

**COMP\_SCI\_396 Data Science Pipeline**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**(Group F) Final Report:**

**A closer look at Critical Health-Related Factors using BRFSS data-2020**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

June 7th, 2022

**Professor**

Huiling hu

**Students**

Sai Indraneel Amara

Amritanshu Ray

Vaibhav Sancheti

Shubham Shahi

# 

# Introduction & Motivation

Every year the Behavioral Risk Factor Surveillance System (BRFSS) sends out a series of questions to each of the states in the United States to record and measure the behavioral risk factors in the population aged 18 or older. This is a collaboration between the US states and the Center for Disease Control and Prevention(CDC) which has been going on since 1984. BRFSS has been collecting data extensively on the preventive health practices, risk behaviors linked to chronic diseases and other factors relating to habits and the access to health care across the US.

The dataset being huge has a lot of scope in exploring the habits, health and lifestyle of the people in different states.

As part of our project, we want to leverage this data to analyze and compare the health care access, habits and lifestyle choices and draw possible links to diseases such as the number of cases of Cancer in the country, the Asthma rate and cases of Diabetes.

Our goal is to compare the statewide data and see how the simple choices and access to health care facilities are connected to the disease rates in a state. This can also help us to come up with answers on which states are prone to more cases and why? What is the demographic that is mostly being affected and do lifestyle choices make a difference.

# Dataset

The [BRFSS 2020 Survey Data](https://www.cdc.gov/brfss/annual_data/annual_2020.html) has been obtained from Kaggle. It is extensive and includes attributes related to the healthcare system in the United States. Out of 279 columns, we would be focusing on the following for the project:

* On Health status and Demographics:

1. Health Status  
   a. Healthy Days  
   b. Health Care Access  
   c. Exercise  
   d. Inadequate Sleep  
   e. Chronic Health Conditions  
   f. Oral Health
2. Demographics
3. Health Care Access
4. Urban-Rural

* Lifestyle choices:

1. Tobacco Use
2. Alcohol Consumption
3. Immunization
4. E-Cigarettes
5. Marijuana Use

* Cancer Related Info:

1. Breast and Cervical Cancer Screening
2. Prostate Cancer Screening
3. Colorectal Cancer Screening
4. Lung Cancer Screening
5. Cancer Survivorship:  
   a. Type of Cancer  
   b. Course of Treatment  
   c. Pain Management
6. Prostate Cancer Screening Decision Making

* Other diseases related Info:

1. HIV/AIDS
2. Diabetes
3. Hepatitis Treatment
4. Asthma

The original dataset can be found on the CDC website [here](https://www.cdc.gov/brfss/annual_data/annual_2020.html). We would also be referring to this [Codebook](https://www.cdc.gov/brfss/annual_data/2020/pdf/codebook20_llcp-v2-508.pdf) to get an understanding of the values indicated in the dataset and what they mean.

# Data Cleaning

## Handling blank and null values:

This dataset is based on a questionnaire, which means that some columns were value dependant on previous columns, i.e. some questions’ responses were dependent on previous questions. For eg:

* Q1: Do you smoke -> No
* Q2: How often have you smoked in the past week? -> [null]
* As we can see, if someone answers No to Q1, then their blank to Q2 cannot be treated the same as a missing value. Moreover, the null value signifies that these many non-smokers exist, hence a null value cannot simply be dropped as it holds significance.

First step in dealing with the dataset cleaning was to drop completely ineffective columns. For this, we dropped columns which had only null values and no other responses.

## Data Encoding:

To handle the blank values with significance, we assigned a new value to the blanks that were outside the scale of the responses. For eg in a yes/no question with 0/1 values, the blanks were replaced with a -1, which helped when modeling the dataset.

## Feature Selection:

To select the most relevant features for our project, we performed a round of manual screening of almost 300 attributes to identify relevant features for particular diseases(Cancer, Diabetes and Asthma). We created a sheet with features, their codes and a short description, then assigned their importance for each disease.

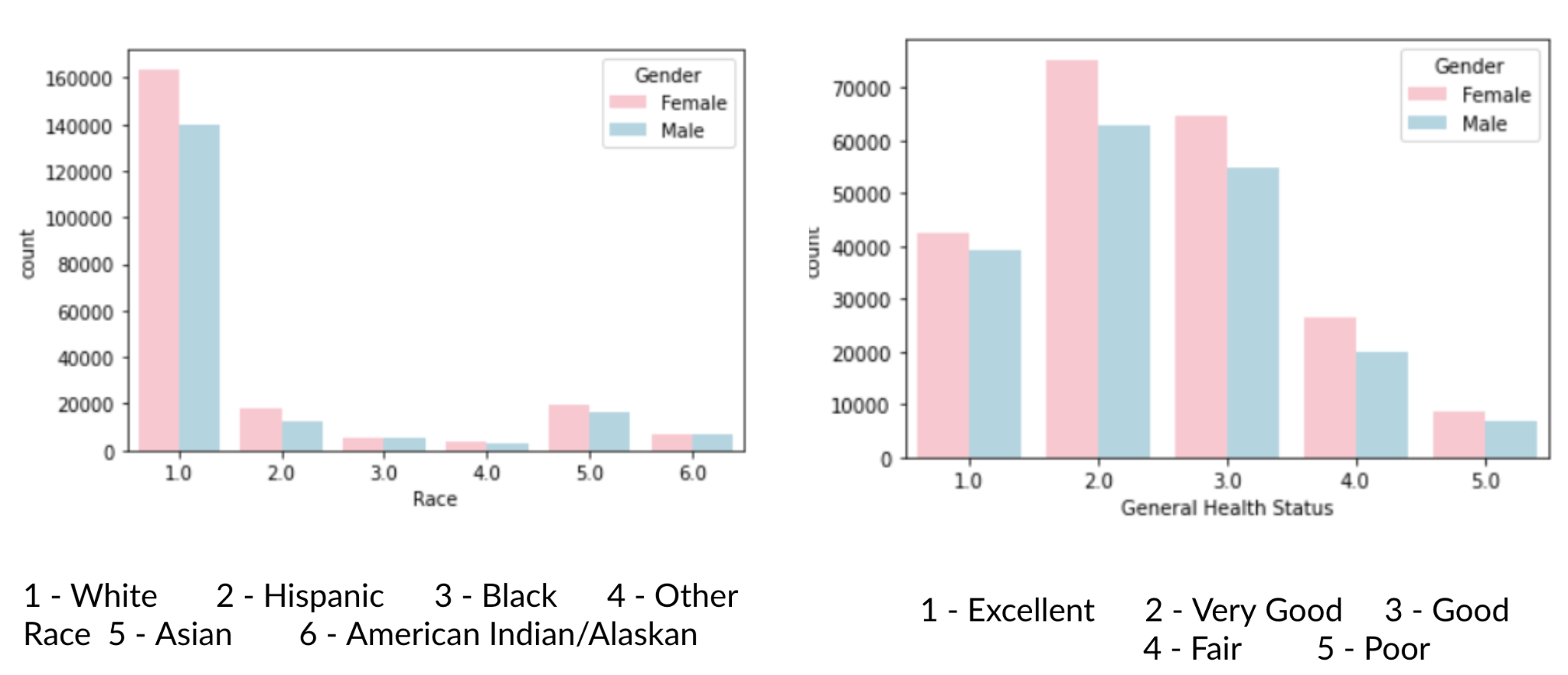
Additionally we ran the model based on the selected features and saw that some of the attributes were giving away the target label. We then did feature importance to find out these attributes and remove them from the dataset to train our model better. Our focus was to identify attributes that are related to symptoms or factors that lead towards a disease.

The codebook has a detailed view of what each response means in the dataset. As an additional criteria, we selected features that had answers in the form of yes/no (1/0), as this helped reduce the dimensionality. This helped to skip features with dependent values and hence, reduce the number of missing values.

# Exploratory Data Analysis

We begin by exploring the general trends of the dataset.:

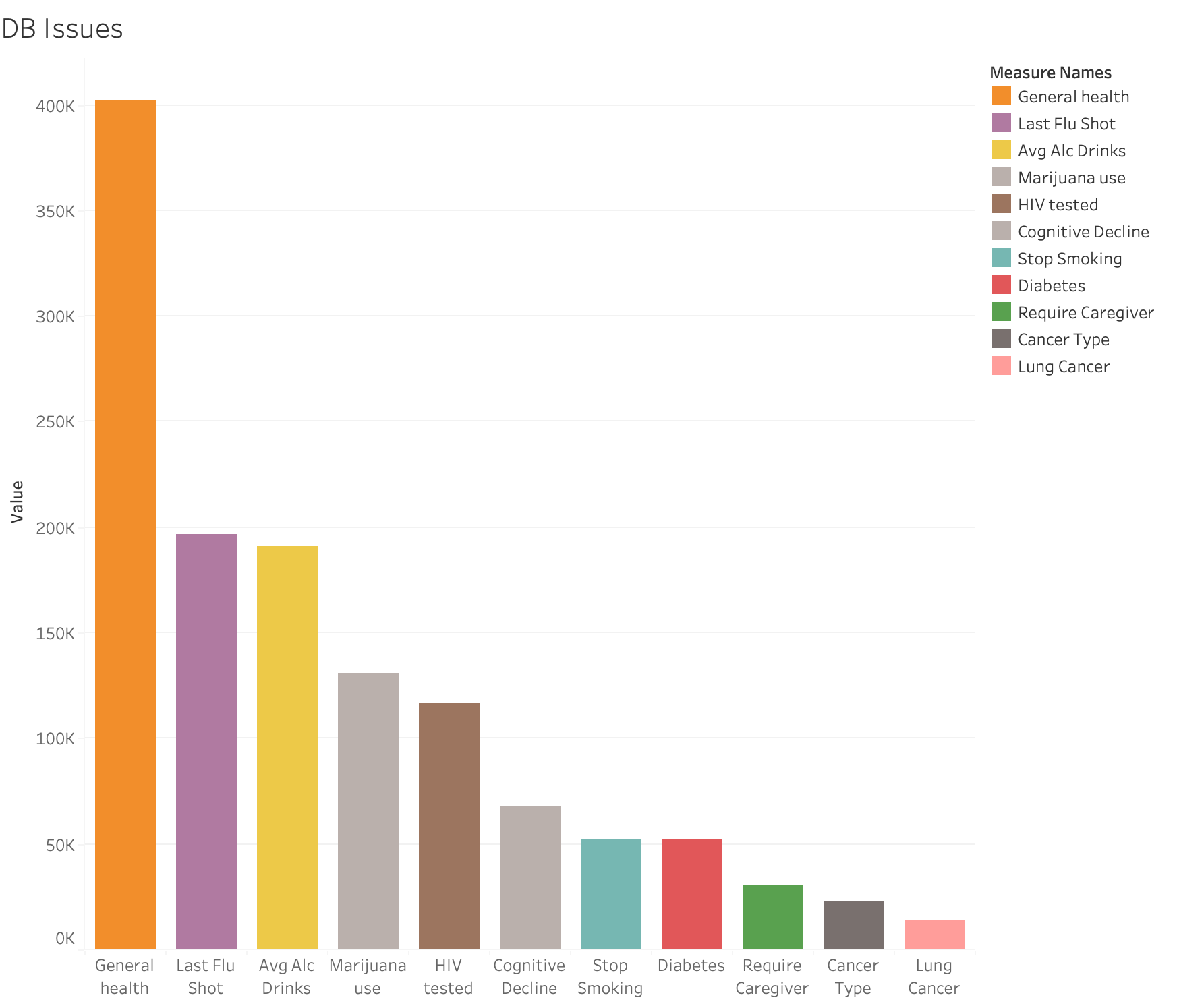
## Bias Based on race:



*Graph 1: Bias in the dataset*

One unique find from this initial EDA showed the data was biased towards white people, who had a lot more representation than any other class. We had to adjust our approach while modeling by being conscious of this bias.

## Missing Values:

Further analysis of the dataset revealed one of the biggest issues in the dataset, the missing values.

*Graph2: Count of Missing values in the specific attributes*

The dataset contains responses from approximately 400,000 people. However, when checking the responses to various columns of interest, we can note the substantial amount of missing data. For personal questions such as alcohol consumption, mariajuana consumption, if they have been HIV tested; the responses fell to only 50% of the total population, while when delving deeper into questions on specific diseases or symptoms, the response rate fell to 10% and even below.

This can be explained by the two facts, the first being people’s natural tendency to hide or share personal habits which may show them in a bad light.

Secondly, as we are dealing with data on diseases, where the percentage of population displaying a history with any disease is low, combined with the dependency of columns resulted in many blank values.

These need to be taken into account when modeling with the dataset.

## State wise health index for Demographic (18 and above)

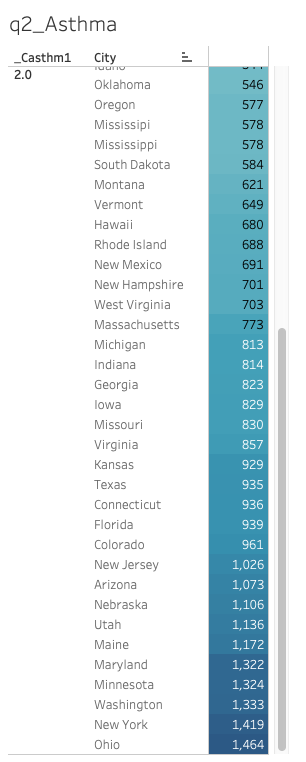
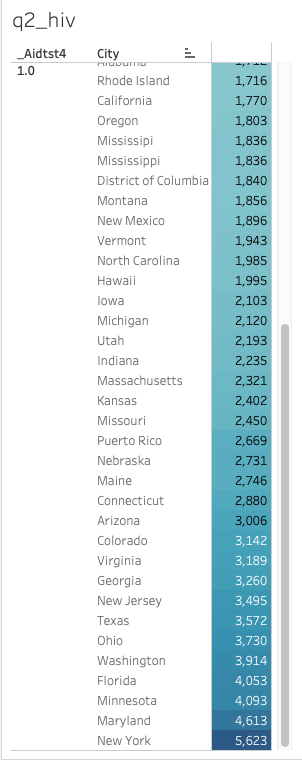
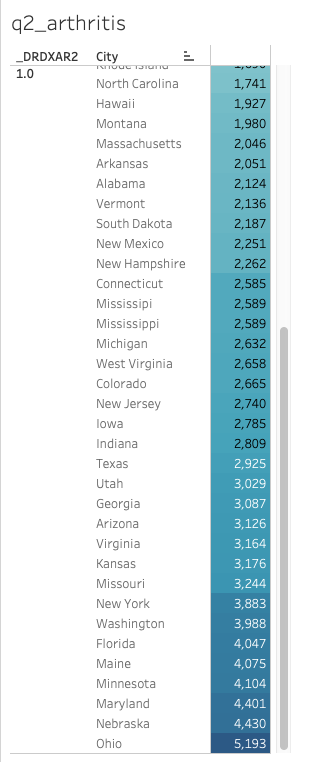
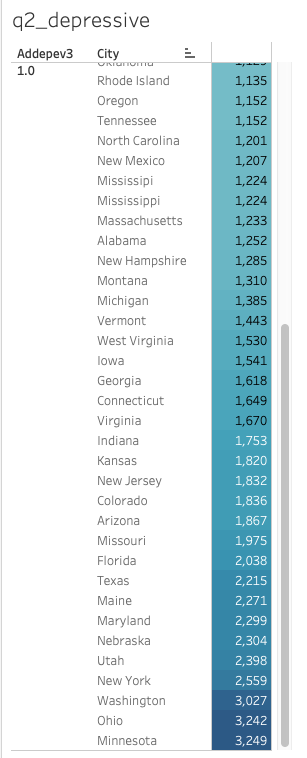
*Graph3: State-wise Good Health status*

## 

## States with more number of overall health issues

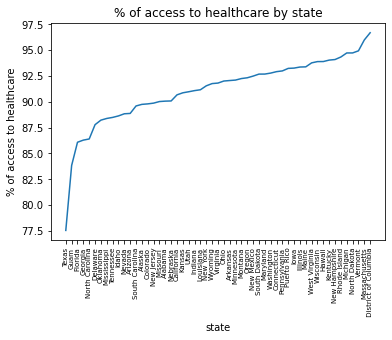
The graphs here show the number of cases in each state for the following health conditions:

* Arthritis
* Asthma
* HIV
* Depressive disorder



*Graph4: Graphs showing state-wise count of people with particular diseases set as true - in descending order(from bottom)*

## Other factors that may affect state-wise diseases

*Graph5 : State-wise Healthcare Access*

We noted that Healthcare Access is not equally distributed among the states, and that many of the southern states had considerably lower access to healthcare as compared to the other states.

This could affect the increase of diseases and cases of poor health in these states over the rest.

Moreover, looking into Healthcare Access amongst the different demographics. We noted the following statistics:

1. All states have greater than 85% population who have health coverage (out of all who answered).
2. Average number of men in households with health coverage is 0.81 and average number of women is 0.99.
3. Average number of men in households with no health coverage is 0.99 and average number of women is 1.02.
4. 94% of cancer patients had their insurance cover the full cost of their treatment.
5. The dataset had an equal proportion of male and female answering the questionnaire. Out of all the people who did not have health coverage, 55% of them were males and the percentage went down to 48% of all people who had health coverage.

## ANOVA Test

Next, we wished to check if states with less number of cancer cases have any other diseases on the rise.

We ran an ANOVA model check.

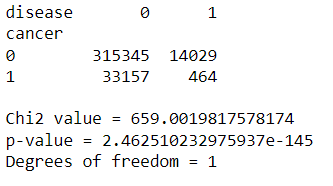
***Null Hypothesis:*** *They are independent*

***Alternate Hypothesis:*** *They are dependent*

Diseases Set:

1. Coronary Heart Disease
2. Pulmonary disease, C.O.P.D., emphysema, chronic bronchitis
3. Kidney Disease
4. HIV/AIDS

Results:



*Image1: ANOVA Test Results*

We find out that the p-value is really small (negligible) and so we can reject this null hypothesis.

# Feature Engineering

For Feature engineering we performed various data transformation techniques to formulate useful features. We used GridSearchCV to find and evaluate the best combination of parameters for our models to fine tune them for best results.

Additionally, PCA was done on the data set for dimensionality reduction which helped us increase our performance.

We used feature\_importances\_ to find the best features based on their relative importance and predict\_proba to compute the prediction probability of the classification.

Using multiple models to fit the data and understand the outcome also helped in feature selection to further improve the model performance.

# Data Modeling

We modeled 3 chronic conditions using the dataset, cancer, diabetes and asthma. For each of them, we removed the attributes that would lead to information leakage and that resulted in around 265-270 columns based on the condition. The attributes were the columns that asked questions like did anyone tell you that you had cancer.

For modeling, we tried models like decision tree, regression, multi layer perceptron, random forest, svm and knn. We then decided based on the results to pick decision trees, MLP and random forest.

We used sklearn for splitting the data into train and test using train\_test\_split. We picked a split size of 0.2 and used stratified sampling because of the class imbalance. Without stratified there was a high chance for most of the positive samples to end up in train. Therefore using stratified improved the results drastically.

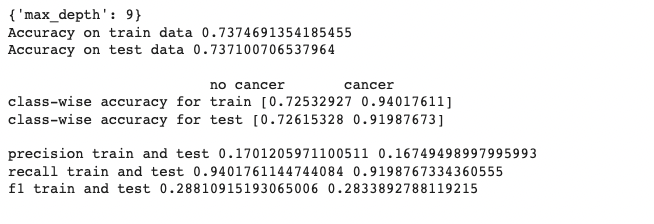
For each model we performed parameter tuning using grid search and cross validation. The parameters we tuned for the decision tree were - *max\_depth, criterion, class\_weight and min\_samples\_split.* For MLP we tuned the number of *hidden layers, number of neurons in each layer and the number of iterations.* Finally for random forest we experimented with - *max\_depth, n\_estimators, ​​criterion, class\_weight and max\_features.*

The metrics we used to compare models were *recall, precision, f1 score and accuracy*. Due to class imbalance and importance of minimum false positives, recall is the best metric.

As part of feature engineering, we performed binarization and binning for various attributes as it made more sense for our use case. We performed standard scaling before training MLP as it is susceptible to range. We also used PCA to work with smaller dimensionality.

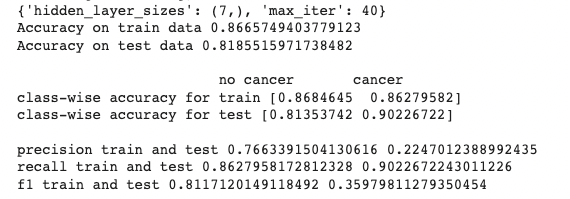
# Results

1. **Cancer**
2. Decision Tree Classifier



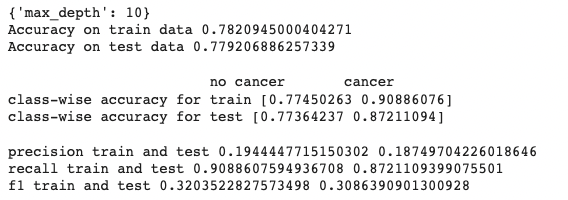
*Image2:Cancer\_ Best Parameters for Decision Tree*

1. MLP with Standard Scaling



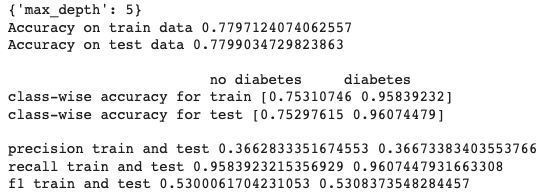
*Image3: Cancer\_Best Parameters for MLP*

1. Random Forest



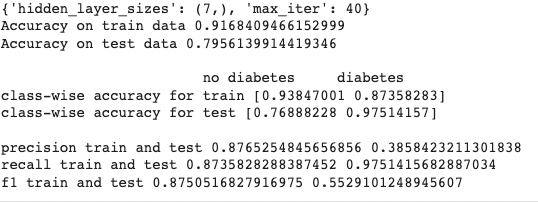
*Image4: Cancer\_Best Parameters for Random Forest*

1. **Diabetes**
2. Decision Tree Classifier:



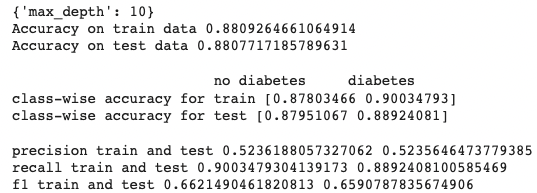
*Image5: Diabetes\_Best Parameters for Decision Tree*

1. MLP with Standard Scaling



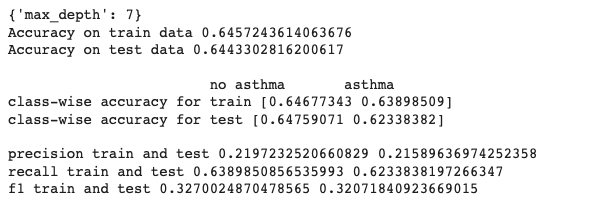
*Image6: Diabetes\_Best Parameters for MLP*

1. Random Forest:



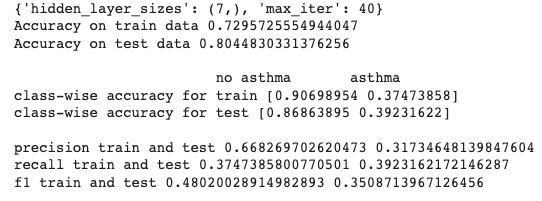
*Image7: Diabetes\_Best Parameters for Random Forest*

1. **Asthma**
2. Decision Tree Classifier:



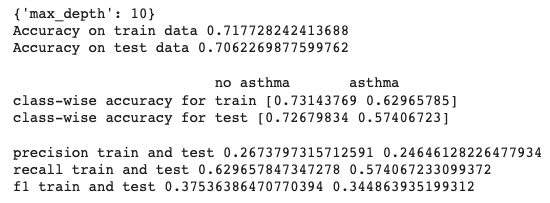
*Image8: Asthma\_Best Parameters for Decision Tree*

1. MLP with Standard Scaling:



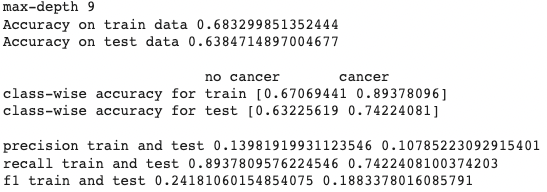
*Image9: Asthma\_Best Parameters for MLP*

1. Random Forest:



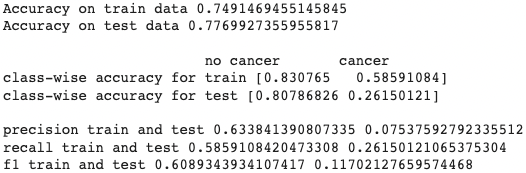
*Image10: Asthma\_Best Parameters for Random Forest*

1. **Cancer**
2. Decision Tree Classifier



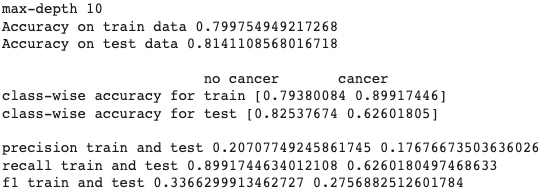
*Image11: Cancer\_PCA for Decision Tree*

1. MLP with Standard Scaling:



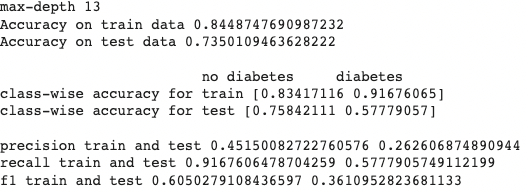
*Image12: Cancer\_PCA for MLP*

1. Random Forest:



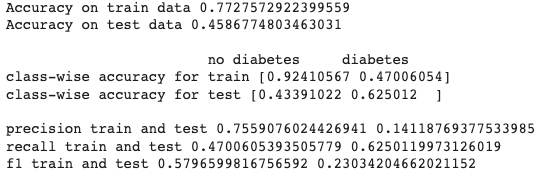
*Image13: Canccer\_PCA for Random Forest*

1. **Diabetes**
2. Decision Tree Classifier:



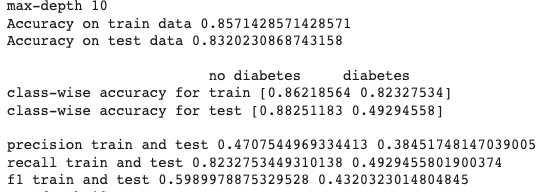
*Image14: Diabetes\_PCA for Decision Tree*

1. MLP with Standard Scaling:



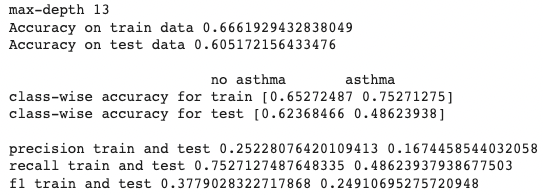
*Image15: Diabetes\_PCA for MLP*

1. Random forest



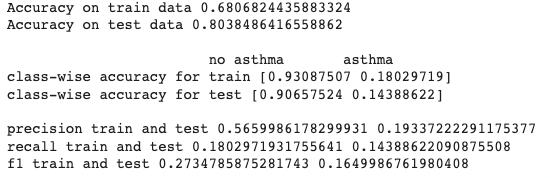
*Image16: Diabetes\_PCA for Random Forest*

1. **Asthma**
2. Decision Tree Classifier



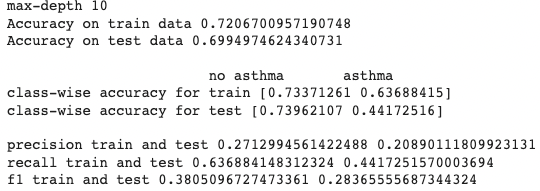
*Image17: Asthma\_ PCA for Decision Tree*

1. MLP with Standard Scaling



*Image18: Asthma\_PCA for MLP*

1. Random forest



*Image19: Asthma\_PCA for Random Forest*

# Findings

Initially, when we used the dataset for modeling we were focusing on the accuracy and we were able to achieve great results. We then looked at class-wise accuracy and realized the positive class had very low accuracy and this was due to a huge class imbalance. We, therefore, used metrics like precision, recall and F1 Score, as they are more suitable for class imbalanced datasets.

We tried different values for the parameters of our model using grid search with cross-validation. We found deeper or more complex models would overfit the training data very quickly and would drastically affect class-wise accuracy even though the overall accuracy improved.

This dataset had a lot of information regarding the typical symptoms of cancer and diabetes and therefore was able to make great predictions, whereas for asthma this was not the case. Similarly, there are a lot of other diseases that could be modeled well with some additional data regarding the individual.

Scaling was very important for MLP as the results were drastically different when the features were passed directly. There was no difference for decision trees and random forest. That could be due to the fact that decision tree algorithms are scale-independent and can work on very large values of data. SVMs and complex neural networks took a long time to train and therefore we could not use them for our experiments.

Using PCA we were able to bring down the dimensions by a third and this dropped the results by a few points as expected. Surprisingly, MLP was able to achieve better results than the other 2 models with only 40 dimensions instead of 100.

During our experimentation, we found looking at feature importance provides relevant information about how to improve the model. If there were attributes that were actually important but were not found important by the model it could be because of an error in feature engineering.

We created a health score for a person that analyses the features of that individual and predicts the risk value or more simply how sustainable is their current lifestyle. If the value is greater than 0.7 then continuing their lifestyle would most likely result in a chronic condition.

# Conclusion

The idea behind the project was to predict health conditions of people across the different states. Though we modeled only for three diseases we saw that there are multiple features available to model for different chronic conditions that can lead to more insight about health conditions.

Moreover, we also found features that are related to financial status and healthcare access of the people hence giving a different perspective on how the health care is affected not only by health conditions but by the access to the basic amenities as well. Also, the data is heavily biased when it comes to race. Hence there should be measures taken so that this can be resolved.

We believe that the dataset can be used to predict and analyze different aspects of healthcare and policies can be tailor made for the states to make it easier for the people to get the required assistance.