Behavioral Risk Factors

Amit Sharma
CMPE Software Engineering
San Jose State University
San Jose, USA
amit.sharma02@sjsu.edu

Swapnil Avinash Parihar
CMPE Software Engineering
San Jose State University
San Jose, USA
swapnilavinash.parihar@sjsu.edu

Pranav Karmalkar
CMPE Software Engineering
San Jose State University
San Jose, USA
pranav.karmalkar@sjsu.edu

***Abstract*—This paper identifies how an individual's features, day-to-day activities, and existing medical conditions can impact their probability of contracting different diseases. An individual is nothing but the sum of his actions. The activities that we do in our day-to-day life, economic standing, gender, and age impact our health and personality. With the help of the BRFSS dataset provided by the Centers for Disease Control and Prevention (CDC), the project explores the correlation between an individual's attributes and their diseases. The paper aims to develop data-driven insights to assist individuals in devising effective and accurate long-term strategies to mitigate the probability of them contracting a particular disease. On a large scale, these insights can help policymakers and communities develop effective policies for all individuals. The paper selects cardiovascular diseases, diabetes, and mental health disorders as the target variables and all other attributes as training data for this classification problem. It proposes different statistical and data-dependent dimensionality reduction techniques for faster model generation. The paper dives deep into many of the pre-processing techniques frequently used and their consequences on the final prediction. It compares multiple classification algorithms best suited for the problem and generates a list of diseases and key-attribute pairs. The paper also gives interesting insights about how various health issues are correlated to various types of categories of people in the dataset. It analyzes the relationship between various features for a particular health issue.**

***Index Terms*—Data Science for Social Good, Centers for Disease Control and Prevention, Behavioral Risk Factor Surveillance System, Dimensionality reduction strategies**

I. Introduction

In medicine, diseases are classified into four broad categories: infectious diseases, deficiency diseases, hereditary diseases, and physiological diseases. Apart from hereditary diseases, almost all other diseases have an environmental factor attached to them. For example, known causes of deficiency diseases are improper eating habits and lack of a particular vitamin in a person's diet. Even the side effects of hereditary diseases like diabetes can be mitigated by including a few essential activities in our daily routine and a proper diet. To this effect, the Behavioral Risk Factor Surveillance System (BRFSS) collects uniform, state-specific data on preventive health practices and risk behaviors linked to chronic diseases, injuries, and preventable infectious diseases in the adult population. BRFSS information includes but is not limited to tobacco use, health care coverage, HIV/AIDS knowledge or prevention, physical activity, and fruit and vegetable consumption. This system collects information of more than 400,000 individuals selected randomly from all the states in the U.S. each year over a telephonic survey.

The dataset serves as an excellent source of information for classification algorithms, which can, in turn, determine if an individual given a set of his attributes has a particular disease or not. The classification algorithm can help generate a list of habits that individuals can avoid to improve their chances of not contracting a disease. This system provides us a unique insight into identifying individuals with these diseases and creating policies that can improve their daily living standards. Another motivation behind this paper is to help policymakers develop data-driven policies to combat preventable infectious diseases by understanding the factors that cause such diseases. In this new age, where policymakers have to present the country's politicians and citizens with vital facts that support the policy change they recommend, this dataset can provide them the required statistical data to support their argument. With such a robust dataset, the paper's objective is to extract these insights from the dataset using data mining and large scale analytic techniques from the world of machine learning.

The process of shortlisting the target disease from the large dataset involved identifying the acute diseases that are currently on the rise, and the result generated in this paper can help policymakers make informed decisions. As per Wikipedia, "In developed countries, the diseases that cause the most sickness overall are neuropsychiatric conditions, such as depression and anxiety."[2]. In the current COVID pandemic situation, many individuals stay away from their loved ones and cannot cope with stress. Relaxing activities like going out with friends and camping are also not recommended as they might increase the virus's spread. Thus, mental health-related issues are one of the top growing concerns in the healthcare industry. Additionally, it is not freely possible to practice basic methods of mitigation of the issues like group therapy. By selecting mental health-related issues as the target column, the project aims to identify key features that impact an individual's mental health. These features would help policymakers and NGOs identify individuals most susceptible to mental health-related issues.

Diabetes is one of the most significant health issues that America faces. Vulnerable people need to be identified and informed. They need to be advised about lifestyle changes

Special thanks for Professor Dr. Jorjeta Jetcheva for her continous encouragement and support.

and workout habits to be maintained to keep their blood sugar regular. The policymakers must keep the number of people affected with diabetes low as it indirectly leads to loss of economy, unnecessary burden on the healthcare system, and most importantly, puts individuals at risk of other problems like heart diseases and gangrene.

Next, we take a brief look at the previous work on this dataset and how Section II dives into the system design implementation details. Section III presents various experiments done on the data and shines a light on the different data pre-processing steps involved during the model generation. Section IV discusses the results of the paper and concludes all the findings.

II. Related Work

BRFSS conducts this telephonic survey every year, and the data for the years 2011-2015 is public almost three years back. Many individuals have tried to apply diverse strategies to extract information from this data, but they haven't found much success due to the dataset's unique nature. Not only is the dataset relatively sparse and skewed, but it also contains many derived columns. All previous work manually selected columns related to the targeted disease and performed a binary classification. This approach suffers from one serious drawback commonly referred to as implicit bias. The experimenter makes assumptions based on his mental models and personal experiences that may not generally apply. To this effect, we automated choosing the relevant columns using statistical analysis and classifiers to identify essential features from the selected dataset. Another approach often used is the approximation of input data in a particular column. Most columns in the dataset have at least these five values 0-1 for yes-no,77-99 for 'don't know-refused,' and a few nan values. Most previous works often merged all these values into one value like two or added them to the column's mode. But such an approximation, according to us, may not be the right approach as for a targeted column such as mental health issues, the psychology behind a person saying 'don't know' and 'refused to answer' is quite different. So all column values were considered "unique" and directly fed into the XGBoost classifier. Another issue seen with the previous works is that many of them have overfitted the final prediction model without caring about pruning strategies. While with this approach, they were able to get very high statistical accuracy, but the model may not apply effectively in the real world. Many other algorithms were also experimented with, like random forests, logistic regression, and clustering models. Still, after observing the input data, a lot of imputing techniques were required for these algorithms and skewed the data and caused a lot of false information. We believe our approach to solving the classification problem results in a more unbiased and accurate model and reliable insights.

III. System Design & Implementation Details

The limitations of the previous work highlighted the need for a two-stage system design. In the first stage, we work with the complete dataset and apply dimensionality reduction techniques to get only the columns most related to the target column. The comprehensive data for a single year consisting of 300+ columns is extensive, with the document size exceeding 500MB. This massive size results in prolonged processing of information, and local hardware often runs into Memory Allocation errors. To perform this computation, we used the faster Google Collab Pro cloud hardware. After completing the required data pre-processing, we applied multiple dimensionality reduction techniques to the dataset. They reduced the columns selected for feeding the dimensionality reducer classifier. The project experimented with and set XGBoost and Random Forest algorithm as dimensionality reducer(DR) classifiers. This classifier provides a sorted list of features according to their importance. The features, i.e., columns with the most relative importance, are extracted from the overall dataset and shortlisted for further processing.

The shortlisted column CSV file generated is then used as input for the next stage. This file consists of less than 60 columns, with the file's size limited to 60 MB. Thus the objective of the first step is to reduce the computation requirements for the next stage. While completing this objective, the pruning method should have a minimum impact on the project's overall accuracy. In the second stage, we carry out algorithm dependent data pre-processing on the input data. For example, XGBoost can handle NaN values in the training data, but the same does not apply to the Random forest algorithm. One of the key reasons for selecting the XGBoost algorithm lies in its ability to handle these NaN values, as most of the input columns are sparse. The algorithm ensures no increase in the model's bias added due to the imputation strategy. At this stage, we carried out a comparison and evaluation of different classifier algorithms. GridSearchCV is used for hyperparameter tuning. The individuals who did not respond to the question or were unsure about the answer formed the prediction dataset. The second stage's objective involved finding which of these individuals in the prediction dataset have a high chance of suffering from the target disease. Figure 1 shows the detailed architecture diagram for the project.

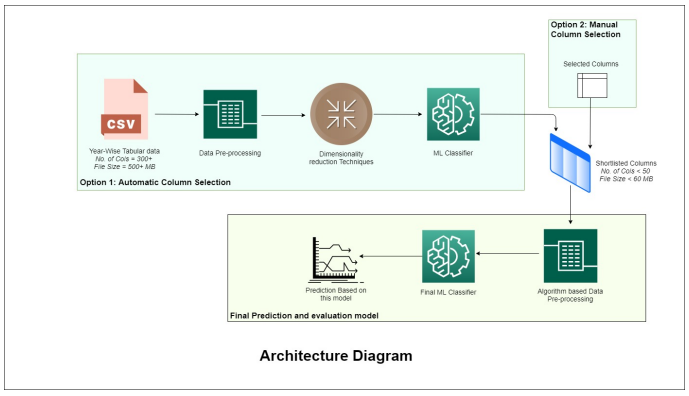

Fig. 1. Architecture diagram for the Project

Developed in Python 3, the project utilizes Scikit learn libraries, including but not limited to Random forest, tree,

GridSearchCV, and confusion_matrix. The project uses Matplotlib and Seaborn libraries for visualization. Pandas and Numpy libraries perform pre-processing and data manipulation duties.

IV. Experiments / Proof of concept evaluation

A. About Dataset

BRFSS collects data in all 50 states as well as the District of Columbia and three U.S. territories[2]. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world[2]. The comprehensive data for a single year consisting of 300+ columns is extensive, with the document size exceeding 500MB. Apart from a few columns that consist of date entries all the other columns contain categorical values. Dataset also consists of derived columns from the original column to facilitate processing.

Source: https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system

B. Data Pre-processing

The primary objective of the first stage of this project is column shortlisting. This shortlisting process forms the base of our data pre-processing model. In the next section, we evaluate various techniques used in the model to this effect. Please note that not each technique applies to all the targeted diseases as the size of test and training data govern this decision:

1) Dropping Derived columns: The dataset consists of derived columns marked with an '_' at the column name's start. Such columns are quite useful while performing linear regression. Still, if the choice of algorithm is a decision tree or random forest, such columns interfere with the process of determining the relative importance of a non-derived column. For example, for mental health-related issues, when we performed XGBoost classification on data with derived columns for the data from the year 2011, we realized that _FLUSHOT came out as one the most critical features in determining the mental health of a person. _FLUSHOT was a combination of a person with/without a flu shot over 65 years of age. One may incorrectly conclude that flu shots can be related to mental health issues. When the column bifurcated, we realized that mental health issues are more dependent on the age parameter. Older people are more susceptible to mental health issues compared to younger ones. To avoid such incorrect conclusions, we removed the derived columns from the original dataset.

2) Dropping columns with too many Nan values and low standard deviation: Many columns in the data do not have much variation. An example of this could be when 10000 individuals' prediction data for mental health-related issues had more than 30 columns with zero variation. Since these columns do not vary on the prediction set, any classification made on these columns would be incorrect, and the model won't fit the prediction dataset well. Similarly, we optionally used strategies like dropping columns with more than 95

3) Principal component analysis: PCA is one of the most common dimensionality reduction techniques used. The approach remodels the input into components that capture the most variance and are named as principal components. The only issue with this technique is that determining each principal component's composition and then determining which of the input columns have the essential features is not intuitive. So while we used PCA for evaluation purposes, it was not used in the final model as the results were not intuitive.

4) Correlation Based dimensionality reduction: If the correlation between two columns is very high, then data from these columns is almost the same and should not be considered for evaluation. To have a minimum impact on the model's accuracy, we set the limit of this value to 0.95.

5) Feedback Based manual reduction: As previously stated that the input dataset consists of many columns that have an obvious correlation to the target data. One example of such a correlation is the column DECIDE which asks the user if they have any difficulty deciding due to their physical and mental health-related issues. For this column, only the person with the physical and mental issues shall respond, thus giving the classifier pre-hints about the target column. Such column reduction techniques fall in the domain of knowledge-based dimensionality reduction.

The model used to predict mental health issues uses 5 fold validation and GridSearchCV for hyperparameter tunning and avoiding overfitting for tree-based algorithms.

C. Mental health related results

In this section, we will list the various insights relating to mental health-related issues. Figure 2 shows how females are more likely to have mental health-related issues as compared to men.

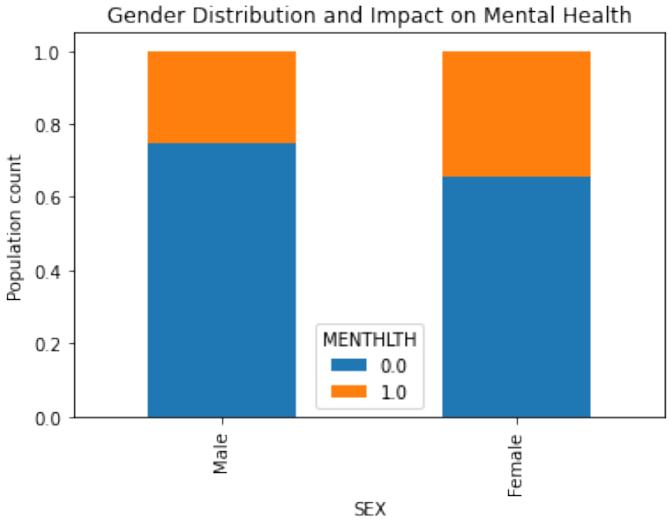

Fig. 2. Gender Distribution and Impact on Mental Health

Figures 3 and 4 show the variation of an individual's income on mental health-related issues. Figure 3 is a stacked bar graph, while figure 4 converts the data into a percentile stacked bar

graph. As it can be seen, percentile stacked bar graphs are more intuitive and can generate more accurate insights than stacked bar graphs. We can conclude from the results that while mental health-related cases reduce as the income of a person increases, the cases are prevalent across all economic classes.

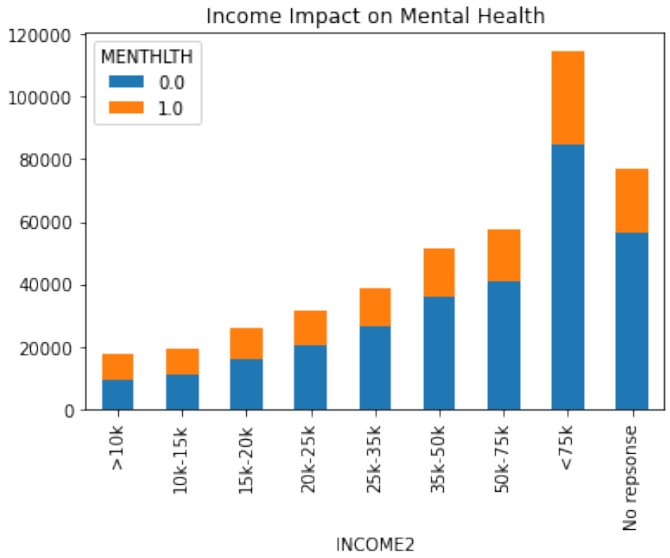

Fig. 3. Stacked Bar Graph of Income and Mental Health Issues

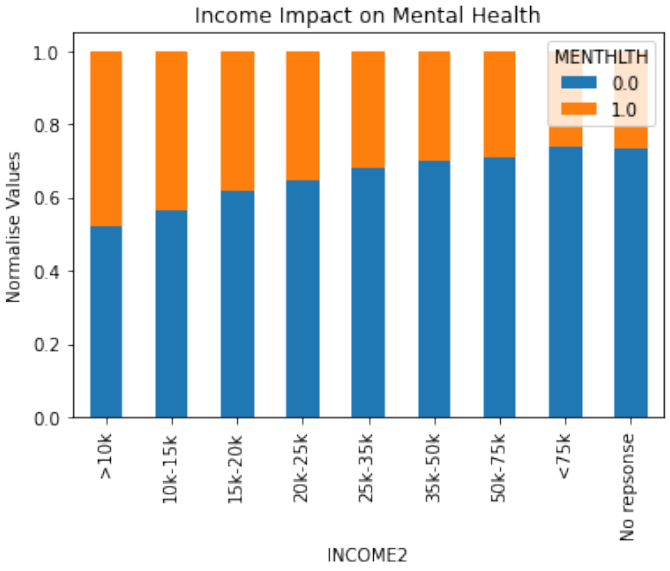

Fig. 4. Stacked Percentage Bar Graph of Income and Mental Health Issues

Figure 5 shows an individual's responses to the question, have you ever been told that you have a mental health-related issue. This figure shows that people around individuals suffering from mental health issues often know that he has some problems.

The factors affecting mental health were also analyzed over the years 2011 to 2015. It showed that the critical factors related to the diseases change with time, highlighting the

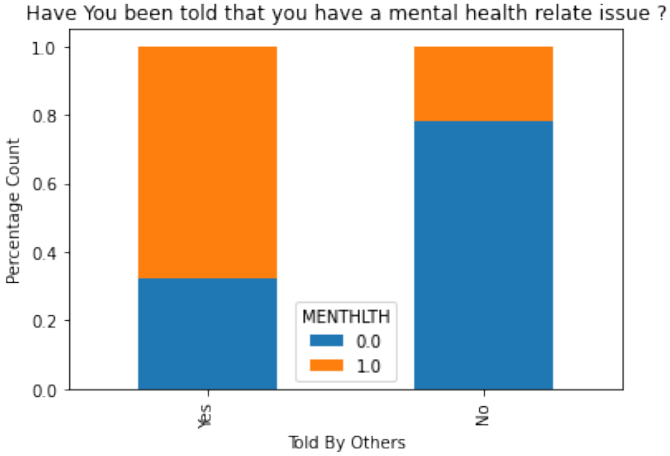

Fig. 5. Have people told you that you have Mental Health related issues

issue's dynamic nature. Figure 6 shows the most re-occurring critical factors in this span.

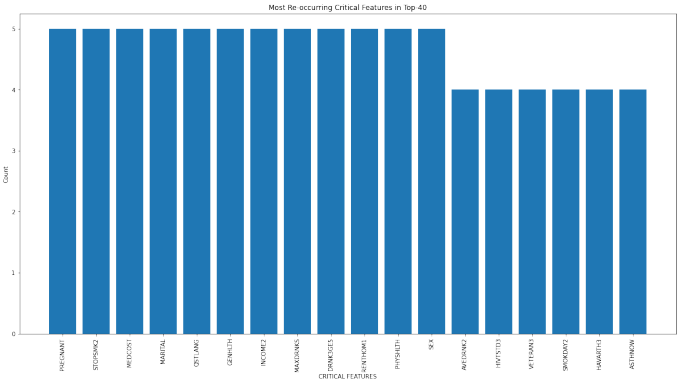

Fig. 6. Key Features across years

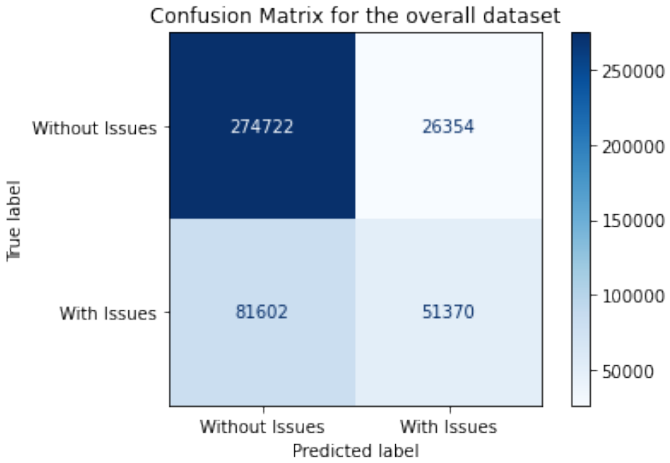

Fig. 7. Confusion Matrix for XGBoost Algorithm

Figure 7 shows the confusion matrix for mental health related issues using the XGBoost algorithm. Random Forest algorithm also showed promising results giving the same

accuracy as XGBoost algorithm. Decision tree algorithm gave an accuracy slightly lower than the two.

D. Diabetes Related Results

Various plots relating to diabetes were drawn before modeling the algorithm. It can be observed from Figure 8 that more women are susceptible to diabetes between the ages of 20 to 45. This also proves the stats that women are more susceptible to getting diabetes during their pregnancy. Also it can be observed that the density of men getting diabetes is more as compared to women after 50s.

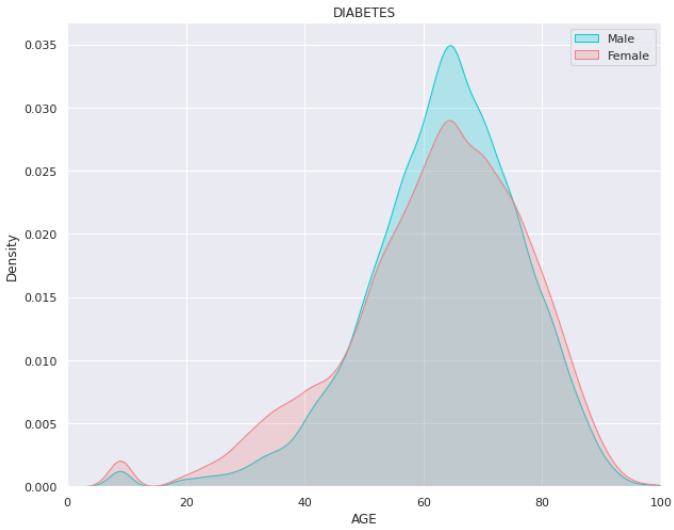

Fig. 8. Age Density vs Diabetes

It can be observed from Figure 9 vs Figure 10 that in both cases of men and women in case general physical health is bad then it is making the individuals susceptible to diabetes.Note that DIABETE3 is diabetes class with 1 - Positive and 3 - Negative.

While experimenting with feature importance for both male and female separately. It was observed that features affecting diabetes for males are different than that for women. Features like general health, education, employment and number of drinks etc. are the same for both. But features like pregnancy age and no of children are only found in women.

After using the XGBoost algorithm it can be observed in Figure 11 that the accuracy, recall and f1 score of the algorithm is pretty accurate.

E. Cardiovascular Health

There are 3 types of Cardiovascular diseases presented in the dataset: Heart attack, Angina and stroke. As positive cases for these diseases are very low in the dataset, all these 3 heart diseases are merged to a single target column. As per figure 12 there is high imbalance between cardiovascular disease detected (class 1) and not detected (class 0) classes. Only 12.5% of cardiovascular disease detected cases..

As per figure 13 males have more percentage of cardiovascular diseases compared to females.

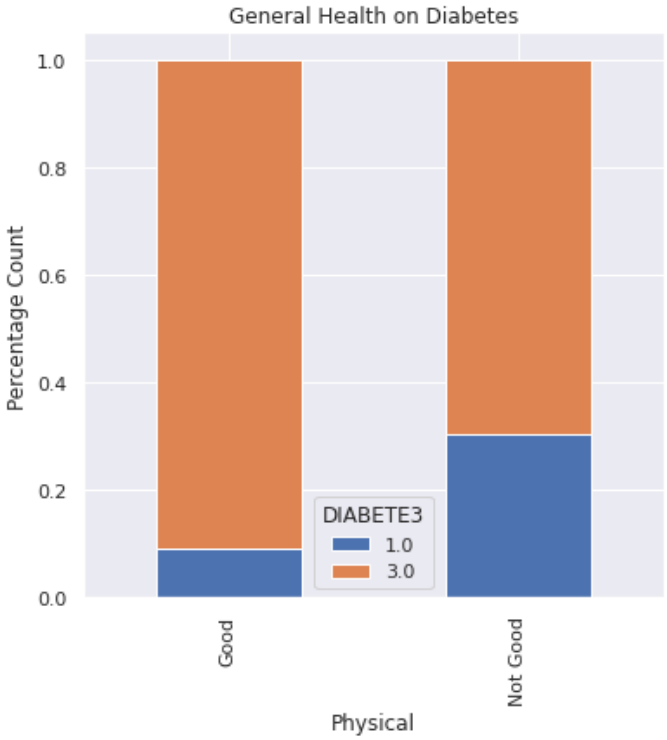

Fig. 9. General health vs Diabetes for Male

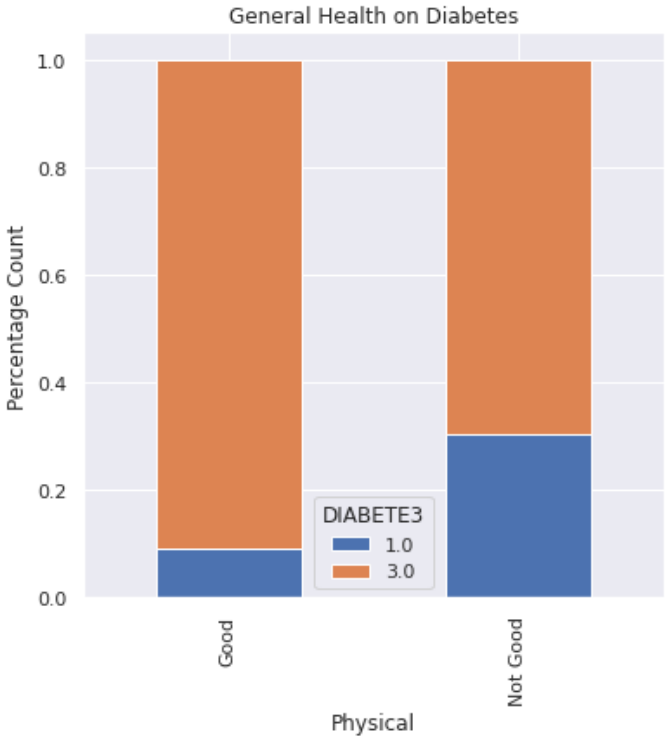

Fig. 10. General health vs Diabetes for Female

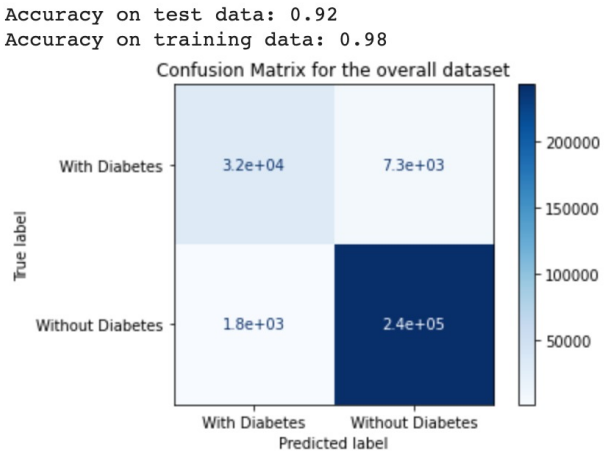

Fig. 11. Confusion Matrix for XGBoost for Female

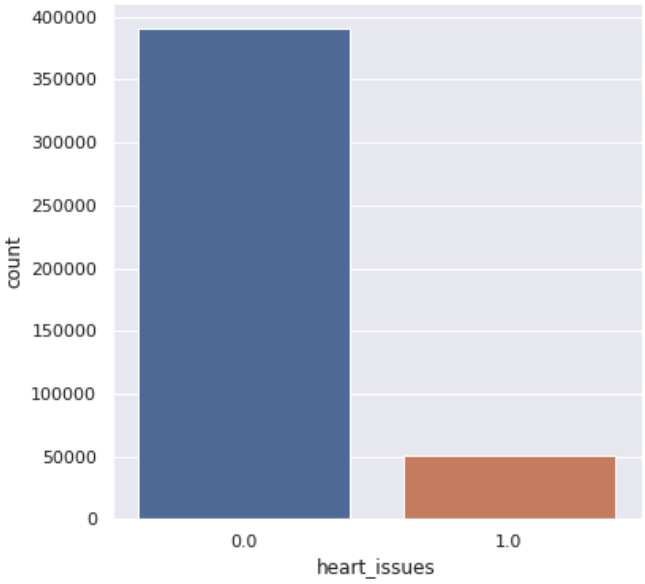

Fig. 12. Imbalance between positive and negative classes

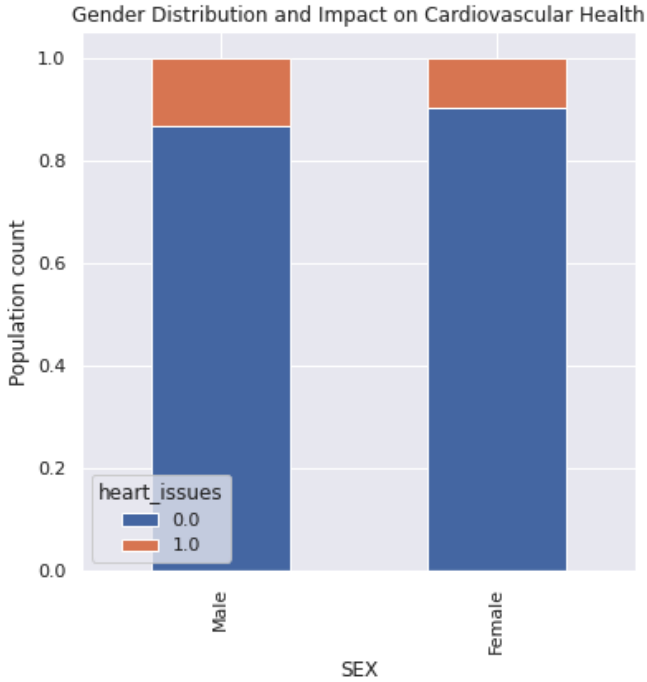

Fig. 13. Gender distribution for Cardiovascular health

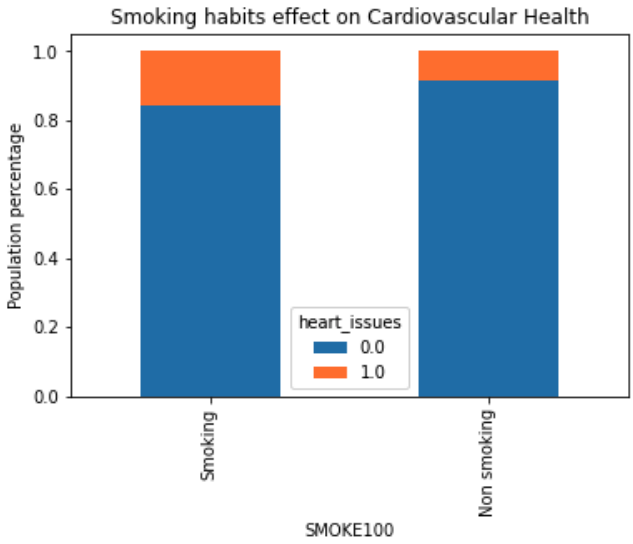

Fig. 14. Smoking causes Cardiovascular health problems

Figure 14 shows one of the known facts that smoking causes Cardiovascular health problems.

Figure 15 Shows overweight and obese people have more percentage of Cardiovascular health issues.

XGBoost accuracy is highest for Cardiovascular health prediction but recall is low and it is missing detection of many positive cases.

Note that, while we have summarised many key features and their impact on various diseases, kindly refer to our github repository for many more such insights and visualization.

V. Discussion & Conclusions

A massive dataset spanning over five years and consisting of attributes corresponding to more than 400,000 individuals for each year was analyzed, and new insights were generated based on the analysis. Every individual had more than 300 attributes corresponding to his/her daily activities, health, income, and social life. The project dealt with a massive, sparse, and complex dataset riddled with inconsistencies like inconsistent column names across each year, attribute duplication with derived columns, random categorical code selection, and many more. All these inconsistencies were studied and mitigated, resulting in many novel column-reduction and data pre-processing techniques. Another critical problem faced was the various bias, including reporting bias and selection bias seen in the dataset. The impact of these biases was studied and mitigated. Carefully thought out data pre-processing ensured that the dataset did not add any new biases to it. An example

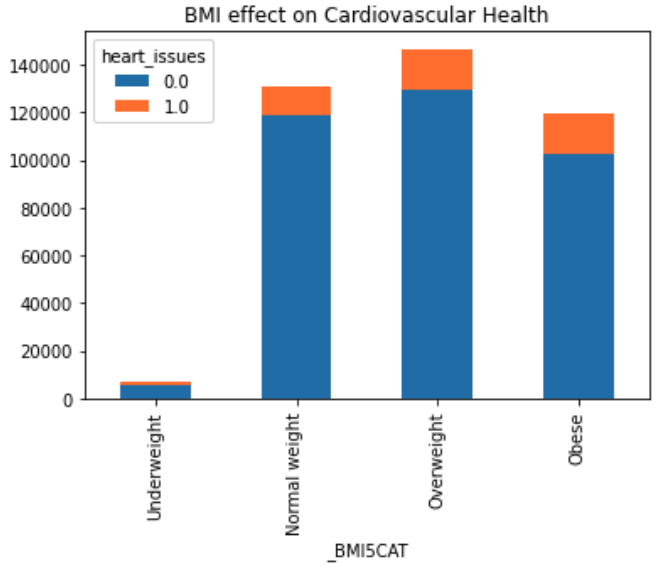

Fig. 15. Obesity and Cardiovascular health issues

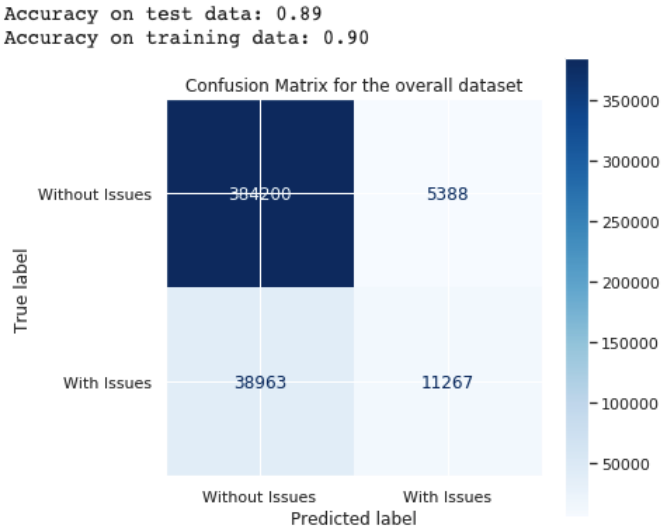

Fig. 16. Confusion matrix for Cardiovascular health issues

of such a step was selecting the XGBoost algorithm for classification, which avoided adding experimenter bias to the dataset. The project's central idea is that our daily habits and our lifestyle can impact the category of diseases that we contact. Selected cardiovascular diseases, diabetes, and mental health disorders, as target columns and various machine learning models, including XGBoost, Random forest and, Decision tree, were compared for this classification problem. Attributes that impact these disorders were derived, and we computed the statistical importance of these derived attributes. The project used a two-stage process that involved the development of two machine learning models. The first one derived the critical attributes that have the most impact on a particular disorder. The second one predicted whether a group of individuals with a known set of attributes were susceptible to the selected disorder. The model evaluation process results showed high accuracy and f1 score, as seen from the confusion matrix shared. We found many surprising correlations and documented them. Many of them showed seemingly innocuous habits impacting the occurrence of the disease.

Mental health-related issues were studied, and critical attributes such as an individual's physical health, economic standing, age, and gender significantly impacted an individual's mental health. This impact was well documented and visualized using stacked percentile bar graphs and counterplots. As it can be seen from the relatively low recall value, the project highlighted the need for a more focused questionnaire for a study relating to mental health issues. We made many attempts to mitigate these shortcomings. One attempt included feeding the entire array of columns to the classifier, allowing it to overfit using the random forest algorithm to determine columns with the highest feature importance. However, overall recall still could not be improved. Like the ensemble method to superimpose the results of various algorithms, other techniques were studied but not selected since they negatively impacted the project's overall accuracy.

VI. Project Plan / Task Distribution

Table 1 shows task distribution:

TABLE I
Task distribution

Task	Assign To	Status
Time Series Analysis of Key columns(Mental Health)	Swapnil	Complete
Data pre-processing using Column reduction techniques (Mental Health)	Swapnil	Complete
Algorithm comparison (Mental Health)	Swapnil	Complete
Stacked bar-graphs and percentage stacked bar-graph visualization (Mental Health)	Swapnil	Complete
Project Documentation: Report preparation, Powerpoint, Github repository Readme	All	Complete
Project Proposal and finalization	All	Complete
Algorithm comparison (Diabetes)	Pranav	Complete
Data pre-processing (Diabetes)	Pranav	Complete
Seaborn Charts and other Visualization (Diabetes)	Pranav	Complete
Architecture diagram	Pranav	Complete
Algorithm comparison (Cardiovascular Disease)	Amit	Complete
Data pre-processing (Cardiovascular Disease)	Amit	Complete
Visualization (Cardiovascular Disease)	Amit	Complete
Common Project Visualization	Amit	Complete

REFERENCES

[1] Centers for Disease Control and Prevention (2017, November). Behavioral Risk Factor Surveillance System, Version 1. Retrieved October 20, 2020 from https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system.

[2] Disease.Wikipedia.Available at https://en.wikipedia.org/wiki/Disease.(Accessed: 02 October 2020).

[3] 'A Gentle Introduction to XGBoost for Applied Machine Learning'.Machinelearningmastery.Available at https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.(Accessed: 10 October 2020)

GITHUB REPOSITORY

https://github.com/swapnilparihar14/Behavioral-Risk-Factors