TEAM A

Tram NGOC Le | BAO NGOC DINH | YANQI SHI | NICHOLAS MERCOGLIANO

GENERAL HEALTH RATING ANALYSIS REPORT

BYBG 7967 | DATA MINING FOR BUSINESS | SECTION: WEDNESDAY 5:45PM – 7:45PM | PROFESSOR FRANCIS MACCRORY

DATE: DECEMBER 6th, 2019

2019

1. **ABSTRACT**

Health consciousness is a significantly psychological element as it plays a vital role in affecting different human attitudes, behavior, and conception pertaining to health issue. Based on the influence of behavioral risk factors and lifestyle on human’s health consciousness, this study aims to help people improve their health awareness by building a model to predict the accuracy of human perceived knowledge about health and which factors affect people’s health consciousness, specifically the factors make people consider themselves healthy. By analyzing the different factors listed above, we hope to determine whether a person is knowledgeable in these areas, or if their opinions of these health issues may need further education. This report starts with introduction of health consciousness matter, following by discussion about the dataset, methodology, problem statement, result analysis and conclusion.

1. **INTRODUCTION**

It is well recognized that health consciousness is a serious and important matter that affects people’s lives. In the Information Age, online health information enhances patient’s knowledge of health issues. However, infinite sources of information mean that it is difficult to select the information with good quality to absorb.

Today, the enormous availability and consumption of false health information is a serious problem that our society is facing currently (Alam, 2018). With a click on Google, we can find any information about anything. Social media platforms, including Twitter, Facebook, WeChat, Instagram, or Reddit, allow people to express their ideas about anything freely. Therefore, there is likelihood of people having misleading understanding about the factors that affect health, which can lead to serious health problems in the future. People also have a tendency to reinforce their own knowledge with anything they can find that agrees with their current ideology, rather than what can be proven or what the overwhelming majority believe to be true.

Because of the ability to freely express ideas, the abundance of misinformation, and the unwillingness to accept a fact if it is contrary to what they initially believe, it can be very easy for anyone to accidentally believe that something that they might consider healthy is actually not very healthy, or vice-versa.

1. **DATA DESCRIPTION**

The BRFSS dataset contained information collected from 53 states and the US territories (“BRFSS”, 2018). The BRFSS is the nation’s premier system of health-related telephone surveys that collect data about U.S residents pertaining their health-related risk behaviors, chronic health conditions, access to health care and use of preventive services related to the leading causes of death and disability. Every year, BRFSS completes more than 400,000 adult interviews through landline and cellphone questionnaire. From the landline survey, BRFSS collects data from a randomly selected adult in a household. From the cellular survey, BRFSS collects data from adults answering the cellular telephones residing in a private residence or college housing (CDC). The project used 2018 survey dataset. The database consists of about 300 variables in total, but we narrowed down to seven inputs. Our selected attributes focused on demographic and health condition awareness of the surveyed people.

The US pollution data is provided by U.S. EPA, United States Environmental Protection Agency (“Pollution in the US since 2000”). It recorded the concentration of main air pollution particles based on a city level since 2000. To make the pollution dataset correspond to the BRFSS dataset, we grouped the US pollution data by the state code and calculated the average concentration of pollution particles in each state. The dataset contained 4 pollution particles, but we only used carbon monoxide (CO) as our input since CO is the major pollutant produced by fossil fuel.

We used Merge, Filler, Select and Type node to process our data. We merged the two datasets by using merge node and included only matching records because the pollution data file collected data from more territories than the BRFSS dataset. For every field, we had some survey participants who refused to answer the question or was unsure, or a null value. Thus, we had to use Select to remove them since those answers would not contribute predictive value to our model. We used Filler to replace values automatically. We had nine different classes of General Health Rating, but we ended up only using the first three which were “Excellent”, “Very Good” and “Good”. In the codebook, “How many days of poor mental/ physical health” field, abbreviated as “poor health”, contained answers varying from 1-30 days and another value of “88”, which indicates that the person had 0 days of poor mental and physical health. The field was a continuous field and we replaced “88” with “0” so that the model would not split at a weird number between 30 and 88 and would read values as a continuous data type from 0-30 days. We used Type node to correctly classify the data type for the fields we used. Our final cleaned dataset had a total of 68,851 records. Overall, most people would consider themselves “Good” or “Very Good” while not a lot of people considered themselves “Excellent”.

The chosen inputs and target are listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Input Type** | **Input Variable** | **Values** |
| Poor Health | Continuous | 0-30 | Number of days with poor physical or mental health in the last 30 days |
| Exercise in last 30 days | Flag | 1,2 | 1=Yes  2=No |
| Heart Attack | Flag | 1,2 | Ever Diagnosed with Heart Attack?  1= Yes  2= No |
| Have Kidney Disease | Flag | 1,2 | (Ever told) you have kidney disease?  1= Yes  2= No |
| Diabetes | Nominal | 1-4 | (Ever told) you have diabetes?  1=Yes  2=Yes, but female told only during pregnancy  3= No  4= No, pre-diabetes or borderline diabetes |
| Income Level | Nominal | 1-8 | 1= Less than $10,000  2=$10,000-$14,999  3= $15,000-$19,999  4=$20,000-$24,999  5=25,000-$34,999  6=$35,000-$49,999  7=$50,000-$74,999  8=$75,000 or more |
| Avg Alc Drinks per Day | Continuous | 0-30 | Average alcoholic drinks per day in the past 30 days |
| CO\_Mean | Continuous |  | Average CO emission in each state |
| General Health | Target | 1,2,3 | 1= Excellent  2= Very Good  3= Good |

1. **PROBLEM STATEMENT**

Every year, Centers for Disease Control and Prevention conducts a BRFSS Survey to collect health data from people across the United States. People answer multiple questions and identify whether they are healthy or not by rating their health based on their own knowledge. In this project, we propose to build a model to predict which health factors influence people’s health consciousness, specifically classify the factors that make people think they are healthy. This analysis will help people improve their health awareness in certain health areas, prevent possible illnesses in the future, and live a healthier life.

1. **METHODOLOGY**

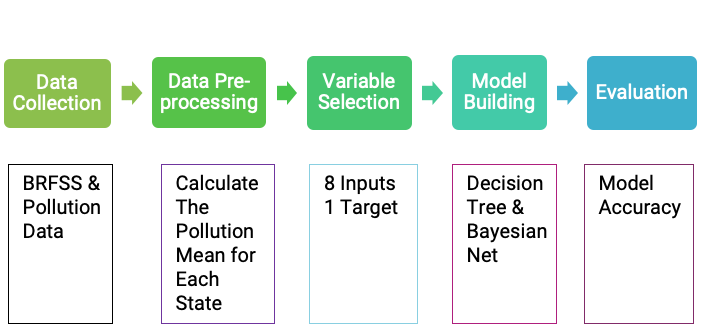
This study applied supervised machine learning methodology, specifically Bayesian and Decision Tree, to train and classify factors influencing people’s health consciousness. The methodology includes data collection, data preprocessing, variable selection, model building and model evaluation.

**Data Collection:** We collected The Behavioral Risk Factor Surveillance System (BRFSS) Annual Survey Data from The Center for Disease Control and Preparation (CDC). The dataset contained information collected from 53 states and the US territories. The BRFSS collected data about U.S residents pertaining their health-related risk behaviors, chronic health conditions, access to health care and use of preventive services related to the leading causes of death and disability. We picked seven fields out of the 300+ fields in order to reduce noise and produce a better model with good interpretive results. For a smaller dataset, we used the pollution dataset from Kaggle, which contained pollution measured from every state and territories in the US for everyday of the year.

**Data preprocessing:** In the second phase, we conducted data pre-processing by cleaning the datasets and eliminating unnecessary information to ensure accuracy for the analysis. Originally, our BRFSS dataset included almost 300 variables, however, we reduced the number of variables to seven for the BRFSS dataset. We calculated the mean of CO, SO2, O3 and NO2 in the pollution.csv file for each state. We merged the data by state code. The state code of both files corresponded to the alphabetical order of state names, so we simply used the node Merge. Since CO was a big pollution factor that led to global warming, we decided to incorporate that field into our model from the smaller dataset. We set the General Health Rating as our target variable.

**Variable Selection:** We chose to focus on general health information and pollution information that we consider to be helpful to our analysis. The general health information included data about whether the person have Kidney Disease, Diabetes, Heart Attack, Average number of drinks in the last 30 days, whether they exercised or not in the last 30 days. We had income level as a demographic field of input and CO mean as an environmental factor input.

**Model Building**: We partitioned the data into 50% for testing data and 50% for training data. We discarded null values, missing data or “Don’t know/ Not sure” replies. We used filler node to replace the numeric value “88” in “How many days of poor physical or mental health (Poor Health)” as 0 following the notation in the CDC BRFSS Codebook. The Poor Health field is a continuous field, and if we leave the notation as 88 it might be mistaken as 88 days of poor health. We created a decision tree model and a Bayes network model.

**Evaluation:** In the final phase, evaluation of the results and models were conducted. We evaluated the accuracy of the model, importance predictors, and reasoning behind the test with histogram and distribution plots. We used the “Analysis” node and “Matrix” node to analyze our data. Based on the result of model evaluation, we picked the model with higher level of accuracy to analyze the final result.

1. **RESULT ANALYSIS**

Regarding model evaluation, Bayes Net and Decision Tree resulted in the same top three important indicators, including Diabetes, Exercise in Past 30 Days, and Poor Health in the same order. Pollution indicator, CO\_Mean, has low level of influence on the people’s health consciousness and the result of classification as well. Pertaining model precision, Decision Tree, with 50.79% of accuracy, performed better than Bayes Net, with 40.1% of accuracy, did on testing set by roughly 10%. Based on the result, we decided to use Decision Tree for further result analysis.

Considering Decision Tree Confusion Matrices, the model performed most correctly at classifying health factors for people who rate their health as “Very Good” (General Health Rating 2.0) with 72.888% of accuracy on testing data. However, the model failed to classify health factors for people who considered their health as “Excellent” (General Health Rating 1.0) and performed poorly on people who rated their health as “Good” (General Health Rating 3.0), with 0.071% and 47.445% of accuracy on testing data, respectively. The accuracy rates for three general health rates varied significantly because of the unequally distributed data. Considering the sample size of General Health Rating, there are about 16% of “Excellent” records, 30% of “Good” records and 30% of “Very Good” records.

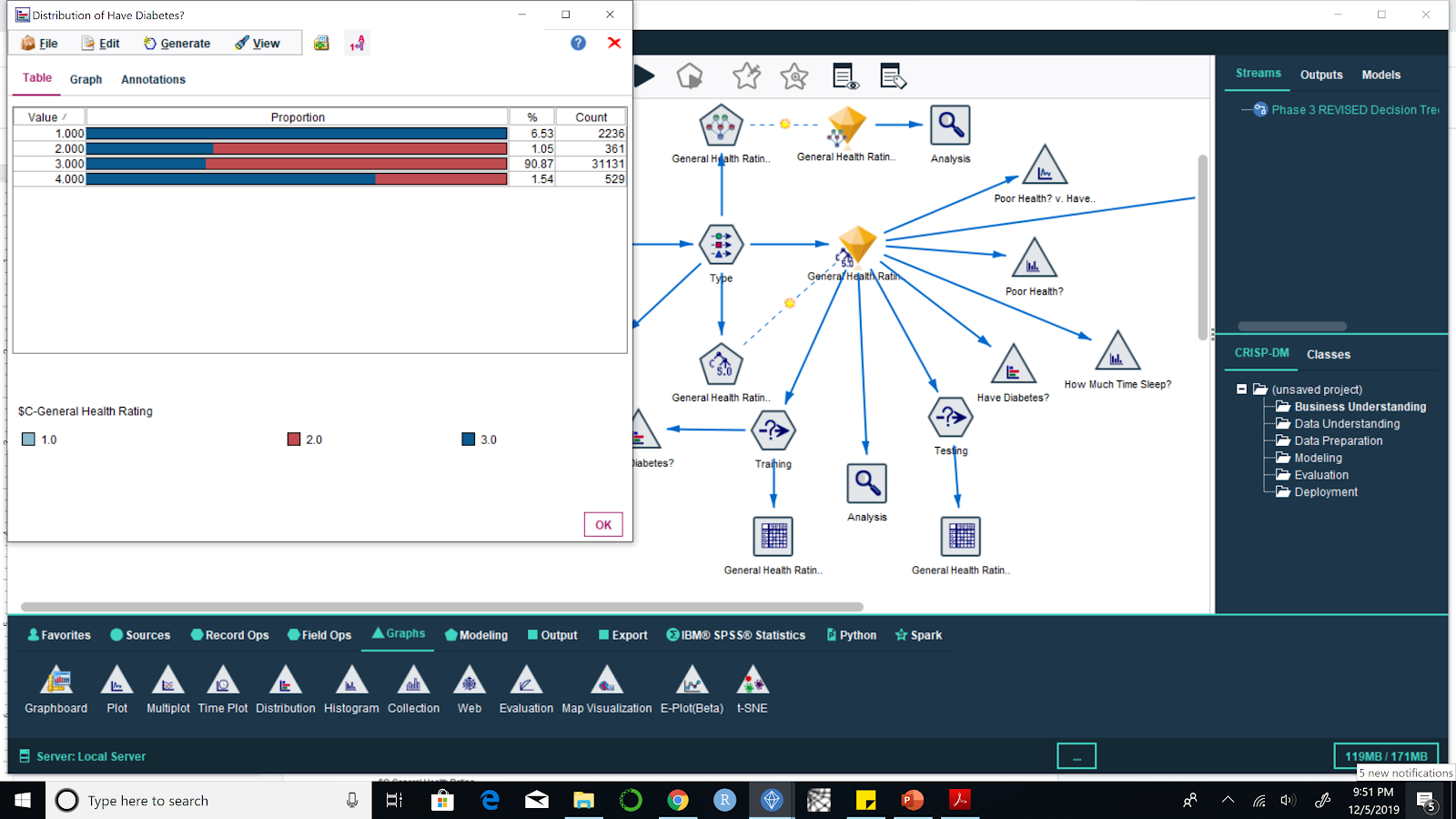
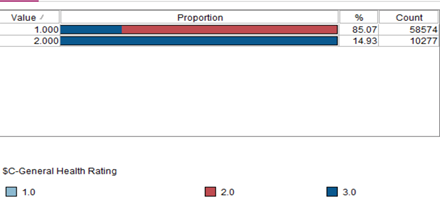
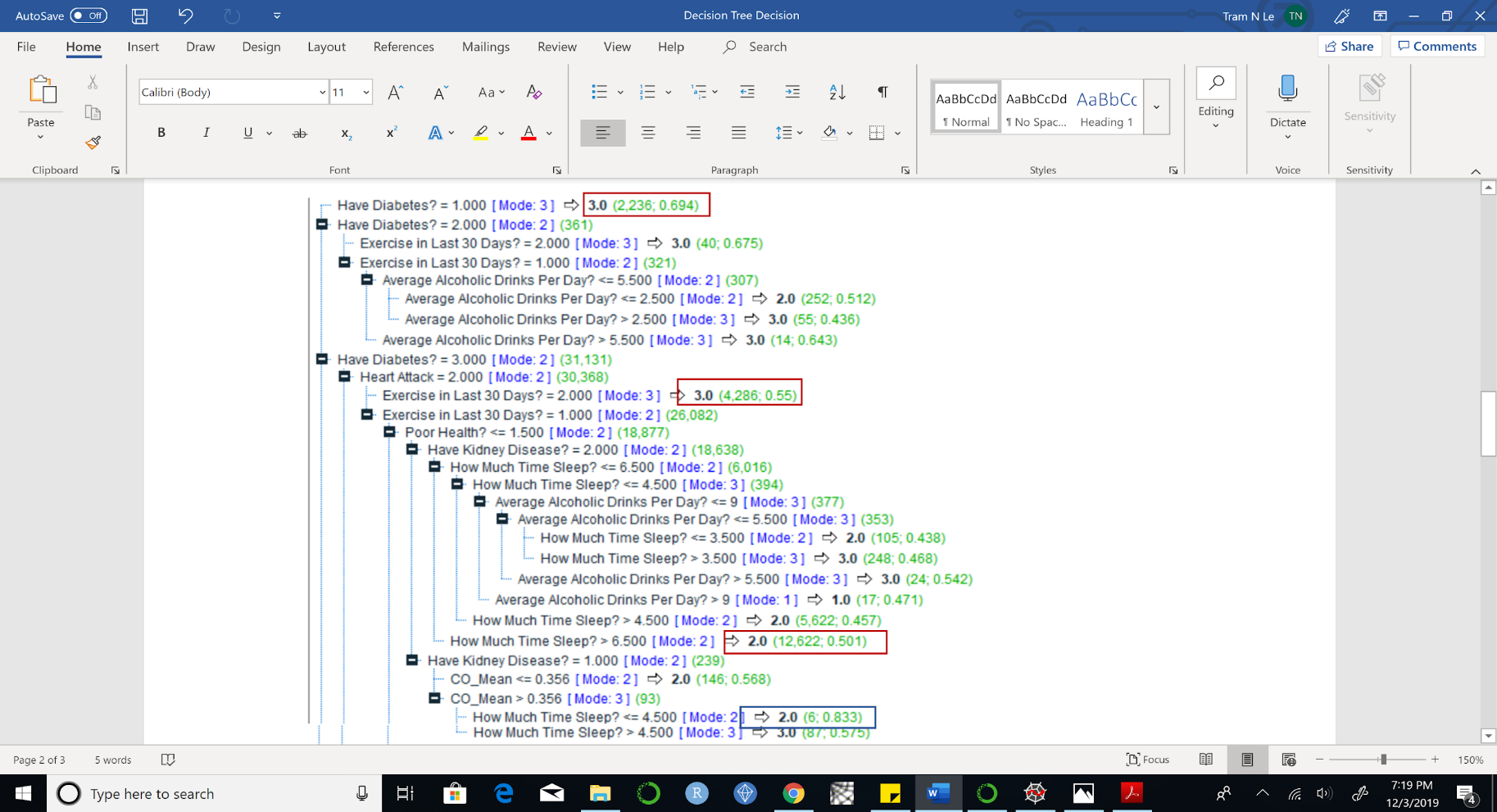
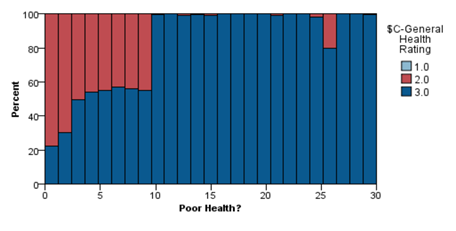
Based on Decision Tree model summary, the majority of rules are about 50% to 60% of accuracy with large sample size while rules of high accuracy rate (more than 80%) are set for tiny sample size.  For example, one rule that people who rate their health as “Very Good” (General Health Rating 2.0) tend to have no history of diabetes, heart attack or kidney disease, experience less than one and a half-day of poor physical and mental health, do exercise in the last 30 days, and sleep more than 6.5 hours is 50.1% accurate for 12,622 records. Another rule, with an accuracy rate of 69.4%, that people who considered their health as “Good” (General Health Rating 3.0) tend to have diabetes is based on a smaller sample size of 2,236 records. In general, these rules are not significant and accurate enough for further analysis.

Regarding top three important indicators and decision tree confusion matrices, further conclusions can be drawn from visualization of important indicator analysis. Overall, people who consider their health as “Very good” (General Health Rating 2.0) tend to exercise in the last 30 days, have less than 10 days of poor physical and mental health, and have diabetes during pregnancy or don’t have diabetes at all. On the other side, people who consider their health as “Good” (General Health Rating 1.0) tend to not exercise in the last 30 days, have more than 10 days of poor physical and mental health, and have diabetes.

1. **CONCLUSION**

In conclusion, Decision Tree is a better model than Bayes Nets in general health rating indicators classification. The top three important indicators of people’s consciousness about health are diabetes, poor health, and exercises in the last 30 days. However, the limitations on the dataset, specifically undistributed sample size, reduce model accuracy rate to 50%. Therefore, Decision Tree is not a good model to apply for further analysis.

From a managerial perspective, we would not recommend apply the model to predict factors that influence people’s consciousness about their health unless more data are collected or any change in strategy is applied. For further improvement of the model, we would recommend execute another strategy and look at the problem from different angles. Instead of predicting what makes people think their health is good from General Health Rating 1.0, 2.0 and 3.0, we can identify what makes people think their health is excellent, good, or poor. We can execute the strategy by classifying the target from the original dataset into three new categories: General Health Rating 1.0 and 2.0 as “Excellent”, General Health Rating 3.0 as “Good”, and General Health Rating 4.0 and 5.0 as “Fair”. The new strategy can deliver a more meaningful view on the topic while increasing the sample sizes” quality with more evenly distributed data for each rate. After improving the model, we can apply the result into marketing campaign to educate people about health awareness along with multiple factors might influence their health quality.

**Figure 1. Exercised in The Last 30 Days Figure 2. Have Diabetes Distribution**

**A screenshot of a cell phone

Description automatically generated Figure 3. Poor Health Histogram Figure 4. Decision Tree Rule Examples**

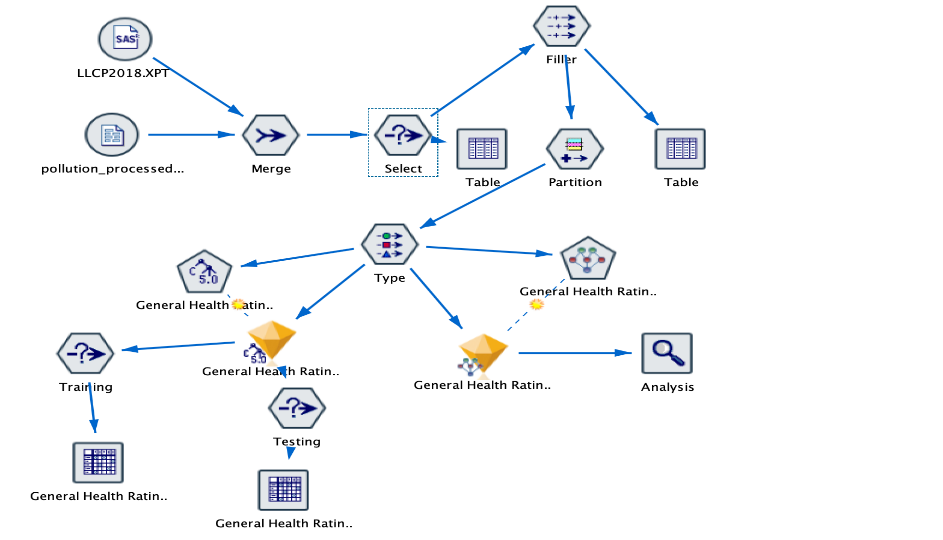
**Figure 5. Decision Tree Confusion Matrices**

A screenshot of a social media post

Description automatically generated**Figure 6. Bayes Net and Decision Tree Predictor Importance Evaluation**

**A screenshot of a cell phone

Description automatically generated**

**Figure 7. Bayes Net and Decision Tree Model Accuracy Evaluation**

**Figure 8. Model Stream of Decision Tree and Bayesian**

**CITATION**

1.     Hoque, M. Z., Alam, M. N., & Nahid, K. A. (2018, September 7). Health Consciousness and Its Effect on Perceived Knowledge, and Belief in the Purchase Intent of Liquid Milk: Consumer Insights from an Emerging Market. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6164144/.

2. The Behavioral Risk Factor Surveillance System (BRFSS) OVERVIEW: BRFSS 2018 . (2019, July 26). Retrieved October 29, 2019, from <https://www.cdc.gov/brfss/annual_data/2018/pdf/overview-2018-508.pdf>.

3.  U.S. Pollution Data: Pollution in the US since 2000. Retrieved November 13, 2019, from <https://www.kaggle.com/sogun3/uspollution>**.**

1. **APPENDIX 1: GROUP MEMBER CONTRIBUTION**

|  |  |
| --- | --- |
| **PROCESS** | **CONTRIBUTION** |
| Data Collection | * BRFSS 2018: found by Jane * Pollution.csv: found by Chris |
| Phase 1: Executive Summary | * Write-up done by all four |
| Phase 2: Presentation | * Introduction: Nick * Problem Statement: Jennie * Methodology: Jane, Chris * Result Analysis: Jennie * Conclusion: Jennie * Appendix: Chris, Jane |
| Data Preprocessing | * Pollution.csv processed by Chris |
| Data Processing | * Processed by Jane & Jennie |
| Model Building | * Decision Tree: Jennie * Bayes Net: Jane |
| Phase 3: Report | * Abstract: Nick * Introduction: Nick * Data: Jane, Chris * Methodology: Jane * Problem Statement: Jennie * Results Analysis: Jennie * Conclusion: Jennie |