Predicting Heart Disease Death Rates in U.S. Counties Using Multiple Risk Factors

Random Forrest Classification

Jordan Bennett, Dept. of Computer Science, Hood College, Frederick Maryland

ABSTRACT

In this paper, I explore heart disease death rates from every county in the United States, and cluster them by those with the highest category of deaths. I then cluster counties with high rates of several heart disease risk factors, including smoking, obesity, population over 65, and median income. Random forest classification is then used to determine an accuracy rate for each individual risk factor to predict whether a county will have the highest category of heart disease deaths. The data suggests that the risk factor most closely related to a heightened rate of heart disease deaths is not a health-related statistic, but in fact median income for the county in question. It was also determined that a county’s obesity rate and percentage of residents over 65 years old were both unable to provide an accurate classification of high heart disease death rate.

INTRODUCTION

Heart disease is the leading cause of death in the United States, with over 659,000 occurrences in 2019 [1]. The Centers for Disease Control and Prevention (CDC) hosts an Interactive Atlas of Heart Disease and Stroke. This provides an interactive map of the US over which users can apply different queries and filters related to various health and environmental factors related to heart disease and stroke cases. Applying a query for heart disease death rates for residents 35 and over (from 2017-2019), it can be seen that the highest rates of deaths seem to be centered around the Southern region of the country, reaching from the East to the central part of Texas.

Graphical user interface, text, application

Description automatically generated

Figure 1. Output from the CDC Interactive Atlas of Heart Disease and Stroke, displaying concentration of heart disease deaths.

Several risk factors were identified that had a similar density of occurrence in the South. These factors were selected to be included in the project, and include obesity rate, active smoker percentage, over 65 years old percentage, and median income for each county in the country. The data for these factors, along with the county death rates, were acquired directly from the CDC statistics map, using an included export feature that allows for the download of csv files.

COMPUTING ENVIRONMENT

This project was completed on a single Windows 10 desktop PC. The system has an Intel 11600k Core i5 processor, with 6 physical cores (12 thread) and a 3.90GHz base frequency. It also has 32 GB of DDR4 3200 memory, and 1 TB of NVMe SSD storage. The GPU is an Nvidia GTX 970. Installed on the PC is RGui, which is a basic development environment for the R programming language. Also utilized for the data exploration tasks was a Docker Desktop deployment of two containers, hosting instances of Elasticsearch and Kibana. These tools were also used for generating some of the visualizations present in this paper. The paper itself and accompanying slideshow presentation were prepared in Microsoft Office.

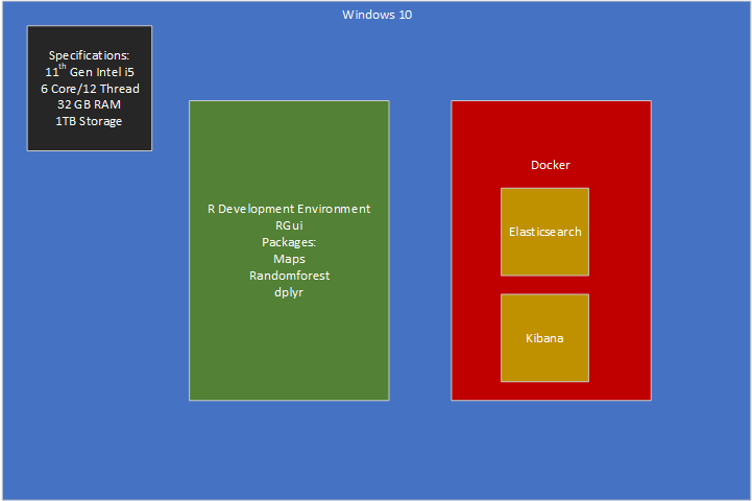


Figure 2. A representation of the computing resources used for the project

METHODOLOGY

The data was obtained using the previously mentioned export function of the Interactive Atlas of Heart Disease and Stroke. Each piece of required information (death rate, smoking, etc.) involved its own export, resulting in 5 separate csv files. Each of these files included columns for both the county names, and a county-specific identifier referred to as “FIPS Code”. These are 5-digit numerical values. I was able to merge the csv files into one dataset by matching each row on the corresponding FIPS code in the other files and appending the columns that represented a delta between the other files. This way the dataset did not contain duplicate columns for county names, etc.

With the dataset properly acquired and the relevant columns appended, I ingested the csv data into my Docker Elastic deployment and began with data exploration.

To further determine whether heart disease deaths in certain counties could be related to high risk factors in nearby areas, I decided to create a visualization showing the counties with the highest rates of death, as well as highest rates of the chosen risk factors. As seen in figures 3 and 4, heightened rates of active smoking and low median incomes seem to be the most severe in very similar areas compared to heart disease death rates.

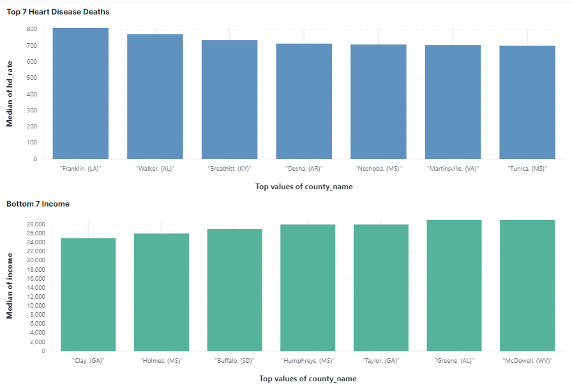


Figure 3. Top 7 counties for heart disease death rate (blue) vs. bottom 7 for median income (green).

A picture containing chart

Description automatically generated

Figure 4. Top 7 counties for heart disease death rate (blue) vs. top 7 for percentage of active smokers (green).

In fact, the counties from Mississippi and Alabama shown in these graphs are within approximately 90 minutes of each other by car (within their respective state, and according to Google Maps). However, the opposite can be said about the graphs for obesity rate and percentage over 65 years old.

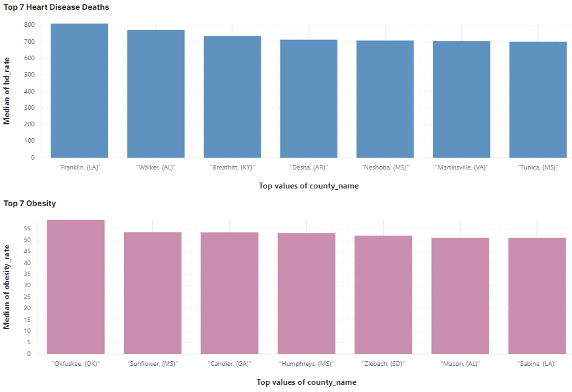


Figure 5. Top 7 counties for heart disease death rate (blue) vs. top 7 for obesity (pink).

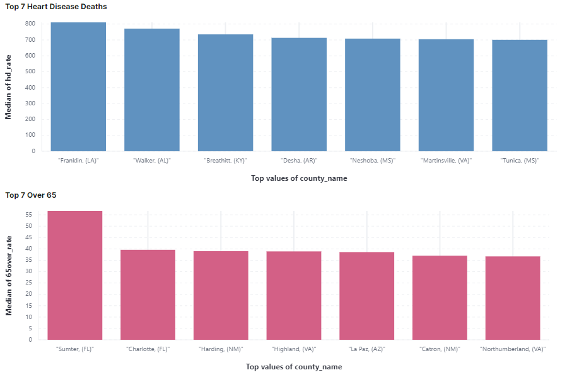


Figure 6. Top 7 counties for heart disease death rate (blue) vs. top 7 for population over 65 (salmon).

In both graphs, there are few clear similarities involving counties within the same state, or even many involving counties within neighboring states. They seem to be more spread out or most prominent in areas different than those of the death rates we are comparing to. For example, Florida is the clear leader in states with over 65 population, yet that isn’t a state that we see at all on the graph for leading death rates.

These graphs are good for an initial view into the data, but they are still only looking at a handful of counties that have the very top values of their respective features. In other words, simply using these visualization graphs to compare geographic occurrences may be over-simplified and unable to give an entirely accurate understanding. To get a more precise understanding of where the risk factors were likely to occur versus where heart disease deaths occur, I used k-means clustering to get a higher-level view of where deaths and risk factors occur.

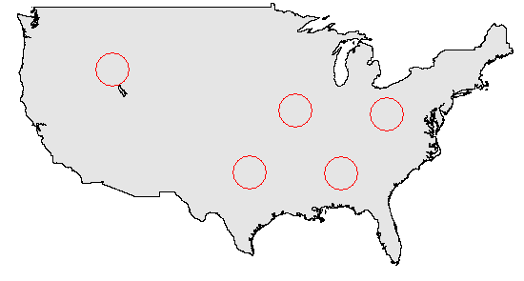


Figure 7. Clustering for high rates of heart disease deaths.

Figure 7 shows the clustering performed for high rates of heart disease death rates and will be used as reference for the cluster maps created for each risk factor.

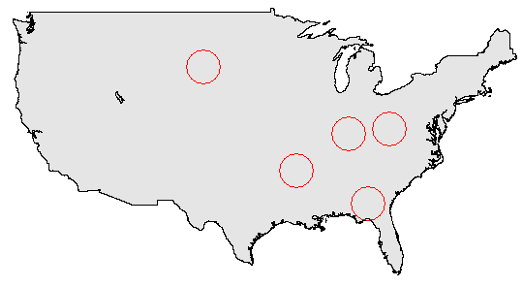


Figure 8. Clustering for high rates of active smokers.

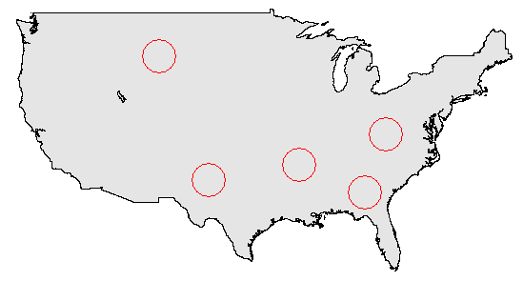


Figure 9. Clustering for lowest median incomes.

As was the case with the graph visualizations, it is immediately evident that low median incomes and high rates of smoking occur in the same geographic areas as high rates of heart disease death.

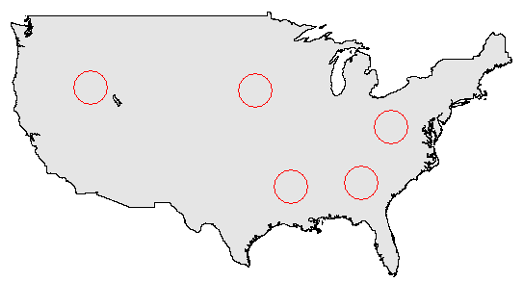


Figure 10. Clustering for high obesity rates.

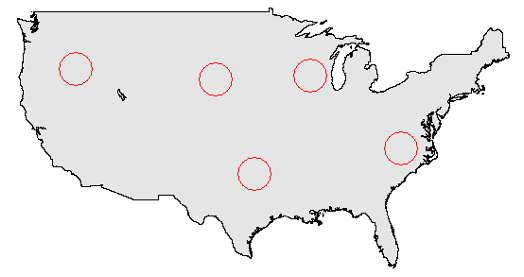


Figure 11. Clustering for high over 65 rates.

As seen in figure 10, high obesity rates also seem to occur alongside the high death rates, despite lacking similarities in the top county graphs. However, the clustering for counties with high rates of population over 65 years old continued to show a lack of relation for that risk factor to death rates geographically.

With data exploration completed, I moved forward with applying random forest classification to each risk factor individually to see which could, on its own, most accurately predict high death rate counties from test data. This involved adding a new feature for each of the 4 risk factors, with a Boolean value declaring whether each county was in the highest category for smoking, obesity, low income, or over 65.

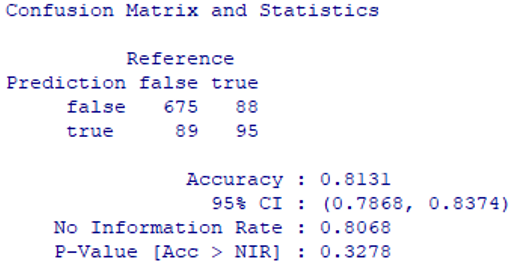


Figure 12. Random forest predict() accuracy for smoking rate.

A state’s rate of active smokers was 81.31% accurate at predicting high rates of heart disease death, as seen in figure 12.

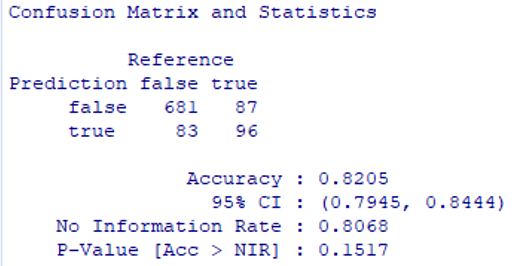


Figure 13. Random forest predict() accuracy for income.

Median income was slightly more accurate, at 82.05%.

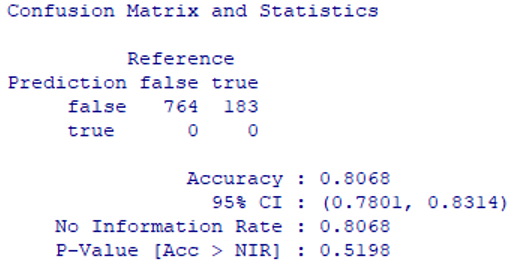


Figure 14. Random forest predict() accuracy for obesity rate.

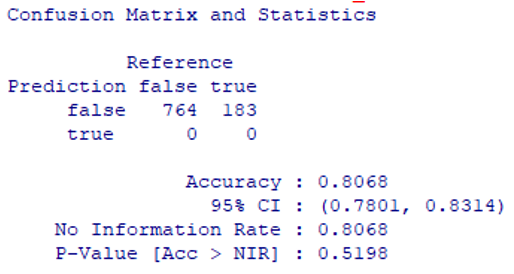


Figure 15. Random forest predict() accuracy for over 65 percentage.

Obesity rate and population over 65 were 80.68%, but neither were an effective predictor for high death rates. Both predicted ‘false’ for every county, suggesting that the training data did not show enough correlation for proper classification. 80.68% of the data contained a ‘false’ value for the high death rate feature, hence the result for these two risk factors.

With smoking rate and low median income appearing to be the only effective features for high death rate classification, the last test performed during this project was to combine the smoking and income features into a final, two-feature random forest classifier.

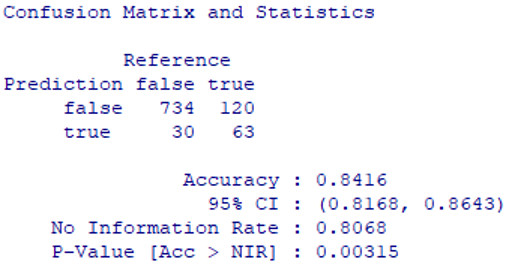


Figure 16. Random forest predict() accuracy for smoking and median income features.

When using the median income category along with county rate of active smokers, the random forest classifier was able to achieve 84.16% accuracy.

CONCLUSION

A county’s obesity rate and percentage of population over 65 years old were not effective classification features for determining high rates of heart disease deaths in the US. However, smoking rates and median income were both effective features with income being the most accurate. When combining the two, random forest classification was 84.16% accurate in determining high heart disease death rate counties.

In hindsight, there were a few improvements that could be made to the methodology in future work. The CDC has a massive amount of data available on the county level, and there may still be undiscovered features in the dataset that could be effective for classification. Also, the category limits for what is considered a ‘high’ death rate county could be tuned, so that a lower percentage of the counties are assigned ‘false’, thus making ineffective classification features less misleading.

REFERENCES

[1] 2021. Leading Causes of Death. CDC: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

[2] 2021.CDC Interactive Atlas of Heart Disease and Stroke: https://nccd.cdc.gov/DHDSPAtlas/?state=Cou nty

[3] Jabbar M.A., Deekshatulu B.L., Chandra P. (2016) Prediction of Heart Disease Using Random Forest and Feature Subset Selection. In: Snášel V., Abraham A., Krömer P., Pant M., Muda A. (eds) Innovations in Bio-Inspired Computing and Applications. Advances in Intelligent Systems and Computing, vol 424. Springer, Cham. https://doi.org/10.1007/978-3-319-28031-8\_16

[4] Singh Y.K., Sinha N., Singh S.K. (2017) Heart Disease Prediction System Using Random Forest. In: Singh M., Gupta P., Tyagi V., Sharma A., Ören T., Grosky W. (eds) Advances in Computing and Data Sciences. ICACDS 2016. Communications in Computer and Information Science, vol 721. Springer, Singapore. https://doi.org/10.1007/978-981-10-5427-3\_63

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00