CX4242 Final Report: Health Data Inequality Prediction & Visualization

Team 28: Health Hackers. Emily Adams, Monica De Alvaro Mena, Juan David Perdomo, Sriya Surapaneni, Jaxon Turner, Huy Trinh

Introduction

In the United States, healthcare is not equal for all. Minorities are more likely to die as infants, suffer diseases and disabilities, and have a shorter life expectancy [1,2]. Health inequality is estimated to cost the United States $310 billion annually, including lost opportunity costs [2]. From a humanitarian and economic perspective, it is extremely beneficial to address healthcare inequality. Using multiple data visualization techniques, artificial intelligence can identify patterns in large amounts of medical data that can then be statistically tested [3,4,5]. The landscape of integrating public health surveillance is still in its early stages — one recent study used machine learning to identify factors that predict frequent smoking in Californian youth [6,7]. If harnessed effectively, relevant surveillance data can be utilized to potentially improve the health of marginalized individuals and boost the economy.

Problem Definition

This project attempts to build an interactive tool for policymakers to identify and predict the effects of changing economic and health policies on health outcomes. County-level data that incorporates socio-demographic and socio-economic factors as predictive variables will be used to assess the effect on health outcomes such as mortality rate and chronic disease [8].

There is evidence that suggests an ongoing healthcare inequality present in today's society that affects millions of people in the United States. It has been observed that individuals and communities within a higher socio-economic status have a better life quality and longevity than those in lower statuses. There are databases with important information that can be used by health entities to make informed decisions to help stabilize healthcare across different socioeconomic statuses. However, these datasets can be very extensive and complex and they are not being used in a productive way to generate useful insights. The issue we are set to tackle is this inability to generate useful insights from the current data available. With our solution, we expect to solve this issue by providing actionable insights out of the complex datasets and thus enhancing the higher level decision makers to take better and more informed courses of action.

Literature Survey

Many studies regarding data visualization have been carried out. Austin et al. [3] and Gotz and Borland [4] emphasize the benefits and challenges of interactive data visualization in detecting patterns in whole-person health data. Liu et al.[9] introduces a new data visualization and digitization method for electronic health records, and Leung et al. [10], Groseclose and Buckeridge [6], and Menon et al.[11] describe big data analysis, visualization, and public health surveillance services for healthcare analytics. Sopan et al. [12], Polychronidou et al. [13], and Dixon et al. [14] present interactive web-based platforms for healthcare data analysis and visualization. These sources are highly relevant as we are greatly focusing on data visualization for healthcare public surveillance data, but they lack exploration of predictive modeling techniques. Our project will expand upon this as our primary goal is to ultimately integrate data visualization with predictive modeling to create a more holistic tool that not only helps people visualize healthcare data but also helps policymakers make predictions based on it.

Additionally, many recent studies are related to machine learning and predictive modeling. Fu et al. [7] describe a machine-learning approach for predicting vaping behavior, and Feng and Jiao[15] and Sivabalaselvamani et al. [5] discuss using machine learning to predict and map health outcomes, at neighborhood and individual levels respectively. Aghdaee et al. [16] explores and maps the relationship between socioeconomic status and access to healthcare. Leist et al. [17], Wichmann et al. [18], and Greenwell et al. [19] discuss improvements in machine learning algorithms for health outcomes predictions. Contrastingly to the sources above, these provide valuable insights into machine learning and predictive modeling techniques for healthcare data but don't fully address the data visualization aspects. Once again, we aim to expand on both predictive modeling and data visualization aspects with our

project, creating a more comprehensive healthcare surveillance tool that benefits the public and policymakers combined.

Many studies discuss health inequalities. Singh et al. [1] and LaVeist et al. [2] provide information on health inequalities and their economic impact, and Hosseinpoor et al. [20] discuss measuring health inequalities in the context of sustainable development goals. McMaughan et al. [8] explores the relationship between socioeconomic status and healthcare access. These sources are greatly beneficial in increasing understanding of healthcare disparities. However, they provide little to no information regarding how to analyze, visualize, or model any data related to healthcare inequalities. To improve upon this, we will integrate insights from these sources into our visualization and modeling components to provide a clearer picture of healthcare inequalities and their consequences.

Finally, there are many ethical and security considerations we need to keep in mind, which some papers describe in detail. Gotz and Borland [21] and Obenshain [22] highlight ethical and security concerns in healthcare data, emphasizing the importance of data validation processes and security measures like encryption and access restrictions. They also describe the potential pitfalls of interactive visualization tools, such as selection bias. In our project, we will work to incorporate data validation processes to ensure our data is ethically and securely handled.

Proposed Method

Innovations & approaches

Our project and its end product encompass different innovations. We will use both a learned random forest and a neural network to create predictions of health outcomes based on user input of health factor variables (independent variables). This is innovative because it simulates policy change in the health space and provides information about how the potential health outcome might change, and there is currently little to no health analytic project that allows this. For example, let's say the user wants to increase the median income of the county; then the model may tell that as we increase the median income, the rating of health outcome (ranging from 0 to 1) increases. Additionally, we can extract the weights of the learned models to help the user see which variable has the strongest impact on the health outcomes variables. These two features are critical for local to federal municipalities because they reveal underlying correlations between health factors and health outcomes, and allow simulating health outcomes in response to health factor change.

As mentioned above for the algorithm approaches, we chose to work with a Random Forest and a Neural Network. The reason why we chose to work with a Random Forest, apart from the good results it usually gives compared to simpler models, is that it will also allow us to get an insight into which of the other factors are more relevant when it comes to predicting the outcome column. This means we can analyze the main factors that cause the health problems/outcomes we are interested in analyzing to help solve the existing inequalities. Moreover, the appeal of the Neural Network lies in the large capability it has of detecting other kinds of non-linear relationships between the independent variables and the target column, allowing us to possibly get better results in our predictions, despite it being a black box where it would be hard to see which are the most important features.

Our visualization will display several data perspectives that users can navigate and input data into. The visualization innovation lies in the implementation of responsive mapping for health outcome prediction and unique dashboards for county comparison. The main user interface is centered on a choropleth map of a given US state's counties that will display historical values of a selected county's health factors and outcome, and the predicted future outcomes from user input of health factors. Additional dashboards will have a similar choropleth map and allow the user to filter on conditions and identify commonalities with other counties in a given year. Lastly, we will provide a dashboard that allows direct comparison of any two counties across different years. These innovations give lawmakers the ability to look at more nuance than our model when identifying what policies have worked in other counties and whether they would be effective in their county. To accomplish our visualization goals, we have opted to use Tableau due to its versatility, sharp visuals, and ability to connect to jupyter notebook ML models through TabPy.

Experiments/Evaluation

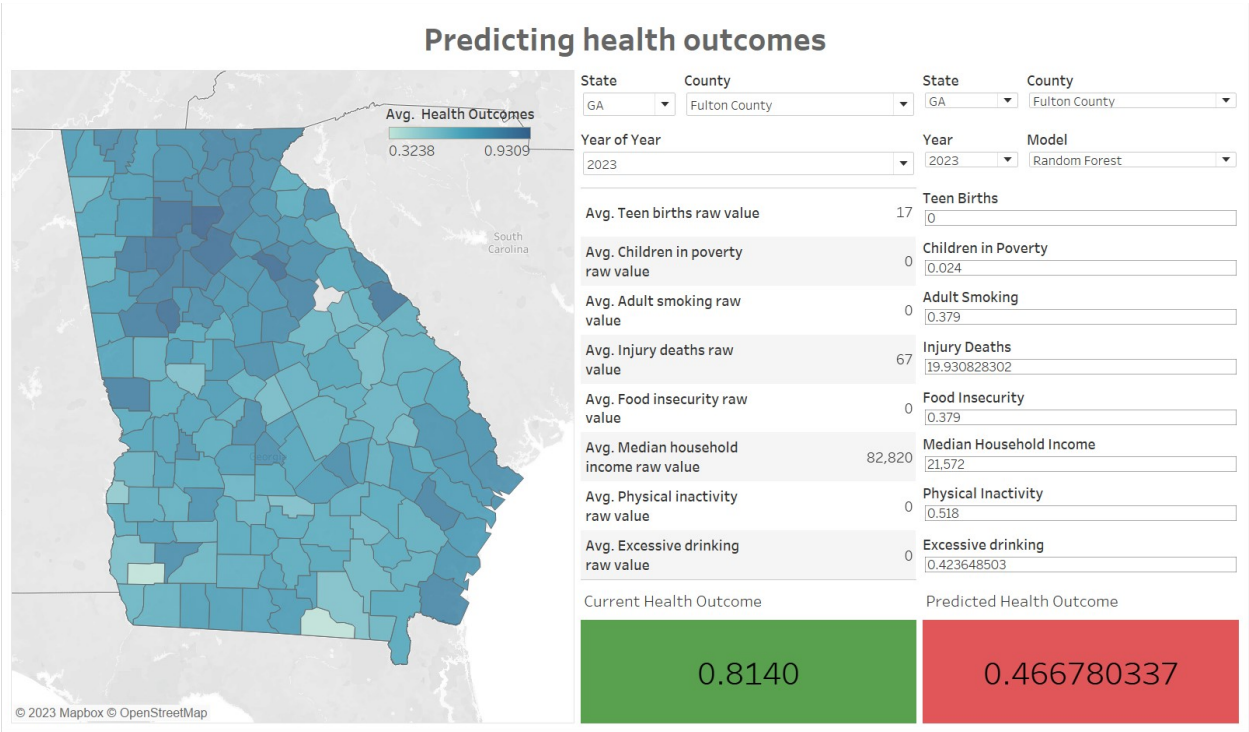
Data Cleaning and Preprocessing

For the data cleaning, we used R to join the CSV files (from source [23]) for the individual years of health data to produce a panel dataset. We kept the columns that we had in common, checking they all correspond, with the same labels and units. We got rid of the columns we couldn't use, due to missing values, as well as filled in some of the other missing values for each county with the average per state and year. The columns we want to use for predictions will be selected in the model code. The predicted column will be created in the model as a representative of health outcomes

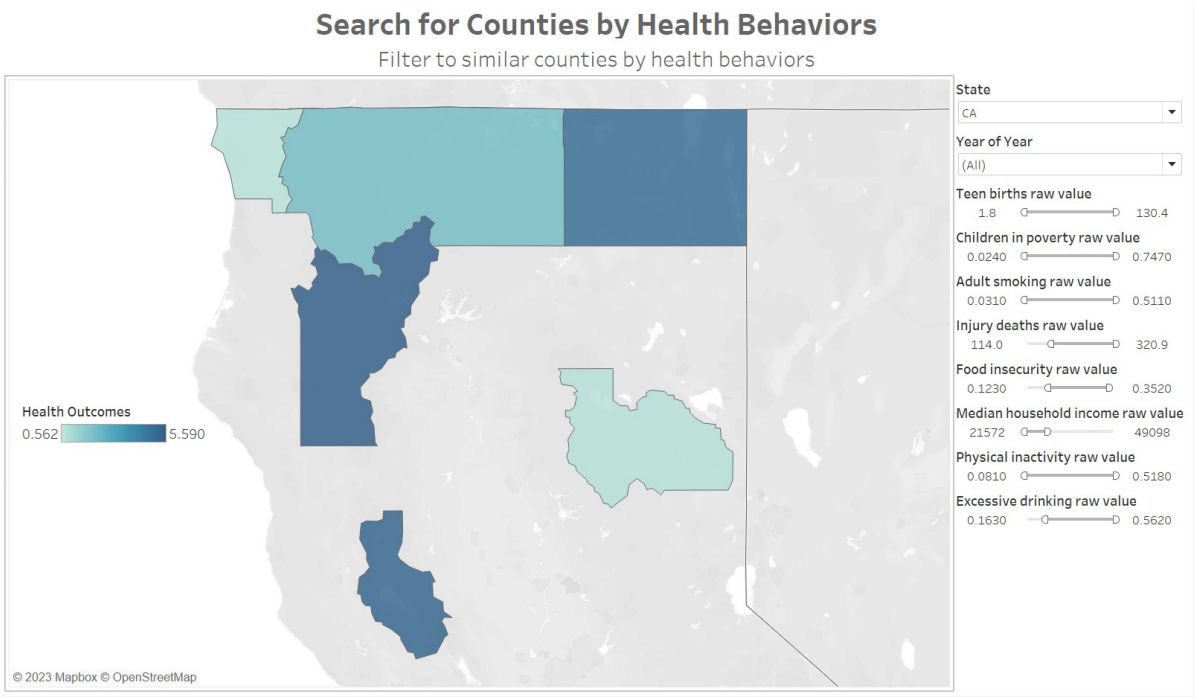
Visualization

Our visualization is produced in Tableau and connected to our model using Tabpy. TabPy allows Tableau to run Python code and external functions in a jupyter notebook file. Currently, TabPy and Jupyter Notebook are both run locally to create the visualization. However, in a future revision, we would aim to move everything to a Tableau server for easier access. Using TabPy and Tableau, we produced three dashboards that allow a wide range of functionality for the user. These dashboards allow the intended user, policymakers, to gain insights into how changing health behaviors and socioeconomic factors in the county will affect health outcomes and identify counties facing similar issues to potentially learn how policy adoption has affected health outcomes in those counties. Our visualizations span the US looking at the county, state, and national levels.

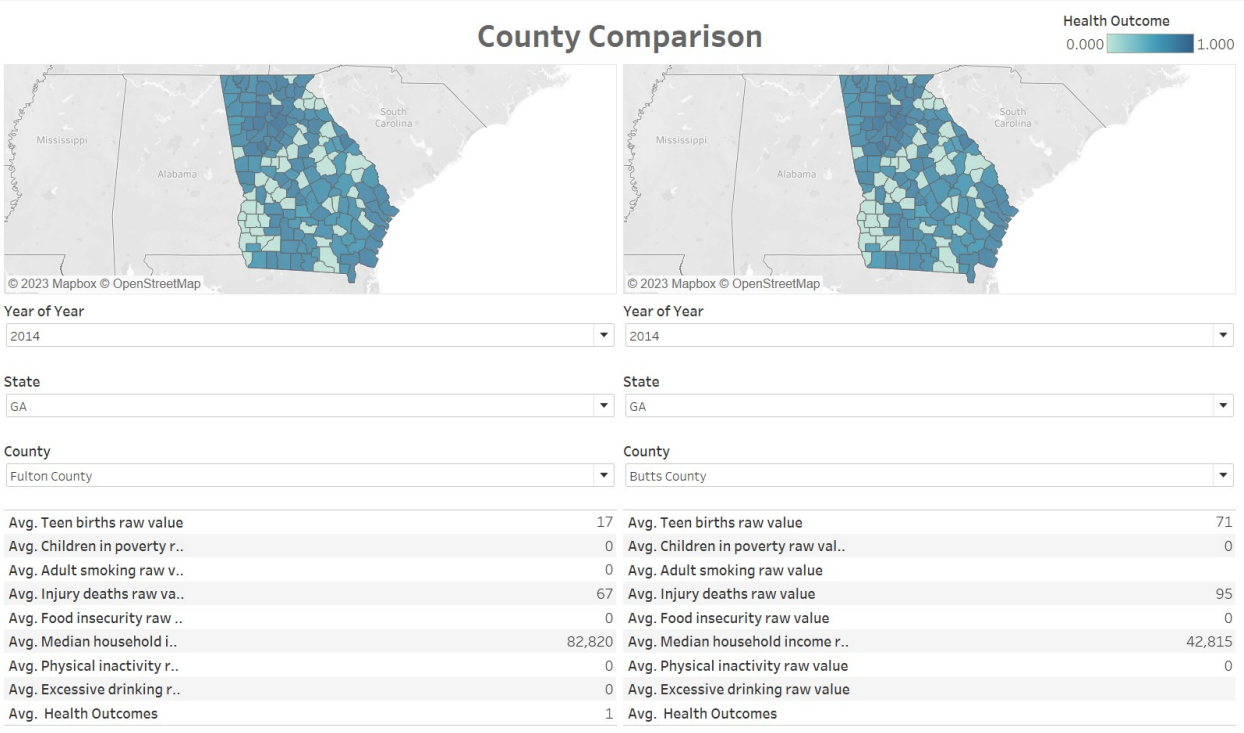
This **first dashboard** is the prediction dashboard. This is meant to help policymakers understand how if they adopted policies that change certain health behaviors in the county, we would expect the effect to be on general health outcomes. Policymakers can see the current values for each health behavior and input new values to understand how that impacts health outcomes in the bottom left corner, the red. They can see how much the health outcomes in the county would change as a result of that policy change.



This **second dashboard** helps policymakers identify counties facing similar conditions. This is intended to facilitate policy outcome observation to help understand what has and hasn't worked for counties facing similar conditions, to lead to more informed policy adoption.



The **third dashboard** allows for direct comparison between two counties to see what the health behaviors and outcomes look like for each of the counties. This can also be used to understand change over time and to compare a county to its previous self.



Model

For the **Random Forest model,** we experimented with what we wanted to do, regarding our two models based on the data from 2019 and 2023, to get a pre and post-pandemic view of the health situation. Firstly, we wanted to figure out how to properly create the output column, taking into account the health outcome variables. According to the page we got our dataset from [23], the following variables are the ones that constitute the outcome variables: Length of Life, Premature Death, Quality of Life, Poor or Fair Health, Poor Physical Health Days, Poor Mental Health Days, Low Birthweight, Length of Life, Life Expectancy, Premature Age-Adjusted Mortality, Child Mortality, Infant Mortality, Quality of Life, Frequent Physical Distress, Frequent Mental Distress, Diabetes Prevalence, HIV Prevalence. Out of this list of 17 features, we chose to work with the following 5 (scaled), in this formula to get our output column: **(Premature Age-Adjusted Mortality * 5) + (Poor or Fair Health * 1) + (Poor Physical Health Days * 1) + (Low Birthweight * 2) + (Poor Mental Health Days * 1)** based on the paper [24], this column is later scaled and flipped so that **0 represents a bad value and 1 being the best**, making it easier to understand for the visualization, given that the ranges of the columns varied.

To train our model, we got rid of all 17 features that, even if they aren't part of the formula, are still considered outcomes. We also got rid of the columns related to race, since we didn't want them to bias our result. We did some experiments without taking them out, and it turned out that one or two of the other outcome variables would always be almost perfectly correlated with the output, so we had to get rid of them. We first did some experiments as a classification problem, since we thought it might be easier for visualization purposes. We tried several values for the number of classes, but despite getting very good values for the accuracy and the OOB error, we found that the error would always be concentrated in the bins right next to the "correct" one, meaning most of the errors were from the points that were right in the border between neighboring bins. Therefore, we ended up changing our regression approach.

After tuning our **RF regression** model, we found that we could get a very accurate prediction, with an mse error of only around 0.21% (2019) and 0.24% (2023) after performing cross-validation with the best hyperparameters we got after tuning the model (we talk about percentages for the mse since the range is from 0 to 1, so it is easy to interpret it). After discovering we needed to get rid of all those outcome columns, we learned that we could get the insights we were looking for from our models. Looking at the most important features list, we got the variables that properly defined and caused that final health outcome indicator. In 2019 we found that the most important factors were Children in Poverty, Adult Smoking, Excessive Drinking, Median Household Income, Teen Births, Insufficient sleep, Children eligible for free or reduced-price lunch, and Injury Deaths. Similarly, for 2023 we got: Physical Inactivity, Median Household Income, Children in Poverty, Food Insecurity, Insufficient Sleep, Injury Deaths, Homicides, and Teen Births. Both lists are in order, with the 8 most important features. These are the results we were looking for. We wanted to gain insights into what factors caused all those health issues that are part of the outcome variable list mentioned before. For visualization purposes we wanted to get some significant parameters users could modify to see how this would change the predictions, therefore we chose 8 common to both that would have visible weight on the predictions, modifying the result, we went with the following: Physical Inactivity, Children in Poverty, Adult Smoking, Median Household Income, Injury Deaths, Food Insecurity, Teen Births, and Excessive Drinking. We will then give our users the possibility to choose some values for those parameters among a certain range so they can see how that will affect the health outcome.

In addition to the random forest, we also implement a **multilayered neural network** that performs regression to predict the health outcome score. Even though neural networks are known to underperform in tabulated data, we found that our neural network did a very good job in predicting the outcome index. We built two different neural networks: one for data in 2019 and one for 2023; both have an MSE of roughly 0.2 for the test case, implying a very small amount of error between predicted and actual values. The 2019-data neural network is shown as the following layered setup: 144x64x32x1 and the 2023-data neural network has 113x64x32x1. However, the model has the ReLu activation function for the hidden layers and the linear activation function for the singular output neuron.

Though the neural network provided great results, unfortunately, we could not incorporate it into our visualization due to limitations with TabPy and time constraints. TabPy is unable to use key functions from the numpy and pandas libraries. The model prediction function was able to run perfectly in Jupyter Notebook when called but was unable to run or returned wildly different results in Tableau when given the same input parameters. TabPy cannot process the "numpy.float" data type, which egregiously inhibits the implementation of the neural network into the visualization. With additional time, we would pursue a different method, to create a new neural network compatible with TabPy.

Conclusions and Discussion

To reiterate, with this project, we aim to address healthcare inequalities in the US through the integration of data visualization and machine learning techniques. Our goal is to empower policymakers to make more informed decisions by simulating the impact of various health factors at the county level on overall health outcomes. Utilizing our Random Forest model and the Dashboards, we aspire to help policymakers and other stakeholders comprehend how principal health determinants influence regional health. By analyzing these results and employing our visualization tools, we facilitate a deeper understanding of the relationships between these factors and health outcomes, thereby enabling the execution of more effective health policies aimed at mitigating existing health inequalities.

It's important to acknowledge certain limitations: the reliance on the quality and availability of county-level data, which may affect model accuracy; the challenges in model interpretability, especially with complex algorithms like neural networks; and ethical and privacy concerns in handling sensitive health data. Additionally, the generalizability of the model's results beyond the specific context of US counties may be limited. Recognizing these limitations is crucial for a realistic assessment of our project’s capabilities and for future enhancements.

All team members have contributed a similar amount of effort.

Bibliography

Literature Review Incorporated in Report

[1] G. K. Singh et al., “Social Determinants of Health in the United States: Addressing Major Health Inequality Trends for the Nation, 1935-2016,” Int. J. MCH AIDS, vol. 6, no. 2, pp. 139–164, 2017, doi: 10.21106/ijma.236.

[2] T. A. LaVeist, D. J. Gaskin, and P. Richard, “The Economic Burden of Health Inequalities in the United States”, [Online]. Available: https://hsrc.himmelfarb.gwu.edu/cgi/viewcontent.cgi?article=1224&context=sphhs_policy_facpubs

[3] R. R. Austin, M. A. Mathiason, and K. A. Monsen, “Using data visualization to detect patterns in whole-person health data,” Res. Nurs. Health, vol. 45, no. 4, pp. 466–476, 2022, doi: 10.1002/nur.22248.

[4] D. Gotz and D. Borland, “Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization,” IEEE Comput. Graph. Appl., vol. 36, no. 3, pp. 90–96, May 2016, doi: 10.1109/MCG.2016.59.

[5] D. Sivabalaselvamani, D. Selvakarthi, J. Yogapriya, M. P. Thiruvenkatasuresh, M. Maruthappa, and A. S. Chandra, “Artificial Intelligence in Data-Driven Analytics for the Personalized Healthcare,” in 2021 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2021, pp. 1–5. doi: 10.1109/ICCCI50826.2021.9402703.

[6] S. L. Groseclose and D. L. Buckeridge, “Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation,” Annu. Rev. Public Health, vol. 38, no. 1, pp. 57–79, 2017, doi: 10.1146/annurev-publhealth-031816-044348.

[7] R. Fu, J. Shi, M. Chaiton, A. M. Leventhal, J. B. Unger, and J. L. Barrington-Trimis, “A Machine Learning Approach to Identify Predictors of Frequent Vaping and Vulnerable Californian Youth Subgroups,” Nicotine Tob. Res., vol. 24, no. 7, pp. 1028–1036, Jul. 2022, doi: 10.1093/ntr/ntab257.

[8] D. J. McMaughan, O. Oloruntoba, and M. L. Smith, “Socioeconomic Status and Access to Healthcare: Interrelated Drivers for Healthy Aging,” Front. Public Health, vol. 8, p. 231, Jun. 2020, doi: 10.3389/fpubh.2020.00231.

[9] N. Liu et al., A New Data Visualization and Digitization Method for Building Electronic Health Record. 2020, p. 2982. doi: 10.1109/BIBM49941.2020.9313116.

[10] C. K. Leung, Y. Zhang, C. S. H. Hoi, J. Souza, and B. H. Wodi, “Big Data Analysis and Services: Visualization on Smart Data to Support Healthcare Analytics,” in 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Jul. 2019, pp. 1261–1268. doi 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00212.

[11] A. Menon, A. M. S, A. Maria Joykutty, A. Y. Av, and A. Y. Av, “Data Visualization and Predictive Analysis for Smart Healthcare: Tool for a Hospital,” in 2021 IEEE Region 10 Symposium (TENSYMP), Aug. 2021, pp. 1–8. doi: 10.1109/TENSYMP52854.2021.9550822.

[12] A. Sopan, A. S.-I. Noh, S. Karol, P. Rosenfeld, G. Lee, and B. Shneiderman, “Community Health Map: A geospatial and multivariate data visualization tool for public health datasets,” Gov. Inf. Q., vol. 29, no. 2, pp. 223–234, 2012, doi: https://doi.org/10.1016/j.giq.2011.10.002.

[13] E. Polychronidou, I. Kalamaras, K. Votis, and D. Tzovaras, “Health Vision: An Interactive Web-based platform for healthcare data analysis and visualization,” in 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Jul. 2019, pp. 1–8. doi: 10.1109/CIBCB.2019.8791462.

[14] B. E. Dixon, S. J. Grannis, U. Tachinardi, J. L. Williams, C. McAndrews, and P. J. Embí, “Daily Visualization of Statewide COVID-19 Healthcare Data,” in 2020 Workshop on Visual Analytics in Healthcare (VAHC), 2020, pp. 1–3. doi: 10.1109/VAHC53729.2020.00007.

[15] C. Feng and J. Jiao, “Predicting and mapping neighborhood-scale health outcomes: A machine learning approach,” Comput. Environ. Urban Syst., vol. 85, p. 101562, Jan. 2021, doi: 10.1016/j.compenvurbsys.2020.101562.

[16] M. Aghdaee et al., “An examination of machine learning to map non‐preference based patient-reported outcome measures to health state utility values,” Health Econ., vol. 31, no. 8, pp. 1525–1557, Aug. 2022, doi: 10.1002/hec.4503.

[17] A. K. Leist, M. Klee, J. H. Kim, D. H. Rehkopf, S. P. A. Bordas, and S. Wade, “Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences,” Sci. Adv., vol. 8, no. 42, p. eabk1942, 2022, doi: 10.1126/sciadv.abk1942.

[18] R. M. Wichmann, F. T. Fernandes, and A. D. P. Chiavegatto Filho, “Improving the performance of machine learning algorithms for health outcomes predictions in multicentric cohorts,” Sci. Rep., vol. 13, no. 1, Art. no. 1, Jan. 2023, doi: 10.1038/s41598-022-26467-6.

[19] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, “A Simple and Effective Model-Based Variable Importance Measure.” arXiv, May 12, 2018. Accessed: Oct. 11, 2023. [Online]. Available: http://arxiv.org/abs/1805.04755

[20] A. R. Hosseinpoor, N. Bergen, A. Schlotheuber, and J. Grove, “Measuring health inequalities in the context of sustainable development goals,” Bull. World Health Organ., vol. 96, no. 9, pp. 654–659, Sep. 2018, doi: 10.2471/BLT.18.210401.

[21] D. Gotz and D. Borland, "Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization," IEEE, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7466736

[22] M. K. Obenshain, “Application of Data Mining Techniques to Healthcare Data,” Infect. Control Amp Hosp. Epidemiol., vol. 25, no. 8, pp. 690–695, 2004, doi: 10.1086/502460.

[23] County Health Rankings & Roadmaps, “Rankings Data Documentation 2010-2019,” County Health Rankings & Roadmaps, [Online]. Available: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019.

[24] “Explore Health Rankings | Rankings Data & Documentation,” County Health Rankings & Roadmaps. Accessed: Nov. 03, 2023. [Online]. Available: https://www.countyhealthrankings.org/sites/default/files/media/document/CHR%20Ranked%20Measures%202010-2023.pdf