Project detail and status update

Purpose

This document is the final report that project team134 completed. This document explains a. the methods used for preparing the data b. project goal and explanation for how the idea stands out c. design of the project, and e. project published implementation.

Recap

The Behavioral Risk Factor Surveillance System (BRFSS), administered by Center for Disease Control (CDC), is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. The data used in our project is a result of the BRFSS 2021 survey.

Project detail

Project Goal and Uniqueness

A large 30 factor 438,693 observation dataset which captured the results of a health survey conducted by the CDC on individuals in the United States in 2021. The survey has 303 unique questions with mostly categorical answers to the questions and cover's every aspect of a participant's life data – societal, lifestyle, demographics, economical, food habit, health care practices, past medical conditions, and upbringings.

Researchers and data enthusiasts have conducted plenty of studies and have produced models in predicting the possibility of CHD and primary factors contributing to it. By design these models choose the statistically significant factors which can explain – say more than 85% of both negative and positive CHD cases. Readers of such reports, who are concerned about controlling CHD contributing factors but also would like to continue other habits with certainty, often do not get insight of such factors that were consciously ruled out. For example, the study may conclude eating red meat regularly causes higher probability of CHD, however, it does not mention the impact of regularly eating white meat, which at this point is subject to one's further diligence to find out. Another area where these studies lack is failing to bring a comparative study among these secondary factors which can be of interest for many people who want to consider them because either primary factors are not applicable or already under control.

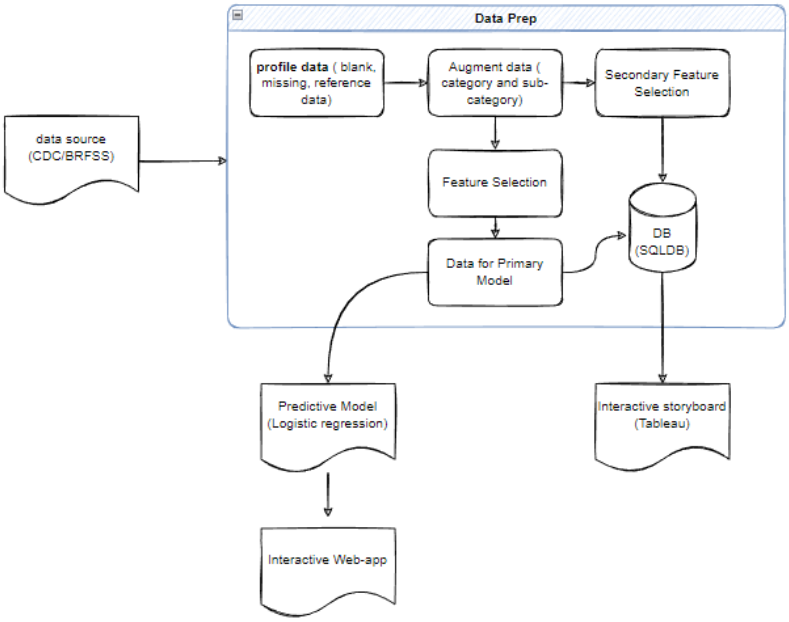
Below are the goals of our project –

1. Develop a predictive model for CHD
2. Determine and develop visualization of the *secondary factors* of CHD
3. Create Interactive visualization of all factors and factor categories. Show inter-category comparison on CHD
4. Likelihood of contracting diseases due to comorbidities
5. Build an interactive web-application for finding out what-if probabilities of CHD

Methodologies

High level design and flow

Primary and Secondary Indicator effect on CHD - Team134



Data Prep

Data Load and Transformations

The health survey data sets were downloaded from CDC Web Site as a SAS Transport File (XPT). In Python, the XPT was loaded into a data frame and saved as a CSV. A Postgres database was created to load the raw csv data as text data types and then transformed into a table with all answers to the survey questions as columns with a numeric Postgres datatype. With the survey data persisted and structured appropriately in a database, data dictionary tables were constructed for each survey question/response variable and the possible numeric response values. Some basic count statistics were run on each variable and persisted in the database. Based on the number of distinct response values for a given survey question, each variable was labelled as categorical or continuous which was determined by the number of distinct values being above or below 20. These reference tables would subsequently be extracted into Excel so the team could provide additional labelling and manual merge of the CDC's Code Book's relevant survey meta data.

Exploratory Data Analysis and Labelling

With the data loaded and prepared in the Postgres database, the variable and variable value data was extracted for the project team to further categorize the data with relevant labels. 13 high level categories were identified with the largest being health condition which identifies the survey participant as having a particular health condition such as diabetes, coronary heart disease (CHD), and cancer. These high-level categories were then expanded into subcategory labels. 68 subcategories were identified and associated with the high-level parent category.

Each survey response categorical variable was tagged with a positive, negative, neutral, or NA if the context of the survey question was applicable. For example, the health condition category question had a 1-Yes reply to a question "Ever Told Cholesterol Is High" then that would have a positive response label.

Data cleanup

The health survey data set contained many columns that would not be consumed by our visualizations or modeling. Hence, we removed those columns to improve performance for queries and running any modeling algorithms. For example, there were a dozen columns that captured many aspects for how the survey interview was conducted and that was not of any concern in the model construction or data visualizations. Also, there were several stratification columns

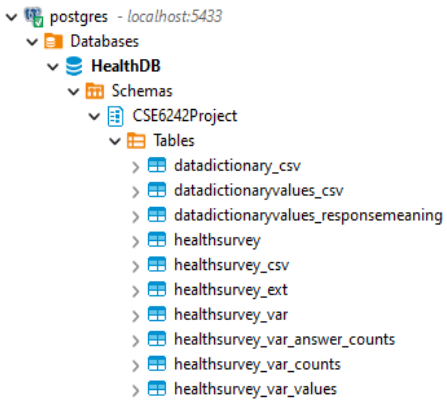
such as age of the participant with columns that stratified the age into different age band intervals. In those instances, we retained the most granular column (i.e., age) instead of the stratified columns. If we needed to stratify, we would do so in the visualizations.

Data Augmentation

With all the survey questions categorized and responses normalized, new columns were added to the health survey dataset with a binary label of 1-true/positive or 0-false/negative for all relevant survey questions. In some instances, the response of multiple survey question responses would yield a positive or negative response in a single column. For example, the survey question "Have you ever been told you had diabetes" has 2 yes responses; 1-Yes and 2- Yes, but female told only during pregnancy. For the new column health condition of diabetes, hc_diabetes, the multiple yes responses would normalize to a binary 1-True/Pos as the visualizations are only concerned that the participant has health condition of diabetes or not. The use of binary factor variables would also allow for building the prediction model more effectively since both the CHD response variable and these new augmented variables are both binary.

Data Storage

PostgresSQL is used as the data storage and management technology. DBeaver is being used as the IDE for managing the Postgres database and authoring tool for the SQL scripts. Scripts were created to manage the import and export of data to and from csv/xls files and for building the data dictionary including some basic statistics on the data set variables. The schema tables are as follows –



Model development

Feature Selection for Primary model

We started with splitting the data into training and testing data sets. Using the data dictionary of the CDC file, we have categorized the predictors on the file as Categorical and continuous variables. We have imputed the missing data for categorical variables with a value of 999.0 (we initially tried using 'missing', but sklearn module is throwing errors, so we ended up imputing with 999.0). For continuous variables, we have imputed with 0 value. Then we encoded all the categorical variables using Label Encoder from sklearn module. The data is split into training (70%) and testing (30%) sets. Removed predictors that are correlated with each other (leaving one of them). Using sklearn library, we performed chi square test and selected the predictors that has a p-value < 0.05 (95% confidence interval). We have a lot of predictors (114) that have a p-value less than 0.05. So, we have used SelectKBest option to select the best 15 predictors. Kept only these best columns in both training and testing sets.

Primary Predictive model

Used Logistic Regression from sklearn to create a model with 50 iterations and fit it for training dataset. We have tested the model with the test set and validated using accuracy score and confusion matrix. Here are the model metrics:

Accuracy - measures the proportion of values the model classified correctly over the total number of data points tested. Accuracy for the model is 93.68%

Precision - measures how good the model is when the prediction is positive. Precision of the model is 94.69%

Recall - measures how good our model is at correctly predicting positive classes. Recall for the model is 99.88%

Metrics show that the model's performance is particularly good and there is no need to explore any other models. We have stored the model in a pickle file so that it can be used for prediction in the next steps. Below are score and accuracy of the model along with confusion matrix:

```
confusion matrix
[[     0   6989      0      0]
 [     0 123355      0      0]
 [     0   1172      0      0]
 [     0     92      0      0]]
Model Score(training data): 93.84173008600281%
Model Score(test data): 93.7291046137013%
Accuracy=93.7291046137013%
```

Secondary Feature Selection

Leveraging feature selection methodology, the primary model retains statistically significant features which can explain an acceptable variance range of response variable – between 80% and 90%. As a result, the rest of the features are ignored as their effect on response variables is deemed not of significance. However, as this data has potential impact on life threatening conditions, our project goes a little beyond and explores the meaning of these secondary attributes in addition to the primary ones. Applying our subject matter expertise and running further feature selection technique, we choose the second set of predictors to explore their and associated categories' degree of impact on CHD using different appropriate visualizations. We hope this will offer more awareness of such well-studied factors on the disease.

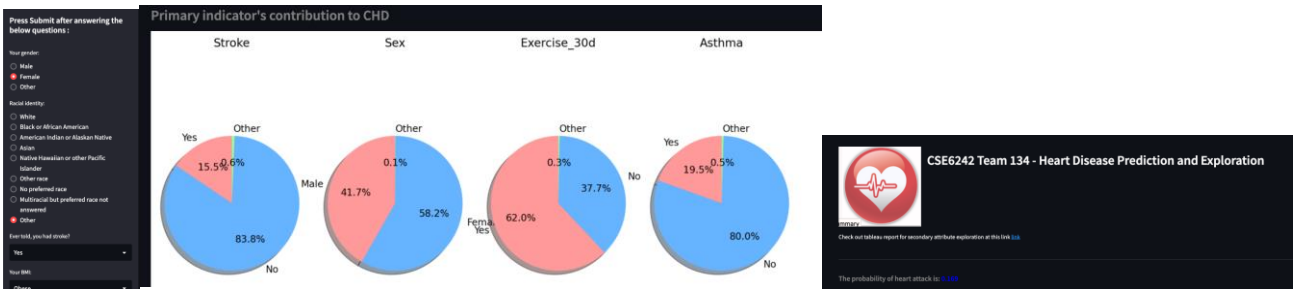
Prediction webapp and data visualization

Prediction and visualization app

This is an interactive app where users can enter their health-related data and find out their chances of having coronary heart disease. Additionally, this app offers charts for all indicators showing how those indicator values contribute to heart attack, such as "Within gender attribute 55% are male out of all surveyed adults".

The predictive logistic model is stored in an AWS s3 bucket. Once certain values are entered the app executes the model to generate probability of having CHD. However, for the reports, the heart disease data is stored in a S3 bucket and uses it to generate the charts. The app is developed in python and leveraging primarily pandas and streamlit and can be accessed here.

Input panel Visualization Prediction results

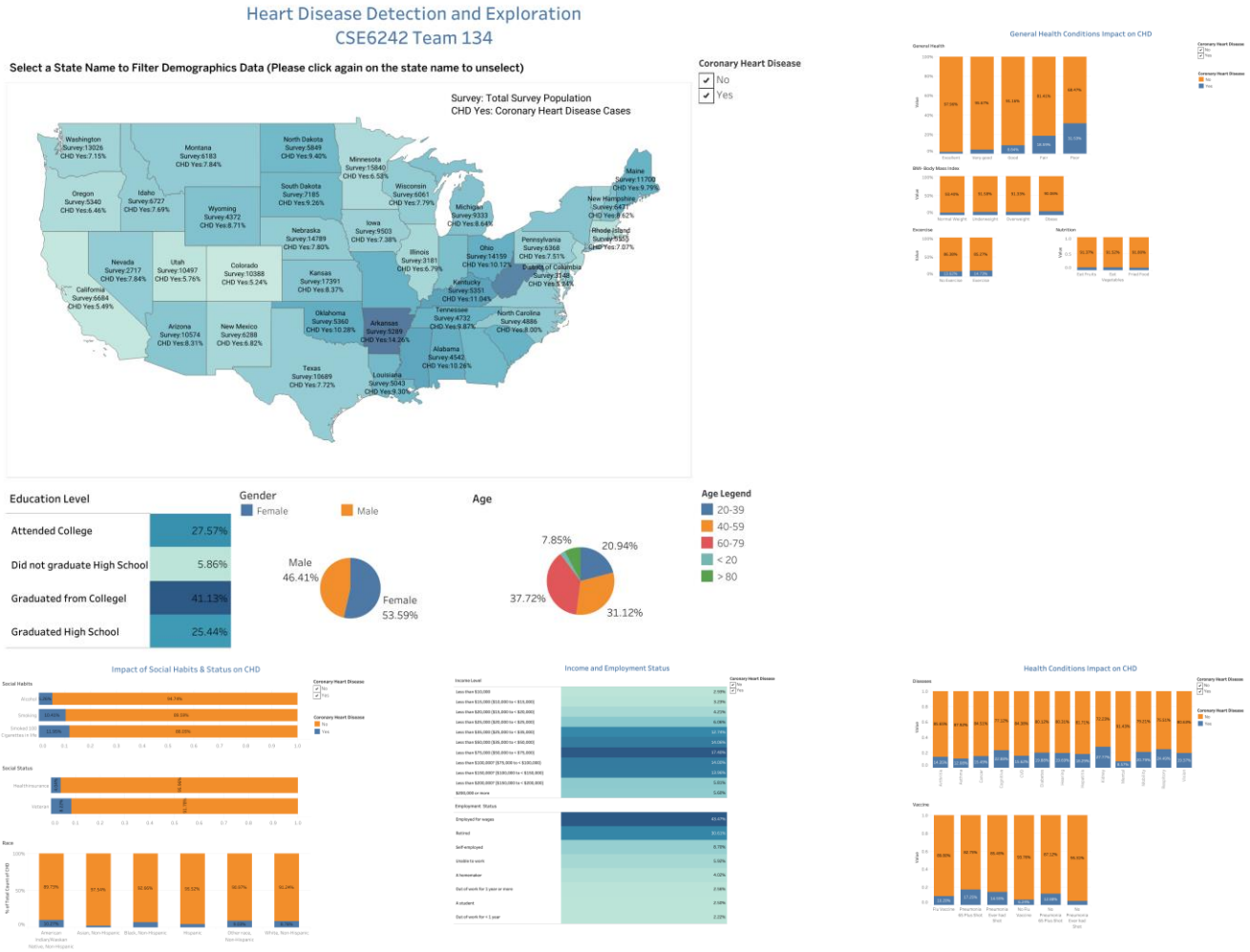


Visualization

We have used Tableau to create visualizations for this project. We will be demonstrating data patterns for numerous factors and factor categories such as health condition, social habits, social status, current treatment, demographic, and their impact on the outcome variable CHD. We provided the ability to filter various attributes and generate interactive graphs including heat map, bar chart, pie chart and bubble chart to display the data trends.

The tableau dashboards that were developed for data exploration of the primary and the secondary factors are available here. The dashboard featuring the primary factors starts off with the state wise representation of the CHD cases. It also represents the distribution of the cases with respect to the Health Conditions of the people who have taken then survey.

The landing page includes distribution of the cases gender-wise, Age bucket wise and educational level. The secondary factors were represented based on the categories that were added during the data exploration. Income and Employment Status, Health Conditions, General health, Social Habits & Social Status details entered during the survey responses.



Conclusions, Observations and Next Steps:

Observations from Visualizations:

- ◊ Arkansas, Virginia, Kentucky, Oklahoma, Ohio, Mississippi, Alabama, South Carolina, Tennessee, and Maine are the Top 10 states with the highest number of CHD cases. These cases are almost equally distributed in Male and Female populations and are seen among age groups >40.
- ◊ Cases are mostly seen among income levels of 50k – 150k and Employment Status of Employed for wages and Retired categories.
- ◊ Kidney, Respiratory, Cognitive, Mobility, Diabetes and Hearing related diseases. Mental related diseases have the lowest rate of CHDs.
- ◊ Poor General Health conditions also are causes for CHD
- ◊ Normal Weight, Overweight or Obese, Exercise & No-Exercise and Eating Fruits, Vegetables and Fried Foods do not affect your chances of having CHD.
- ◊ Drinking alcohol has much lesser impact than smoking.

Learning from the Predictive Model:

- ◊ Out of the 300+ survey questions, we found almost 130+ attributes are highly correlated and redundant for the modelling. This underscores the fact that many of our habitual attributes are tied either by cause and effect or has similar effect on heart health.
- ◊ The 13 primary factors have huge, combined effect on coronary heart disease probability while rest add marginal risks as found from elbow method.
- ◊ The result of the model was predicting that 13 primary factors to be the most influencing ones, however from the data exploration we have observed that in addition to the model output- income level, social habits, social status, education background, nutrition and underlying health conditions could also be contributing to the disease.
- ◊ The secondary features, that are further categorized into different buckets, should be the next set of habits or lifestyles that one can go after to reduce the CHD risk, or one may choose to indulge in as those factors add only marginal risk.

Next Steps:

- ◊ As a next step, along with the primary factors, we would want the researchers to consider the other secondary factors that could be managed over a period can lower the chances. For example, the income level >150k seems to be less in population around 10% including >250k, but more analysis can be done on their Nutrition and Social habits to reduce any future chances of having the disease.
- ◊ As our project offers a unique look at CHD data, an interesting find would be if combination of effects reduces the CHD risk. For example, while smoking itself can be leading cause for CHD but would smoking, regular health checkup and maintaining good BMI reduces the chance overall?
- ◊ Work with subject matter expert to make the feature selection more ‘daily-life’ friendly and more regionalized – such as while drinking is common in western countries and smoking is middle eastern countries, can we have different model per geography so that ‘ dos and don’t’s ’ are more easier to follow.
- ◊ Also, we would like to link the prediction app and the dashboard where we would like to enhance the web app to add in the secondary features that can be altered to reduce the chances of having the CHD.
- ◊ In the prediction app, introduce recommendation engine to curate real time studies and researches from web to user’s feed based on the ‘top 3 ‘ most detrimental options they chose about their health.

Team Member Effort Distribution

All team members have contributed a similar amount of effort.

References

i Personal Key Indicators of Heart Disease
https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download
ii Prediction of Coronary Heart Disease Using Risk Factor Categories. Peter WF Wilson, Ralph B Dr'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, William B Kenne. May 12 1998. Circulation. 1998;97:1837–1847
iii Prediction of Coronary Heart Disease Using Risk Factor Categories. Peter WF Wilson, Ralph B Dr'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, William B Kenne. May 12 1998. Circulation. 1998;97:1837–1847
iv Gender/Sex as a Social Determinant of Cardiovascular Risk.
https://pubmed.ncbi.nlm.nih.gov/29459471/
v Sex/Gender Differences in Cardiovascular Disease Prevention What a Difference a Decade Makes.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3362050/
vi https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3028955/
vii https://link.springer.com/article/10.1007/s00125-014-3260-6
viii https://pubmed.ncbi.nlm.nih.gov/26567979/
ix https://www.aad.org/media/stats-skin-cancer
x https://www.nature.com/articles/s41467-020-15639-5
xi https://www.niddk.nih.gov/health-information/kidney-disease/heart-disease
xii https://www.cdc.gov/kidneydisease/publications-resources/ckd-nationalfacts.
html#:~:text=CKD%20is%20more%20common%20in,Hispanic%20Asian%20adults%20(13%25)
xiii High-Value Use Cases for Predictive Analytics in Healthcare
https://healthitanalytics.com/news/10-high-value-use-cases-for-predictive-analytics-in-healthcare