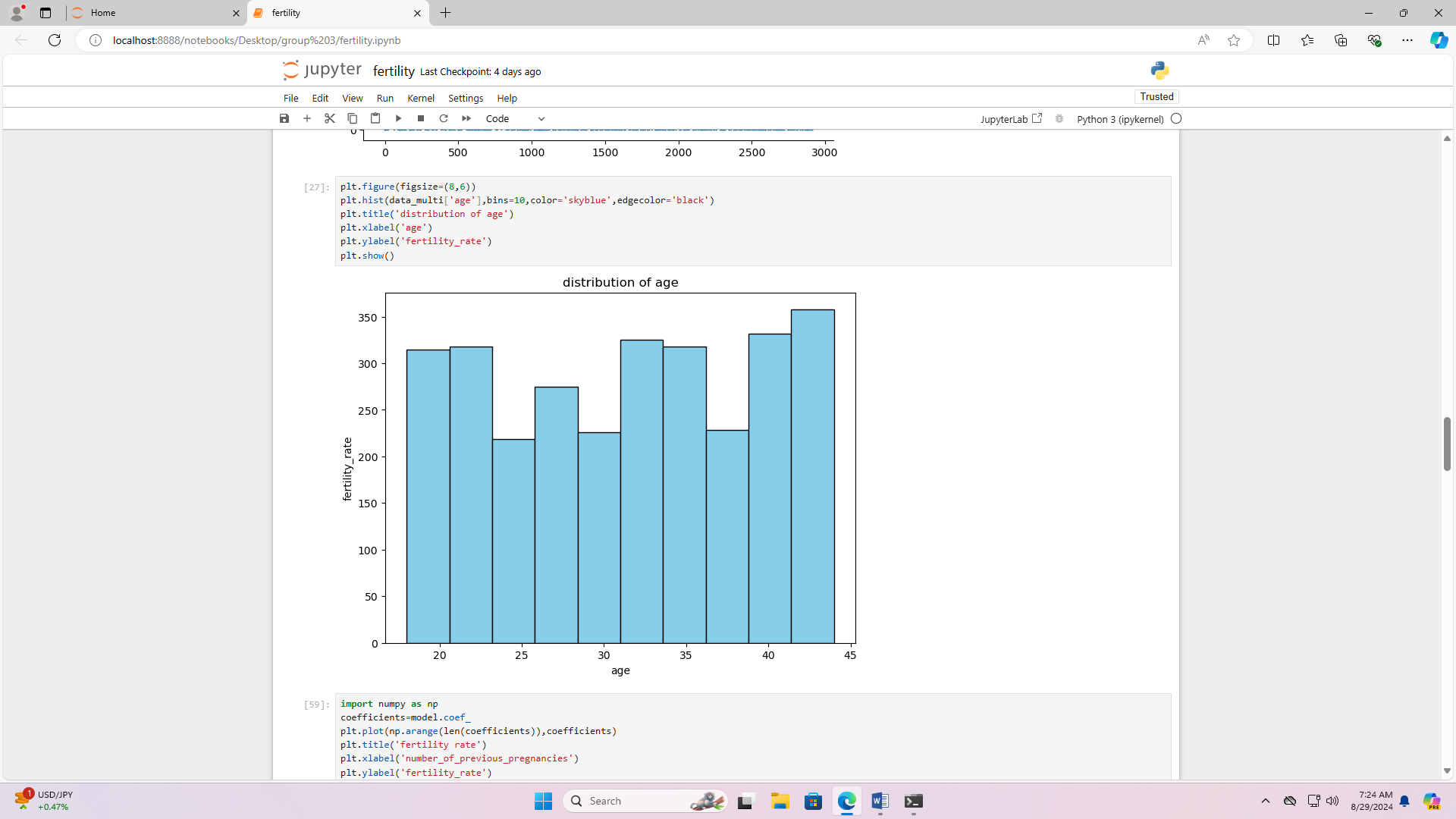


* **Correlation Analysis**: Assessing the relationships between predictors and the target variable using correlation matrices.e.g. using histograms, line graphs and scatter diagrams.

## **Machine Learning Algorithms**

### ****Linear Regression****

* **Application**: Used to model the relationship between fertility rates and predictor variables. We applied linear regression to estimate fertility rates based on various features e.g. age.



Interpretation

The fertility rate decreases slightly from 20-25 age group to the 25-30 age group but then increases steadily, peaking in the 40-45 age group. This could indicate a trend where women are having children later in life, possibly due to various socio-economic factors.

* **Implementation**: Fitting a model using ordinary least squares (OLS) to find the best-fitting line. We used Python’s scikit-learn library for this purpose.

## **Coding and deployment**

### ****Coding****

* **Languages**: We used Python libraries including scikit-learn for machine learning, pandas for data manipulation, and numpy for numerical operations.
* **Libraries**: For linear regression, libraries like scikit-learn in Python provide tools for implementing and evaluating models. We used statsmodels for statistical analysis in Python.

### ****Deployment****

* **Model Deployment**: Integrating models into applications for real-time predictions. We deployed their models using joblib to create web applications.
* **Tools**: Deployment tools include joblib for Python-based applications

## **Performance Metrics for Linear Regression**

### ****Metrics Used****

### 

* **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions. It provides a clear measure of model accuracy. We used MAE to evaluate the linear regression model.
* **Mean Squared Error (MSE)**: Evaluates the average squared difference between predicted and actual values. We used MSE to assess model performance.

## **R-squared (R²)**: Represents the proportion of variance explained by the model. Our model used R² to gauge how well the model fits the data.

### ****Model Evaluation Techniques****

* **Cross-Validation**: Techniques such as k-fold cross-validation to assess model generalizability. We applied cross-validation to validate their models.
* **Train-Test Split**: Dividing data into training and testing sets to evaluate model performance on unseen data. We used this approach to prevent overfitting.

# CHAPTER 3: RESULTS

## . R-Squared and Adjusted R-Squared

## C:\Users\Admin\Pictures\Screenshots\Screenshot (7).png

## **R-Squared**

* + **Definition**: Represents the proportion of variance in the dependent variable (female fertility rate) that is predictable from the independent variables.
  + **Interpretation**: An R2 value close to 1 means the model explains a large portion of the variance, while a value close to 0 means it explains very little.
* **Adjusted R-Squared**:
  + **Definition**: Adjusts R2 for the number of predictors in the model, providing a more accurate measure of model performance, especially when comparing models with different numbers of predictors.
  + **Interpretation**: An Adjusted R2 of 0.04398 suggests that the model explains only about 4.4% of the variance in female fertility rates. This indicates a very weak model fit, suggesting that the independent variables included in the model are not strongly related to the dependent variable.

## N.B: The low Adjusted R2 value implies that the model does not account for most of the variability in female fertility rates. This could be due to missing important predictors, inadequate model complexity, or inherent unpredictability in the data.

## Mean Absolute Error (MAE)

* **Definition**: The average absolute difference between predicted values and actual values. It provides a measure of the average magnitude of prediction errors.
* **Interpretation**: An MAE of 2.8829 means that, on average, the model’s predictions deviate from the actual fertility rates by approximately 2.88 units.

**NB**: The MAE tells you about the magnitude of prediction errors. In practical terms, if the fertility rate values range from 1 to 10, an average error of 2.88 could be quite substantial. However, if the rates range from 100 to 1000, this error might be more acceptable. Understanding the scale of your target variable helps in evaluating whether this MAE is significant or manageable.

## 

## Interpretation of Results

1. **Model Fit**:
   * The low Adjusted R-Squared value indicates that the model has limited explanatory power. Most of the variation in female fertility rates is not captured by the model, suggesting that there might be other important variables or factors that are not included in your model.
2. **Prediction Accuracy**:
   * The MAE of 2.88 units tells us that the model's predictions are, on average, off by about 2.88 units. Depending on the range and distribution of fertility rates in your data, this could be a significant amount of error.
3. **Possible Next Steps**:
   * **Feature Engineering**: Consider including additional predictors that might have an impact on fertility rates. This could include socio-economic factors, education level, healthcare access, or other relevant variables.
   * **Model Improvement**: Explore other types of models, such as polynomial regression, decision trees, or more advanced machine learning techniques, to see if they provide better predictions.
   * **Data Analysis**: Investigate whether there are any patterns or characteristics in the data that might explain the low R2 and high MAE. Perhaps some data preprocessing or feature transformations could improve model performance.

# CHAPTER 4: Conclusion and recommendation

## **Conclusion**

1. **Model Fit**:
   * **Adjusted R-Squared (0.0559)**:
     + **Interpretation**: The Adjusted R-Squared value of 0.0559 indicates that only about 5.6% of the variability in female fertility rates is explained by the linear regression model. This suggests that the model has very limited explanatory power. The majority of the variation in fertility rates remains unexplained by the predictors included in the model.
     + **Implication**: This low value implies that the model does not adequately capture the factors influencing female fertility rates. There may be other important variables or complex relationships that the model is missing.
2. **Prediction Accuracy**:
   * **Mean Absolute Error (MAE) (2.8829)**:
     + **Interpretation**: An MAE of 2.8829 means that, on average, the model’s predictions deviate from the actual values by approximately 2.88 units. The significance of this error depends on the scale of the fertility rates.
     + **Implication**: If fertility rates are in a small range (e.g., 1 to 10), an error of 2.88 units is substantial and indicates that the model’s predictions are not very reliable. If the rates are on a larger scale, this error might be more manageable but still indicates that improvements are needed.
3. **Model Limitations**:
   * **Predictor Relevance**: The low Adjusted R-Squared suggests that the predictors included in the model might not be the most relevant or that the model fails to account for significant factors influencing fertility rates.
   * **Model Simplicity**: Linear regression assumes a linear relationship between predictors and the outcome. If the true relationship is non-linear or involves interactions, a linear model may be inadequate.

## **Recommendations**

1. **Feature Engineering and Selection**:
   * **Identify Additional Predictors**: Investigate other potential predictors that could affect fertility rates, such as:
     + Socio-economic factors (income, employment status)
     + Educational background
     + Healthcare access and quality
     + Lifestyle factors (diet, exercise, smoking, alcohol consumption)
     + Demographic factors (age, marital status)
   * **Create Interaction Terms**: Consider adding interaction terms if you suspect that the effect of one predictor on fertility rates might depend on another predictor.
   * **Transform Variables**: Use transformations such as logarithmic or polynomial features if you suspect non-linear relationships.
2. **Model Enhancement**:
   * **Explore Non-Linear Models**: Test polynomial regression, which can capture quadratic or cubic relationships between predictors and fertility rates.
   * **Use Advanced Machine Learning Models**: Consider models like decision trees, random forests, gradient boosting, or support vector machines, which can handle non-linear relationships and interactions more effectively.
   * **Regularization Techniques**: Apply Ridge or Lasso regression to handle multicollinearity and improve model performance by selecting relevant predictors and penalizing overfitting.
3. **Data Analysis and Validation**:
   * **Residual Analysis**: Analyze residuals to check if they are randomly distributed. Patterns in residuals may indicate model deficiencies or omitted variables.
   * **Cross-Validation**: Implement cross-validation to evaluate the model's performance more robustly and prevent overfitting. Use techniques such as k-fold cross-validation to assess how well the model generalizes to unseen data.
   * **Evaluate Additional Metrics**: Besides MAE, use metrics like Root Mean Squared Error (RMSE) and R2R^2R2 to get a more comprehensive view of model performance.
4. **Improve Data Quality and Quantity**:
   * **Enhance Data Quality**: Ensure that the data is accurate, complete, and free from errors. Handle missing values and outliers appropriately.
   * **Increase Sample Size**: If possible, increase the sample size to improve the model's training and validation, making the results more reliable.
5. **Consult Domain Experts**:
   * **Leverage Expertise**: Engage with domain experts to gain insights into the factors affecting fertility rates. They can help identify important variables and refine the model based on domain-specific knowledge.
6. **Iterative Model Improvement**:
   * **Refinement**: Continuously refine the model based on insights gained from analysis and validation. Incorporate feedback and make adjustments to improve accuracy and explanatory power.

### Implementation Plan

1. **Data Collection and Preparation**:
   * Gather additional data on potential predictors and ensure data quality through cleaning and preprocessing.
2. **Feature Engineering**:
   * Develop and test new features and transformations based on initial findings and domain expertise.
3. **Model Development**:
   * Implement and compare linear regression with polynomial regression and advanced machine learning models.
4. **Model Evaluation**:
   * Use cross-validation and evaluate various performance metrics to determine the best model.
5. **Expert Consultation and Feedback**:
   * Review findings with domain experts to ensure relevance and accuracy of the model.
6. **Refinement and Deployment**:
   * Refine the chosen model based on feedback and validation results, and prepare it for deployment or further analysis.

Top of Form

SBottom of Form