Predicting COVID-19 Death Rates Using Machine Learning Techniques

Kelly Li, Yana Xu

Abstract

The COVID-19 pandemic has led to significant global mortality, with the United States experiencing the highest number of reported deaths. Accurately predicting COVID-19 death rates is critical for guiding public health responses and optimizing healthcare resource allocation. This project aimed to develop a predictive model using machine learning techniques to estimate COVID-19 death rates at the population level across U.S. states. We compiled and integrated publicly available data from the CDC and U.S. Census Bureau, including variables such as confirmed cases, hospitalizations, ICU occupancy, and vaccination rates. Data were aggregated at both monthly and weekly levels, and standardized per 100,000 population. When building predictive models, we used four models, including linear regression, K-nearest neighbors, random forest, and XGBoost. Predictive models were then being compared and tested using RMSE and R^2 for us to select the one with the highest accuracy. Results indicated that monthly aggregation produced more accurate predictions than weekly aggregation. Among the models, KNN demonstrated the highest predictive accuracy, followed closely by Random Forest. Visual analyses further supported the robustness of KNN predictions. Our findings suggest that simple, interpretable models with appropriate temporal aggregation can effectively predict COVID-19 mortality trends and may support future public health decision-making.

Introduction

Coronavirus disease (COVID-19) is an infected disease caused by the SARS-Cov-2 virus. Since the first report of a novel upper respiratory disease of unknown etiology on December 12, 2019, in Wuhan, China, the COVID-19 pandemic has spread rapidly around the world. To date, there have been over 778 millions cumulative COVID-19 cases, resulting in more than 7.1 million deaths worldwide. Compared to the SARS virus, the SARS-Cov-2 virus spreads more easily from person to person at a faster rate, leading the World Health Organization to declare COVID-19 as a global pandemic. Among all countries, the United States has been severely affected. After the first confirmed case of COVID-19 was reported in California on Jan 26, 2020, the United States has recorded a cumulative total of 103 million cases. Of these cases, approximately 1.2 million have resulted in deaths, making the United States the country with the highest number of COVID-19 deaths globally.

Hence, predicting the COVID-19 death rate at a population level is crucial for understanding trends in disease severity, informing public health strategies, and guiding resource allocation. A predictive model that accurately predicts the poor outcome for COVID-19 patients could assist in efficiently allocating limited medical resources, improving the quality of healthcare, and ultimately optimize patient management (Mamandipoor et al., 2022). By anticipating the periods of higher mortality risk, policymakers and healthcare providers can better prepare for and respond to ongoing and future public health challenges.

Several factors have been associated with COVID-19 death rates across different populations. Demographic variables, such as older age, presence of comorbidities, and obesity have been consistently linked to higher mortality risk. Healthcare system capacity, including the availability of intensive care unit (ICU) beds and ventilators, has also played a critical role, particularly during periods of surge. Studies have shown that decreased ICU beds availability and increasing community case burden have been implicated as risk factors for poor COVID-19 outcomes (Kadri et al., 2021). Vaccination has also been shown to significantly reduce the risk of severe outcomes and death, with 75.31% decrease in COVID-19 cases and 74.89% reduction in the death rate upon getting fully vaccinated (Rustagi et al., 2022). Furthermore, geographic and social determinants of health, such as regional public health policies, population density, and access to healthcare, have contributed to differences in COVID-19 mortality between areas.

Given the available datasets and the focus of this project on predicting COVID-19 death rates at a population level, we selected the number of confirmed cases, new hospitalizations, ICU beds occupancy, complete vaccination series percentage, number of booster, and percentage of bivalent booster as covariates. The purpose of this project is to develop an efficient model using machine learning techniques for predicting COVID-19 death rates. The project seeks to answer one question. Using all the relevant predictors, which ML model is more effective for death rates prediction of COVID-19 patients?

Methods: Data

The primary data sources for this study were publicly available datasets from the United States Census Bureau and the Centers for Disease Control and Prevention (CDC). State population

estimates from 2020 to 2024 were obtained from the U.S. Census Bureau dataset. Specifically, we selected "NST-EST2024-ALLDATA" as it includes detailed components of population change and estimates by year, which are more suitable for building a predictive model. COVID-19-related data, including cases, hospitalizations, deaths, and vaccinations, were retrieved from CDC APIs. Additionally, U.S. region data by state was retrieved from GitHub and used to better visualize the distinctions and changes in death rates between states across time.

To ensure the validity and usability of the predictive model, we first filtered the population data from 2020 to 2024 to retain only U.S. states, the District of Columbia, and Puerto Rico. Next, region data was joined to create a final population dataset with matched states, respective regions, and population estimates for each year. For each COVID-19 dataset, we wrangled the data frames and created timely aggregations by summing counts or averaging percentages. Dates were standardized to the Morbidity and Mortality Weekly Report (MMWR) calendar system to align different datasets on a weekly basis. However, we also standardized yearly and monthly data to ensure the machine learning models we build have the flexibility to achieve the highest possible accuracy.

Population estimates were joined to the COVID-19 datasets by matching on state and calendar year. After merging, smaller territories such as Guam and the Virgin Islands were excluded to focus on states and major jurisdictions. Notably, missing values were converted to zero because some data were not available during the periods of interest. For example, COVID-19 vaccines or boosters were not available when the pandemic began (early 2020), but important data like death rates and hospitalizations are crucial for building a predictive model for COVID-19 death rates.

Methods: Machine Learning

We conducted our analysis using the R programming language, utilizing the dataset previously mentioned. The dataset includes state-level data on COVID-19 cases, deaths, hospitalizations, vaccination, booster percentages, and ICU hospitalizations spanning a specified period. To prepare the dataset for analysis, we standardized key variables, cases, deaths, and hospitalizations to rates per 100000 population, which is a more meaningful comparison across

states. Missing data for vaccination and booster percentages were imputed as zeros to ensure consistency and completeness.

Our primary outcome of interest was the number of COVID-19 deaths per 100,000 individuals, with predictor variables including cases per 100,000, hospitalizations per 100,000, vaccination and booster percentages, ICU hospitalizations, and total booster doses administered. To explore temporal effects, we aggregated data at both monthly and weekly intervals, calculating mean values for each period to determine the optimal frequency for predictive modeling.

Before model development, we employed a feature cleaning procedure to eliminate predictors exhibiting zero variance or perfect collinearity, as these could negatively impact the model.For the machine learning part, we implemented four models to predict COVID-19 deaths: a linear regression model, a K-nearest neighbors model, a Random Forest model, and an XGBoost model.

Beginning with the Linear Regression model, this assesses the direct linear relationships between predictors and the outcome. Then there is the K-Nearest Neighbors model, which predicts outcomes based on similarities with neighboring data points, optimized through 5-fold cross-validation to determine the optimal number of neighbors. Random Forest model is an ensemble approach that aggregates predictions from multiple decision trees to enhance predictive accuracy Lastly, the XGBoost model is a more advanced gradient-boosting algorithm that effectively captures complex interactions within tabular data.

To evaluate each model's performance, we calculated the Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error, and R-squared. Models were trained using data from 2021, while testing and validation utilized data from 2022, ensuring temporal generalizability of results. We then compared the predictive accuracy across monthly and weekly aggregations.

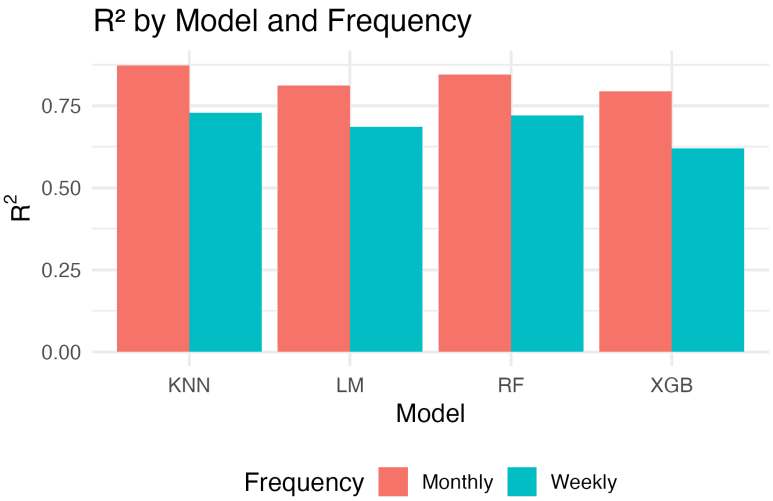
For visual interpretation, we generated different graphical representations, including bar charts to compare R^2 values, boxplots to illustrate prediction error distributions, and time series plots

contrasting actual and predicted death rates. These visualizations provided clear and intuitive insights into model performances and helped in the comprehensive evaluation of our predictive approaches.

Results

In our analysis, we compared the performance of four different predictive machine learning methods: Linear regression model, KNN model, Random Forest model, and XGBoost, to forecast COVID-19 deaths per 100000 individuals. We evaluated the models using both monthly and weekly aggregated data to identify the most accurate approach.

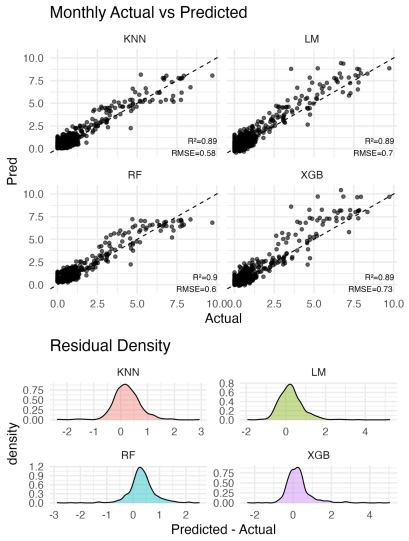
Overall, the monthly aggregated data provided better prediction accuracy compared to weekly data. Among the monthly models, KNN had the best predictive performance, with the lowest Root Mean Squared Error (RMSE) of 0.57 and the highest R-squared (R^2) value of approximately 0.87. Random Forest followed closely, with an RMSE of 0.63 and an R^2 of about 0.84. The Linear Regression model showed slightly higher error with an RMSE of 0.70 and an R^2 of around 0.81, while XGBoost had the highest error among monthly models, with an RMSE of 0.73 and an R^2 of roughly 0.79.



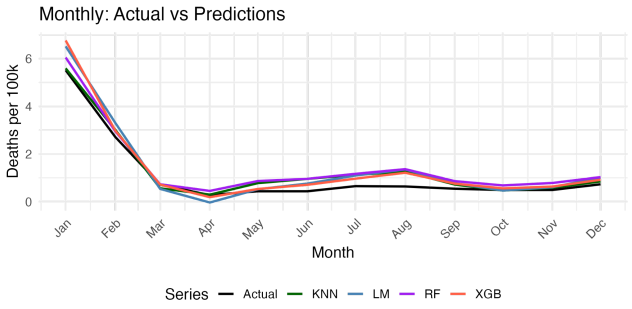
When evaluating weekly aggregated data, we noticed an increase in prediction errors across all models. Again, KNN performed best among the weekly predictions, with an RMSE of 0.89 and

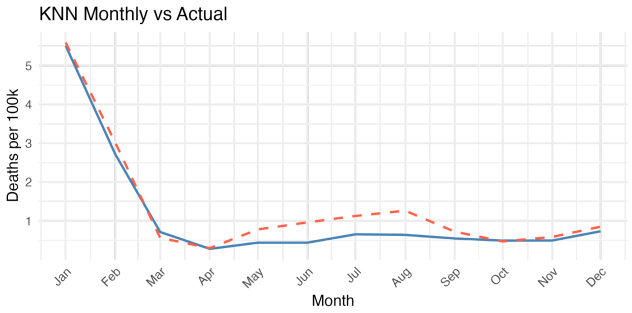
an R^2 of approximately 0.73. Random Forest was very close, with a slightly higher RMSE of 0.90 and an R^2 around 0.72. Linear Regression and XGBoost showed greater prediction errors with RMSE values of 0.96 and 1.05, respectively, and lower R^2 scores of 0.68 and 0.62.

The difference in performance between monthly and weekly predictions is clearly reflected in the comparative analysis. Monthly predictions consistently outperformed weekly predictions, particularly in lower RMSE and higher R^2 values across all models. This difference highlights that monthly aggregation may better capture overall trends and reduce noise in the data compared to weekly aggregation.



We further explored individual model predictions through scatterplots and residual density analyses. These visualizations demonstrated that KNN and RF predictions closely aligned with actual observed values, suggesting balanced and accurate predictions. Residuals for these models were distributed evenly around zero, indicating minimal systematic error. On the other hand, residuals for LM and XGB were slightly broader, suggesting larger deviations from actual observations.





Additionally, the time series analysis of monthly data provided deeper insights into model performance over the entire year. The KNN model closely matched the actual trend, especially during significant decreases early in the year, suggesting that this model effectively captures the general pattern of COVID-19 deaths. Similar patterns were observed for RF, although slight deviations occurred in later months. LM and XGB displayed a broader deviation from actual values throughout the year, particularly noticeable from April onward.

Discussion

Our study's primary aim is to answer the question: Which readily available forecasting model works best for predicting COVID-19 deaths using up-to-date state-level data? We looked at four common algorithms: Linear Regression, K-Nearest Neighbors, Random Forest, and XGBoost—using data from 2021 and evaluating how they performed with data from 2022. What we found was pretty clear.

Across all the tests we ran, models that worked with monthly data did better than those based on weekly figures. Among the monthly models, KNN showed the tightest error band (with an RMSE of about 0.57 deaths per 100,000 and an R^2 of around 0.87), followed closely by RF. Linear Regression and XGBoost didn't perform as well. This pattern held up across different data slices, suggesting our findings weren't just random glitches.

The benefits of using monthly data were very noticeable. While weekly figures might seem tempting because they offer quicker updates, they often suffer from issues like holiday backlogs, irregular data releases, and inconsistent weekend reporting at the state level. Monthly aggregation smooths out these bumps and allows models to focus on real epidemiological trends

instead of administrative hiccups. For public health planners, this is practical: when there's a need for actionable forecasts, like staffing hospitals or ordering antivirals, a monthly forecast, although slower, is much more dependable and worth the wait.

The KNN model has the best performance. Intuitively, KNN works by asking, "Which past state-month looked most like the current one in terms of cases, hospitalizations, vaccines, boosters, and ICU strain?" If a clear analogue exists, KNN simply copies the analogue's death rate. During the pandemic states rarely peaked together; waves rolled region by region. That staggered pattern means every state typically has a handful of recent "neighbors" it can learn from, making KNN's logic surprisingly well matched to the problem. Random Forest came a close second, which makes sense: like KNN it captures nonlinearities and interactions without demanding huge sample sizes, but it adds a healthy dose of ensembling that buffers against outliers. In contrast, Linear Regression forces one global line through all states and months, missing localized twists, while XGBoost, with its many boosting rounds and hyperparameters, appears to have overfit the limited 2021 training set and then stumbled on the new-look Omicron era of 2022.

When we plotted predicted deaths against actual deaths, KNN and Random Forest predictions lined up well with reality, showing balanced forecasts. In contrast, Linear Regression and XGBoost had wider dispersions, especially in months with low mortality, indicating a systematic bias. The results for XGBoost were particularly unexpected; despite its reputation for performance, it showcased that even sophisticated algorithms can struggle with short, noisy data.

These findings are crucial, especially since pandemic response strategies prioritize quick forecasting. Our research indicates that a basic KNN model could already meet this need if the data is gathered appropriately. This is important for smaller health departments that don't have dedicated data science teams but still require reliable death projections to guide their policies. Plus, since KNN is straightforward, officials can easily trace predictions back to specific past neighbors, allowing for verification.

However, our analysis has limitations. We only covered one year of training data, skipping 2020 due to its chaotic reporting.. This means we missed out on the full spectrum of variant dynamics. Additionally, we treated all predictors as if they were happening at the same time, ignoring the noted lag, usually two to four weeks between cases, hospitalizations, and deaths. We also assumed missing vaccination data was zero, which could have skewed the metrics. Finally, we focused only on officially reported COVID-19 deaths instead of excess mortality, meaning we didn't account for deaths misclassified as something else.

Future research should aim to extend the training period, introduce lagged features, and test models that can draw on information from multiple states while retaining local characteristics. Incorporating trends in mobility, variant prevalence, and social distancing data could improve forecasts, especially for early waves. Analyzing excess death data could also reveal if the same models apply in that context.