Reflection

47238348

2024-05-19

Reflective Report on Portfolio 4: COVID-19 Vaccination Analysis

Introduction

In this report, I reflect on the process of analyzing the COVID-19 vaccination data to understand factors influencing vaccination rates.

Problem-Solving Process and Learning

Process

I began by loading and cleaning the dataset, handling missing values, and filtering the data for the year 2021. Next, I performed exploratory data analysis (EDA) to visualize relationships between variables such as GDP per capita, median age, and vaccination rates. I selected important features for modeling and built regression models to predict vaccination rates. Throughout the process, I encountered challenges such as data imputation and feature selection, which I addressed by iterating on my approach and consulting online resources.

Learning

I have progressed significantly from the start of the unit, gaining a deeper understanding of data cleaning, EDA, and regression modeling. I became proficient in using tools like Pandas, Seaborn, and Scikit-learn. The iterative nature of the data science process taught me the importance of refining models and validating results.

Future Interests

I am interested in applying the skills I gained in this project to other domains, such as healthcare and finance, where data-driven decision-making is crucial. I plan to explore advanced modeling techniques and deep learning to enhance my analytical capabilities.

Discussion Points

Choice of Dataset

I chose the COVID-19 vaccination dataset because it provides a rich set of features and a well-defined target variable, making it suitable for regression modeling. The dataset's relevance to real-world applications in public health also motivated my choice.

Identifying the Problem

The primary problem was to predict vaccination rates based on various features. This problem is significant because understanding factors influencing vaccination rates can help policymakers design more effective vaccination campaigns. I focused on understanding the impact of key features like GDP per capita, median age, and healthcare factors on vaccination rates.

Model Selection

I selected linear regression models due to their simplicity and interpretability. These models are suitable for identifying relationships between features and the target variable. Additionally, I experimented with regularized regression techniques like Ridge and Lasso to handle multicollinearity and improve model performance.

Insights and Conclusions

The analysis revealed that features such as GDP per capita and median age are strong predictors of vaccination rates. The model's performance, measured by R-squared, was consistent with my expectations, indicating a good fit. However, there is room for improvement by incorporating more advanced modeling techniques and external data sources.

Conclusion

This project has been a valuable learning experience, enhancing my data analysis and modeling skills. I look forward to applying these skills in future projects and continuing to learn and grow in the field of data science.