Background Information

The COVID-19 pandemic has caused significant disruptions across the world, with varying fatality rates across regions. Understanding the relationship between factors like confirmed cases, deaths, and the resulting fatality rate is crucial for public health planning. Predicting fatality rates can help policymakers allocate resources more effectively and intervene in areas that are more likely to experience severe outcomes. While raw case and death counts provide insight into the scope of the pandemic, they do not always explain differences in fatality rates across states due to demographic, healthcare, and policy variations. This project aims to predict the fatality rate of COVID-19 using cases, deaths, and temporal factors (like month) to better understand what factors might drive differences across states. The results of this analysis can highlight patterns that will inform decision-makers for future health emergencies.

Linear regression is used as the baseline model to capture the linear relationship between COVID-19 fatality rates and predictor variables such as cases and deaths. However, given the possibility of non-linear interactions, other models such as Random Forest Regressor or XGBoost may also be explored to improve predictive performance. The focus is on identifying key relationships between cases, deaths, and fatality rates across states and understanding if time-based factors (e.g., months) influence the trends.

Problem Statement

This project aims to predict COVID-19 fatality rates using cases, deaths, and state data to identify which regions are more likely to have higher fatality rates.

Hypothesis

It is hypothesized that states with higher case and death counts will have higher fatality rates, and that the month of the year will also have an impact, with later months showing a decline in fatality rates due to improvements in healthcare and vaccine rollouts.

Methods

The primary dataset includes columns for cases, deaths, and fatality rate across different counties and states from March to December 2020. The features (independent variables) include:

cases: Number of COVID-19 cases
deaths: Number of deaths attributed to COVID-19
month and day of the year: Temporal indicators
Model Selection and Data Preparation

The data was cleaned by removing any rows with missing values using .dropna(). The features were split into training and testing datasets (70% train, 30% test) using train_test_split from scikit-learn. The initial model used was Linear Regression to capture the linear relationship between the features and fatality rate.

Handling Missing Data:

Missing values in the feature matrix were imputed using the mean strategy through SimpleImputer. This ensured that the training and testing data were consistent and aligned.

Evaluation Metrics:

Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.
R-squared (R^2): Explains the proportion of variance in the fatality rate explained by the model.
Additional Models Considered:
Random Forest Regressor: Explored to capture potential non-linear relationships in the data.
XGBoost Regressor: Used as another non-linear model to improve predictive power.
Despite testing these models, Linear Regression was chosen for simplicity and interpretability in reporting results.
Results and Discussion

The linear regression model produced the following results on the test set:

Mean Squared Error (MSE): 0.0187
R-squared (R^2): 0.0008
The MSE of 0.0187 indicates that the model's predictions, on average, deviate by a small margin, though this is expected given the small range of the fatality rate (between 0 and 1). However, the R^2 value of 0.0008 suggests that the model explains less than 0.1% of the variance in fatality rates. This indicates that cases and deaths alone are not sufficient predictors of fatality rate, and additional factors—such as population density, healthcare access, vaccination rates, or state-specific policies—might be needed to improve the predictions.

The coefficients of the linear regression model revealed that the month variable had a small but negative coefficient (-0.0006), indicating a slight decline in fatality rates over time, possibly due to improvements in treatment protocols or vaccine rollouts. However, the coefficients for cases and deaths were close to zero, suggesting these variables had no significant direct impact on the predicted fatality rate, possibly because they are already captured in the calculation of the fatality rate itself.

Given the poor performance of the linear regression model, future work could involve using non-linear models such as Random Forest or XGBoost, which are better suited for capturing complex relationships. Additionally, adding more relevant features (e.g., vaccination rates, population density) could significantly improve the model's predictive power.

Conclusion

The linear regression model reveals several insights into the relationships between COVID-19 fatality rates and various predictors, such as cases, deaths, and state-specific characteristics. However, the small coefficients for **cases (-0.0000)** and **deaths (0.0000)** indicate that these variables have **minimal direct impact** on the fatality rate, possibly because the fatality rate itself already captures this relationship through its definition (deaths/cases). Similarly, **month (-0.0006)** shows a slight but negative impact, suggesting that the **fatality rate decreases marginally over time**, which aligns with the expectation that public health interventions, treatments, and vaccination efforts improved throughout the pandemic.

The state-level coefficients reveal that certain states have **higher or lower predicted fatality rates** compared to the baseline state. For instance, **Maryland (0.1018)**, **Connecticut (0.0336)**, and **New Jersey (0.0387)** have notably higher coefficients, suggesting that these states experienced **higher fatality rates**, likely due to factors such as **population density, healthcare strain, or demographic vulnerabilities**. Conversely, states such as **Alaska (-0.0177)** and **Hawaii (-0.0091)** exhibit negative coefficients, implying **lower fatality rates**, which might reflect **geographic isolation, lower population density, or effective pandemic management**.

Overall, while the model provides some meaningful insights into how **location and time** might influence fatality rates, the **very small coefficients for cases and deaths** suggest that these variables alone are **insufficient predictors**. The **intercept (0.0354)** indicates that, on average, the model predicts a baseline fatality rate of **3.54%**. These results highlight the need for **additional features**—such as **vaccination rates, healthcare access, or socioeconomic factors**—to improve the model’s predictive power. Non-linear models like **Random Forest** or **XGBoost** may also capture more complex relationships that linear regression fails to detect.