Characteristics of Asthma Self-Management on Adults 2016 Behavioral Risk Factor Surveillance System Asthma Call-Back Survey.

ABSTRACT

Asthma is a chronic respiratory disease of our bronchial tubes. People with asthma can develop complication if it is not professionally managed. Our project was to find a way to well controlled asthma to avoid subsequent complications. The data came from CDC BRFSS Asthma Call-Back Survey. It contained 899 variables and 13, 922 cases. The different steps were data exploration, data preparation, and regression. Logistic regression was performed to examine the incidence of predictors on asthma self- management. Regression step resumed by the well-controlled asthma can be predicted at 64.2% with a precision of 65,10%. The model was successful at predicting well controlled

Alain Kuiete Tchoupou, Farhana Zahir, Shovan Biswas, Scott reed, Habib K, Vijaya Cherukuri.

DATA 621 – Business Analytics and Data Mining

Table of Contents

Characteristics of Asthma Self-Management on Adults 2016 Behavioral Risk Factor Surveillance System Asthma Call-Back Survey.

Alain Kuiete Tchoupou, Farhana Zahir, Shovan Biswas, Scott reed, Habib K, Vijaya Cherukuri.

Abstract

Asthma is a chronic respiratory disease of our bronchial tubes. People with asthma can develop complication if it is not professionally managed. Our project was to find a way to well controlled asthma to avoid subsequent complications. The regression method was used to extract the characteristics of an asthma well controlled and to address an asthma poorly controlled. The data came from CDC BRFSS Asthma Call-Back Survey. It contained 899 variables and 13, 922 cases. The different steps were data exploration, data preparation, and regression. The related variables to the subject were selected and explored. There were many steps back and forth used to transform variables to factors, collapsing the factors with many levels, clustering a group of variables on self-management to extract a binary response variable. Logistic regression was performed to examine the incidence of predictors on asthma self-management. Regression step resumed by the well-controlled asthma can be predicted at 64.2% with a precision of 65,10%. The model was successful at predicting well controlled asthma. Low income and people in all ethnicity group could self-manage their asthma as well as high income levels.

Keywords: asthma management, asthma education, adults, self-management, asthma episode, asthma attack.

Introduction

According to American Academy of Asthma Allergy and Immunology (AAAAI)[13], asthma is a chronic disease involving the airways in the lungs. These airways, or bronchial tubes, allow air to come in and out of the lungs. The most common symptom is wheezing. This is a scratchy or whistling sound when you breathe. Other symptoms include, shortness of breath, chest tightness or pain, chronic coughing, trouble sleeping due to coughing or wheezing.

Asthma symptoms, also called asthma flare-ups or asthma attacks, are often caused by allergies and exposure to allergens such as pet dander, dust mites, pollen or mold. Non-allergic triggers include smoke, pollution or frigid air or changes in weather. People with asthma are at risk of developing complications from respiratory infections such as influenza and pneumonia. That is why it is important for asthma sufferers, especially adults, to get vaccinated annually. There is no cure for asthma, but symptoms can be controlled with effective asthma treatment and management. This involves taking your medications as directed and learning to avoid triggers that cause your asthma symptoms. With proper treatment and an asthma management plan, you can minimize your symptoms and enjoy a better quality of life.

Literature Review

We search the web by combining the keywords, asthma, mortality, morbidity, self-management, education, machine learning, logistic regression, data mining, risk factors, etc. Among scholarly journals that we found, we focus on data science project using machine learning, logistic regression,

One of the journals written by Zahran et al [6], attempted to improve the self-management care among persons with asthma. They look for characteristics that procure better education on self-management to people with asthma. The use logistic regression to found that people with asthma episode who regularly reported to the doctor and sometimes get hospitalized, are more likely to receive multiple self-management education components. But old adult with education less than high school diploma who smoke, are less likely to have asthma education. Thakur, et al [11], showed that socioeconomic status is a key factor of predicting asthma. Its effect varied in term of race and ethnicity. They found that "African American children had 23% greater odds of asthma with each decrease in the socioeconomic index (adjusted odds ratios (AOR), 1.23; 95% CI, 1.09–1.38). Conversely, Mexican American children have 17% reduced odds of asthma with each decrease in the socioeconomic index (AOR, 0.83; 95% CI, 0.72–0.96) "(Thakur et al [11]). Our study focused only on adults and we bounded all response variables in one instead of developing one model for each response variable as it had been done in other studies.

Methodology

We were trying to answer these questions in our project.

What were the characteristics of a good asthma management?

How could a bad asthma management be addressed?

Data Exploration

The Exploration of each variable in the data set allowed us to determine if the variable was categorical or numeric, the distribution is skewed, normal or uniform, the correlation between variable was close to 1 and need action. We also look for missing values.

Data Preparation

On the 899 variables in the data set, the variables used for the study were selected base on the BRFSS questionnaire related to asthma education, asthma management or individual with asthma episode. The categories of each variable were shrunk to an acceptable number. The missing values were removed. To validate the model with an unknown data, the data set was split into training and testing set.

The response variable was built with seven variables collected from the BRFSS questionnaire section concerning knowledge of asthma and management plan. A classification using K-means Clustering was conducted to determine whether an individual asthma is well controlled. This implied the analysis of clustering to form a binary response variable.

Logistic Regression Modeling

The models were built on generalized linear model (GLM) associated with stepwise selection, penalized logistic regression with tune parameter, and partial least squared(PLS). Ridge Regression and Lasso Regression were the model used for penalized logistic regression. The performance of each model was measure using AIC (Akaike Information Criteria), AUC (Area Under the Curve), accuracy, precision, specificity, sensitivity, F1 score, MSE (Mean Square Error) with Lasso and Ridge Regression.

The best model predicted whether an individual asthma was well controlled base on response he gave to the questionnaire. The characteristics of well controlled asthma could be extract. Advices for not well controlled asthma could be given.

Experimentation and Results

Data Exploration

The investigation of the BRFSS/ASTHMA SURVEY ADULT QUESTIONNAIRE – 2016 allowed us to select variables related to asthma education, management, and action plan base on the suggestion of Zahran et al. [6]. Each variable contained a certain number of categories corresponding to the type of response given by the participant of Behavior Risk Factor Surveillance System (BFRSS) Asthma Call-Back Survey. For example, with the question:

"Have you ever taken a course or class on how to manage your asthma?" The value was one of the following: 1 = YES, 2 = NO, 7 = DON'T KNOW, 9 = REFUSED. All numeric variables were transformed to categorical variables.

The response variables were selected from the section Knowledge of asthma and management plan. These variables explained whether the individual asthma was well controlled. The predictors related to asthma and management plan were selected in different sections. There were section concerning:

- Recent history on individual asthma management. The variables stated whether in the recent past the individual knew how to manage asthma.
- History of asthma. The variables explained how the individual handled symptoms and episodes of asthma attack in the past year.
- Health care utilization. The variables explained if the individual used insurance and visits hospital because of asthma.
- Modification of environment. The variables explained weather the individual was educated on how to modify his environment for better live.

All the variables have been collected in the table 1 below:

Table 1: Data Dictionary

Variable	Description	Comment
TCH_SIGN	Has a doctor or other health professional ever taught you how to recognize early signs or symptoms of an asthma episode?	Response Variable Management Plan Knowledge of Asthma
TCH_RESP	Has a doctor or other health professional ever taught you what to do during an asthma episode or attack?	Response Variable Management Plan Knowledge of Asthma
TCH_MON	A peak flow meter is a handheld device that measures how quickly you can blow air out of your lungs. Has a doctor or other health professional ever taught you how to use a peak flow meter to adjust your daily medications?	Response Variable Management Plan Knowledge of Asthma
MGT_PLAN	An asthma action plan, or asthma management plan, is a form with instructions about when to change the amount	Response Variable Management Plan Knowledge of Asthma

	or type of medicine, when to call the doctor for advice, and when to go to the emergency room. Has a doctor or other health professional EVER given you an asthma action plan?	
MOD_ENV	Now, back to questions specifically about you. Has a health professional ever advised you to change things in your home, school, or work to improve your asthma	Response Variable Management Plan
MGT_CLAS	Have you ever taken a course or class on how to manage your asthma?	Response Variable Management Plan Knowledge of Asthma
INHALERW	Did a doctor or other health professional watch you use the inhaler	Response Variable Management Plan
INCIDNT	How long ago was that when you were first told by a doctor or other health professional that you had asthma?	Explanatory Variable Knowledge of Asthma
LAST_MD	How long has it been since you last talked to a doctor or other health professional about your asthma? This could have been in your doctor's office, the hospital, an emergency room or urgent care center.	Explanatory Variable Recent History Knowledge of Asthma
LAST_MED	How long has it been since you last took asthma medication?	Explanatory Variable Recent History
LASTSYMP	Symptoms of asthma include coughing, wheezing, shortness of breath, chest tightness or phlegm production when you do not have a cold or respiratory infection. How long has it been since you last had any symptoms of asthma?	Explanatory Variable Recent History Knowledge of Asthma
DUR_30D	Do you have symptoms all the time? "All the time" means symptoms that continue throughout the day. It does not mean symptoms for a little while each day.	History of Asthma
EPIS_12M	Asthma attacks, sometimes called episodes, refer to periods of worsening asthma symptoms that make you limit your activity more than you usually do, or make you seek medical care. During the past 12 months, have you had an episode of asthma or an asthma attack?	Symptom and Episode in Past Year
COMPASTH	Compared with other episodes or attacks, was this most recent attack shorter, longer, or about the same?	Symptom and Episode in Past Year
INS1	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare or Medicaid?	Health Care Utilization
INS2	During the past 12 months was there any time that you did not have any health insurance or coverage?	Health Care Utilization

ER_VISIT	An urgent care center treats people with illnesses or injuries that must be addressed immediately and cannot wait for a regular medical appointment. During the past 12 months, have you had to visit an emergency room or urgent care center because of your asthma?	Health Care Utilization
HOSP_VST	During the past 12 months, that is since [1 YEAR AGO TODAY], have you had to stay overnight in a hospital because of your asthma? Do not include an overnight stay in the emergency room.	Health Care Utilization
ACT_DAYS30	During just the past 30 days, would you say you limited your usual activities due to asthma not at all, a little, a moderate amount, or a lot?	Health Care Utilization
ASMDCOST	Was there a time in the past 12 months when you needed to see your primary care doctor for your asthma but could not because of the cost) Was there a time in the past 12 months when you were referred to a specialist for asthma care but could not go because of the cost?	Cost of Care
ASRXCOST	Was there a time in the past 12 months when you needed to buy medication for your asthma but could not because of the cost?	Cost of Care
WORKENV5	Things in the workplace such as chemicals, smoke, dust or mold can make asthma symptoms worse in people who already HAVE asthma or can actually CAUSE asthma in people who have never had asthma before. Are your asthma symptoms MADE WORSE by things like chemicals, smoke, dust or mold in your CURRENT job?	Work Related Asthma
WORKENV6	“Some examples of things in the workplace that may cause asthma or make asthma symptoms worse include: flour dust in a bakery, normal dust in an office, smoke from a manufacturing process, smoke from a co-worker’s cigarette, cleaning chemicals in a hospital, mold in a basement classroom, a coworker’s perfume, or mice in a research laboratory.” Was your asthma first CAUSED by things like chemicals, smoke, dust or molding your CURRENT job?	Work Related Asthma
WORKENV7	Were your asthma symptoms MADE WORSE by things like chemicals, smoke, dust or mold in any PREVIOUS job you ever had?	Work Related Asthma
WORKENV8	Was your asthma first CAUSED by things like chemicals, smoke, dust or molding any PREVIOUS job you ever had?	Work Related Asthma
WORKQUIT1	“Some examples of things in the workplace that may cause asthma or make asthma symptoms worse include flour dust in a bakery, normal dust in an office, smoke from a manufacturing process, smoke from a co-worker’s cigarette, cleaning chemicals in a hospital, mold in a	Work Related Asthma

	basement classroom, a coworker's perfume, or mice in a research laboratory." Did you ever lose or quit a job because things in the workplace, like chemicals, smoke, dust or mold, caused your asthma or made your asthma symptoms worse?	
WORKTALK	Did you and a doctor or other health professional ever DISCUSS whether your asthma could have been caused by, or your symptoms made worse by, any jobyou ever had?	Work Related Asthma
WORKSEN3	Have you ever been TOLD BY a doctor or other health professional that your asthma was caused by, or your symptoms made worse by, any job you ever had?	Work Related Asthma
WORKSEN4	Have YOU ever TOLD a doctor or other health professional that your asthma was caused by, or your symptoms made worse by, any job you ever had?	Work Related Asthma
. COPD	Have you ever been told by a doctor or health professional that you have chronic obstructive pulmonary disease also known as COPD?	Comorbid Conditions
EMPHY	Have you ever been told by a doctor or other health professional that you have emphysema?	Comorbid Conditions
BRONCH	Have you ever been told by a doctor or other health professional that you have Chronic Bronchitis?	Comorbid Conditions
DEPRESS	Chronic Bronchitis is repeated attacks of bronchitis over a long period of time. Chronic Bronchitis is not the type of bronchitis you might get occasionally with a cold. Have you ever been told by a doctor or other health professional that you were depressed?	Comorbid Conditions

Data Preparation

Cleaning

Categorized the Variables All the variables were turned from numeric type to factor in R.

Collapsing Number of Categories in Variables. Some that classes appeared in the testing set but not in the training set were either collapsed in the corresponding variable or removed. Variables with large among of classes were removed from the dataset, it also appears that those variables were mixed types.

Table 2. Summary of the dataset before categorizing the variables

	TCH.SIGN	TCH.RESP	TCH.MON	MGT.PLAN	MGT.CLAS	INHALERW	MOD.ENV	SEX
Min.	1	1	1	1	1	1	1	1
1st Qu.	1	1	1	1	2	1	1	1
Median	1	1	2	2	2	1	2	2
Mean	1.493	1.403	1.671	1.853	1.936	1.796	1.725	1.656
3rd Qu.	:2.000	:2.000	:2.000	:2.000	:2.000	:2.000	:2.000	:2.000
Max.	:9.000	:9.000	:9.000	:9.000	:9.000	:9.000	:9.000	:2.000
	NA	NA	NA	NA	NA	NA	NA	NA

	AGEG.F7	X_RACEGR3	EDUCAL	X_INCOMG	X_RFBMI5	SMOKE100	COPD	EMPHY
Min.	1	1	1	1	1	1	1	1
1st Qu.	4	1	4	2	1	1	2	2
Median	5	1	5	4	2	2	2	2
Mean	4.645	1.672	4.965	4.107	2.058	1.554	1.868	1.959
3rd Qu.	:6.000	:1.000	:6.000	:5.000	:2.000	:2.000	:2.000	:2.000
Max.	:7.000	:9.000	:9.000	:9.000	:9.000	:9.000	:9.000	:9.000
	NA	NA	NA	NA	NA	NA	NA's :44	NA's :44

	DEPRESS	BRONCH	DUR.30D	INCINDT	LAST.MD	LAST.ME D	LAST.SYM P	EPIS.12M	COMPAS TH
Min.	1	1	1	1	4	1	1	1	1
1st Qu.	1	1	6	3	4	1	1	1	3
Median	2	2	10	3	4	3	3	2	6
Mean	1.654	1.803	9.263	2.9	6.504	6.495	5.436	2.921	6.233
3rd Qu.	:2.000	:2.000	:12.000	:3.0	: 6.000	: 7.000	: 6.000	:6.000	:11.000
Max.	:9.000	:9.000	:99.000	:9.0	:99.000	:99.000	:99.000	:9.000	:11.000
	NA's :44	NA's :44	NA	NA	NA	NA	NA	NA	NA

	INS1	INS2	ER.VISI T	HOSP.V ST	ASMDC OST	ASRXC OST	ASSPC OST	WORKT ALK	ACT.DA Y30
Min.	1	1	1	1	1	1	1	1	1
1st Qu.	1	2	2	2	2	2	2	2	1
Median	1	2	2	2	2	2	2	2	2
Mean	1.071	2.138	3.397	3.602	2.551	2.49	2.565	1.965	2.415
3rd Qu.	:1.000	:2.000	:5.000	:5.000	:2.000	:2.00	:2.000	:2.000	:4.000
Max.	:9.000	:9.000	:9.000	:7.000	:9.000	:9.00	:9.000	:9.000	:9.000
	NA	NA	NA	NA	NA's :5	NA's :5	NA's :5	NA's :12	NA

There were few variables with missing values. As expected, the minimum of each variable was 1. The maximum was either 9 or 99. The exception was the variable SEX that had two values 1 and 2.

Table 7. Structure of Data Before Turning to Categorical Data

'data.frame': 11494 obs. of 33 variables:
$ TCH.SIGN : num 1 2 1 2 2 1 1 2 1 2 ...
..- attr(*, "label")= chr "EVER TAUGHT RECOGNIZE EARLY SIGN OR SYMPTOMS"
..- attr(*, "format.sas")= chr "TCH_SIGN"
$ TCH.RESP : num 1 1 1 2 1 1 1 1 1 1 ...
..- attr(*, "label")= chr "EVER TAUGHT WHAT TO DO DURING ASTHMA EPISODE OR ATTACK"
..- attr(*, "format.sas")= chr "TCH_RESP"
$ TCH.MON : num 2 2 2 2 2 1 1 2 2 2 ...
..- attr(*, "label")= chr "EVER TAUGHT HOW TO USE A PEAK FLOW"
..- attr(*, "format.sas")= chr "TCH_MON"
$ MGT.PLAN : num 2 2 2 2 2 2 1 2 2 2 ...
..- attr(*, "label")= chr "EVER GIVEN AN ASTHMA ACTION PLAN"
..- attr(*, "format.sas")= chr "MGT_PLAN"
$ MGT.CLAS : num 2 2 2 2 2 2 2 2 2 2 ...
..- attr(*, "label")= chr "EVER TAKEN A COURSE TO MANAGE ASTHMA"
..- attr(*, "format.sas")= chr "MGT_CLAS"
$ INHALERW : num 2 2 1 1 1 1 1 1 1 1 ...
..- attr(*, "label")= chr "INHALER USE WATCHED"
..- attr(*, "format.sas")= chr "INHALERW"
$ MOD.ENV : num 2 2 2 2 1 2 2 2 1 2 ...
..- attr(*, "label")= chr "EVER ADVISED CHANGE THINGS IN YOUR HOME"
..- attr(*, "format.sas")= chr "MOD_ENV"
$ SEX : num 1 2 2 2 2 2 1 2 2 2 ...
..- attr(*, "label")= chr "RESPONDENTS SEX"
..- attr(*, "format.sas")= chr "SEX"
$ AGEG.F7 : num 4 5 5 3 6 5 4 6 6 7 ...
..- attr(*, "label")= chr "AGE COLLAPSED TO 7 GROUPS FOR ASTHMA CALL-BACK"
..- attr(*, "format.sas")= chr "AGEG_F7Z"
$ X_RACEGR3: num 3 1 1 5 1 5 1 1 1 1 ...
..- attr(*, "label")= chr "COMPUTED FIVE LEVEL RACE/ETHNICITY CATEGORY."
..- attr(*, "format.sas")= chr "_3RACEGR"
$ EDUCAL : num 6 4 4 5 6 6 6 6 6 5 ...
..- attr(*, "label")= chr "EDUCATION LEVEL"
..- attr(*, "format.sas")= chr "EDUCA"
$ X_INCOMG : num 5 1 1 5 5 5 5 5 3 9 ...
..- attr(*, "label")= chr "COMPUTED INCOME CATEGORIES"
..- attr(*, "format.sas")= chr "_INCOMG"
$ X_RFBMI5 : num 2 2 2 2 2 2 1 2 2 1 ...
..- attr(*, "label")= chr "OVERWEIGHT OR OBESE CALCULATED VARIABLE"
..- attr(*, "format.sas")= chr "_5RFBMI"
$ SMOKE100 : num 2 1 1 2 1 2 1 1 2 2 ...
..- attr(*, "label")= chr "SMOKED AT LEAST 100 CIGARETTES"
..- attr(*, "format.sas")= chr "SMOK100_"
$ COPD : num 2 1 2 2 2 2 2 2 2 1 ...
..- attr(*, "label")= chr "EVER TOLD HAVE CHRONIC OBSTRUCTIVE PULMONARY DISEASE"
..- attr(*, "format.sas")= chr "COPD"
$ EMPHY : num 2 2 2 2 2 2 2 2 2 2 ...
..- attr(*, "label")= chr "EVER TOLD HAVE EMPHYSEMA"
..- attr(*, "format.sas")= chr "EMPHY"
$ DEPRESS : num 2 1 2 2 2 2 2 2 1 1 ...
..- attr(*, "label")= chr "EVER TOLD DEPRESSED"
..- attr(*, "format.sas")= chr "DEPRESS"
$ BRONCH : num 2 1 2 2 1 2 2 2 1 2 ...
..- attr(*, "label")= chr "EVER TOLD HAVE CHRONIC BRONCHITIS"
..- attr(*, "format.sas")= chr "BRONCH"
$ DUR.30D : num 10 2 12 6 12 10 12 6 1 6 ...
..- attr(*, "label")= chr "CONSTANT SYMPTOMS"

..- attr(*, "format.sas")= chr "DUR_30D"
$ INCINDT : num 3 2 3 3 3 2 3 3 3 3 ...
..- attr(*, "label")= chr "TIME SINCE DIAGNOSIS"
..- attr(*, "format.sas")= chr "INCIDNT"
$ LAST.MD : num 5 4 4 7 4 4 4 5 4 5 ...
..- attr(*, "label")= chr "LAST TALKED TO A DOCTOR"
..- attr(*, "format.sas")= chr "LAST_MD"
$ LAST.MED : num 4 1 3 7 3 1 1 6 1 5 ...
..- attr(*, "label")= chr "LAST TOOK ASTHMA MEDICATION"
..- attr(*, "format.sas")= chr "LAST_MED"
$ LAST.SYMP: num 4 1 3 7 3 4 3 5 1 5 ...
..- attr(*, "label")= chr "LAST HAD ANY SYMPTOMS OF ASTHMA"
..- attr(*, "format.sas")= chr "LASTSYMP"
$ EPIS.12M : num 1 1 1 6 1 2 1 6 2 6 ...
..- attr(*, "label")= chr "ASTHMA EPISODE OR ATTACK"
..- attr(*, "format.sas")= chr "EPIS_12M"
$ COMPASTH : num 1 3 1 6 3 11 3 6 11 6 ...
..- attr(*, "label")= chr "TYPICAL ATTACK"
..- attr(*, "format.sas")= chr "COMPASTH"
$ INS1 : num 1 1 1 2 1 1 1 1 1 1 ...
..- attr(*, "label")= chr "INSURANCE"
..- attr(*, "format.sas")= chr "INS1Z"
$ INS2 : num 2 2 2 5 2 2 2 2 2 2 ...
..- attr(*, "label")= chr "INSURANCE OR COVERAGE GAP"
..- attr(*, "format.sas")= chr "INS2Z"
$ ER.VISIT : num 6 2 2 5 2 2 2 5 2 6 ...
..- attr(*, "label")= chr "EMERGENCY ROOM VISIT"
..- attr(*, "format.sas")= chr "ER_VISIT"
$ HOSP.VST : num 6 2 2 5 2 2 2 5 2 6 ...
..- attr(*, "label")= chr "HOSPITAL VISIT"
..- attr(*, "format.sas")= chr "HOSP_VST"
$ ASMDCOST : num 2 2 2 5 2 2 2 5 2 2 ...
..- attr(*, "label")= chr "COST BARRIER: PRIMARY CARE DOCTOR"
..- attr(*, "format.sas")= chr "ASMDCOST"
$ ASRXCOST : num 2 2 2 5 2 2 2 5 1 2 ...
..- attr(*, "label")= chr "COST BARRIER: MEDICATION"
..- attr(*, "format.sas")= chr "ASRXCOST"
$ ASSPCOST : num 2 2 2 5 2 2 2 5 2 2 ...
..- attr(*, "label")= chr "COST BARRIER: SPECIALIST"
..- attr(*, "format.sas")= chr "ASSPCOST"
$ WORKTALK : num 2 2 2 2 2 2 2 2 2 2 ...
..- attr(*, "label")= chr "DOCTOR DISCUSSED WORK ASTHMA"
..- attr(*, "format.sas")= chr "WORKTALK"

Each variable name was followed by its definition.

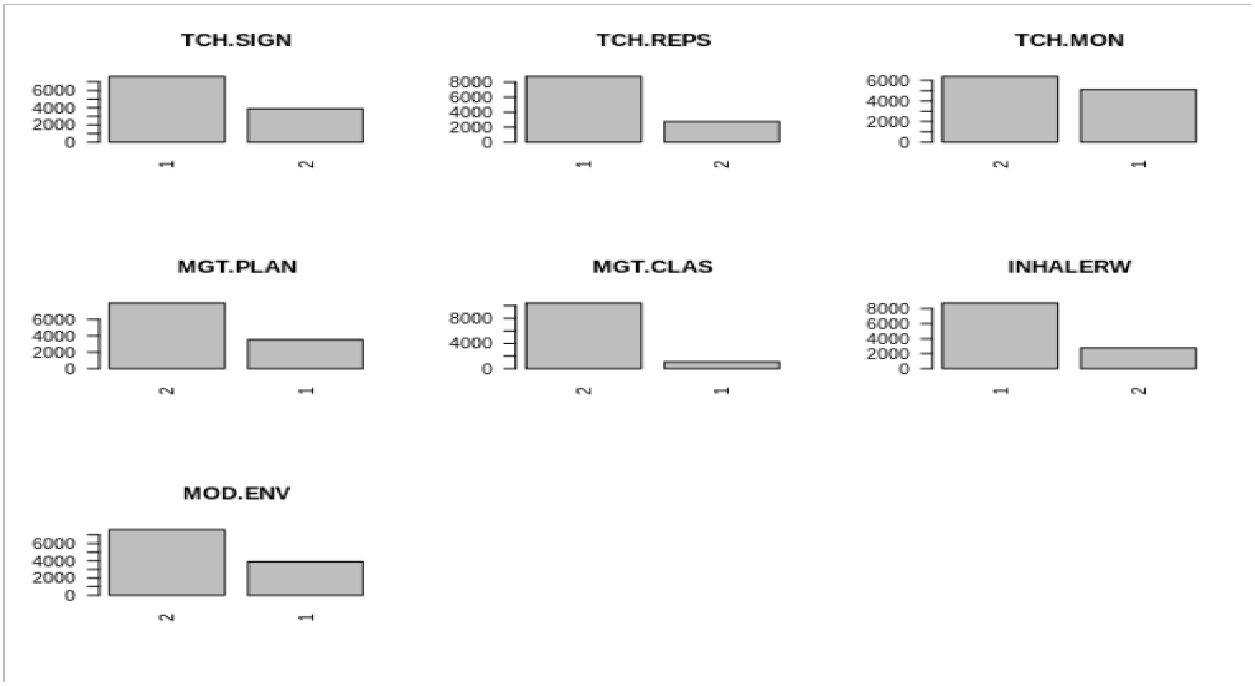
Table 9 Structure of the Variables after Turning to Categorical Variable

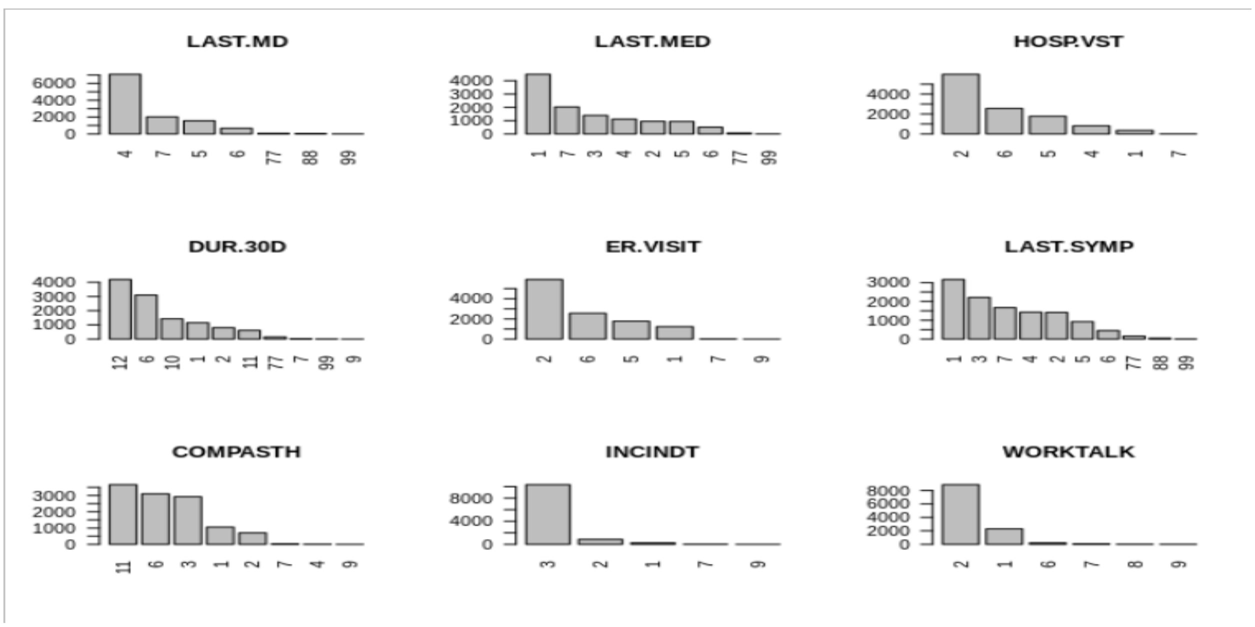
'data.frame': 13922 obs. of 33 variables:
$ TCH.SIGN : Factor w/ 4 levels "1","2","7","9": 1 2 1 2 2 1 2 1 2 1 ...
$ TCH.RESP : Factor w/ 4 levels "1","2","7","9": 1 1 1 2 1 1 2 1 1 1 ...
$ TCH.MON : Factor w/ 4 levels "1","2","7","9": 2 2 2 2 2 1 2 1 2 2 ...
$ MGT.PLAN : Factor w/ 4 levels "1","2","7","9": 2 2 2 2 2 2 2 1 2 2 ...
$ MGT.CLAS : Factor w/ 4 levels "1","2","7","9": 2 2 2 2 2 2 2 2 2 2 ...
$ INHALERW : Factor w/ 6 levels "1","2","5","6",..: 2 2 1 1 1 1 3 1 1 1 ...
$ MOD.ENV : Factor w/ 4 levels "1","2","7","9": 2 2 2 2 1 2 2 2 2 1 ...
$ SEX : Factor w/ 2 levels "1","2": 1 2 2 2 2 2 1 1 2 2 ...
$ AGEG.F7 : Factor w/ 7 levels "1","2","3","4",..: 4 5 5 3 6 5 5 4 6 6 ...

$ X_RACEGR3: Factor w/ 6 levels "1","2","3","4",..: 3 1 1 5 1 5 1 1 1 1 ...
$ EDUCAL : Factor w/ 6 levels "1","2","3","4",..: 6 4 4 5 6 6 6 6 6 6 ...
$ X_INCOMG : Factor w/ 6 levels "1","2","3","4",..: 5 1 1 5 5 5 5 5 5 3 ...
$ X_RFBMI5 : Factor w/ 3 levels "1","2","9": 2 2 2 2 2 2 2 1 2 2 ...
$ SMOKE100 : Factor w/ 4 levels "1","2","7","9": 2 1 1 2 1 2 2 1 1 2 ...
$ COPD : Factor w/ 4 levels "1","2","7","9": 2 1 2 2 2 2 2 2 2 2 ...
$ EMPHY : Factor w/ 4 levels "1","2","7","9": 2 2 2 2 2 2 2 2 2 2 ...
$ DEPRESS : Factor w/ 4 levels "1","2","7","9": 2 1 2 2 2 2 2 2 2 1 ...
$ BRONCH : Factor w/ 4 levels "1","2","7","9": 2 1 2 2 1 2 2 2 2 1 ...
$ DUR.30D : Factor w/ 7 levels "1","10","11",..: 2 5 4 6 4 2 1 4 6 1 ...
$ INCINDT : Factor w/ 4 levels "1","2","3","7": 3 2 3 3 3 2 3 3 3 3 ...
$ LAST.MD : Factor w/ 5 levels "4","5","6","7",..: 2 1 1 4 1 1 3 1 2 1 ...
$ LAST.MED : Factor w/ 5 levels "4","5","6","7",..: 4 1 3 4 3 1 5 1 3 1 ...
$ LAST.SYMP: Factor w/ 8 levels "1","2","3","4",..: 4 1 3 7 3 4 2 3 5 1 ...
$ EPIS.12M : Factor w/ 4 levels "1","2","6","7": 1 1 1 3 1 2 1 1 3 2 ...
$ COMPASTH : Factor w/ 6 levels "1","11","2","3",..: 1 4 1 5 4 2 4 4 5 2 ...
$ INS1 : Factor w/ 4 levels "1","2","7","9": 1 1 1 2 1 1 1 1 1 1 ...
$ INS2 : Factor w/ 5 levels "1","2","5","7",..: 2 2 2 3 2 2 2 2 2 2 ...
$ ER.VISIT : Factor w/ 5 levels "1","2","5","6",..: 4 2 2 3 2 2 4 2 3 2 ...
$ HOSP.VST : Factor w/ 6 levels "1","2","4","5",..: 5 2 2 4 2 2 5 2 4 2 ...
$ ASMDCOST : Factor w/ 5 levels "1","2","5","7",..: 2 2 2 3 2 2 2 2 3 2 ...
$ ASRXCOST : Factor w/ 5 levels "1","2","5","7",..: 2 2 2 3 2 2 2 2 3 1 ...
$ ASSPCOST : Factor w/ 5 levels "1","2","5","7",..: 2 2 2 3 2 2 2 2 3 2 ...
$ WORKTALK : Factor w/ 6 levels "1","2","6","7",..: 2 2 2 2 2 2 2 2 2 2 ...

For the sake of having same levels of one variable in the training and testing set, the levels of factors were reduced to a maximum of 8. On 33 variables there were 14 with 4 levels, 7 with 5 levels, 7 with 6 levels, 2 with 7 levels, 1 with 8 levels, 1 with 3 levels, and 1 with 2 levels.

The figures below gave the frequency of each level in the variable. The levels of the response variables were reduced to 2.


TCH.SIGN
0
2000
4000
6000
1
2
TCH.REPS
0
2000
4000
6000
8000
1
2
TCH.MON
0
2000
4000
6000
2
1
MGT.PLAN
0
2000
4000
6000
2
1
MGT.CLAS
0
4000
8000
2
1
INHALERW
0
2000
4000
6000
8000
1
2
MOD.ENV
0
2000
4000
6000
2
1


LAST.MD
0
2000
4000
6000
4 7 5 6 77 88 99
LAST.MED
0
1000
2000
3000
4000
1 7 3 4 2 5 6 77 99
HOSP.VST
0
2000
4000
2 6 5 4 1 7
DUR.30D
0
1000
2000
3000
4000
12 6 10 1 2 11 77 7 99 9
ER.VISIT
0
2000
4000
2 6 5 1 7 9
LAST.SYMP
0
1000
2000
3000
1 3 7 4 2 5 6 77 88 99
COMPASTH
0
1000
2000
3000
11 6 3 1 2 7 4 9
INCINDT
0
4000
8000
3 2 1 7 9
WORKTALK
0
2000
4000
6000
8000
2 1 6 7 8 9

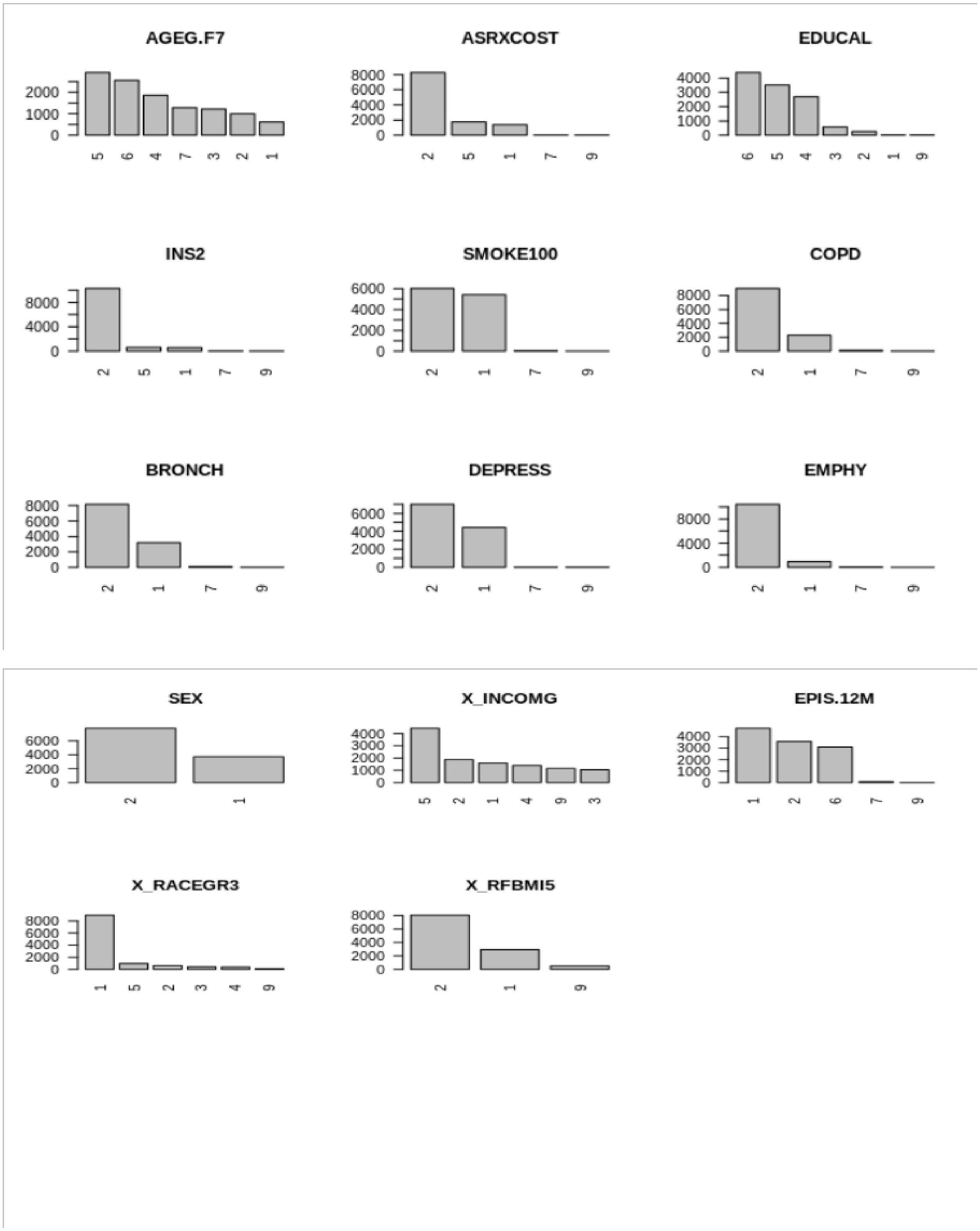


Figure 1-30. Frequency of Categories in each Variable

Correlation Between Variables

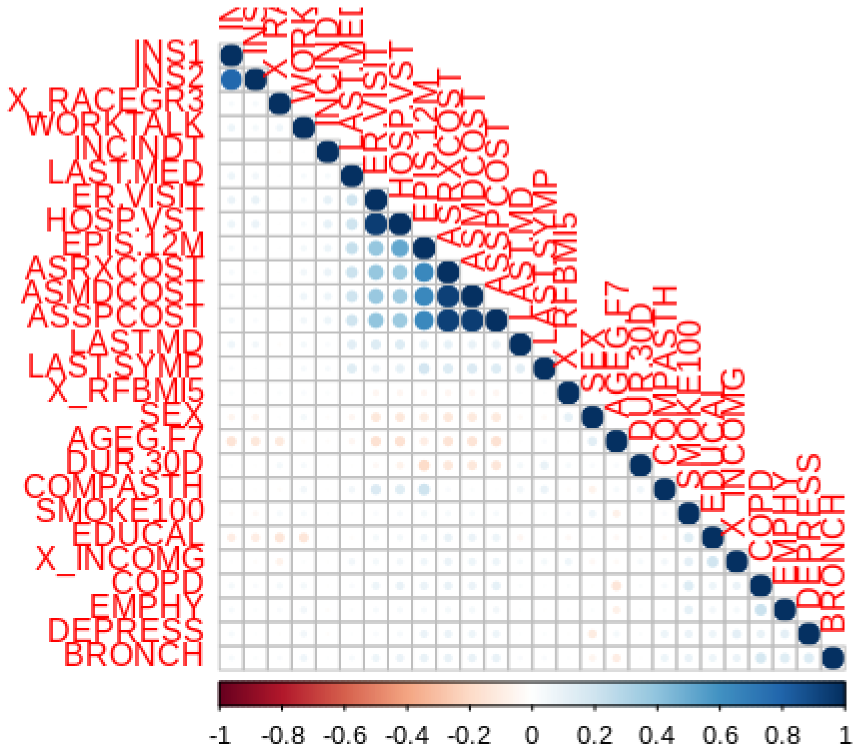


Figure 35. Correlation Matrix

The matrix correlation showed that the group of variables ASRXCOST, ASMDCOST, ASSPCOST, were highly correlated. Only one variable of the three remained in the final data frame. For the same raison other variables were removed from the data set.

Clustering Analysis

Select the Response Variables to Cluster. In the BFRSS questionnaire, the variables related to asthma education, asthma self-management, knowledge of asthma, and management plan were selected as part of response variable or further clustering.

Analyze the Clustering. There are seven clusters that correspond to each level of asthma management skill. The final response variable was built by taken clusters 2 and 3 corresponding to asthma well controlled as Yes and the rest of clusters as NO. But it gave bad response on the confusion matrix. Particularly, the sensitivity was null for most of model. The clusters 1 and 4 corresponding to asthma not well controlled were added to change the metrics. There was significant amelioration in True positive of the confusion matrix.

The figures (37-46) below, were used to explore the clustering. The elbow method determined the number of clusters with the Scree plot. The next figures visualized the proportion of yes or no of one variable in each cluster.

Distribution of different response variables in each cluster

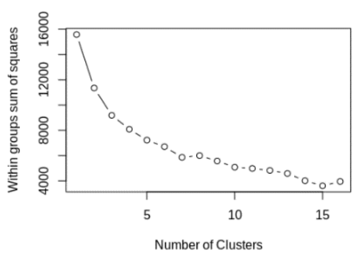


Fig 37. Scree Plot

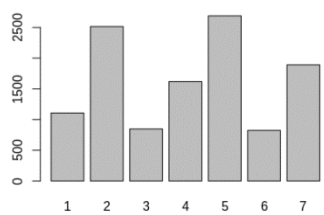


Fig 38. Clusters

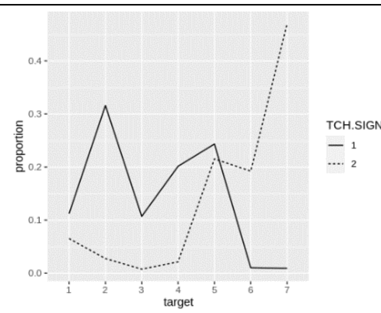


Fig. 39 Proportion of TCH.SIGN in each Cluster

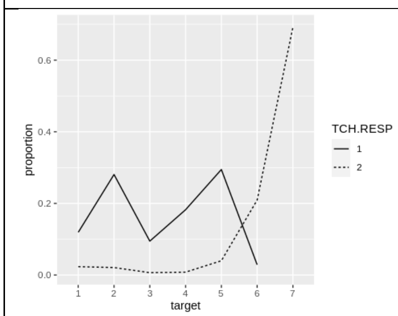


Fig. 40 Proportion of TCH.RESP in each Cluster

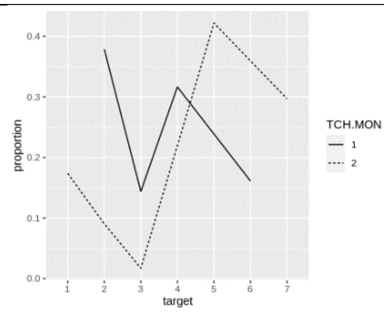


Fig. 41 Proportion of TCH.MON in each Cluster

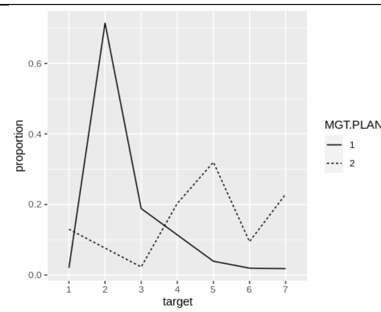


Fig. 42 Proportion of MGT.PLAN in each Cluster

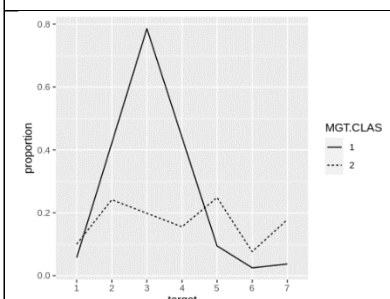


Fig. 43 Proportion of MGT.CLAS in each Cluster

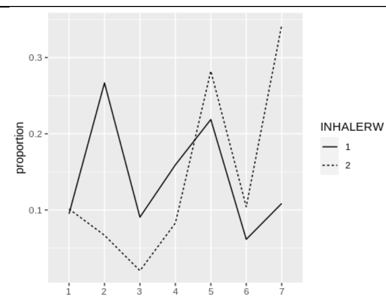


Fig. 44 Proportion of INHALERW in each Cluster

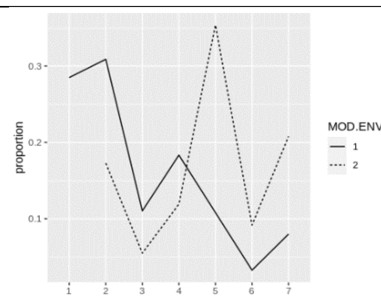


Fig. 46 Proportion of MOD.ENV in each Cluster

Fig.37-46 Clustering results

In each cell in the table below, the number of cases of YES and NO were compared and the response with the higher number of cases were selected.

Table 11. Deduct the Binary response Variable

	RESPONSE	TCH.SIGN	TCH.RES	TCH.MON	MGT.PLAN	MGT.CLAS	INHALERW	MOD.ENV
1	1=YES	2	5	2	2	3	2	2
2	2=NO	7	7	5	5	5	7	5

The clusters 2 and 3 are chosen as well controlled asthma.

Table 13. Table Interpretation of Clusters

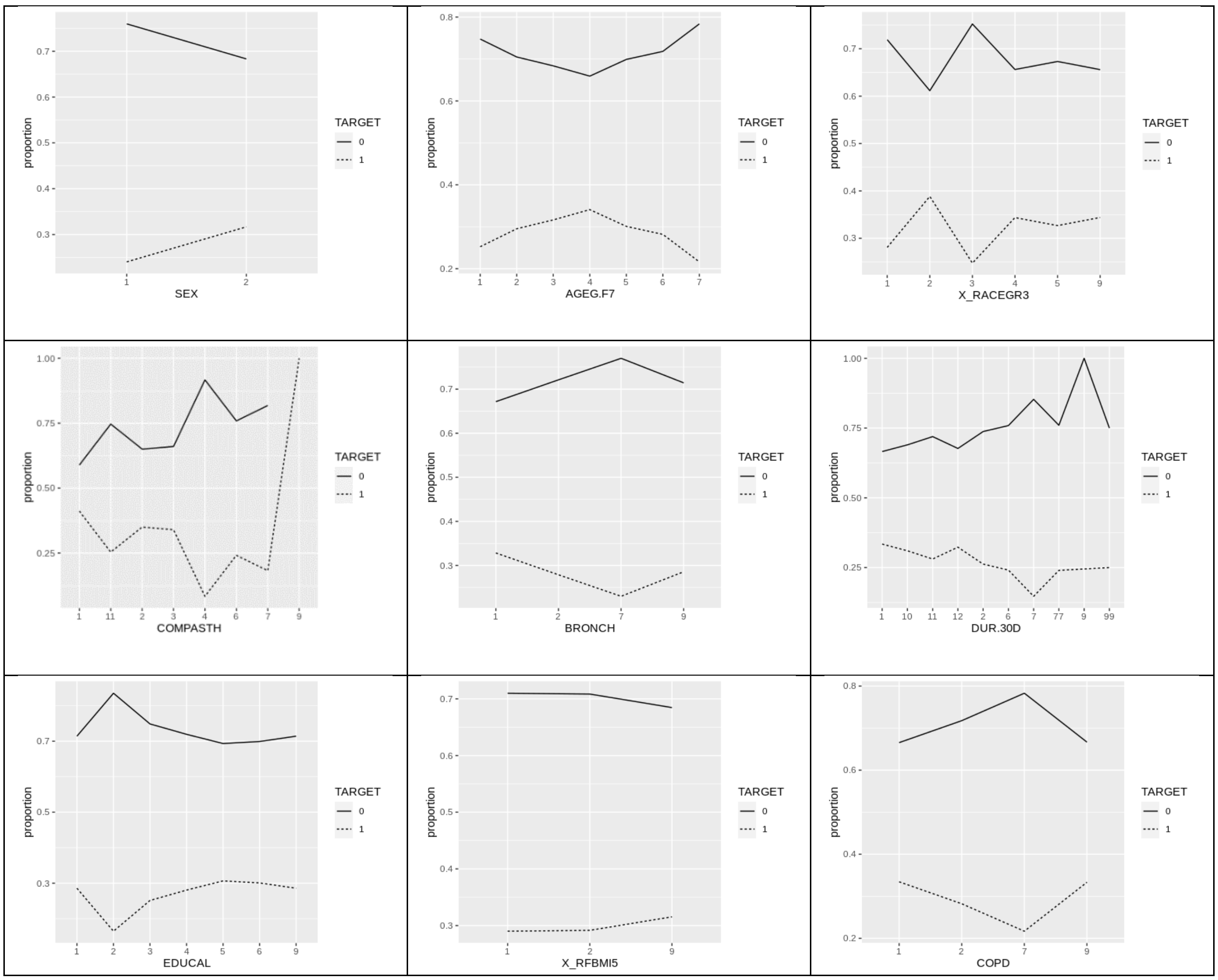
Cluster	TCH. SIGN	TCH. RES	TCH. MON	MGT. PLAN	MGT. CLAS	INHAL ERW	MOD. ENV	NUM YES	%	Asthma Management Level
1	YES	YES	NO	NO	NO	YES	YES	4	8.16	Not Well Controlled
2	YES	YES	YES	YES	YES	YES	NO	6	12.2	Well Controlled
3	YES	YES	YES	YES	YES	YES	YES	7	14.3	Very Well Controlled
4	YES	YES	YES	NO	NO	YES	NO	4	8.16	Not Well Controlled
5	YES	YES	NO	NO	NO	YES	NO	3	6.12	Poorly Controlled
6	NO	NO	YES	NO	NO	YES	NO	2	4.08	Poorly Controlled
7	NO	NO	NO	NO	NO	YES	NO	1	2.04	Very Poorly Controlled
NUM YES	5	5	4	3	2	7	2			
%	10.2	10.2	8.16	6.12	4.08	14.3	2.04	55.1	57.1	

The contribution or weight of the YES response of each response variable to the final response variable were calculated and interpreted as follow.

On the response variable ,14.3% had been watched by a health professional using an inhaler (INHALERW). 10.2% had been taught how to recognize and how to do in case of asthma episode (TCH.RESP). 8.16% had taught how to use peak flow (TCH.MON). 6.12% had received an action management plan (MGT.PLAN). 4.08%

had taken any class or course on how to manage asthma (MGT.CLAS). 2.04% had been asked to a health care professional to modify the environment to ameliorate asthma condition (MOD.ENV). The YES response contributed for 55.1% in the final response. The YES response also contributed for 57.1 % in the clustering. Based on the result above, the clusters 1, and 4 were added to the asthma well controlled

The figures (47-65) below examined the relationship between the TARGET variable and some predictors. It gave the proportion of YES or NO of the TARGET in term of factors in the predictor



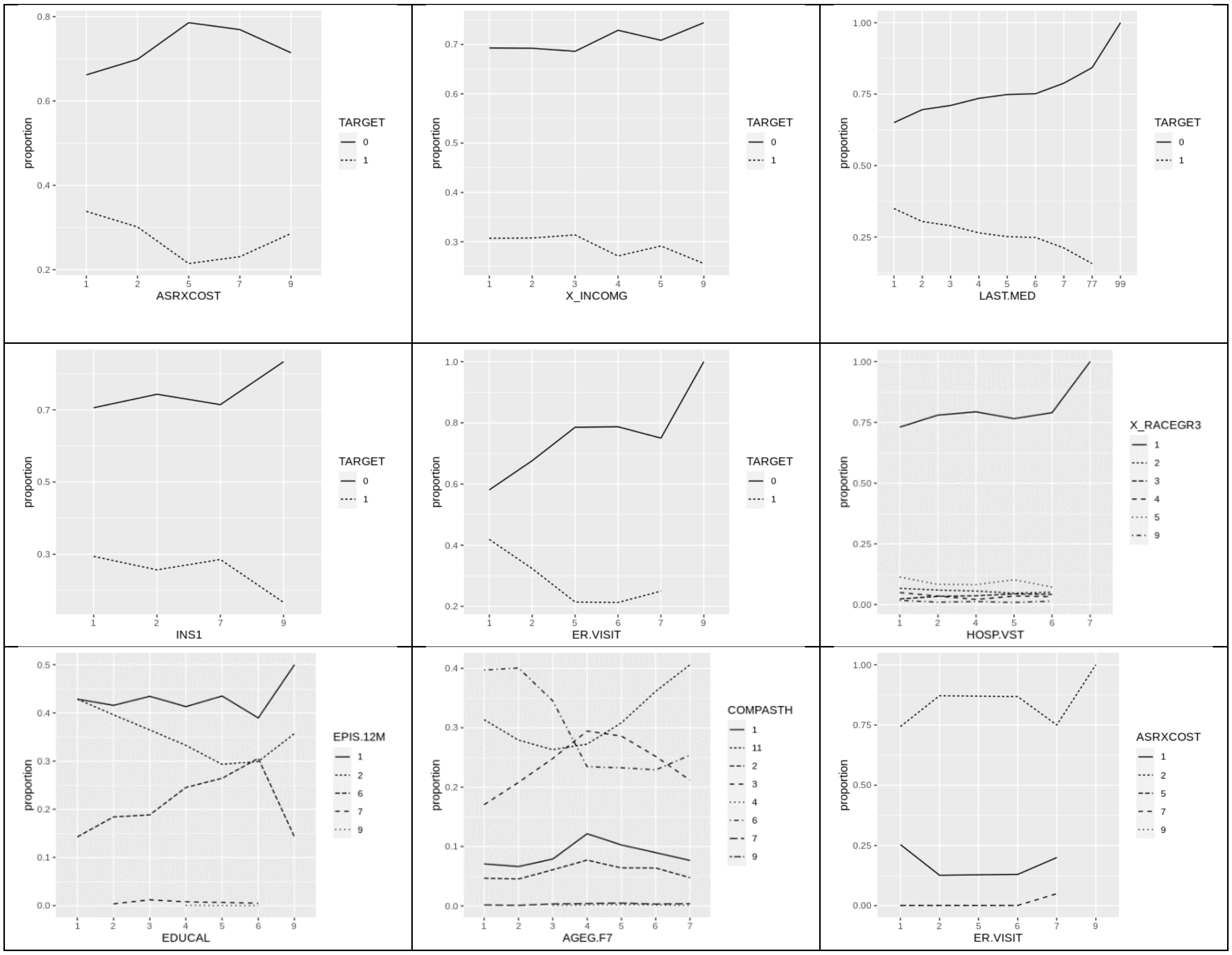


Figure 47-65. Relationship between the response variable-predictor, and predictor-predictor

Logistic Regression

Best Models

Two models were built base on generalized linear model and step wise selection was performed to select the model with the lowest AIC. Four models were built on penalized least squared. Two were on ridge regression model and two others were on lasso regression. The shrinkages parameter lambda was used to tune the models in the sake of lambda min and best lambda that minimized the cross-validation error. There were also two models on partial least squared. The best model was selected by tuning parameters, preprocessing with scaling and centering the predictors, and using the metric ROC for model selection.

To compare the models obtained, confusion matrices were built to extract the accuracy, precision, sensitivity, specificity, F1 score, and AUC. These metrics were run with training and testing sets.

Table 17. Metrics with the Training Set

	glm.train11.	glm.train12	ridge.train1	ridge.train2	lasso.train1	lasso.train2	pls.train1	pls.train2
Accuracy	0.641	0.640	0.615	0.587	0.642	0.634	0.640	0.639
Precision	0.652	0.650	0.746	0.767	0.652	0.639	0.651	0.650
Sensitivity	0.688	0.690	0.414	0.317	0.695	0.709	0.689	0.690
Specificity	0.587	0.583	0.842	0.892	0.583	0.549	0.584	0.582
F1	0.670	0.670	0.533	0.448	0.673	0.672	0.669	0.669
AUC	0.698	0.696	0.697	0.689	0.698	0.687	0.699	0.699

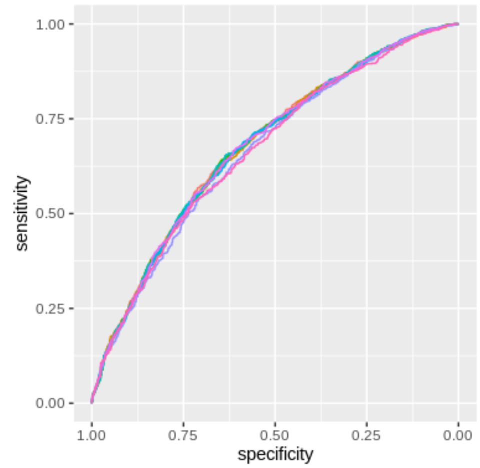
Table 19. Metrics with the Testing Set

	glm.mod11	glm.mod12	ridge.mod1	ridge.mod2	lasso.mod1	lasso.mod2	pls.mod1	pls.mod2
Accuracy	0.634	0.633	0.595	0.565	0.635	0.619	0.637	0.639
Precision	0.642	0.642	0.709	0.718	0.642	0.624	0.644	0.646
Sensitivity	0.699	0.695	0.399	0.297	0.704	0.709	0.703	0.706
Specificity	0.561	0.564	0.815	0.868	0.558	0.517	0.562	0.563
F1	0.669	0.667	0.511	0.420	0.672	0.664	0.672	0.675
AUC	0.680	0.679	0.677	0.666	0.679	0.667	0.679	0.679

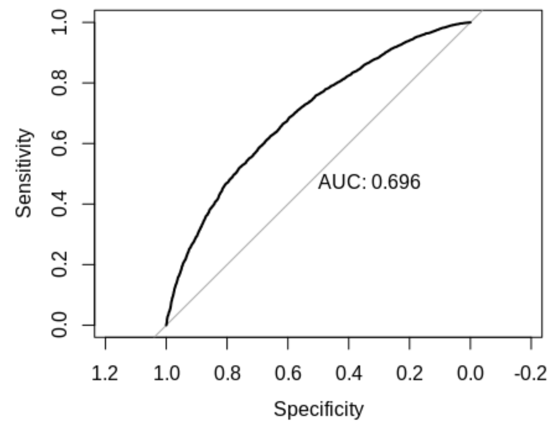
The values of metrics on the testing set were little less than those of the training set. These metrics from the confusion matrices were in the same range for the different models

On the training set the partial least squared (pls1 model) performed better with 0.699 on AUC but perform less than glm1 model on the testing set. To avoid overfitting, the lasso model was selected.

The figures below gave the AUC of all the models in the first graph and the AUC of the selected model. All the models had quasi-similar AUC. This means that each model could have been taken for the final model. The second graph was the final model AUC.



AUC of Different Models



AUC of the Final Model

Figure 67-69. Area Under the Curves

Table 21. Confusion Matrix of the Best Model

Confusion Matrix and Statistics
Reference
Prediction 0 1
0 3114 1817
1 2280 4253
Accuracy : 0.6426
95% CI : (0.6338, 0.6514)
No Information Rate : 0.5295
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.2793
Mcnemar's Test P-Value : 5.281e-13
Sensitivity : 0.7007
Specificity : 0.5773
Pos Pred Value : 0.6510
Neg Pred Value : 0.6315
Prevalence : 0.5295
Detection Rate : 0.3710
Detection Prevalence : 0.5699
Balanced Accuracy : 0.6390
'Positive' Class : 1

Table 23. Coefficients of the Best Model

115 x 1 sparse Matrix of class "dgCMatrix"	DEPRESS7 -2.072517e-01	LAST.SYMP88 -4.452216e-01
s0	DEPRESS9 4.736412e-02	LAST.SYMP99 1.351633e+00
(Intercept) 7.516123e-01	BRONCH2 -1.208318e-01	EPIS.12M2 .
(Intercept) .	BRONCH7 -1.706140e-01	EPIS.12M6 .

SEX2 3.237876e-01	BRONCH9 .	EPIS.12M7 .
AGEG.F72 2.481293e-01	DUR.30D10 2.479306e-01	EPIS.12M9 .
AGEG.F73 9.722777e-02	DUR.30D11 .	COMPASTH11 -4.255287e-01
AGEG.F74 5.848666e-02	DUR.30D12 4.038546e-02	COMPASTH2 -2.544247e-01
AGEG.F75 -1.943397e-01	DUR.30D2 -1.884803e-01	COMPASTH3 -1.527074e-01
AGEG.F76 -3.626295e-01	DUR.30D6 -5.673635e-02	COMPASTH4 -1.245726e+00
AGEG.F77 -6.372159e-01	DUR.30D7 -7.693035e-01	COMPASTH6 .
X_RACEGR32 9.638874e-02	DUR.30D77 -9.908753e-02	COMPASTH7 -9.171228e-01
X_RACEGR33 -1.976777e-01	DUR.30D9 -1.964870e+00	COMPASTH9 2.909219e-01
X_RACEGR34 9.315882e-02	DUR.30D99 1.049682e-01	INS12 -5.866362e-02
X_RACEGR35 1.684440e-01	INCINDT2 2.395983e-01	INS17 .
X_RACEGR39 8.039592e-02	INCINDT3 9.161978e-01	INS19 .
EDUCAL2 -5.371539e-01	INCINDT7 .	INS22 .
EDUCAL3 -3.465724e-01	INCINDT9 6.080227e-01	INS25 -4.216524e-16
EDUCAL4 -1.630750e-01	LAST.MD5 .	INS27 -3.909976e-01
EDUCAL5 .	LAST.MD6 -2.698725e-01	INS29 3.798357e-01
EDUCAL6 9.982970e-02	LAST.MD7 -5.651505e-01	ER.VISIT2 -4.022512e-03
EDUCAL9 5.456655e-01	LAST.MD77 -9.768712e-01	ER.VISIT5 .
X_INCOMG2 1.784381e-02	LAST.MD88 -6.503356e-01	ER.VISIT6 -2.729847e-01
X_INCOMG3 .	LAST.MD99 -2.765027e-01	ER.VISIT7 .
X_INCOMG4 .	LAST.MED2 -2.454941e-01	ER.VISIT9 1.435276e+00
X_INCOMG5 1.344841e-01	LAST.MED3 -4.591047e-01	HOSP.VST2 -1.454704e-01
X_INCOMG9 -1.163555e-01	LAST.MED4 -5.211568e-01	HOSP.VST4 -7.131029e-02
X_RFBMI52 .	LAST.MED5 -4.428678e-01	HOSP.VST5 -1.568415e-01
X_RFBMI59 .	LAST.MED6 -2.328288e-01	HOSP.VST6 -1.351478e-01
SMOKE1002 1.768927e-02	LAST.MED7 -3.408649e-01	HOSP.VST7 3.575437e+00
SMOKE1007 3.860349e-01	LAST.MED77 -5.143170e-01	ASRXCOST2 .
SMOKE1009 3.242238e-01	LAST.MED99 -2.462502e+00	ASRXCOST5 -3.371539e-02
COPD2 -1.052063e-01	LAST.SYMP2 1.058153e-01	ASRXCOST7 -6.240170e-01
COPD7 -2.379751e-01	LAST.SYMP3 2.101672e-01	ASRXCOST9 .
COPD9 .	LAST.SYMP4 1.388828e-01	WORKTALK2 -8.425751e-01
EMPHY2 .	LAST.SYMP5 -2.701455e-02	WORKTALK6 -8.531304e-01
EMPHY7 .	LAST.SYMP6 1.986131e-02	WORKTALK7 -6.686409e-01
EMPHY9 -6.026120e-02	LAST.SYMP7 .	WORKTALK8 -4.826395e-01
DEPRESS2 -5.743822e-02	LAST.SYMP77 -2.395556e-01	WORKTALK9 -1.617558e-01

Looking at the coefficients of the final model, the following assumptions could be made.

GENDER The women were more likely to well manage their asthma than men.

AGE GROUP Middle age participants were more likely to well control their asthma than young adults, but elderlies were less likely to control their asthma by themselves.

LEVEL OF EDUCATION EDUCAL had an impact on asthma well controlled. The controverting remark was that participants with High School degree and some college educations decreased the likelihood of asthma well controlled. But participants with elevated level of education (Bachelor, Master, PhD) increased the likelihood of asthma well controlled with baseline been no less than high school diploma.

INCOME LEVEL had positive effect on asthma self-management. On unit change in income group 2 increase the odds ratio of asthma well controlled by 1.9%.

HOSP.VST Participants who did not remember if they had stayed overnight in a hospital because of your asthma were more likely to have positive effect on asthma well controlled than those who went in the past twelve months. But participants who did not go, did not have any symptom to visit a healthcare professional, had a negative effect on asthma self-management.

LAST TIME TALK TO A DOCTOR OR A HEALTH PROFESSIONAL LASTMD had a negative effect on asthma self-management in the way that it decreased the likelihood of asthma well controlled, based on participants who talked to a doctor about their asthma not far than last year, on participants who communicated with a doctor or a health care professional more than 3 to 5 years, or never.

LAST MEDICATION The impact of LASTMED is significant on asthma well controlled. The farther the last medication was taken, the less likelihood the asthma was well controlled.

LAST SYMPTOM The more the last symptom of episode of attack is far from the date 1, the better the impact on the good asthma self-management.

COMPARATIVE LENGHT OF EPISODE ATTACK(COMPASTH) asthma episodes that last long had negative effect on asthma self-management.

WORKTALK is a key feature influencing an asthma well controlled among adults. The less an adult has discussed the cause the asthma related to job with a doctor or other health professional, it is less likely that the adult has a good management of the asthma.

Table 27. Variable Importance

Variable	Importance
HOSP.VST7	3.58
LAST.MED99	2.46
DUR.30D9	1.96
ER.VISIT9	1.44
LAST.SYMP99	1.35
COMPASTH4	1.25
LAST.MD77	0.98
COMPASTH7	0.92
INCINDT3	0.92
WORKTALK6	0.85
WORKTALK2	0.84
DUR.30D7	0.77
(Intercept)	0.75
WORKTALK7	0.67
LAST.MD88	0.65
AGEG.F77	0.64
ASRXCOST7	0.62
INCINDT9	0.61
LAST.MD7	0.57
EDUCAL9	0.55
EDUCAL2	0.54
LAST.MED4	0.52
LAST.MED77	0.51
WORKTALK8	0.48
LAST.MED3	0.46
LAST.SYMP88	0.45
LAST.MED5	0.44
COMPASTH11	0.43
INS27	0.39
SMOKE1007	0.39
INS29	0.38
AGEG.F76	0.36
EDUCAL3	0.35
LAST.MED7	0.34
SMOKE1009	0.32
SEX2	0.32

Variable	Importance
COMPASTH9	0.29
LAST.MD99	0.28
ER.VISIT6	0.27
LAST.MD6	0.27
COMPASTH2	0.25
AGEG.F72	0.25
DUR.30D10	0.25
LAST.MED2	0.25
INCINDT2	0.24
LAST.SYMP77	0.24
COPD7	0.24
LAST.MED6	0.23
LAST.SYMP3	0.21
DEPRESS7	0.21
X_RACEGR33	0.20
AGEG.F75	0.19
DUR.30D2	0.19
BRONCH7	0.17
X_RACEGR35	0.17
EDUCAL4	0.16
WORKTALK9	0.16
HOSP.VST5	0.16
COMPASTH3	0.15
HOSP.VST2	0.15
LAST.SYMP4	0.14
HOSP.VST6	0.14
X_INCOMG5	0.13
BRONCH2	0.12
X_INCOMG9	0.12
LAST.SYMP2	0.11
COPD2	0.11
DUR.30D99	0.10
EDUCAL6	0.10
DUR.30D77	0.10
AGEG.F73	0.10
X_RACEGR32	0.10
X_RACEGR34	0.09
X_RACEGR39	0.08
HOSP.VST4	0.07

Variable	Importance
EMPHY9	0.06
INS12	0.06
AGEG.F74	0.06
DEPRESS2	0.06
DUR.30D6	0.06
DEPRESS9	0.05
DUR.30D12	0.04
ASRXCOST5	0.03
LAST.SYMP5	0.03
LAST.SYMP6	0.02
X_INCOMG2	0.02
SMOKE1002	0.02
ER.VISIT2	0.00
INS25	0.00
(Intercept)	0.00
EDUCAL5	0.00
X_INCOMG3	0.00
X_INCOMG4	0.00
X_RFBMI52	0.00
X_RFBMI59	0.00
COPD9	0.00
EMPHY2	0.00
EMPHY7	0.00
BRONCH9	0.00
DUR.30D11	0.00
INCINDT7	0.00
LAST.MD5	0.00
LAST.SYMP7	0.00
EPIS.12M2	0.00
EPIS.12M6	0.00
EPIS.12M7	0.00
EPIS.12M9	0.00
COMPASTH6	0.00
INS17	0.00
INS19	0.00
INS22	0.00
ER.VISIT5	0.00
ER.VISIT7	0.00
ASRXCOST2	0.00
ASRXCOST9	0.00

Important variables concerned more the state of the participant health and action taken in the past on behalf of the asthma self-management. For example, the participant had been hospitalized in the past 12 months, length of time the participant did not take medication, frequency of effective symptom of asthma, the patient went to an emergency room during the past 12 months, length of time since the last episode of asthma attack. On the second place, come the demographic variables such as age group, educational level, and gender. The income group and ethnicity group did not have perceptible influence on asthma self-management.

Discussion and Conclusion

The questions we wanted to answer were:

What are the characteristics of an asthma well controlled?

How to address an asthma poorly controlled?

The approach to give a solution was to build a regression model on a CDC data from BRFSS Asthma Call-Back Survey. This data set contains 899 variables and 13, 922 cases. Data exploration, data preparation, and regression were proceeded. The data exploration step involved transforming numeric variables to factors to extract the count of each category in the variable. The numbers of levels were reduced in certain variables. Over 11494 cases, 66.30 % of participants had been taught by a doctor or other health professional how to recognize early signs or symptoms of an asthma episode 76.21 % of participants had been taught by a doctor or other health professional what to do during an asthma episode or attack, 44.52 % of participants had been taught by a doctor or other health professional how to use a peak flow meter, 30.64 % of participants had received an asthma action plan from a doctor or other health. Professional, 9.38 % of participants had taken a course or class on how to manage your asthma, 76.02 % had received advices by a health professional on improvement of their environment, 33.83 % of participants had been watched by a doctor or other health professional using the inhaler. The data preparation step started with the clustering of 7 variables to build the response variable. The 7 clusters resulted were classified as: Very well controlled for cluster 3, well controlled for cluster 2, not well controlled for clusters 1 and 4, poorly controlled for cluster 5 and 6, and very poorly controlled for cluster 7. Regression step resumed by the well-controlled asthma can be predicted at 64.2% with a precision of 65,10%. The model was successful at predicting well controlled asthma. Low income and people in all ethnicity group can self-manage their asthma.

References

[1] Belgrave, D., Henderson, J., Simpson, A., Buchan, I., Bishop, C., & Custovic, A. (2017). “Disaggregating asthma: Big investigation versus big data”. *The Journal of allergy and clinical immunology*, *139*(2), 400–407. https://doi.org/10.1016/j.jaci.2016.11.003. [Accessed: Nov-18-2020]

[2] A. Agarwal, C. Baechle, R. S. Behara, and V. Rao, “Multi-method approach to wellness predictive modeling,” *Journal of Big Data*, 01-Jan-1970. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0049-0. [Accessed: Nov-18-2020].

[3] Quirt, J., Hildebrand, K. J., Mazza, J., Noya, F., & Kim, H. (2018). Asthma. *Allergy, asthma, and clinical immunology : official journal of the Canadian Society of Allergy and Clinical Immunology*, *14*(Suppl 2), 50. https://doi.org/10.1186/s13223-018-0279-0 [Accessed: Nov-18-2020].

[4] Agache, I., & Akdis, C. A. (2016). Endotypes of allergic diseases and asthma: An important step in building blocks for the future of precision medicine. *Allergology international: official journal of the Japanese Society of Allergology*, *65*(3), 243–252. https://doi.org/10.1016/j.alit.2016.04.011.

[5] ahran, H. S., Bailey, C. M., Qin, X., & Moorman, J. E. (2015). Assessing asthma control and associated risk factors among persons with current asthma - findings from the child and adult Asthma Call-back Survey. *The Journal of asthma: official journal of the Association for the Care of Asthma*, *52*(3), 318–326. https://doi.org/10.3109/02770903.2014.956894. [Accessed: Nov-18-2020].

[6] Zahran, H. S., Person, C. J., Bailey, C., & Moorman, J. E. (2012). Predictors of asthma self-management education among children and adults--2006-2007 behavioral risk factor surveillance system asthma call-back survey. *The Journal of asthma: official journal of the Association for the Care of Asthma*, *49*(1), 98–106. https://doi.org/10.3109/02770903.2011.644012. [Accessed: Nov-18-2020].

[7] Pandey, G., Pandey, O. P., Rogers, A. J., Ahsen, M. E., Hoffman, G. E., Raby, B. A., Weiss, S. T., Schadt, E. E., & Bunyavanich, S. (2018). A Nasal Brush-based Classifier of Asthma Identified by Machine Learning Analysis of Nasal RNA Sequence Data. *Scientific reports*, *8*(1), 8826. https://doi.org/10.1038/s41598-018-27189-4. [Accessed: Nov-18t-2020].

[8] Yaghoubi, M., Adibi, A., Safari, A., FitzGerald, J. M., & Sadatsafavi, M. (2019). The Projected Economic and Health Burden of Uncontrolled Asthma in the United States. *American journal of respiratory and critical care medicine*, *200*(9), 1102–1112. https://doi.org/10.1164/rccm.201901-0016OC. [Accessed: Nov-18-2020].

[9] Xie, L., Gelfand, A., Delclos, G. L., Atem, F. D., Kohl, H. W., 3rd, & Messiah, S. E. (2020). Estimated Prevalence of Asthma in US Children With Developmental Disabilities. *JAMA network open*, *3*(6), e207728. https://doi.org/10.1001/jamanetworkopen.2020.7728. [Accessed: Nov-18-2020].

[10] Ahmed A. Arif & Purva Korgaonkar (2016) The association of childhood asthma with mental health and developmental comorbidities in low-income families, Journal of Asthma, 53:3, 277-281, DOI: 10.3109/02770903.2015.1089277. [Accessed: Nov-18-2020].

[11] Thakur, N., Oh, S. S., Nguyen, E. A., Martin, M., Roth, L. A., Galanter, J., Gignoux, C. R., Eng, C., Davis, A., Meade, K., LeNoir, M. A., Avila, P. C., Farber, H. J., Serebrisky, D., Brigino-Buenaventura, E., Rodriguez-Cintron, W., Kumar, R., Williams, L. K., Bibbins-Domingo, K., Thyne, S., … Burchard, E. G. (2013). Socioeconomic status and childhood asthma in urban minority youths. The GALA II and SAGE II studies. *American journal of respiratory and critical care medicine*, *188*(10), 1202–1209. https://doi.org/10.1164/rccm.201306-1016OC.

[12] Wise, S. K., Lin, S. Y., Toskala, E., Orlandi, R. R., Akdis, C. A., Alt, J. A., Azar, A., Baroody, F. M., Bachert, C., Canonica, G. W., Chacko, T., Cingi, C., Ciprandi, G., Corey, J., Cox, L. S., Creticos, P. S., Custovic, A., Damask, C., DeConde, A., DelGaudio, J. M., … Zacharek, M. (2018). International Consensus Statement on Allergy

and Rhinology: Allergic Rhinitis. *International forum of allergy & rhinology*, *8*(2), 108–352. https://doi.org/10.1002/alr.22073. [Accessed: Nov-18-2020].

[13] CDC - BRFSS - 2016 BRFSS Asthma Call-back Survey (ACBS) https://www.cdc.gov/brfss/acbs/2016_documentation.html

[14] https://www.cdc.gov/brfss/acbs/2016/pdf/acbs_2016_adult_questionnaire-final-508.pdf

[15] *Jeff Nieman, Nidhi Kao, Vineet Labru, Corinne Dickey, Clark Austin, Laura* Clarke Behavior of Service Contract Renewals *July 21, 2016*

[16]

Appendices

Supplementary Figures and Tables

Table 28: Exponential coefficient of the best model

EXP(COEF)	EXP(COEF)	EXP(COEF)
(Intercept) 1.0000000	DEPRESS9 1.0000000	LAST.SYMP77 0.9516277
SEX2 1.3715424	BRONCH2 0.9497260	LAST.SYMP88 0.7044179
AGEG.F72 1.1353642	BRONCH7 0.8022356	LAST.SYMP99 2.1353510
AGEG.F73 1.1383840	BRONCH9 0.8986916	EPIS.12M2 1.0000000
AGEG.F74 1.1272107	DUR.30D10 1.1558258	EPIS.12M6 1.0000000
AGEG.F75 0.9174573	DUR.30D11 0.9794923	EPIS.12M7 0.7353709
AGEG.F76 0.8369060	DUR.30D12 1.0000000	EPIS.12M9 1.8642546
AGEG.F77 0.6516000	DUR.30D2 0.8266066	COMPASTH11 0.7245937
X_RACEGR32 1.5458413	DUR.30D6 0.9428379	COMPASTH2 0.8569885
X_RACEGR33 0.8811561	DUR.30D7 0.6393818	COMPASTH3 0.9002239
X_RACEGR34 1.2053924	DUR.30D77 0.9236939	COMPASTH4 0.4189712
X_RACEGR35 1.2711658	DUR.30D9 0.4502718	COMPASTH6 1.0000000
X_RACEGR39 1.3059020	DUR.30D99 1.0000000	COMPASTH7 0.4823005
EDUCAL2 0.4823598	INCINDT2 1.0000000	COMPASTH9 4.3270731
EDUCAL3 0.7373260	INCINDT3 1.8136867	INS12 0.9913896
EDUCAL4 0.9211749	INCINDT7 0.9283848	INS17 1.0387907
EDUCAL5 1.0000000	INCINDT9 1.4563394	INS19 1.0000000
EDUCAL6 1.0000000	LAST.MD5 1.0000000	INS22 1.0000000
EDUCAL9 1.0000000	LAST.MD6 0.8301914	INS25 1.0000000
X_INCOMG2 1.0187322	LAST.MD7 0.5882414	INS27 0.9115631
X_INCOMG3 1.0744569	LAST.MD77 0.4851028	INS29 1.0000000
X_INCOMG4 0.9233082	LAST.MD88 0.9959058	ER.VISIT2 0.9047477
X_INCOMG5 1.0000000	LAST.MD99 0.7931066	ER.VISIT5 0.9976700
X_INCOMG9 0.9144733	LAST.MED2 0.8532456	ER.VISIT6 0.6665296
X_RFBMI52 0.9735209	LAST.MED3 0.7249611	ER.VISIT7 0.9634864
		ER.VISIT9 0.4671836
		HOSP.VST2 0.9679800
		HOSP.VST4 1.0000000

X_RFBMI59 1.0000000	LAST.MED4 0.6746924	HOSP.VST5 0.8511923
SMOKE1002 1.0232549	LAST.MED5 0.7327481	HOSP.VST6 0.9442016
SMOKE1007 1.5909494	LAST.MED6 0.8040649	HOSP.VST7 1.0000000
SMOKE1009 1.0000000	LAST.MED7 0.8293982	ASRXCOST2 1.0000000
COPD2 0.8113806	LAST.MED77 0.6840154	ASRXCOST5 0.9836202
COPD7 0.6948123	LAST.MED99 0.2998855	ASRXCOST7 0.9531719
COPD9 1.0000000	LAST.SYMP2 1.0708459	ASRXCOST9 1.0000000
EMPHY2 1.0000000	LAST.SYMP3 1.2038517	WORKTALK2 0.5347354
EMPHY7 0.9070408	LAST.SYMP4 1.1195110	WORKTALK6 0.5529522
EMPHY9 1.0000000	LAST.SYMP5 1.0000000	WORKTALK7 0.5359874
DEPRESS2 1.0944140	LAST.SYMP6 1.0293480	WORKTALK8 1.0000000
DEPRESS7 0.8646488	LAST.SYMP7 0.9209834	WORKTALK9 0.8686183

Model Diagnostic

Table 29 Significance of Predictors Selected in the model

Single term deletions
Model:
TARGET ~ SEX + AGEG.F7 + X_RACEGR3 + EDUCAL + X_INCOMG + BRONCH +
DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP + COMPASTH +
HOSP.VST + WORKTALK
Df Deviance AIC
<none> 11520 11674
SEX 1 11576 11728
AGEG.F7 6 11639 11781
X_RACEGR3 5 11534 11678
EDUCAL 6 11558 11700
X_INCOMG 5 11534 11678
BRONCH 3 11532 11680
DUR.30D 7 11537 11677
INCINDT 4 11616 11762
LAST.MD 6 11571 11713
LAST.MED 8 11588 11726
LAST.SYMP 7 11541 11681
COMPASTH 6 11569 11711
HOSP.VST 5 11553 11697
WORKTALK 5 11763 11907
The deletion of one predictor increased the AIC and the Deviance.

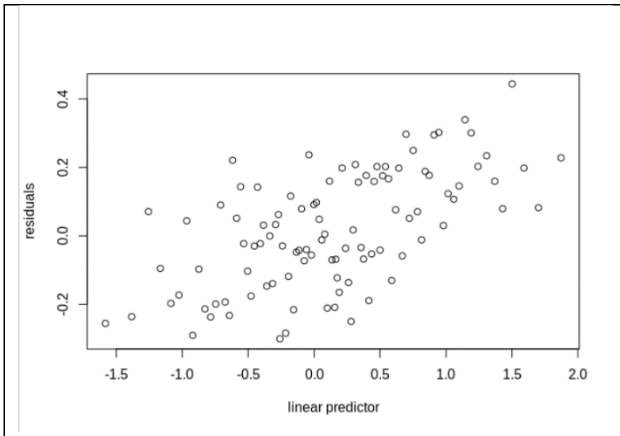


Fig. 71: Binned plot of the residuals against the predictors. Glm model

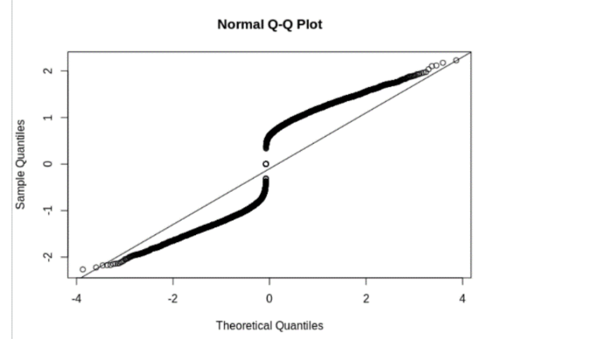


Fig. 73: Normal quantile of the glm model. The plot is close to the linear relationship.

R Statistical Programming Code

https://raw.githubusercontent.com/AlainKuiete/DATA621-FINAL-PROJECT/main/data621FinalProject_g53.Rmd

```
---
title: "DATA 621 Final Project"
author: Farhana Zahir, Vijaya Cherukuri, Scott Reed, Shovon Biswas, Habib Khan, Alain
  Kuiete Tchoupou
date: "11/18/2020"
output:
  word_document: default
  html_document: default
  pdf_document: default
---
## OVERVIEW

The self-management of asthma help improve patient health.
Asthma self-management provide to the patient and caregivers the skills to understand the
disease and its treatment.
It teaches them to take medications appropriately, recognize early signs and symptoms of
asthma episodes, seek medical care as appropriate, and identify and avoid environmental
asthma allergens and irritants
In this project, we study the characteristics that influence asthma self-management.


```{r eval=TRUE, echo=FALSE, message=FALSE, warning=FALSE, results='hide'}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(psych)
library(GGally)
library(corrplot)
library(DMwR)
library(caret)
library(VIM)
library(glmnet)
library(doParallel)
```
```

```
library(xgboost)
library(mice)
library(data.table)
library(kableExtra)
library(mlbench)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
library(haven)
asthma.adult <- read_sas("acbs_2016_adult_public_llcp.sas7bdat")
#View(asthma.adult)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
#write.csv(asthma.adult, "asthma_adult.csv")
#cn <- colnames(asthma.adult)
#write.csv(cn, "asthma_column_name.csv")
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
dim(asthma.adult)
```

The data set come from CDC with ulr =
"https://www.cdc.gov/brfss/acbs/2016_documentation.html". It is a survey study.
The download file is "2016 ACBS Adult Data SAS [ZIP - 3.10 MB]"
The unzip file has 899 variables and 13,922 cases.
We have selected the variables to use on our studies.

EXPLORATORY DATA ANALYSIS

Meaning of variables used in the dataset

Response Variables

ASTHNOW Have you ever been told by a doctor or other health professional that you have asthma?

TCH_SIGN Has a doctor or other health professional ever taught you...
a. How to recognize early signs or symptoms of an asthma episode?

TCH_RESP Has a doctor or other health professional ever taught you...
b. What to do during an asthma episode or attack?

TCH_MON A peak flow meter is a hand held device that measures how quickly you can blow air
out of your lungs. Has a doctor or other health professional ever taught you…
c. How to use a peak flow meter to adjust your daily medications?

MGT_PLAN An asthma action plan, or asthma management plan, is a form with instructions about when to change the amount or type of medicine, when to call the doctorfor advice, and when to go to the emergency room.
Has a doctor or other health professional EVER given you an asthma action plan?

MOD_ENV (7.13) INTERVIEWER READ: Now, back to questions specifically about you. Has a health professional ever advised you to change things in your home, school, or work to improve your asthma

MGT_CLAS Have you ever taken a course or class on how to manage your asthma?

INHALERH (8.3) Did a doctor or other health professional show you how to use the inhaler?

INHALERW (8.4) Did a doctor or other health professional watch you use the inhaler?

Responses types
(1) YES
(2) NO
(7) DON'T KNOW
(9) REFUSED

Possible Predictors

MISS_DAY = "NUMBER OF MISSED DAYS"

MOD_ENV = "EVER ADVISED CHANGE THINGS IN YOUR HOME"

AGEDX = "AGE AT ASTHMA DIAGNOSIS"

AGEG_F6_M = "MODIFIED SIX AGE GROUPS USED IN ASTHMA ADULT POST-STRATIFICATION"

AIRCLEANER = "AIR CLEANER USED"

ASMDCOST = "COST BARRIER: PRIMARY CARE DOCTOR"

ASRXCOST = "COST BARRIER: MEDICATION"

ASSPCOST = "COST BARRIER: SPECIALIST"

CATTMPTS_F = "DISPOSITION CODES FOR CALL ATTEMPTS 1 THROUGH 20 ..."

EMP_STAT = "CURRENT EMPLOYMENT STATUS"

EPIS_12M = "ASTHMA EPISODE OR ATTACK"

EPIS_TP = "NUMBER OF EPISODES / ATTACKS"

ER_TIMES = "NUMBER OF EMERGENCY ROOM VISITS"

ER_VISIT = "EMERGENCY ROOM VISIT"

EVER_ASTH = "EVER HAVE ASTHMA INCONSISTENT WITH BRFSS"

HOSPPLAN = "HOSPITAL FOLLOW-UP"

HOSPTIME = "NUMBER OF HOSPITAL VISITS"

HOSP_VST = "HOSPITAL VISIT"

QSTLANG_F = "LANGUAGE IDENTIFIER"

SCR_MED3 = "HAVE ALL THE MEDICATIONS"

UNEMP_R = "REASON NOT NOW EMPLOYED"

URG_TIME = "NUMBER OF URGENT VISITS"

WORKENV5 = "ASTHMA AGGRAVATED BY CURRENT JOB"

WORKENV6 = "ASTHMA CAUSED BY CURRENT JOB"

WORKENV7 = "ASTHMA AGGRAVATED BY PREVIOUS JOB"

WORKENV8 = "ASTHMA CAUSED BY PREVIOUS JOB"

WORKQUIT1 = "EVER CHANGE OR QUIT A JOB"

WORKSEN3 = "DOCTOR DIAGNOSED WORK ASTHMA"

WORKSEN4 = "SELF-IDENTIFIED WORK ASTHMA"

WORKTALK = "DOCTOR DISCUSSED WORK ASTHMA"

INS1 = "INSURANCE"

INS2 = "INSURANCE OR COVERAGE GAP"

LASTSYMP = "LAST HAD ANY SYMPTOMS OF ASTHMA"

LAST_MD = "LAST TALKED TO A DOCTOR"

LAST_MED = "LAST TOOK ASTHMA MEDICATION"

COMPASTH = "TYPICAL ATTACK"

```
### Constructing the Data Frame by Selecting variables
We select all possible variable that we can use in our dataset.
We also start to clean the dataset

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asthma.mgt.adult1 <- data.frame(TCH.SIGN = asthma.adult$TCH_SIGN,
                                TCH.RESP = asthma.adult$TCH_RESP,
                                TCH.MON = asthma.adult$TCH_MON,
                                MGT.PLAN = asthma.adult$MGT_PLAN,
                                MGT.CLAS = asthma.adult$MGT_CLAS,
                                INHALERW = asthma.adult$INHALERW,
                                MOD.ENV = asthma.adult$MOD_ENV,
                                SEX = asthma.adult$SEX,
                                AGEG.F7 = asthma.adult$AGEG_F7,
                                "RACE.GR3" = asthma.adult[, "_RACEGR3"],
                                #"EDUCA" = asthma.adult[, "_EDUCAG"],
                                EDUCAL = asthma.adult$EDUCA,
                                #INCOMEL = asthma.adult$INCOME2,
                                "INCOMG" = asthma.adult[, "_INCOMG"],
                                #"BMISCAT "= asthma.adult[, "_BMI5CAT"],
                                "RFBMIS" = asthma.adult[, "_RFBMI5"],
                                SMOKE100 = asthma.adult$SMOKE100,
                                COPD = asthma.adult$COPD,
```
```

```
                                EMPHY = asthma.adult$EMPHY,
                                DEPRESS = asthma.adult$DEPRESS,
                                BRONCH = asthma.adult$BRONCH,
                                #SYMP.30D = asthma.adult$SYMP_30D,
                                DUR.30D = asthma.adult$DUR_30D,
                                #ASLEEP30 = asthma.adult$ASLEEP30,
                                #SYMPFREE = asthma.adult$SYMPFREE,
                                INCINDT = asthma.adult$INCIDNT,
                                LAST.MD = asthma.adult$LAST_MD,
                                LAST.MED = asthma.adult$LAST_MED,
                                LAST.SYMP = asthma.adult$LASTSYMP,
                                EPIS.12M = asthma.adult$EPIS_12M,
                                #EPIS.TP = asthma.adult$EPIS_TP,
                                #DUR.ASTH = asthma.adult$DUR_ASTH,
                                COMPASTH = asthma.adult$COMPASTH,
                                INS1 = asthma.adult$INS1,
                                INS2 = asthma.adult$INS2,
                                #ER.TIMES = asthma.adult$ER_TIMES,
                                ER.VISIT = asthma.adult$ER_VISIT,
                                #URG.TIMES = asthma.adult$URG_TIME,
                                HOSP.VST = asthma.adult$HOSP_VST,
                                #HOSPTIME = asthma.adult$HOSPTIME,
                                #HOSPPLAN = asthma.adult$HOSPPLAN,
                                ASMDCOST = asthma.adult$ASMDCOST,
                                ASRXCOST = asthma.adult$ASRXCOST,
                                ASSPCOST = asthma.adult$ASSPCOST,
                                WORKTALK = asthma.adult$WORKTALK
                                )
```

```
#### summary of the data set
Here we categ
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asthma.mgt.adult2 <- data.frame(apply(asthma.mgt.adult1, 2, as.factor ))
summary(asthma.mgt.adult2)
write.csv(summary(asthma.mgt.adult1), "summary12.csv")
```

Here we collapse certain variables with to many classes, and factors with few cases.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asthma.mgt.adult2 <- asthma.mgt.adult2 %>%
  mutate(EDUCAL = fct_collapse(EDUCAL,
    "1" = "1",
    "2" = "2",
    "3" = "3",
    "4" = "4",
    "5" = "5",
    "6" = c("6", "9")
  ),
  LAST.MD = fct_collapse(LAST.MD,
    "4" = "4",
    "5" = "5",
    "6" = "6",
    "7" = "7",
    "9" = c("77", "88", "99"),
  ),
  LAST.MED = fct_collapse(LAST.MED,
    "4" = "1",
```
```

```
    "5" = "2",
    "6" = "3",
    "7" = "4",
    "9" = c("77", "88", "99"),
  ),
  INCINDT = fct_collapse(INCINDT,
    "1" = "1",
    "2" = "2",
    "3" = "3",
    "7" = c("7", "9")
  ),
  LAST.SYMP = fct_collapse(LAST.SYMP,
    "1" = "1",
    "2" = "2",
    "3" = "3",
    "4" = "4",
    "5" = "5",
    "7" = "7",
    "9" = c("77", "88", "99")
  ),
  DUR.30D = fct_collapse(DUR.30D,
    "1" = "1",
    "2" = "2",
    "6" = "6",
    "7" = c("7", "9","77" ,"99"),
    "10" = "10",
    "11" = "11",
    "12" = "12"
  ),
  EPIS.12M = fct_collapse(EPIS.12M,
    "1" = "1",
    "2" = "2",
    "6" = "6",
    "7" = c("7", "9")
  ),
  ER.VISIT = fct_collapse(ER.VISIT,
    "1" = "1",
    "2" = "2",
    "5" = c("5", "7", "9"),
    "7" = c("7", "9")
  ),
  COMPASTH = fct_collapse(COMPASTH,
    "1" = "1",
    "2" = "2",
    "3" = c("3","4"),
    "6" = "6",
    "7" = c("7", "9"),
    "11" = "11"
  ),
  )
summary(asthma.mgt.adult2)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# asthma.mgt.adult2 <- asthma.mgt.adult2 %>%
#   mutate(TCH.SIGN = fct_collapse(TCH.SIGN,
#                                   "1" = "1",
```

```
#                                   "2" = "2",
#                                   "7" = c("7", "9")))
# summary(asthma.mgt.adult2$TCH.SIGN)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.mgt.ad.min <-  asthma.mgt.adult1 %>% filter(TCH.SIGN == 1 | TCH.SIGN == 2,
                                                 TCH.RESP == 1 | TCH.RESP == 2,
                                                 TCH.MON == 1 | TCH.MON == 2,
                                                 MGT.PLAN == 1 | MGT.PLAN == 2,
                                                 MGT.CLAS == 1 | MGT.CLAS  == 2,
                                                 INHALERW == 1 | INHALERW == 2,
                                                 MOD.ENV == 1 | MOD.ENV == 2)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
dim(asth.mgt.ad.min)
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.mgt.ad.min2 <- asth.mgt.ad.min
# asth.mgt.ad.min2 <- asth.mgt.ad.min %>% filter(LAST.MD != 77 & LAST.MD != 99,
#                                                LAST.MED != 77 & LAST.MED != 99,
#                                                LAST.SYMP != 77 & LAST.SYMP !=
99,LAST.SYMP != 88,
#                                                INCINDT != 7 & INCINDT != 9,
#                                                SYMP.30D != 77 & SYMP.30D != 99,
#                                                DUR.30D != 9 & DUR.30D != 99,
#                                                ASLEEP30 != 99, ASLEEP30 !=66, ASLEEP30
!= 100, ASLEEP30 != 111,
#                                                EPIS.12M != 7 & EPIS.12M != 9,
#                                                COMPASTH != 7 & COMPASTH != 9,
#                                                INS1 != 7 & INS1 != 9,
#                                                ER.VISIT != 7 & ER.VISIT != 9,
#                                                ER.TIMES != 777 & ER.TIMES != 999,
#                                                URG.TIMES != 777 & URG.TIMES != 999,
#                                                HOSP.VST != 9,
#                                                HOSPTIME != 777
#                                                )
```



```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# asth.mgt.ad.min %>% select(DUR.ASTH) %>% filter(DUR.ASTH==0)
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
#dim(asth.mgt.ad.min2)
```


Structure of the data

```{r}
str(asthma.mgt.adult2)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
str(asth.mgt.ad.min2)
attach(asth.mgt.ad.min2)
```

Summary of the Data
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
sum.data <- summary(asth.mgt.ad.min2)
sum.data
#write.csv(sum.data, "summary_data.csv")
```

Distribution of the Variables in the Data

Histograms
Histograms tell us how the data is distributed in the dataset (numeric fields).

```{r, message = FALSE, warning = FALSE, echo = F}
multi.hist(asthma.mgt.adult1[1:9])
multi.hist(asthma.mgt.adult1[10:18])
multi.hist(asthma.mgt.adult1[19:27])
multi.hist(asthma.mgt.adult1[28:33])
```

The correlations betweeen predictors

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
cor_asthma.adult <- cor(asth.mgt.ad.min2[,-c(1:7)], use = "na.or.complete")
corrplot(cor_asthma.adult, order = 'hclust', type = 'lower')
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# cor(asthma.mgt.adult1[,-1], use = "na.or.complete")
# write.csv(cor(asthma.mgt.adult1[,-1], use = "na.or.complete"), "predictors_cor.csv")
```

There are highly correlated predictors. We are going to remove some of them.

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.mgt.ad.min21 <- select(asth.mgt.ad.min2, -ASSPCOST, -ASMDCOST)
colnames(asth.mgt.ad.min21)
```

CONSTRUCT THE RESPONSE VARIABLE
We first extract variables related to education,
Selection of variables
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
```

```
responses <- data.frame(
                        TCH.SIGN = asth.mgt.ad.min21$TCH.SIGN,
                        TCH.RESP = asth.mgt.ad.min21$TCH.RESP,
                        TCH.MON = asth.mgt.ad.min21$TCH.MON,
                        MGT.PLAN = asth.mgt.ad.min21$MGT.PLAN,
                        MGT.CLAS = asth.mgt.ad.min21$MGT.CLAS,
                        INHALERW = asth.mgt.ad.min21$INHALERW,
                        MOD.ENV = asth.mgt.ad.min21$MOD.ENV
                        )
head(responses)
```

```
#### Exploration of the clustering
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
responses.cat <- data.frame(apply(responses, 2, as.factor))
summary(responses.cat)
```

Elbow method to find the number of clusters
We run kmeans with different clusters from 1 to 16 and we produce a
scree plot to determine the number of cluster at the elbow.
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Elbow method Scree Plot",
echo=FALSE, results='show'}
set.seed(25)
# Initialize total within sum of squares error: wss
wss <- 0

# Look over 1 to 15 possible clusters
for (i in 1:16) {
  # Fit the model: km.out
  km.out <- kmeans(responses.cat, centers = i, nstart = 20, iter.max = 50)
  # Save the within cluster sum of squares
  wss[i] <- km.out$tot.withinss
}

# Produce a scree plot
plot(1:16, wss, type = "b",
     xlab = "Number of Clusters",
     ylab = "Within groups sum of squares"
)
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# Select number of clusters
k <- 7
```
 The number of cluster is 3

Now we do the clustering and extract the centers of resulting model
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
set.seed(25)
# Build model with k clusters: km.out
km.out <- kmeans(responses.cat, centers = k, nstart = 20, iter.max = 50)

# View the resulting model
km.out$centers
```
```

````
```


#### We add the point classification to the original data
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
resp.asthma <- cbind(responses.cat, target = km.out$cluster)
head(resp.asthma)
write.csv(resp.asthma, "response_interpret.csv")
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
resp.asthma$target <- as.factor(resp.asthma$target)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
summary(resp.asthma)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "View of the clustering result",
echo=FALSE, results='show'}
plot(resp.asthma$target)
```


### Interpretation of the Selft-Management Response clustering
#### TCH.SIGN
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, TCH.SIGN, target) %>% summarise(count=n()) %>%
  group_by(TCH.SIGN) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
 asth.res1 <- egt %>% group_by(TCH.SIGN) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res1
```



```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=TCH.SIGN, linetype=TCH.SIGN))+geom_line()
```
In  the target response, 8 is the positive answer, 3 is the negative answer, 5 is don't
know and 6 is refused for the question:
TCH_SIGN  Has a doctor or other health professional ever taught you...
a. How to recognize early signs or symptoms of an asthma episode?



#### TCH.RESP
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, TCH.RESP, target) %>% summarise(count=n()) %>%
  group_by(TCH.RESP) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```
````
````

````
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.res2 <- egt %>% group_by(TCH.RESP) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res2
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=TCH.RESP, linetype=TCH.RESP))+geom_line()
```
In  the target response, 8 is the positive answer, 3 is the negative answer, 1 is don't
know and 1 is refused for the question:
TCH_RESP Has a doctor or other health professional ever taught you...
b. What to do during an asthma episode or attack?


#### TCH.MON
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, TCH.MON, target) %>% summarise(count=n()) %>%
  group_by(TCH.MON) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.res3 <- egt %>% group_by(TCH.MON) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res3
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=TCH.MON, linetype=TCH.MON))+geom_line()
```
In  the target response, 8 is the positive answer, 7 are the negative answers, 2 is don't
know and 2 is refused for the question:
TCH_MON A peak flow meter is a hand held device that measures how quickly you can blow
air
out of your lungs. Has a doctor or other health professional ever taught you…
c. How to use a peak flow meter to adjust your daily medications?



#### MGT.PLAN
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, MGT.PLAN, target) %>% summarise(count=n()) %>%
  group_by(MGT.PLAN) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.res4 <- egt %>% group_by(MGT.PLAN) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res4
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=MGT.PLAN, linetype=MGT.PLAN))+geom_line()
````
````

~~~
```
~~~

In the target response, 8 is the positive answer, 3 is the negative answer, 9 is don't know and 9 is refused for the question:
MGT_PLAN An asthma action plan, or asthma management plan, is a form with instructions about when to change the amount or type of medicine, when to call the doctor for advice, and when to go to the emergency room.
Has a doctor or other health professional EVER given you an asthma action plan?

MGT.CLAS

~~~
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, MGT.CLAS, target) %>% summarise(count=n()) %>%
  group_by(MGT.CLAS) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.res5 <- egt %>% group_by(MGT.CLAS) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res5
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=MGT.CLAS, linetype=MGT.CLAS))+geom_line()
```
~~~

In the target response, 8 is the positive answer, 8 or(3,7) is the negative answer, 8 is don't know and 6 is refused for the question:
MGT_CLAS Have you ever taken a course or class on how to manage your asthma?

INHALERW

~~~
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, INHALERW , target) %>% summarise(count=n()) %>%
  group_by(INHALERW ) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.res6 <- egt %>% group_by(INHALERW) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res6
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=INHALERW , linetype=INHALERW))+geom_line()
```
~~~

In the target response, 8 is the positive answer, 3 is the negative answer, 4 is don't know and 1 is refused for the question:
INHALERW (8.4) Did a doctor or other health professional watch you use the inhaler?

MOD.ENV

~~~
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
egt <- group_by(resp.asthma, MOD.ENV , target) %>% summarise(count=n()) %>%
~~~

````
  group_by(MOD.ENV) %>% mutate(etotal=sum(count), proportion=count/etotal)
tibble::as.tibble(egt)
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.res7 <- egt %>% group_by(MOD.ENV) %>% mutate(group.max = max(count)) %>%
group_by(target) %>% filter(count==group.max)
asth.res7
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ggplot(egt, aes(x=target, y=proportion, group=MOD.ENV, linetype=MOD.ENV))+geom_line()
```


#### Summary  of the response variables
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# response.var <- data_frame(RESPONSE = c("1=YES", "2=NO"),
#                            TCH.SIGN = asth.res1$target,
#                            TCH.RES  = asth.res2$target,
#                            TCH.MON = asth.res3$target,
#                            MGT.PLAN = asth.res4$target,
#                            MGT.CLAS = asth.res5$target,
#                            INHALERW = asth.res6$target,
#                            MOD.ENV = asth.res7$target)
# response.var
```

#### Asthma controlled levels (Weather the Individual Asthma Is Well Controlled or Not)
```{r}
asth.edu11 <- merge(asth.res1, asth.res2 ,by.x = "target", by.y = "target", all = TRUE,
no.dups =TRUE) %>%
  merge(., asth.res3 ,by.x = "target", by.y = "target", all = TRUE, no.dups =TRUE) %>%
  merge(., asth.res4 ,by.x = "target", by.y = "target", all = TRUE, no.dups =TRUE) %>%
  merge(., asth.res5 ,by.x = "target", by.y = "target", all = TRUE, no.dups =TRUE) %>%
  merge(., asth.res6 ,by.x = "target", by.y = "target", all = TRUE, no.dups =TRUE) %>%
  merge(., asth.res7 ,by.x = "target", by.y = "target", all = TRUE, no.dups =TRUE) %>%
  select(., target, TCH.SIGN, TCH.RESP, TCH.MON, MGT.PLAN, MGT.CLAS, INHALERW, MOD.ENV)
asth.edu11
write.csv(asth.edu11, "asthma_edu_level2.csv")
```
#### 1, 4 = Very Poorly Controlled
#### 6 = Poorly Controlled
#### 5,7 = Not Well Controlled
#### 2,3 = Well Controlled


##### For the response variable TARGET, an excellent management skill has number 2 but a
poor management skill has number 7 and 5.
##### We can build a logistics regression on the dataset.

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
resp.asthma2 <- resp.asthma
resp.asthma2$target <- if_else(resp.asthma2$target==2 |resp.asthma2$target==3, 1, 0)
```
````
````

````
## !!!! Please, check the values of yes in the response. var and change if condition of
resp.asthma2$taget
## !!!! above are different. Remove "Break" in the chunk below!

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
#break
```


### Here we remove the varibles used to calculate the target variable and reformat the
data frame.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.mgt.ad.min31 <- asth.mgt.ad.min21 %>%
  select( -TCH.SIGN,-TCH.RESP, -TCH.MON, -MGT.PLAN, -MGT.CLAS, -INHALERW, -MOD.ENV) %>%
  mutate(TARGET = resp.asthma2$target) %>% relocate(TARGET, .before = SEX)
str(asth.mgt.ad.min31)
```



### PREPARE THE DATA FOR MODELISATION

#### We remove the rows with missing values.
Here were are going to drop missing data because they are only 12  over 13,922 rows.
We also transform all predictors to categorical.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
asth.mgt.ad.min33 <- drop_na(asth.mgt.ad.min31)
asth.mgt.ad.min35 <- asth.mgt.ad.min33
asth.mgt.ad.min33[,-1] <- data.frame(apply(asth.mgt.ad.min33[,-1], 2, as.factor))
asth.mgt.ad.min35 <- data.frame(apply(asth.mgt.ad.min35, 2, as.factor))
summary(asth.mgt.ad.min35)
```

#### Visualization of some combine variables
#### Target and predictors
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., SEX, TARGET) %>%
  summarise(count=n()) %>%
  group_by(SEX) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=SEX, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```



```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., AGEG.F7, TARGET) %>%
  summarise(count=n()) %>%
  group_by(AGEG.F7) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=AGEG.F7, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```
````

````
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., COMPASTH, TARGET) %>%
  summarise(count=n()) %>%
  group_by(COMPASTH) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=COMPASTH, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., BRONCH, TARGET) %>%
  summarise(count=n()) %>%
  group_by(BRONCH) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=BRONCH, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
egt <- group_by(asth.mgt.ad.min35, EDUCAL, TARGET) %>% summarise(count=n()) %>%
  group_by(EDUCAL) %>% mutate(etotal=sum(count), proportion=count/etotal)
ggplot(egt, aes(x=EDUCAL, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., COPD, TARGET) %>%
  summarise(count=n()) %>%
  group_by(COPD) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=COPD, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., ASRXCOST, TARGET) %>%
  summarise(count=n()) %>%
  group_by(ASRXCOST) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=ASRXCOST, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```
````

````
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., X_INCOMG, TARGET) %>%
  summarise(count=n()) %>%
  group_by(X_INCOMG) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=X_INCOMG, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., LAST.MED, TARGET) %>%
  summarise(count=n()) %>%
  group_by(LAST.MED) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=LAST.MED, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., INS1, TARGET) %>%
  summarise(count=n()) %>%
  group_by(INS1) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=INS1, y=proportion, group=TARGET, linetype=TARGET))+geom_line()
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good skill
management in terme of Duration of Asthma Attack", echo=FALSE, results='show'}
egt <- summarize(group_by(asth.mgt.ad.min35, LAST.SYMP, TARGET), count = n())
egt <- mutate(egt, etotal =sum(count), proportion= count/etotal)
ggplot(data=egt, aes(x=LAST.SYMP, y=proportion, group=TARGET,
linetype=TARGET))+geom_line()
```

### Correlation between two predictors
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Proportion of Good Skill
Management in terme of Education Level", echo=FALSE, results='show'}
library(dplyr)
asth.mgt.ad.min35 %>%
  group_by(., X_INCOMG, INS2) %>%
  summarise(count=n()) %>%
  group_by(X_INCOMG) %>%
  mutate(etotal=sum(count), proportion=count/etotal)%>%
  ggplot(., aes(x=X_INCOMG, y=proportion, group=INS2, linetype=INS2))+geom_line()
```

High proportion of no insurance in all income groups.

#### Splitting the data into train and test sets
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
````
````

````
library(caret)
set.seed(25)
asth.mgt.ad.min33$TARGET <- as.numeric(asth.mgt.ad.min33$TARGET)
inTraining <- createDataPartition(asth.mgt.ad.min33$TARGET, p = .80, list = FALSE)
training1 <- asth.mgt.ad.min33[ inTraining,]
testing1  <- asth.mgt.ad.min33[-inTraining,]
```

### BUILDS MODELS



#### Model using full predictors with glm
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
glm.all <- glm(TARGET~., data=training1, family=binomial)
glm.all
```




#### Confusion Matrix with the testingset
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# glm.pred <- predict(glm.all, newdata = testing1[,-1], type = "response")
# predicted <- as.factor(ifelse(glm.pred>.5,1,0))
# glm.cm <- confusionMatrix(data = predicted, testing1$TARGET, positive = '1')
# glm.cm
```


#### First glm model using backward elimination of step function

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
glm.mod11 <- step(glm.all, trace = 0)
glm.mod11
```
Call:  glm(formula = TARGET ~ SEX + AGEG.F7 + X_RACEGR3 + EDUCAL + BRONCH +
    DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP + COMPASTH +
    HOSPTIME + ASRXCOST + WORKTALK, family = binomial, data = training1)

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
# glm.mod11 <- glm(formula = TARGET ~ SEX + AGEG.F7 + X_RACEGR3 + EDUCAL + BRONCH +
#     DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP + COMPASTH +HOSPPLAN+
#    + ASRXCOST + WORKTALK, family = binomial, data = training1)

```

#### Confusion Matrix with the testingset
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
glm11.pred <- predict(glm.mod11, newdata = testing1[,-1], type = "response")
predicted <- as.factor(ifelse(glm11.pred>.5,1,0))
glm11.cm <- confusionMatrix(data = predicted, factor(testing1$TARGET), positive = '1')
glm11.cm
```


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
````

````
```

#### Second glm model
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
glm.mod12 <-  glm(formula = TARGET ~ SEX + AGEG.F7 + X_RACEGR3 + EDUCAL + X_INCOMG +
    BRONCH + DUR.30D + INCINDT + LAST.MD + LAST.MED + LAST.SYMP +
    COMPASTH + WORKTALK, family = binomial, data = training1)
glm.mod12
```


#### Confusion Matrix with the testingset
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
glm12.pred <- predict(glm.mod12, newdata = testing1[,-1], type = "response")
predicted <- as.factor(ifelse(glm12.pred>.5,1,0))
glm12.cm <- confusionMatrix(data = predicted, factor(testing1$TARGET), positive = '1')
glm12.cm
```

#### Lasso and Ridge model

Since our dataset has multiple variable, we can use penalized logistic regression to find
an optimal performing model.
Ridge Regression and Lasso Regression have two different approaches.
Ridge Regression incorporates all variables in the model and gives the coefficients of
variables with minor contribution close to zero
Lasso Regression keeps only the most significant variables and gives zero to the
coefficient of the rest of variables.

#### Split the data into trainset and testingset, Dumy code categorical predictors
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
set.seed(25)
inTraining <- createDataPartition(asth.mgt.ad.min33$TARGET, p = .80, list = FALSE)
training2 <- asth.mgt.ad.min33[ inTraining,]
testing2  <- asth.mgt.ad.min33[-inTraining,]
x <- model.matrix(TARGET ~., data = training2)
y = training2$TARGET
xt <- model.matrix(TARGET ~., data = testing2)
yt <- as.factor(testing2$TARGET)

```



#### Ridge Regression
We fit and obsrve the coefficients of rigde regression against the log of lambda.
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Variation of Ridge Model
Coefficient by Log Lambda", echo=FALSE, results='show'}
fit.ridge <- glmnet(x = x,y=y, alpha=0, family="binomial")
plot(fit.ridge, xvar= "lambda", label=TRUE)
```
The coefficients are significative for negative log lambda and start stabilize around -4

```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Lambda that Minimises MSE",
echo=FALSE, results='show'}
cv.ridge <- cv.glmnet(x = x, y = y, alpha=0)
plot(cv.ridge)
```
````
````

The plot shows that the log of the optimal value of lambda (i.e. the one that minimises the root mean square error) is approximately -3. The exact value can be viewed by examining the variable lambda_min in the code below. In general though, the objective of regularisation is to balance accuracy and simplicity. In the present context, this means a model with the smallest number of coefficients that also gives a good accuracy. To this end, the cv.glmnet function finds the value of lambda that gives the simplest model but also lies within one standard error of the optimal value of lambda.

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
cv.ridge$lambda.min
```

Confusion matrix with lambda min
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ridge.model1 <- glmnet(x = x,y=y, lambda = cv.ridge$lambda.min, alpha=0,
family="binomial")
ridge.pred1 <- predict(ridge.model1, newx = xt)
predicted <- rep(0, length(yt))
predicted[ridge.pred1>0.5] <- "1"
ridge.cm1 <- confusionMatrix(data = as.factor(predicted), yt, positive = '1')
ridge.cm1
```
We observe overfitting with this ridge model

Confusion matrix with best lambda
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ridge.model2 <- glmnet(x = x,y=y, lambda = cv.ridge$lambda.1se, alpha=0,
family="binomial")
ridge.pred2 <- predict(ridge.model2, newx = xt)
predicted <- rep(0, length(yt))
predicted[ridge.pred2>0.5] <- "1"
ridge.cm2 <- confusionMatrix(data = as.factor(predicted), yt, positive = '1')
ridge.cm2
```
We observe overfitting with this second ridge model

Getting the coefficients
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
coef(ridge.model1)
```

Lasso Regression

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
fit.lasso <- glmnet(x =x, y = y, alpha = 1, family = "binomial")
plot(fit.lasso, xvar = "dev", label = TRUE)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
fit.lasso <- glmnet(x,y)
plot(fit.lasso, xvar = "lambda", label = TRUE)
plot(fit.lasso, xvar = "dev", label = TRUE)
```

`````
#### Find the best lambda using cross validation
```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Lambda that minimises MSE in
Lasso",echo=FALSE, results='show'}
cv.lasso <- cv.glmnet(x,y)
plot(cv.lasso)
```
The plot shows that the log of the optimal value of lambda (i.e. the one that minimises
the root mean square error) is approximately -10. The exact value can be viewed by
examining the variable lambda_min in the code below. In general though, the objective of
regularisation is to balance accuracy and simplicity. In the present context, this means
a model with the smallest number of coefficients that also gives a good accuracy.  To
this end, the cv.glmnet function  finds the value of lambda that gives the simplest model
but also lies within one standard error of the optimal value of lambda.


#### Confusion Matrix with lambda min

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
lasso.model1 <- glmnet(x =x, y = y, lambda = cv.lasso$lambda.min, alpha = 1, family =
"binomial")
lasso.pred1 <- predict(lasso.model1, newx = xt, type = "response")
predicted <- as.factor(ifelse(lasso.pred1>.5,1,0))
lasso.cm1 <- confusionMatrix(data = predicted, yt, positive = '1')
lasso.cm1
```

#### Getting the coefficients
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
coef(lasso.model1)
```



#### Confusion Matrix with best lambda

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
lasso.model2 <- glmnet(x =x, y = y, lambda = cv.lasso$lambda.1se, alpha = 1, family =
"binomial")
lasso.pred2 <- predict(lasso.model2, newx = xt, type = "response")
predicted <- as.factor(ifelse(lasso.pred2>.5,1,0))
lasso.cm2 <- confusionMatrix(data = predicted, yt, positive = '1')
lasso.cm2
```


##### Calculating the AICc of Ridge and Lasso Models

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
AICc <- function(fit){
  tLL <- fit$nulldev - deviance(fit)
  k <- fit$df
  n <- fit$nobs
  AICc <- -tLL+2*k+2*k*(k+1)/(n-k-1)
  return (AICc)
}

```
`````

````
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
AICc(ridge.model1)
AICc(lasso.model1)
```

#### Partial Least Squared

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
training3 <- training1
testing3 <- testing1
training3$TARGET <- ifelse(training3$TARGET=="1","T","F")
testing3$TARGET <- ifelse(testing3$TARGET=="1","T","F")
testing3$TARGET <- factor(testing3$TARGET)
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ctrl <- trainControl(method = "repeatedcv", repeats = 3)

plsFit1 <- train(
  TARGET ~ .,
  data = training3,
  method = "pls",
  preProc = c("center", "scale"),
  tuneLength = 15,
  ## added:
  trControl = ctrl
)
```

#### Confusion Matrix with best lambda


```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
pls1.pred <- predict(plsFit1, newdata = testing3[,-1], type = "prob")
pls.pred1 <- predict(plsFit1, newdata = testing3[,-1], type = "raw")
pls.cm1 <- confusionMatrix(data = pls.pred1, testing3$TARGET, positive = 'T')
pls.cm1

```





#### Here we train the model with partial least square using tune parameter.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
ctrl <- trainControl(
  method = "repeatedcv",
  repeats = 3,
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)

set.seed(123)
plsFit2 <- train(
````
````

````
  TARGET ~ .,
  data = training3,
  method = "pls",
  preProc = c("center", "scale"),
  tuneLength = 15,
  trControl = ctrl,
  metric = "ROC"
)
plsFit2
```

#### Confusion Matrix with best lambda
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
pls2.pred <- predict(plsFit2, newdata = testing3[,-1], type = "prob")
pls.pred2 <- predict(plsFit2, newdata = testing3[,-1], type = "raw")
pls.cm2 <- confusionMatrix(data = pls.pred2, testing3$TARGET, positive = 'T')
pls.cm2

```



### SELECT MODELS
#### We compare the models with the accuray, precision, sensitivity, specificity, and F1
score from the confusion matrix
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
cm.metric <- function(cm){
  test = c(cm$overall["Accuracy"],
           cm$byClass["Precision"],
           cm$byClass["Sensitivity"],
           cm$byClass["Specificity"],
           cm$byClass["F1"])
  return(test)
}
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
metrics.mod <- data.frame(#glm.mod = cm.metric(glm.cm),
                          glm.mod11 = cm.metric(glm11.cm),
                          glm.mod12 = cm.metric(glm12.cm),
                          ridge.mod1 = cm.metric(ridge.cm1),
                          ridge.mod2 = cm.metric(ridge.cm2),
                          lasso.mod1 = cm.metric(lasso.cm1),
                          lasso.mod2 = cm.metric(lasso.cm2),
                          pls.mod1 = cm.metric(pls.cm1),
                          pls.mod2 = cm.metric(pls.cm2))
metrics.mod
```

With precision and specificity equal to 1, the ridge.mod2 model is overfitting. But
lasso.mod1 has the best accuracy, precision, sensivity, and specificity.

### Using pROC package.
We can plot the ROC curve and extract the AUC value.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
library(pROC)
library(dplyr)
prediction <- data.frame(TARGET = testing1$TARGET,
                         glm1 = glm11.pred,
````

````
                          gml2 = glm12.pred,
                          pls1 = pls1.pred,
                          pls2 = pls2.pred,
                          rp1 = ridge.pred1,
                          rp2 = ridge.pred2,
                          lp1 = lasso.pred1,
                          lp2 = lasso.pred2)

```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}

```


```{r, eval=TRUE, message=FALSE, warning=FALSE, fig.cap= "Best Model with AUC",
echo=FALSE, results='show'}
## With ggplot2 ##
library(ggplot2)
# Create multiple curves to plot
roc1 <- roc(TARGET ~., data = prediction)
ggroc(roc1)

```
The Lasso model has the best Area Under the Curve.

### We run the lasso.mod1 model with the entire dataset

#### The best model is Lasso  model
#### The Statistic of the best model is given below.

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
set.seed(25)
x <- model.matrix(TARGET ~., data = asth.mgt.ad.min33)
y = asth.mgt.ad.min33$TARGET
```

```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
lasso.model <- glmnet(x =x, y = y, lambda = cv.lasso$lambda.min, alpha = 1, family =
"binomial")
lasso.pred <- predict(lasso.model, newx = x, type = "response")
lasso.predicted <- as.factor(ifelse(lasso.pred>.5,1,0))
lasso.cm <- confusionMatrix(data = factor(lasso.predicted), factor(y), positive = '1')
lasso.cm
```
#### AUC of the best model
```{r, message = FALSE, warning = FALSE, echo = F, results='show'}
plot(roc(y, lasso.pred), print.auc = TRUE)
```




#### Coefficients of the best model
The dot before the coefficient means that the lasso model ignore unimportant class of the
variable.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
coef(lasso.model)
````

````
```


#### We look at the odd ratio off each variable
A value greater than 1 means an increase  effect on the odd ratio compare to baseline.
For example, focusing on SEX variable, Women(SEX2) are more likely to have good Skill on
asthma management than men(SEX1 the baseline). Other variables can be interpret the same
way.
```{r, eval=TRUE, message=FALSE, warning=FALSE, echo=FALSE, results='show'}
exp(coef(lasso.model))
```
````
````