Modeling Deaths for Diabetes & Kidney Diseases, Globally

Authors: Shreyas Dikshit, Amare Diotte, Taylor Goodwin, Aditi Gupta

Summary

The project aims to use historically available data about the death rates for Diabetes and Kidney diseases to predict future death rates so that necessary steps can be taken by government and health officials to mitigate the same.

The death rates are collected for various countries and are based on a set of economic and health-related factors such as government health expenditure, a country's GDP, a country's populations, access to technology and internet, access to sanitation facilities etc.

The death rates would be predicted for a particular year (eg. 2019) by building a machine learning model such as linear regression on previous years' data (eg. 2010 or 2015). The model would be evaluated against the actual available death rates for 2019 and would be iteratively improved until the error in prediction reduces.

The data was collected from two sources:

- Global Burden of Disease (GBD [5] - provides a comprehensive picture of mortality and disability across countries, time, age, and sex. It quantifies health loss from hundreds of diseases, injuries, and risk factors, so that health systems can be improved and disparities eliminated. It is maintained by the Institute for Health Metrics and Evaluation (IHME), a research institute working in the area of global health statistics and impact evaluation at the University of Washington in Seattle.

- World Development Indicators [6] - is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. The database contains 1,400 time series indicators for 217 economies and more than 40 country groups, with data for many indicators going back more than 50 years.

Before building machine learning models, exploratory data analysis was performed to gain insights into the relationships between the various factors and the death rates. This will involve visualizing the data using techniques such as scatter plots, box plots, and world maps, to identify any trends or insights in the data.

Methods

A detailed "*Methods*" section is available separately, please see:
https://docs.google.com/document/d/1Y4MWRae78cLqWdwoEYex4NoBRUJmWwVRWy6I6k-A6nQ/edit?usp=sharing

Results

In examining the impacts of demographic variables upon the national death rates of diabetes and kidney disease globally, first exploring the differences in the distributions of death rates and incidence of disease rates, was essential.

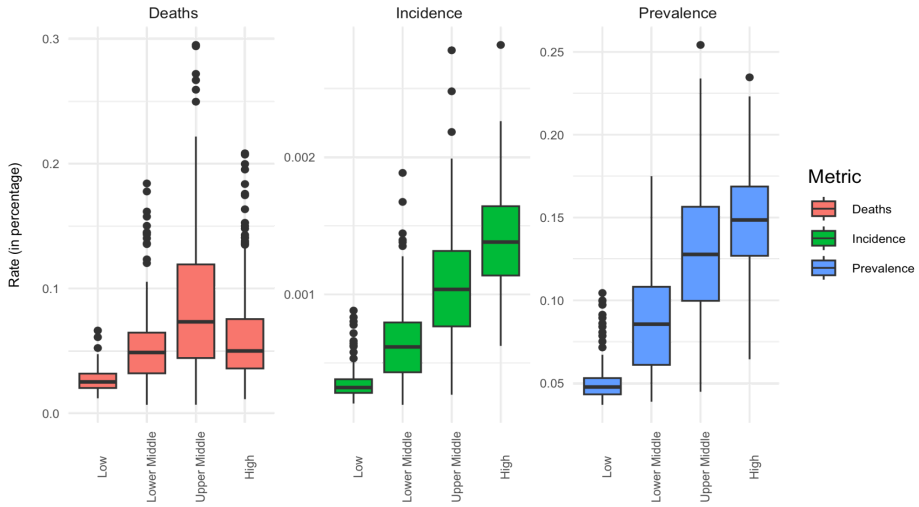

Figure 1: Distribution and incidence of death rates

As can be seen in the plot above, the highest mean death rates occur in upper middle income countries, while interestingly, the highest mean rates of diabetes and kidney disease incidence appear in high income countries. This difference is likely related to other variables in the national environment, including a nation's income level, various health factors and/or other demographically influential measures. Next, in exploring the influential factors which may have an impact on national rates of death from diabetes and kidney disease, the following predictive multivariable linear regression model was developed.

The Model:

$$\textit{Death rate} = b_0 + 70.4332\, x_1 + -0.0176\, x_2 + 0.0013\, x_3 + 0.0202\, x_4 + -0.0009\, x_5$$

y = National Diabetes & Kidney Disease Death Rates

$\mathbf{b_0}$ = Intercept

$\mathbf{x_1}$ = Diabetes & Kidney Disease Incidence Rate

$\mathbf{x_2}$ = Low Income Nation

$\mathbf{x_3}$ = Lower Middle Income Nation

$\mathbf{x_4}$ = Upper Middle Income Nation

$\mathbf{x_5}$ = Prop. of Population Internet Access

With the aim of predicting future rates of death for diabetes and kidney disease, we ultimately assessed our model using comparative RMSE values and by looking at the measurable

difference between the predicted 2019 values and actual 2019 values. In predicting 2019 values, the model achieved a low RMSE score of 0.03387909, indicating the model to be a fairly good predictor of national diabetes and kidney disease death rates, for at least a few years beyond the testing data available.

Discussion and Conclusion

The results of this project have significant meaning and impact, as they provide valuable insights into the future death rates for diabetes and kidney diseases in various countries, across the world. By using historically available data and building a machine learning model, we were able to predict future death rates for these diseases and identify influential factors, such as government health expenditure, a country's GDP, and access to technology and sanitation facilities.

This information can be of immense value to government officials, healthcare providers, and public health experts, who can use the data to inform policy decisions and allocate resources in a more targeted manner. For instance, governments can use this information to prioritize funding for healthcare infrastructure, improve access to healthcare in areas with high death rates, and implement preventative measures to curb the incidence of these diseases.

Moreover, the results of this project can benefit individuals who are at a higher risk of developing diabetes and kidney disease, so as to make them aware of the risks and encourage them to make lifestyle changes that can help prevent or manage the diseases.

Despite the success of our model in predicting future death rates, there is always room for improvement in future work. For instance, the model can be further refined by incorporating more demographic and health-related factors and using more advanced machine learning techniques. Additionally, the project can be extended to include analysis of trends over a longer time period for other diseases such as Cardiovascular diseases and exploring the impact of other factors, such as self-harm and interpersonal violence, on the incidence of these diseases.

Overall, this project provides valuable insights into the future death rates for diabetes and kidney disease in various countries and highlights the importance of investing in preventative healthcare measures and improving access to healthcare infrastructure.

Statement of contributions

Shreyas Dikshit:

- Procured dataset from the GDB website in the required format
- Assisted in pivoting and tidying the World Bank data
- Prepared exploratory data analysis: lineplots and barplots
- Assisted in the stepwise model selection technique to build the linear regression model

- Organized a git repository to share and maintain code
- Prepared the problem statement and some EDA slides of the presentation
- Prepared the Summary and Exploratory Data Analysis sections of the final report

Amare Diotte:
- Database design and implementation
- Database updates (4 total)
- Data tidying and normalization
- Algorithms and visualizations used for data exploration
- Prepared exploratory data analysis plots (world map, box plots)
- “Data exploration and tidying” section of methods
- Image editing/layout for EDA data in presentation

Taylor Goodwin:
- Researched and procured demographic and economic dataset from the World Bank.
- Performed literature reviews of relevant existing models and general research of diabetes and kidney disease death rates’ behavior in light of demographic elements.
- Tidied and prepared consolidated training dataset and testing datasets in R.
- Prepared exploratory data analysis plots: scatterplots, histograms and maps.
- Prepared and evaluated linear regression models.
- Finalized and performed model validation of the regression models, including evaluation of statistical measures and creation and interpretation of validation plots.
- Prepared *Methods, Data Normalization and Modeling* section of final report.
- Prepared *Results* section of final report.
- Created slides and images for presentation of model training, development, validation and prediction.

Aditi Gupta:
- Conducted literature review and implemented various unsupervised machine learning techniques such as Principal Component Analysis and Singular Value Decomposition to come up with the best method to reduce dimensionality and get relevant factors in R.
- Resourced and finalized on the greedy stepwise model selection technique to iteratively add/drop the best/worst variable to the model based on RMSE values.
- Prepared exploratory data analysis plots: bar graphs.
- Preparing the Data Gathering and Preprocessing slides for presentation.
- Prepared *Discussion and Conclusion* section of final report.
- Prepared the *References* section and did figure editing for the final report and methods document.

References

Research References-

[1] Bein, M. A., Unlucan, D., Olowu, G., & Kalifa, W. (2017). Healthcare spending and health outcomes: evidence from selected East African countries. African Health Sciences, 17(1), 247–254. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5636241/

[2] Rabbi, A. M. F., & Mazzuco, S. (2018). Mortality and life expectancy forecast for (comparatively) high mortality countries. Genus, 74, 18. https://genus.springeropen.com/articles/10.1186/s41118-018-0042-x

[3] Centers for Disease Control and Prevention. (n.d.). Diabetes and Chronic Kidney Disease. U.S. Department of Health & Human Services. https://www.cdc.gov/diabetes/managing/diabetes-kidney-disease.html#:~:text=Both%20type%201%20and%20type%202%20diabetes%20can%20cause%20kidney%20disease.&text=Kidney%20 diseases%20are%20the%209th,begin%20treatment%20for%20 kidney%20failure

[4] Adejumo, W. A., Tijani, A. R., & Onatola, S. A. (2019). What Are the Socio-Economic Predictors of Mortality in a Society? Journal of Financial Risk Management, 8(4), 165-176. https://www.scirp.org/journal/paperinformation.aspx?paperid=96881

Data Sources -

[5] Institute for Health Metrics and Evaluation. (2020). Global Burden of Disease Study 1990-2019. University of Washington, Population Health Building/Hans Rosling Center, 3980 15th Ave. NE, Seattle, WA 98195, USA: Institute for Health Metrics and Evaluation. Retrieved March 15, 2023, from https://vizhub.healthdata.org/gbd-results/

[6] World Bank. (2023). World Development Indicators. Retrieved March 15, 2023, from https://data.worldbank.org/

Appendix

R code, datasets and databases can be accessed here:
DS_5110_Project_Team11 (github.com)