COVID-19 Data Analysis and Prediction Using Machine Learning

Introduction

The purpose of this project is to analyze real-world COVID-19 data and apply data science techniques to explore trends, visualize patterns, and build a simple machine learning model. The dataset used in this project contains daily COVID-19 case and death counts for counties across the United States and was obtained from Kaggle, a public data science platform.

The main objective of this analysis is to understand how COVID-19 cases evolved over time within a selected U.S. state and to identify which counties contributed most to total case counts. In addition, a simple machine learning model was developed to predict cumulative COVID-19 cases based on time. This project demonstrates key steps of the data analysis process, including data cleaning, exploratory data analysis, visualization, and basic machine learning modeling.

Dataset Description

The dataset used in this project is titled “US Counties COVID-19 Daily Data” and was downloaded from Kaggle. It contains daily records of COVID-19 cases and deaths reported at the county level across the United States. The main variables in the dataset include:

date – the reporting date

county – the county name

state – the U.S. state

fips – county identification code

cases – cumulative number of confirmed COVID-19 cases

deaths – cumulative number of COVID-19 deaths

The dataset includes a large number of rows due to daily reporting across many counties and states. For the purpose of this project, the analysis was limited to a single state in order to reduce complexity and make trends easier to interpret.

Data Cleaning and Preprocessing

Several preprocessing steps were performed to prepare the data for analysis. First, the dataset was loaded into Python using the Pandas library. The structure of the data was examined using summary functions to identify column types and missing values.

The date column was converted from a text format into a datetime format to enable proper time-based analysis and visualization. Rows containing missing values in key columns such as cases and deaths were removed to ensure accuracy and prevent errors during modeling.

To simplify the analysis, the dataset was filtered to include data from only one U.S. state. County-level data within the selected state were then aggregated by date, producing total daily case and death counts at the state level. Additionally, new features were created by calculating daily new cases and deaths using the difference between consecutive days. These preprocessing steps ensured that the dataset was clean, consistent, and suitable for further analysis.

Exploratory Data Analysis and Visualization

Exploratory Data Analysis (EDA) was conducted to better understand patterns and trends in COVID-19 cases. Several visualizations were created using Matplotlib.

The first visualization showed the total number of COVID-19 cases over time within the selected state. This line chart revealed a strong upward trend, indicating the cumulative nature of reported cases during the pandemic. The overall shape of the curve highlighted periods of rapid growth as well as slower increases.

The second visualization focused on daily new COVID-19 cases. This chart provided a clearer view of fluctuations over time, including peaks and declines that correspond to different waves of the pandemic. Some negative values were observed due to data corrections, which is common in real-world public health datasets.

A third visualization displayed the top ten counties in the selected state with the highest total number of COVID-19 cases. This bar chart showed that a small number of counties contributed a disproportionately large share of total cases, often corresponding to counties with larger populations.

Together, these visualizations helped identify important trends and provided meaningful insights into how COVID-19 spread across regions.

Machine Learning Model

To satisfy the machine learning component of the project, a Linear Regression model was developed. The goal of the model was to predict cumulative COVID-19 case counts based on time.

Because machine learning models require numerical input, the date variable was transformed into a numerical format by converting each date into an ordinal day number. This value represents the number of days elapsed since a fixed reference date and serves as the input feature.

The dataset was split into training and testing sets using an 80/20 ratio. The model was trained on the training data and evaluated on the test data. Model performance was measured using Root Mean Squared Error (RMSE), which represents the average prediction error in the same units as the target variable.

Although the RMSE value was relatively large, this result is expected due to the rapidly increasing and non-linear nature of cumulative COVID-19 case data.

Results and Limitations

The linear regression model was able to capture the general upward trend of cumulative COVID-19 cases over time. A visualization comparing actual values with predicted values showed that the model approximated the overall direction of the data but did not capture sudden changes or complex patterns.

This limitation is expected because linear regression assumes a straight-line relationship, while pandemic dynamics are influenced by many non-linear factors such as public health policies, testing availability, and behavioral changes. As a result, this simple model is best suited for demonstrating basic machine learning concepts rather than producing highly accurate forecasts.

Conclusion

In this project, real-world COVID-19 data were analyzed using data science techniques. The dataset was cleaned, processed, and explored through visualizations that revealed important trends in cumulative and daily case counts. A simple machine learning model was implemented to predict case counts based on time, demonstrating the fundamentals of predictive modeling.

This project highlights the importance of data preprocessing, visualization, and model evaluation when working with real-world data. Future work could include using more advanced models, incorporating additional features such as population size, or applying time-series forecasting techniques to improve predictive performance. moral decision.