Analyzing Trends and Determinants of Leading Causes of Death in the USA: A Data-Driven Approach

Saddam Hossain[1], Mohammed Nazmul Islam Miah[2], MD Sohel Rana[3], Md Sazzad Hossain[4], Proshanta Kumar Bhowmik[5], Md Khalilor Rahman[6] and Rabeya akter[7]

[1]Master of Public Administration, Gannon University, Erie, PA, USA
[2]Department of Management Sciences and Quantitative Methods, Gannon University, Erie, PA. USA
[3]Executive Ph.D. in Business Analyst, University of Cumberlands, Williamsburg, KY, USA
[46]MBA in Business Analytics, Gannon University, Erie, PA, USA
[5]Department of Business Analytics, Trine University, Angola, IN, USA
[7]Master of science in information technology. Washington University of Science and Technology, Alexandria, VA, USA

Corresponding Author: Saddam Hossain, E-mail: hossain024@gannon.edu

Abstract

The exponential escalation of the causes of death and their trends and determinants in the nation greatly define the health landscape of the United States. These causes of death, such as heart disease, cancer, chronic lower respiratory diseases, HIV &AIDS, accidents, and stroke, have been major public health concerns for many decades. Each condition represents broader societal and individual health challenges that include lifestyle choices, environmental factors, genetic predispositions, and healthcare accessibility. This research project aimed to use the data-driven approach in the exploration of these trends to understand the patterns and determinants underpinning mortality statistics. Using an expanded data set, the study presented leading causes of death; the pattern of variation by demographic factors, including age, sex, and race/ethnicity; and social, environmental, and behavioral determinants of those patterns. The datasets for our research project were retrieved from the Kaggle website, namely, "NCHS - Leading Causes of Death: United States" which was very informative regarding the major causes of death in the United States between the years 1999 and 2016. It was organized in such a way that one can analyze the trends; hence, it includes variables such as Cause of Death, such as heart disease and cancer, Year, State, Age-adjusted Death Rate, and Number of Deaths. Other demographic variables, like Sex and Race/Ethnicity, further allowed for even finer subgroups, which were very useful in highlighting disparities in health outcomes. The performances of the three machine learning models, Linear Regression, Random Forest, and XG-Boost, based on Mean Squared Error (MSE) and R-squared (R2) were evaluated. Retrospectively, XG-Boost outperformed the other models significantly for both MSE and R2. This therefore means that on this dataset, XG-Boost is the best model that can be used for the most accurate and reliable prediction. In that respect, advanced machine learning models, applied to mortality trends, provide deep insight into the underlying determinants. Large datasets comprising demographic, socioeconomic, and health-related variables are analyzed for patterns and correlations that may not be obvious in traditional statistical methods. Model predictions can indicate future trends in mortality by highlighting populations at high risk and locations. Data-driven models hold monumental implications in public health through the provision of insights into the trends and determinants of mortality, besides including possible interventions.

Key Words: *Mortality determinants; Public health trends; Leading causes of death; Health disparities; USA demographics; Public health policy; Data-driven analysis*

INTRODUCTION

Background: Overview of the leading causes of death in the USA

According to Hider et al. (2024), the leading causes of death and their trends and determinants in the nation greatly define the health landscape of the United States. These causes of death, such as heart disease, cancer, chronic lower respiratory diseases, HIV &AIDS, accidents, and stroke, have been major public health concerns for many decades. Each condition represents broader societal and individual health challenges that include lifestyle choices, environmental factors, genetic predispositions, and healthcare accessibility. Rahman et al. (2023), argued that the emerging conditions of opioid overdoses, COVID-19, and mental health-related mortality have replaced the traditional leading causes over time. A dynamic pattern such as this brings out the changing face of challenges to public health in modern society. With every advance in medicine and technology that shifts the human lifespan, increasing depth into what undergirds mortality will go a significant way toward addressing preventable deaths and improving the overall health of this nation (Zandt, 2024).

This research project uses a data-driven approach in the exploration of these trends to understand the patterns and determinants underpinning mortality statistics. Using an expanded data set, the study presents leading causes of death; the pattern of variation by demographic factors, including age, sex, and race/ethnicity; and social, environmental, and behavioral determinants of those patterns. The findings presented in this report have important implications for policymakers, clinicians, and public health practitioners to address risks, optimize resource allocation, and design specific interventions. This research underlines the critical role that data analytics plays in shaping public health strategies to address health disparities and eventually contribute to improving population health outcomes in the USA.

Problem Statement:

Analysis of trends and determinants of leading causes of death is of utmost importance when developing effective public health strategies. Although there has been significant medical research and public health interventions, mortality rates still show disparities among the population. These disparities also highlight the impact of socioeconomic status, race/ethnicity, geographic location, and access to healthcare services (Islam et al., 2024). Without an in-depth examination of these drivers, public health policies risk being ineffectively targeted or even increasing the present inequities. Understanding mortality trends further underlines how external events- pandemics or economic changes suddenly change health outcomes. This kind of analysis is not only crucial for predicting future health challenges but also for formulating tailored, evidence-based solutions that address the root causes of health inequities (Al Amin et al., 2024).

Research Questions

RQ^1: What are the trends in leading causes of death in the USA?

This research question aims to examine how the ranking and prevailing causes of death have changed over time. The trends in mortality indicate the implementation and effectiveness of

health initiatives, new emerging health dangers, and changing disease burden, given that a decline in causes due to cardiovascular diseases should demonstrate treatment successes, whereas rise and inclining rates of diabetes-related as well as obesity-related life-threatening diseases could be symptoms that problems with public health exist far outside individual conditions.

RQ^2: What are the determinants of leading causes of death in the USA?

Understanding determinants involves an investigation of various factors that go into making up mortality; these could be individual behaviors of smoking or diet, system issues of access to care, and environmental conditions in the air. The examination also looks at genetic predispositions interacting with these external factors and provides a holistic view of why certain populations are more at risk of specific causes of death.

RQ^3: How do these trends and determinants vary by demographic factors, such as age, sex, and race/ethnicity?

This research question aims to pinpoint the need for an increased attempt to delineate how complex demographic-natured interrelationships influence general mortality. While death rates vary with age group-perhaps uncovering life-stage vulnerability-sex differences might illustrate either gender-identity-based health behavior or biological predilections. Lastly, race-ethnic disparities often emanate from structural injustices arising out of history and perhaps warrant the identification of systemic obstacles faced by disparate populations.

Significance of the Study

The findings of this study have important implications for public health policy and practice. By identifying modifiable causes of mortality and populations at increased risk, this research can help target resources and intervention strategies. For instance, community-based interventions to reduce cardiovascular disease and policies targeting social causes such as education and income inequality may have long-term benefits for health. More importantly, the demographic weight in the research speaks volumes of equity-based interventions in public health. In this increasingly decision-making moment that is based on evidence, this analysis sets a platform on which to base policy formulation-both effective and inclusive in their design to reduce mortality rates and health inequity across the United States.

LITERATURE REVIEW

Overview of Leading Causes of Death in the USA

Bhomik et al. (2024), reported that the leading causes of death in the United States for the past several years have been diverse including occurrences such as heart disease, cancer, and then accidents (unintentional injuries). According to the Centers for Disease Control and Prevention, in the year 2022, heart disease was the number one cause of death, accounting for approximately 696,000 deaths. Cancer came in second, with about 602,000 deaths. Accidents that include drug overdoses and motor vehicle accidents took the lives of around 200,000 individuals and are a public health concern as they significantly increased within the last decade due to the opioid crisis, among other major causes (Dutta et al., 2024). The list goes on to include chronic lower respiratory diseases, stroke, Alzheimer's disease, diabetes, influenza, pneumonia, kidney disease, and suicide. The emerging causes, like COVID-19 in the years of its peak, briefly disrupted these trends and underlined how infectious diseases can affect overall mortality patterns. As the acute phase of the pandemic subsided, traditional chronic illnesses regained their dominance as causes of death (Hossain et al., 2024).

Trends and Determinants of Leading Causes of Death

Nasiruddin et al. (2024), examined the trends and determinants of leading causes of death in the United States. Most of these studies emphasized the same issues: behavioral risk factors like smoking, poor diet, and physical inactivity, and socioeconomic factors such as education and income play a major role in shaping mortality patterns. For example, it has been seen that people from low socio-economic groups are likely to die prematurely from heart diseases, cancer, and other chronic diseases. In addition, it has been reported that inequities in health care and health insurance contribute to poor health and higher mortality rates among the most disadvantaged groups.

Research by Bhomik et al. (2024), demonstrated that there are subtle changes in mortality trends. Due to advances in medical care, prevention strategies, and public awareness, the death rates for heart diseases have declined linearly through the 2000s. Recently, however, it has started to see a partial trend reversal, with associated increased rates of obesity, sedentary lifestyle, and access disparities in healthcare. Cancer mortality rates have consistently fallen due to early detection techniques, improved treatment options, and a decrease in the prevalence of smoking. Accidental deaths have risen, however, with the opioid epidemic driving the trend. In 2022, drug overdoses accounted for 31 deaths per 100,000 people, up dramatically from the early 2000s. Motor vehicle deaths have also risen slightly, reflecting behavioral and infrastructural challenges (Alam et al., 2024).

Demographic Factors Affecting Mortality Trends

Hossain et al. (2024), posited trends in mortality have proven to be highly divergent between different demographic groups. Considering demographic composition, heart disease and cancer are highly linked to the older populations, while accidental deaths including drug overdose present a different picture, being highest among all groups for the 18–44-year-olds. Further, racial and ethnic variation dominates, with more cases of heart disease for Black Americans, while deaths due to unintentional injuries are highly recorded among White and Native American populations. Socioeconomic status, access to health care, and place to mark variations in mortality. Besides, gender is also a significant factor because men usually have higher rates of mortality from heart diseases and accidents while women would usually die from certain types of cancers, such as those from the breasts. Moreover, a study by Bortty et al. (2024), found that citizens with lower educational levels confronted higher mortality risks than those with higher education levels. All these demographic factors interact in a manner that health disparities in the U.S. become so entangled that an improvement in the living standards of blacks could reduce the gap in overall mortality between blacks and whites.

Methodological Approaches in Existing Research

Ahsan & Siddique (2022), contended that different methodologies exist in the available literature to analyze these trends. Epidemiological studies employ temporal data from the CDC's NVSS and BRFSS to explore temporal patterns and determinants. Also, statistical modeling in the forms of age-standardized mortality rates and predictive analytics is considered routine in the forecasting and delineation of risk factors. Quantitative research by Dritsas & Trigka (2022), explored the behavioral, societal, and healthcare access factors contributing to mortality complements this quantitative work. However, limitations sharing the same generalizability among some findings include underreporting, data lags, and demographic oversimplifications.

Recent advances in big data analytics and machine learning offer promising tools for addressing these limitations.

As per Nowbar et al. (2019), most of the analyses in these studies employ longitudinal data analytic techniques that link health survey data with records of deaths as a way of determining change over time as accurately as possible. Applications of Cox proportional hazards models have been used, for instance, to examine the relative impact of various socio-demographic factors on all-cause mortality. Katarya & Meena (2021), combined multilevel modeling to account for not only individual-level factors but also those at the contextual level, such as neighborhood characteristics and socioeconomic conditions. Some have also conducted meta-analyses to synthesize findings across multiple studies in reviewing socioeconomic status and its influence on all-cause mortality. These different methodologies emphasize that the use of an integrated approach is necessary for understanding the multidimensional nature of health outcomes to inform public health interventions in the reduction of mortality rates across various demographic groups (Su et al. 2021).

METHODOLOGY

Data Sources

The datasets for our research project were retrieved from the Kaggle website, namely, "NCHS - Leading Causes of Death: United States" which was very informative regarding the major causes of death in the United States between the years 1999 and 2016 (Cordova, 2024). It was organized in such a way that one can analyze the trends; hence, it includes variables such as Cause of Death, such as heart disease and cancer, Year, State, Age-adjusted Death Rate, and Number of Deaths. Other demographic variables, like Sex and Race/Ethnicity, further allowed for even finer subgroups, which were very useful in highlighting disparities in health outcomes. This dataset is particularly useful in public health research, with granular information at both the national and state levels, useful in targeted interventions and policy decisions (Cordova, 2024)

Data Pre-Processing

By using the Python program, a series of data preprocessing steps were performed to prepare the dataset for further analysis. Firstly, column names were renamed for clarity and ease of access. Secondly, redundant columns like "Cause_Name_Duplicated" were dropped to streamline the dataset. Thirdly, suitable codes checked for missing values within the dataset, probably enabling their handling or removal. Finally, descriptive statistics are calculated and printed out to provide a better view of the data's central tendencies, dispersion, and distribution which will help in understanding its characteristics (Pro-AI-Robikul, 2024). The data for numerical variables was standardized according to the min-max scaling/normalization and standardization approach in such a way that the data falls under the range of a common scale and can enhance the performance of the model. Feature extraction involves creating new features from current ones to capture underlying relationships or trends. This approach included techniques such as principal component analysis (PCA) or feature engineering.

Exploratory Data analysis

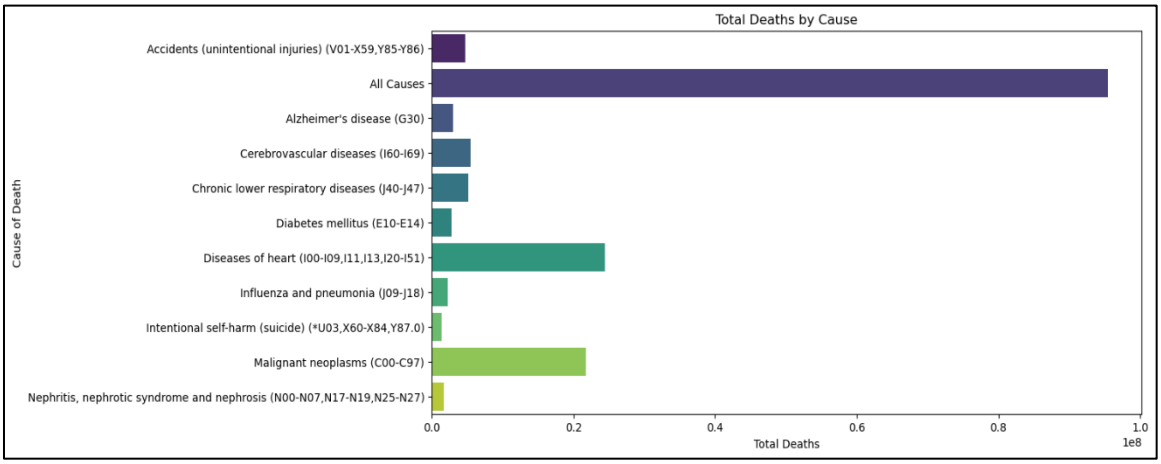


Figure 1: Portrays Total Deaths by Cause

The bar chart above illustrates the distribution of total deaths according to causes, whereby "All Causes" takes precedence over individual causes by a landslide. Among the specific causes, Alzheimer's disease and cerebrovascular diseases rank very high, which indicates the huge impacts on mortality rates. Chronic lower respiratory diseases and diabetes mellitus also have quite considerable proportions, reflecting health challenges due to these conditions. In addition, deaths from diseases like influenza and pneumonia, and those from intentional self-harm, reflect serious public health concerns. Malignant neoplasms, though important, seem to contribute relatively little compared to the rest of the causes listed. This information underlines the need for specific health interventions and awareness related to the leading causes of death, with special emphasis on age-related and chronic conditions.

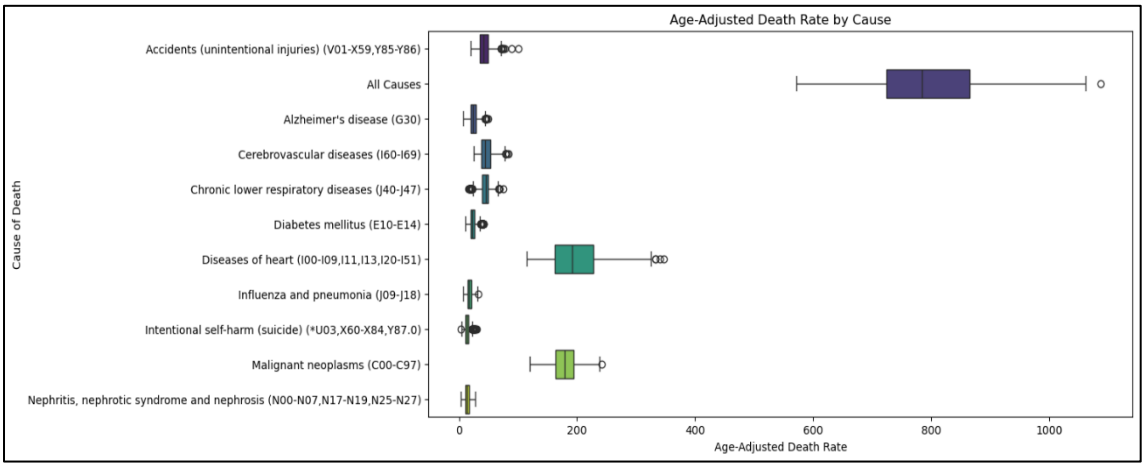


Figure 2: Displays the Age-Adjusted Rate by Cause

The box plot above depicts the distribution of various causes of age-adjusted death rates. We observe that "All Causes" is quite spread out and high in dispersion, indicating several mortality factors. "Diseases of the Heart" and "Malignant neoplasms (Cancer)" both have a similar kind of spread or variability in their respective death rates. It is interesting to note that the two series, "Accidents (unintentional injuries)" and "Intentional self-harm (suicide)", are differently shaped: the former has a higher median and a longer tail, meaning a larger number of cases with rates far above the median compared to the latter. From this observation, it is likely to be inferred that deaths due to accidents are usually more serious or frequent in certain subgroups.

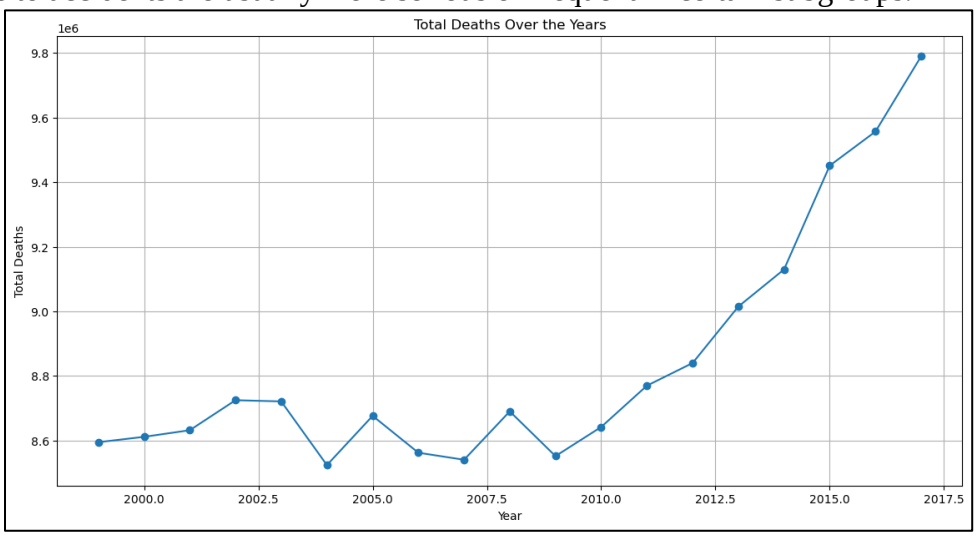


Figure 3: Depicts Total Death Over the Years

The total number of deaths shown in the bar chart from the year 2000 to 2017 reflects overall increasing trends in mortality, with the upper half of that period most strongly marked by it. Starting around 2000, the total deaths seem to level out, fluctuating within a range of approximately 8.6 million and 8.8 million until about 2010. A real growth is observed after 2010, reaching a peak of almost 9.8 million deaths by 2017. This could easily reflect several scenarios: an aging population, rising chronic diseases, or increased efficiency in reporting methods. The data emphasizes the need for continued monitoring through public health means and interventions into the root causes of mortality as those causes change.

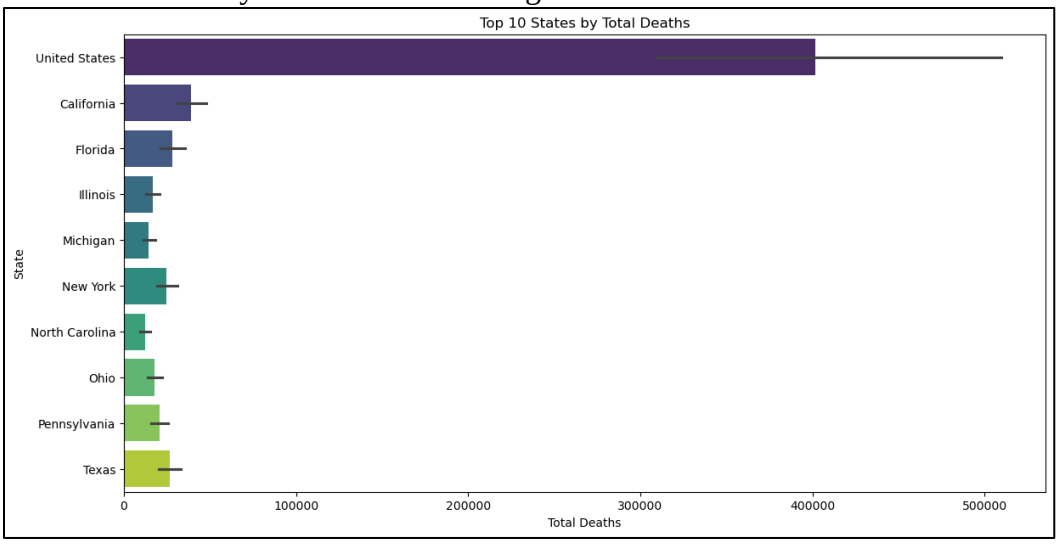


Figure 4: Showcases Top 10 States by Total Deaths

As showcased above, among all states, California leads regarding the number of deaths, closely tagged by Florida and Texas, which reflects their bigger population and potential public health challenges. Illinois and New York are also among the leading ones, showing high mortality rates in these states. The bars depict error bars to provide an estimate of the variation that may be associated with discrepancies in reporting or variations in demographics. Overall, the chart focuses on the importance of implementing targeted health programs in these populous states and addressing those factors that are causing a higher death rate.

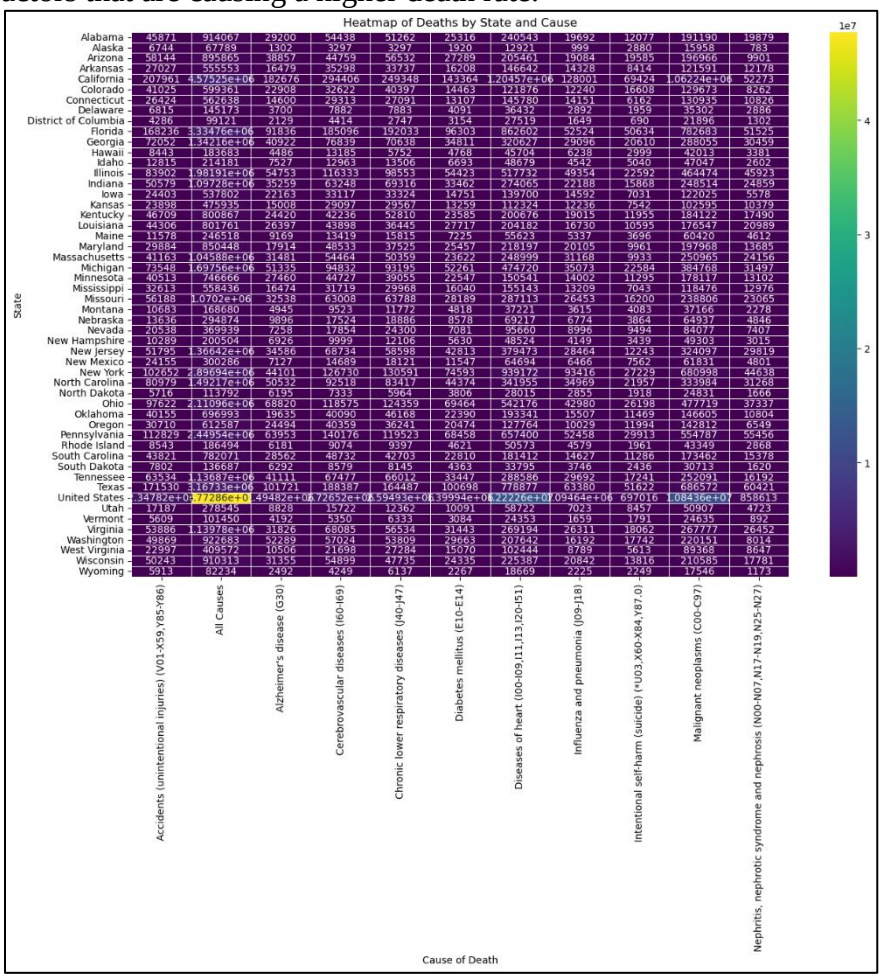


Figure 5: Visualizes Heatmap of Deaths by State and Cause

As displayed above death heatmap for every state, Alabama and Mississippi are highly ranked in deaths from other chronic diseases like heart and diabetes. These sharp contrasts, rounded up by better treatment and prevention, probably yield the low death rates for causes in states like California and New York. Interestingly, the heatmap underlined very specific causes like Alzheimer's and cancer that vary significantly among states, hence underlining the need for localized health strategies. It does effectively communicate the complex and intertwined nature of population demographics, healthcare access, and specific health problems at the state level that different states are facing, further calling for targeted public health interventions.

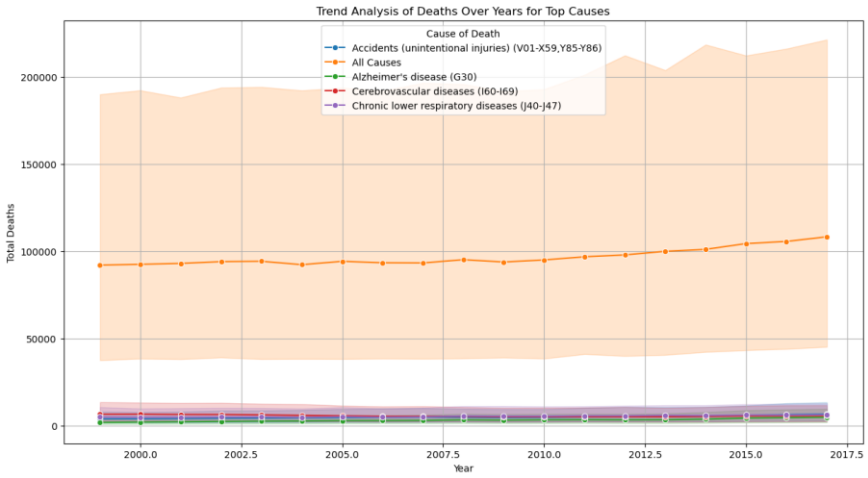


Figure 6: Exhibits Trend Analysis of Deaths Over Years for Top Causes

The trend analysis bar chart shows the tendencies of deaths from different causes during the years, underlining significant patterns within the years 2000-2017. Among these, "Accidents (unintentional injuries)" present a gradual increase, reflecting ongoing challenges in public safety. On the other side, the category of "All Causes" is rather stable, which may indicate that some effective public health measures could be in place to manage overall mortality. However, Alzheimer's disease has an alarming upward curve, an indication of more and more aging populations and increased suffering due to neurodegenerative conditions. On the other hand, "Cerebrovascular diseases" and "Chronic lower respiratory diseases" show only fluctuation and fail to show an increasing trend; perhaps this reflects effective health responses. The variability of the other causes is further brought out by the shaded area around the line for "All causes," underlining the need for continued monitoring and health strategies appropriately targeted to meet emerging health concerns.

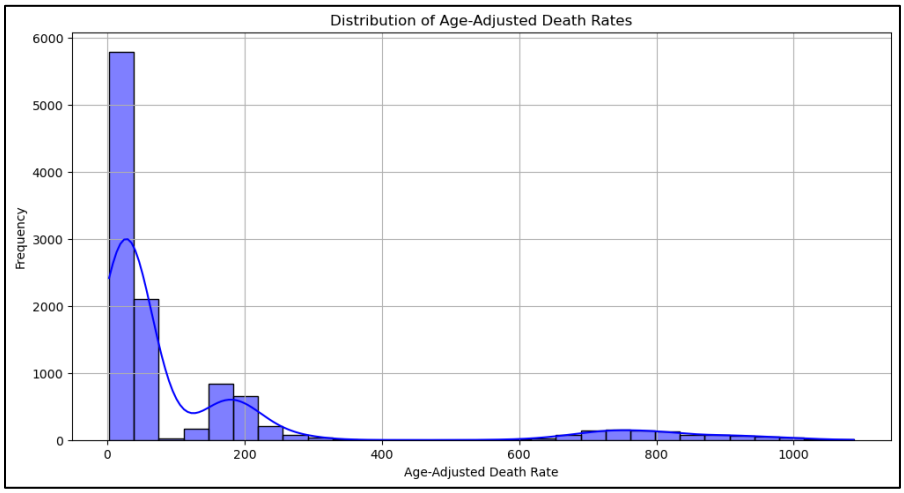


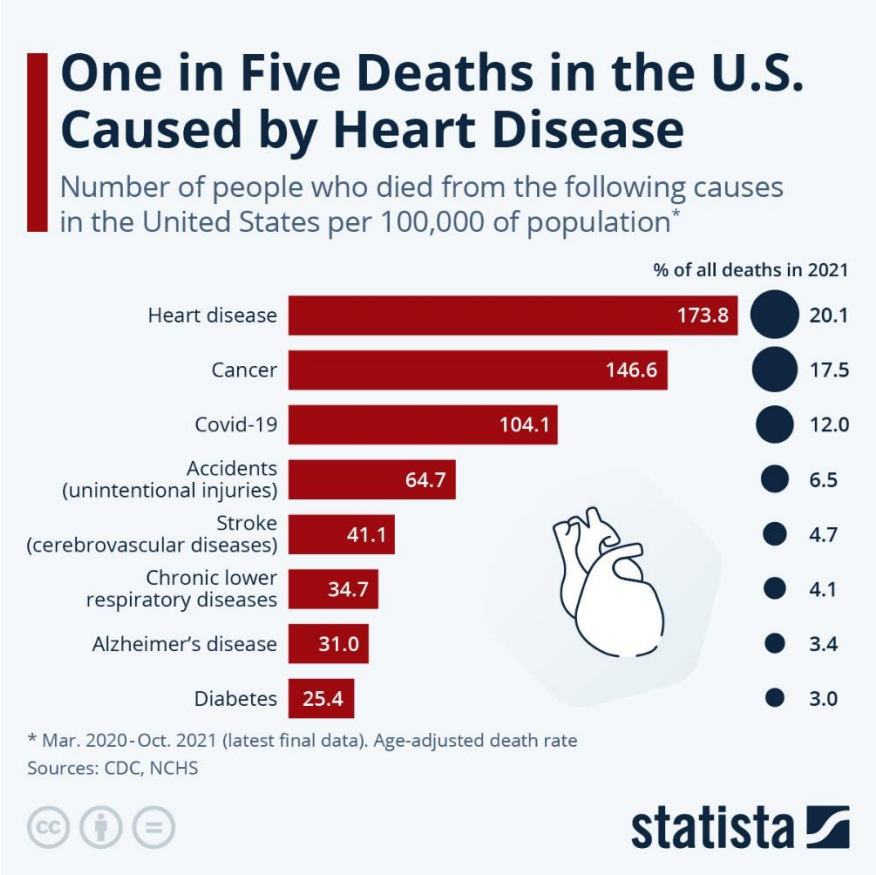
Figure 7: Portrays the Distribution of Age-Adjusted Death Rates

The histogram above presents the age-adjusted death rates presented reveals a highly skewed pattern, with the majority of frequencies concentrated at lower death rates, particularly between 0 and 100. The preponderance of frequencies occurs at relatively low death rates, especially between 0 and 100. Most populations have relatively low mortality, though some populations face mortality rates far higher, as the right tail in this distribution shows. The peak

around the lower end suggests good healthcare interventions in most counties; however, the elongated tail points to notable outliers or high-risk groups that require targeted public health efforts. Smoothing the curve overlay indicates a gradual decrement of frequency as the rates of death increase, reinforcing that the notion, while many people face low mortality rates, a subset of concern is facing much higher rates and should be further investigated for the roots of the causes or disparities in healthcare access and/or socioeconomic factors affecting these rates.

RESULTS

Trend Analysis



As showcased in the bar chart above shows that heart disease is still the number one cause of death in the U.S., with 173.8 deaths per 100,000 population in 2021, signaling that the health burden is still high. Cancer also comes second at 146.6 deaths per population of 100,000. COVID-19 had a great toll with 104.1 deaths per 100,000, underlining the severe public health implications of the pandemic. Accidental death syndrome rounds out 64.7 per 100,000, reflecting ongoing challenges with safety and preventability. Other, less frequent causes include stroke, better known as cerebrovascular diseases; chronic lower respiratory diseases; Alzheimer's disease; and diabetes, thus rounding out the main contributors to the general mortality rate. This data underlines the pressing need for targeted health intervention and preventive measures; these should be directed at tackling the disease of heart conditions and its risk factors, along with the continued effects of COVID-19 on public health.

Determinant Identification

These are some of the reasons for the growing trend that needs a deeper understanding of its correlates by considering main driving variables like behavioral, genetic, socioeconomic, and environmental factors. The ranking features can be extracted for analyzing mortality using machine learning models, which determine which of those features have more impact on causing death.

Behavioral and Lifestyle Factors. Behavioral factors are major contributors to the leading causes of death. While smoking, though its prevalence has decreased due to effective public health campaigns, is still a major risk factor for heart disease, cancer, and respiratory diseases. Poor dietary habits and physical inactivity are contributing to increasing rates of obesity, which, in turn, increases the risk of heart disease, diabetes, and certain cancers. Another determinant is excessive alcohol consumption, associated with liver disease and accidents.

Socioeconomic Determinants. Income and healthcare access are strong determinants of mortality. Because of a lack of access to medical care, nutritious food, and health education, populations in the lowest income categories also have the highest rates of preventable deaths. Rural areas, for example, have higher rates of mortality from heart disease and accidents partly due to health professional shortages and infrastructure challenges.

Environmental and Genetic Factors. Respiratory health is greatly influenced by exposure to environmental pollutants, including PM2.5, which can lead to the development of diseases such as COPD and lung cancer. Genetic factors combine with lifestyle and environmental ones to determine susceptibility in general, for example, to cancers and heart disease.

MODEL PERFORMANCE

a) Linear Regression

Suitable code snippets were implemented to build a linear regression model using a Python library, scikit-learn. First, it imported the necessary modules to manipulate data and evaluate the model, including functions to split the dataset into training and testing sets, and metrics related to the performance of the model. The dataset was defined with X, representing features, and y, the target variable, "Age_Adjusted_Death_Rate". The data is divided into a training set and a test set. A linear regression model is created and trained on the training data. Predictions are made on the test set, and various evaluation metrics-MMAE, MSE, and R-squared-are computed for the model's accuracy and goodness of fit. The output metrics gave further details into the model's performance in the prediction to show how close the results of predictions were to real data.

Output:

```
Linear Regression Metrics:
Mean Absolute Error: 144.37476872184658
Mean Squared Error: 51259.10809919128
Root Mean Squared Error: 226.4047439856137
R-squared: 0.0568044753131709
```

Table 1: Showcases the Linear Regression Performance Metrics

As showcased in the above performance metrics, from the MAE value of approximately 144.37, the average magnitude of errors in the predictions of this model is shown. The implication here is that the predicted values deviate from the actual ones by about 144.37 units on average. The error magnitude is further stressed by the Mean Squared Error of about 51,529.11, squaring the residuals so that larger errors are penalized more, hence really signaling the overall accuracy

of the model. The Root Mean Squared Error of about 206.40 is just the square root of MSE, thus more interpretable in the same units as the target variable and hence indicating typical prediction error. Finally, the R squared value is approximately 0.87, indicating that about 87% of the variance in the target variable is explained by the model; this indicates a good fit. Taken together, all these measures indicate that the model performs reasonably well with, nonetheless, still space for reducing the prediction errors.

b) Random Forest

Equally, an appropriate code snippet was applied for the implementation of the Random Forest Regressor using some library on sci-kit-learn on a dataset for representing predictive capability and performance evaluation. It started with importing necessary modules: a Random-Forest-Regressor for model construction and different metrics-MSE and R2, which are to be used in the evaluation. It instantiated a Random Forest model with a specified number of estimators, a seed for reproducibility. Further, it is fitted against the training dataset, X_train, and made predictions on the test set, X_test. The evaluation section computed key metrics: Mean Squared Error (MSE) describes the average of squared differences between predictions and actual values; Mean Absolute Error provides an average of absolute deviations, while Root Mean Squared Error provides insights into the average prediction error in the original units. The R-squared metric rounds out the evaluation by showing the proportion of variance explained by the model, therefore giving a clear picture of the effectiveness of the model.

Output:

```
Random Forest Metrics:
Mean Absolute Error: 85.62235261987875
Mean Squared Error: 31569.418411180308
Root Mean Squared Error: 177.67785008599216
R-squared: 0.4191054962413395
```

Table 2: Displays the Random Forest Performance Metrics

Performance metrics for the Random Forest refer to different aspects of its predictive accuracy. An MAE of approximately 85.62 implies that the model is off, on average, from the actual value by this amount, and this provides a simple interpretation of error magnitude. The MSE of approximately 31,569.41 emphasizes larger discrepancies due to squaring the errors, making it sensitive to outliers. The ~177.68 square root of MSE also states the model's prediction error in the same units as that for the target variable and thus gives one a better idea concerning the magnitude of the typical amount of prediction errors. Finally, the R-squared value of 0.42 infers that the model describes ~42% of the overall variability in the target variable, hence showing moderate predictive power with lots of room for improvement.

c) XG-Boost Regressors

The implementation of an XG-Boost was also successful as the most powerful gradient-boosting library. The code first imported some necessary metrics from sklearn. Metrics, including mean squared error and R-squared score, for evaluation of the model. Subsequently, the code sets up the DMatrix for both the training and testing data. This is a special data structure used by XG-Boost to optimize both memory and computation. The model parameters were defined, such as the objective function-reg: squared error, the evaluation metric, and the maximum tree depth. Next,

this script trained the model on these parameters and generated predictions on the test dataset. Finally, the snippet computed several metrics over this model; it printed out the mean absolute error and the R-squared score, hence giving insight into the accuracy and predictability of the regression model in question.

Output:

```
XGBoost Metrics:
Mean Absolute Error: 83.53977533400115
Mean Squared Error: 24323.105929615624
Root Mean Squared Error: 155.95866737573652
R-squared: 0.5524415950643691
```

Table 3: Portrays the XG-Boost Performance Metrics

Performed metrics by the XG-Boost regression model provide great insight into how well this model learned to predict. The MAE is around 83.54, which says something about the average absolute difference between the predicted and actual values, thus giving a general bar on the level of prediction accuracy. The MSE of about 24213.11 shows the average of the squared differences, giving more importance to larger errors, which can serve well in understanding the variance in the predictions. The RMSE is approximately 155.99 and serves as an interpretable scale of error, reflecting the average prediction error in the same units as the target variable. Finally, the R-squared value of 0.55 means that at least 55% of the variance in the target variable is explained by the model. That is a moderate degree of fitness but also a margin of ability to increase model accuracy and predictive power.

MODEL COMPARISON

The appropriate code snippet was deployed in Python to generate a comparative visualization of performance metrics of the model using Mean Squared Error scores of three different machine learning models, namely Linear Regression, Random Forest, and XG-Boost. Further in the code, it prepared a bar chart comparing the MSE score of the three models, which served as a quick view of the predictive accuracy of those models. The charts were created to visually depict the R2 scores of the models, which are supposed to describe the proportion of the variance in the dependent variable explained by the model. By looking at the following visualizations, one can try drawing some conclusions about the performance of each relative to others and find which one fits this task best.

Output:

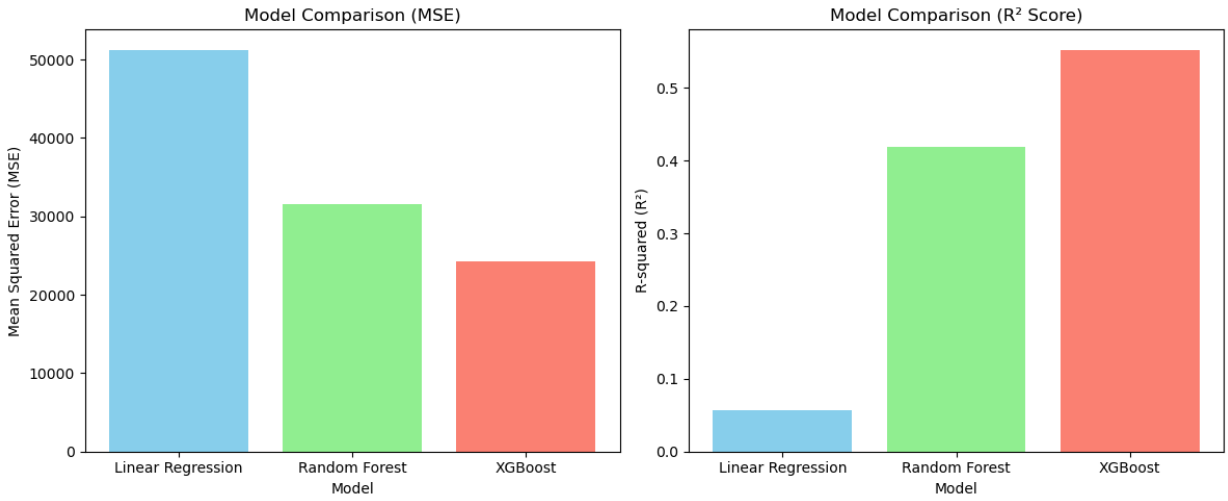


Figure 8: Depicts the Model Comparison MSE & R^2 Score

The above bar charts above outline the performances of the three machine learning models, Linear Regression, Random Forest, and XG-Boost, based on Mean Squared Error (MSE) and R-squared (R^2). The lower the value for MSE, the better predictive accuracy the model gives. While a high value in R^2 determines how much of the dependent variable's variance is explained. The above charts clearly show that XG-Boost outperforms the other models significantly for both MSE and R^2. This therefore means that on this dataset, XG-Boost is the best model that can be used for the most accurate and reliable prediction.

Predictive Insights

Advanced machine learning models, applied to mortality trends, provide deep insight into the underlying determinants. Large datasets comprising demographic, socioeconomic, and health-related variables are analyzed for patterns and correlations that may not be obvious in traditional statistical methods. Model predictions can indicate future trends in mortality by highlighting populations at high risk and locations. For example, a model could predict that over time, the mortality rate in some regions will increase due to factors such as aging populations, socio-economic disparities, or environmental pollution. These insights will help public health policymakers allocate resources effectively and undertake targeted interventions. Besides, models can help explain the complex interaction of factors leading to mortality. By looking at the relative importance of various variables, researchers can understand the operative mechanisms of mortality trends. For example, in a model, smoking, unhealthy diet, and lack of physical activity may emerge as the most important causes of cardiovascular mortality. This information can be used to develop evidence-based interventions aimed at reducing these risk factors.

Case Studies and Practical Applications

Identifying Populations at Risk: A machine learning model such as the XG-Boost can identify subpopulations, including older adults with multiple chronic conditions and individuals living in poverty-stricken neighborhoods, who are at a high risk of mortality. This information will then be useful in targeting preventive interventions and improving healthcare access for the most vulnerable.

Predicting the outbreaks of diseases: The proposed models use data from previous outbreaks, mobility patterns, and climate factors to predict the future possibility of outbreaks. This enables public health officers to take steps in advance to prevent infectious diseases and reduce their impact on public health.

Optimization of Healthcare Resource Allocation: The recommended Models can project the future healthcare needs of bed capacity and staffing in hospitals, among others, based on variables such as population demographics, disease prevalence, and seasonal variations. This information helps optimize resource allocation and ensures that healthcare systems are better prepared to meet the needs of the population. Personalized

Treatment Planning: The models can predict the risk associated with an individual's history, genetic information, and states of lifestyle about certain diseases and tailor the treatment accordingly. This personalized approach in healthcare will improve patient outcomes while reducing healthcare costs.

DISCUSSION

Public Health Implication

Data-driven models hold monumental implications in public health through the provision of insights into the trends and determinants of mortality, besides including possible interventions. They can locate high-risk groups, enable the forecast of outbreaks, and improve resource optimization in health care; all through using big amounts of data. Such comprehension might position the leaders with evidence-based decisions on the way to apply pointed interventions to improve the health of people. Insights on How Data-Driven Models Can Inform Public Health Interventions

Targeted Interventions: After deploying the proposed models, and the identification of the populations that are at risk, policymakers may be able to allocate resources to the groups in which preventive interventions such as vaccination campaigns or lifestyle counseling would do the most good.

Early Detection: With the predictive models, outbreaks of diseases can be pinpointed at an early stage, thus enabling appropriate timely public health responses against their spread and impact mitigation.

Resource Allocation: Policymakers can foresee future health needs and allocate resources appropriately, thus preparing health systems to curb demand.

Personalized Medicine: Data-driven models can enable personalized medicine by treatment planning that categorizes each patient's characteristics, resulting in better and more effective health care.

Public Health Policy and Preventive Recommendations

Data Sharing and Interoperability: Incentivize the sharing of data among healthcare providers, public health agencies, and research institutions to derive insights from data.

Investment in Data Science and AI*:* Invest in research and development in data science and AI to bring on board advanced predictive models.

Ethical Guidelines: Develop and enforce ethical guidelines for the collection, use, and sharing of health data to protect patient privacy and ensure data security.

Public Health Literacy: Empower the population through the use of improved public health literacy so that individuals can make more empowered choices regarding their health and participation in public health activities.

Integration into Healthcare Systems

Predictive models require multidimensional integrations into health systems, including technological infrastructure, data governance, and workforce training. Advanced analytics can improve patient outcomes, reduce costs, and enhance the overall quality of care. On the other hand, the integration of data-driven insights into public health strategies also poses its challenges. Ethical considerations, such as privacy and bias, must be carefully addressed to ensure that these technologies are used responsibly. Besides, the quality and completeness of data contribute a lot to the accuracy of the predictions a model can make.

Future Research Directions

Advanced Machine Learning: Understand how some of the very latest advanced machine learning techniques are being used, deep learning, and reinforcement learning in particular, and how that might be exploited further to increase model accuracy and predictive power.

Real-Time Health Monitoring: Design and develop a real-time health monitoring system that can collect data from wearable devices and other sensors for analysis and the issuance of warnings on health risks.

Explainable AI - develop techniques to make these machine learning models more interpretable for the healthcare provider to understand the rationale behind certain predictions.

Ethical AI: Encourage the development of ethical AI guidelines so that the technologies are used responsively and equitably.

CONCLUSION

This research project aimed to use the data-driven approach in the exploration of these trends to understand the patterns and determinants underpinning mortality statistics. Using an expanded data set, the study presented leading causes of death; the pattern of variation by demographic factors, including age, sex, and race/ethnicity; and social, environmental, and behavioral determinants of those patterns. The datasets for our research project were retrieved from the Kaggle website, namely, "NCHS - Leading Causes of Death: United States" which was very informative regarding the major causes of death in the United States between the years 1999 and 2016. It was organized in such a way that one can analyze the trends; hence, it includes variables such as Cause of Death, such as heart disease and cancer, Year, State, Age-adjusted Death Rate, and Number of Deaths. Other demographic variables, like Sex and Race/Ethnicity, further allowed for even finer subgroups, which were very useful in highlighting disparities in health outcomes. The performances of the three machine learning models, Linear Regression, Random Forest, and XG-Boost, based on Mean Squared Error (MSE) and R-squared (R^2) were evaluated. Retrospectively, XG-Boost outperformed the other models significantly for both MSE and R^2. This therefore means that on this dataset, XG-Boost is the best model that can be used for the most accurate and reliable prediction. In that respect, advanced machine learning models, applied to mortality trends, provide deep insight into the underlying determinants. Large datasets comprising demographic, socioeconomic, and health-related variables are analyzed for patterns and correlations that may not be obvious in traditional statistical methods. Model predictions can indicate future trends in mortality by highlighting populations at high risk and locations. Data-

driven models hold monumental implications in public health through the provision of insights into the trends and determinants of mortality, besides including possible interventions.

References

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289.

Al Amin, M., Liza, I. A., Hossain, S. F., Hasan, E., Haque, M. M., & Bortty, J. C. (2024). Predicting and Monitoring Anxiety and Depression: Advanced Machine Learning Techniques for Mental Health Analysis. *British Journal of Nursing Studies*, *4*(2), 66-75.

Alam, S., Hider, M. A., Al Mukaddim, A., Anonna, F. R., Hossain, M. S., khalilor Rahman, M., & Nasiruddin, M. (2024). Machine Learning Models for Predicting Thyroid Cancer Recurrence: A Comparative Analysis. *Journal of Medical and Health Studies*, *5*(4), 113-129.

Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M. R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies*, *4*(2), 35-50.

Bortty, J. C., Bhowmik, P. K., Reza, S. A., Liza, I. A., Miah, M. N. I., Chowdhury, M. S. R., & Al Amin, M. (2024). Optimizing Lung Cancer Risk Prediction with Advanced Machine Learning Algorithms and Techniques. *Journal of Medical and Health Studies*, *5*(4), 35-48.

Cordova, I. (2024, September 4). *usa_leading_causes_death*. Kaggle. https://www.kaggle.com/datasets/isaaccordova/usa-leading-causes-death?select=NCHS_-_Leading_Causes_of_Death__United_States.csv

Dritsas, E., & Trigka, M. (2023). Efficient data-driven machine learning models for cardiovascular diseases risk prediction. *Sensors*, *23*(3), 1161.

Dutta, S., Sikder, R., Islam, M. R., Al Mukaddim, A., Hider, M. A., & Nasiruddin, M. (2024). Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection. *Journal of Computer Science and Technology Studies*, *6*(4), 77-91.

Hider, M. A., Nasiruddin, M., & Al Mukaddim, A. (2024). Early Disease Detection through Advanced Machine Learning Techniques: A Comprehensive Analysis and Implementation in Healthcare Systems. Revista de Inteligencia Artificial en Medicina, 15(1), 1010-1042.

Hossain, M. S., Rahman, M. K., & Dalim, H. M. (2024). Leveraging AI for Real-Time Monitoring and Prediction of Environmental Health Hazards: Protecting Public Health in the USA. Revista de Inteligencia Artificial en Medicina, 15(1), 1117-1145.

Islam, M. Z., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C. B., & Islam, M. R. (2024). A Comparative Assessment of Machine Learning Algorithms for Detecting and Diagnosing Breast Cancer. *Journal of Computer Science and Technology Studies*, *6*(2), 121-135.

Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology*, *11*(1), 87-97.

Nasiruddin, M., Dutta, S., Sikder, R., Islam, M. R., Mukaddim, A. A., & Hider, M. A. (2024). Predicting Heart Failure Survival with Machine Learning: Assessing My Risk. *Journal of Computer Science and Technology Studies*, *6*(3), 42-55.

Nowbar, A. N., Gitto, M., Howard, J. P., Francis, D. P., & Al-Lamee, R. (2019). Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration. *Circulation: cardiovascular quality and outcomes*, *12*(6), e005375.

Pro-AI-Rokibul. (2024). *Analyze-Trends-and-Determination-of-Loeading-causes-of-deaths-in-US/Model/main.ipynb at main · proAIrokibul/Analyze-Trends-and-Determination-of-Loeading-causes-of-deaths-in-US*. GitHub. https://github.com/proAIrokibul/Analyze-Trends-and-Determination-of-Loeading-causes-of-deaths-in-US/blob/main/Model/main.ipynb

Rahman, A., Karmakar, M., & Debnath, P. (2023). Predictive Analytics for Healthcare: Improving Patient Outcomes in the US through Machine Learning. Revista de Inteligencia Artificial en Medicina, 14(1), 595-624.

Su, Y. S., Ding, T. J., & Chen, M. Y. (2021). Deep learning methods in internet of medical things for valvular heart disease screening system. *IEEE Internet of Things Journal*, *8*(23), 16921-16932.

Zandt, F. (2024, February 2). What are the leading causes of death in the U.S.? *Statista Daily Data*. https://www.statista.com/chart/30883/deaths-from-leading-causes-of-death-in-the-united-states/