# Predicting Asthma with Machine Learning Algorithms: Evidence from 2019 BRFSS Data

**ABSTRACT**

Studies in the past have explored asthma prevalence in the United States using national survey data. However, we did not find literatures specific to Michigan that have used predictive models for asthma prediction in adults and identifying important predictors. Using a secondary dataset from the 2019 Behavioral Risk Factor Surveillance System, we attempt to build machine learning models to predict asthma and identify risk factors of asthma among adults in Michigan. A total of 10,518 participants were surveyed in Michigan. A sample of 10,411 individuals, resulted after data cleaning, was analyzed. Of those, 1127 reported having asthma during the survey time. Various predictive models were built from the training data set with 10-fold cross validation and parameter tuning performed to achieve optimal model and then tested on the test data set. Among the models built, the performances of logistic regression, LASSO, and elastic net were somewhat similar with sensitivity at around 61% and AUC at around 65%. Due to ease of interpretability, logistic regression was picked for further exploration. Chronic pulmonary obstructive disease, Sex, Age group, Diabetes, Employment, Race, and Income were identified as important predictors.

# TABLE OF CONTENTS

## INTRODUCTION

In general terms, the Center for Disease Control and Prevention (CDC) defines asthma as a disease that affects lungs. It is one of the most common long-term diseases of children, but adults can have asthma, too. CDC identifies that asthma causes repeated episodes of wheezing, breathlessness, chest tightness, and nighttime or early morning coughing [1]. Both incidence and prevalence of asthma have been increasing in the United States. Current asthma prevalence increased from 7.2% in 2001 to 9.0% in 2019 [2]. Michigan has higher asthma prevalence rates than the national average. Based on 2019 Behavioral Risk Factor System Surveillance (BRFSS) data, an estimated 11.1% of Michigan adults had current asthma [2]. BRFSS constructs two different asthma measures: Lifetime Asthma and Current Asthma. Lifetime asthma is defined as an affirmative response to the question "Have you ever been told by a doctor (nurse or other health professional) that you have asthma?". Current asthma is defined as an affirmative response to that question followed by an affirmative response to the subsequent question "Do you still have asthma?" [2] Although asthma cannot be cured, it can be managed by avoiding things that trigger asthma attacks and receiving appropriate medical care [1].

### PURPOSE OF THE STUDY

This study will attempt to develop predictive models and evaluate their performance in predicting asthma among Michigan adult population. We will also study important factors associated with asthma. Knowing these factors can help to design appropriate interventions for asthma control. Although studies in the past have identified risk factors of asthma in the US population [6,8,9], no studies have focused specifically on Michigan population. Also, no past studies have employed modern machine learning algorithms for predictive modeling and identification of significant factors specific to Michigan population. The following algorithms will be used to build predictive models and study risk factors associated with asthma.

i) Logistic Regression (LR)

ii) Partial Least Squares (PLS)

iii) Random Forest (RF)

iv) Gradient Boosting (GB)

v) Least Absolute Shrinkage and Selection Operator (LASSO)

vi) Elastic Net (EN)

vii) K-Nearest Neighbors (KNN)

viii) Support Vector Machines (SVM)

## REVIEW OF LITERATURE

Machine learning algorithms have been extensively used for predicting health outcomes. Regarding asthma prediction, Finkelstein and Jeong [3] used data submitted by adult asthma patients during home telemonitoring to predict asthma exacerbations before they occur. Using a 7-day window, a naïve Bayesian classifier, adaptive Bayesian network, and support vector machines were able to predict asthma exacerbations occurring on day 8 with accuracy of 0.77, 1.00, and 0.80 respectively. Using data from the 2016 National Survey of Children's health, Harvey and Kumar [4] developed linear regression, decision trees, random forest, K-nearest neighbors and Naïve Bayes models for prediction of asthma development in children. Of all those classifiers, random forest resulted in highest prediction accuracy of 90.9%.

Zein et.al.[5] predicted asthma exacerbation using data extracted from electronic health records (EHRs) of asthma patients treated at the Cleveland Clinic from 2010 through 2018. The study used demographic information, comorbidities, laboratory values, and asthma medications as covariates. Logistic regression, random forests, and gradient boosting decision trees were used for the prediction. Light gradient boosting machine was found to be the best model with area under the curve (AUC) of 0.71. Risk factors included age, long-acting β agonist, high-dose inhaled glucocorticoid, or chronic oral glucocorticoid therapy. The study also predicted emergency department visits, and hospitalizations.

Previous studies have also explored risk factors of asthma among US population. Gwynn [6] used data from 2000 BRFSS and found female sex, age-group between 18-34 years, lower socioeconomic status, obesity, current and former smokers as potential risk factors. Using BRFSS 2006 – 2010 data, Hsu et.al. [7] found female sex, clinical comorbidities (Chronic Obstructive Pulmonary Disease, Coronary Artery Disease), depression, mold in the home, obesity and financial barriers to asthma-related health care to be significantly associated with asthma-related hospitalizations and emergency departments or urgent care center visits (ED/UCV) among older adults.

Zahran and Bailey [8] studied factors associated with asthma from 2009-2010 BRFSS data. Higher asthma prevalence was found in adults with low income, obesity, current and former smoking habits, and having health insurance. Greenblatt et.al. [9] studied gender specific determinants of asthma among US adults from BRFSS datasets corresponding to years 2007-2012. Important factors identified were gender, obesity, current smoking habits, low income, among others.

Rivera et. al. [10] found US military service members deployed in Iraq and Afghanistan had higher rates of new-onset asthma than those who did not deploy. Study by Ehrlich et.al. [11] revealed that respondents with diabetes had higher asthma rates than those without diabetes.

Besides asthma, predictive models have been developed for detecting other conditions such as diabetes [12], breast cancer [13], coronary artery disease [14], among others.

## METHODOLOGY

### Data Source

Data for this study is taken from 2019 Behavioral Risk Factor Surveillance System (BRFSS). BRFSS is administered and supported by CDC and field operations are managed by state health departments in all the states in the US and participating US territories. It is a telephone survey designed to collect data on health-related risk behaviors, chronic health conditions, health care access, and use of preventive services form the noninstitutionalized adult population (≥18 years) residing in the United States [15]. This study utilizes the 2019 BRFSS data collected only from Michigan.

### Data Cleaning

First of all, BRFSS data which was in .sas format was imported to SAS 9.4. The dataset was huge as it consisted of data from all the states of US. Thus, we only selected data for Michigan state and exported it as excel(.xlsr) format for further data cleaning. The excel format obtained from SAS was imported to R program. Out of 354 variables, only those variables which were found to be relevant based on the past literature reviews were selected using dplyr package [16] in R. New dataset consisting of 17 variables was made and further data cleaning was done.

Table 1: Variable Description

| Variable Type | Variable Names (Labels) |
|---|---|
| Demographic/Socio-economic Characteristics | SEX (Sex), AGE_G (Age group), EDUCAG (Education), INCOMG (Income), IMPRACE (Race), URBSTAT (Urban/Rural Status), VETERAN3 (Veteran Status), EMPLOY1 (Employment Type), MEDCOST (Could not see doctor because of cost) |
| Personal Habits | SMOKER3 (Smoker Status), USENOW3 (Smokeless tobacco products), SMOKE100 (Smoked at least 100 cigarettes) |
| Health Characteristics | DIABETE4 (Diabetes), BMI5CAT (Body Mass Index), CHCCOPD2 (Chronic Obstructive Pulmonary Disease), FLUSHOT7 (Flu shot/spray past 12 months) |

Regarding missing values for the response variable ASTHMA, missing values were removed and rows of other variables with respect to those missing value were also removed. The table below summarizes the distribution of missing values for each predictor considered in the study.

Table 2: Distribution of missing values

| Variable Name | % of Missing Values | Variable Name | % of Missing Values |
|---|---|---|---|
| SEX | 0.0 | MEDCOST | 0.31 |
| AGE_G | 0.0 | SMOKER3 | 3.19 |
| EUCAG | 0.27 | USENOW3 | 2.69 |
| INCOMG | 0 | SMOKE100 | 3.08 |
| IMPRACE | 0 | DIABETE4 | 0.15 |
| URBSTAT | 0 | BMI5CAT | 6.24 |
| VETERAN3 | 0.52 | CHCCOPD2 | 0.53 |
| EMPLOY1 | 1.13 | FLUSHOT7 | 5.96 |

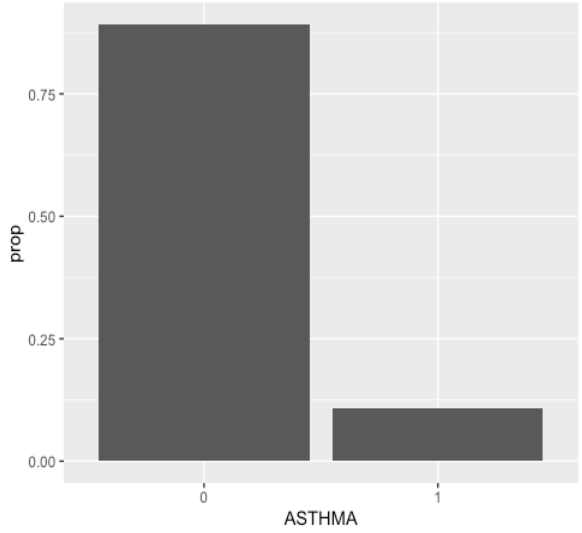
For explanatory variables *mice* package [17] was used to impute missing data. This imputes values for missing data based on the data characteristics. For continuous variable, it uses predictive mean matching (PMM) and for categorical variables with binary class it uses logistic regression for imputation. Regarding categorical variables with multiple classes, it uses multinominal logistic regression. We ran 3 number of imputations with 5 iterations to compare the imputed value and imputation 2 was selected after comparing the imputed values among imputation 1, 2 and 3. After missing value imputation, values imputed in numeric forms were given labels to provide meaning to corresponding categories they belong to. For example, in original dataset “SEX” was given categories of 1 for “male” and 2 for “female”. Accordingly, we converted imputed numbers to “male” and “female” as category.

## Data Analysis Method

### Descriptive and Visualization Analysis

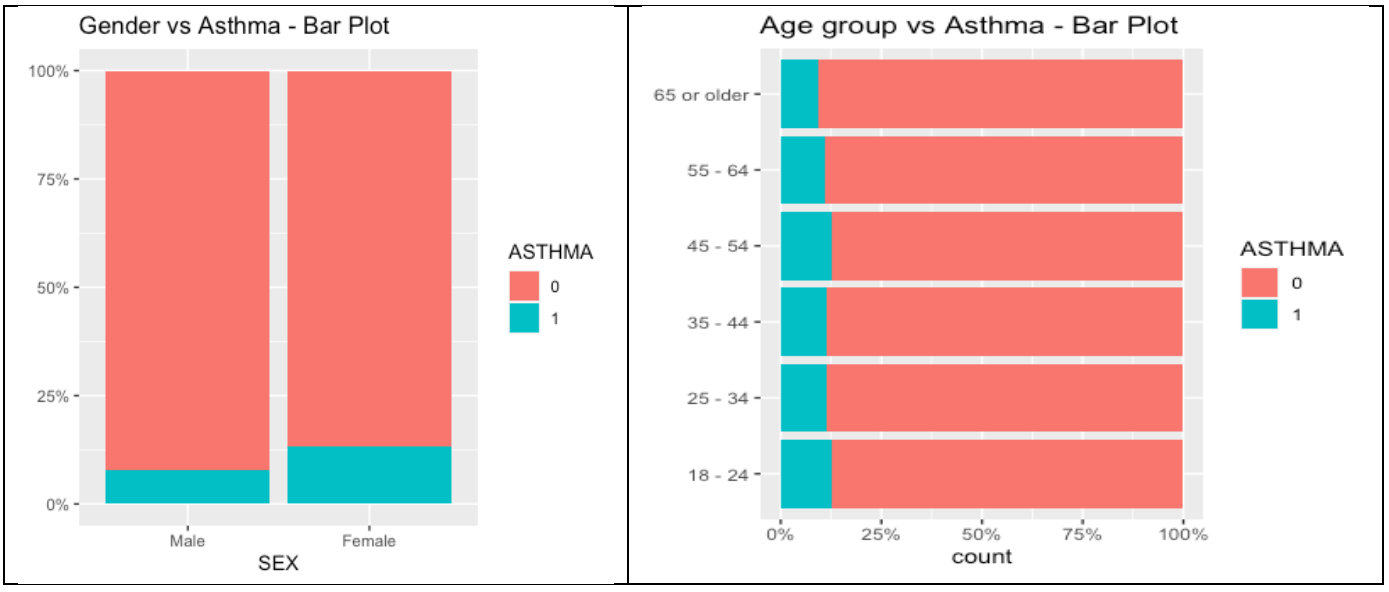
Data have been summarized in graphs and tables in this section. Distribution of predictors with respect to response variable have been presented in graphs. In addition, the chi-square test of association between variables was conducted and the results are summarized in a table later.

*Fig 1: Distribution of the response variable*

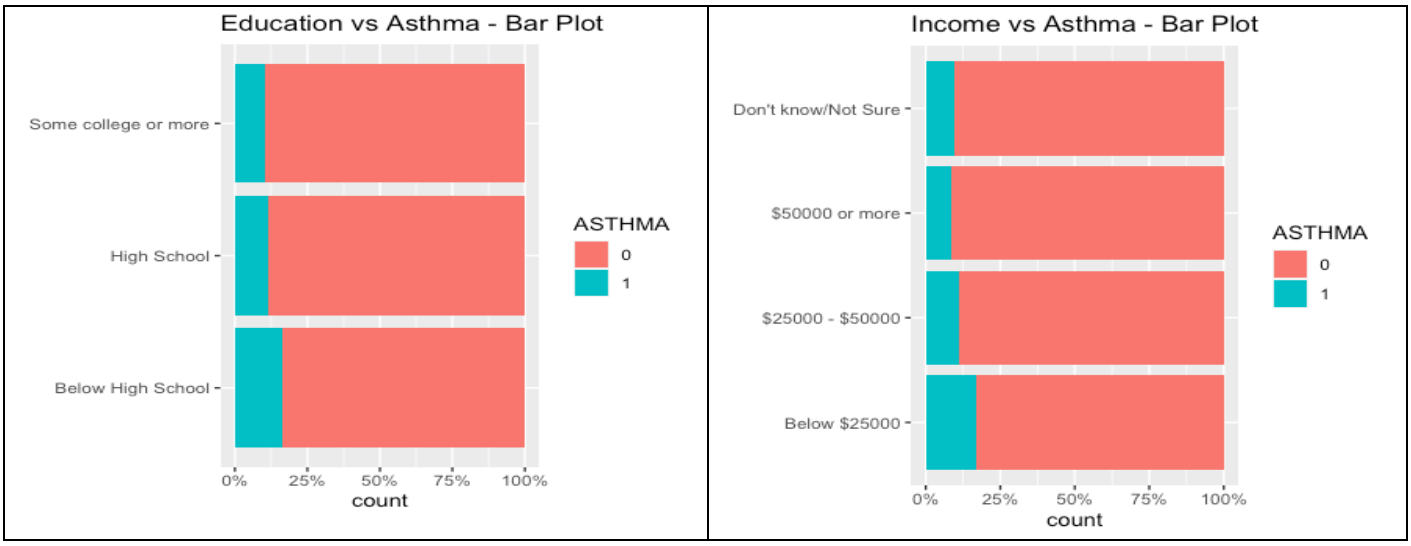



Regarding the distribution of asthma among the survey participants, 10.8% were told by doctor (nurse or other health professional) that they had asthma at some period before the survey and were still having it. Majority (89.2%) of the survey participants reported not having asthma.

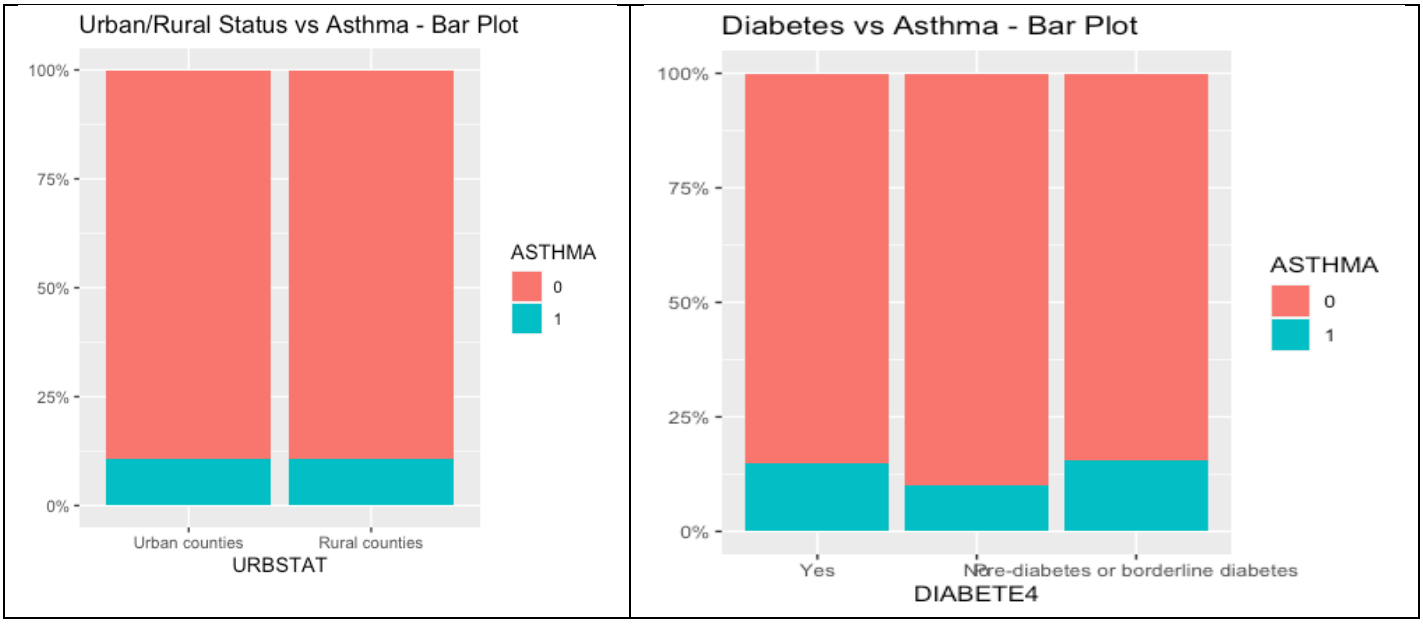
*Fig 2: Distribution of the Response Variable Vs Some Important Predictors*



Asthma was found to be more prevalent among females than in males. Previous studies have found the same regarding asthma prevalence in the two genders [6,9]. The lowest age-group of the study participants: between 18-24 years were the highest prevalent group.

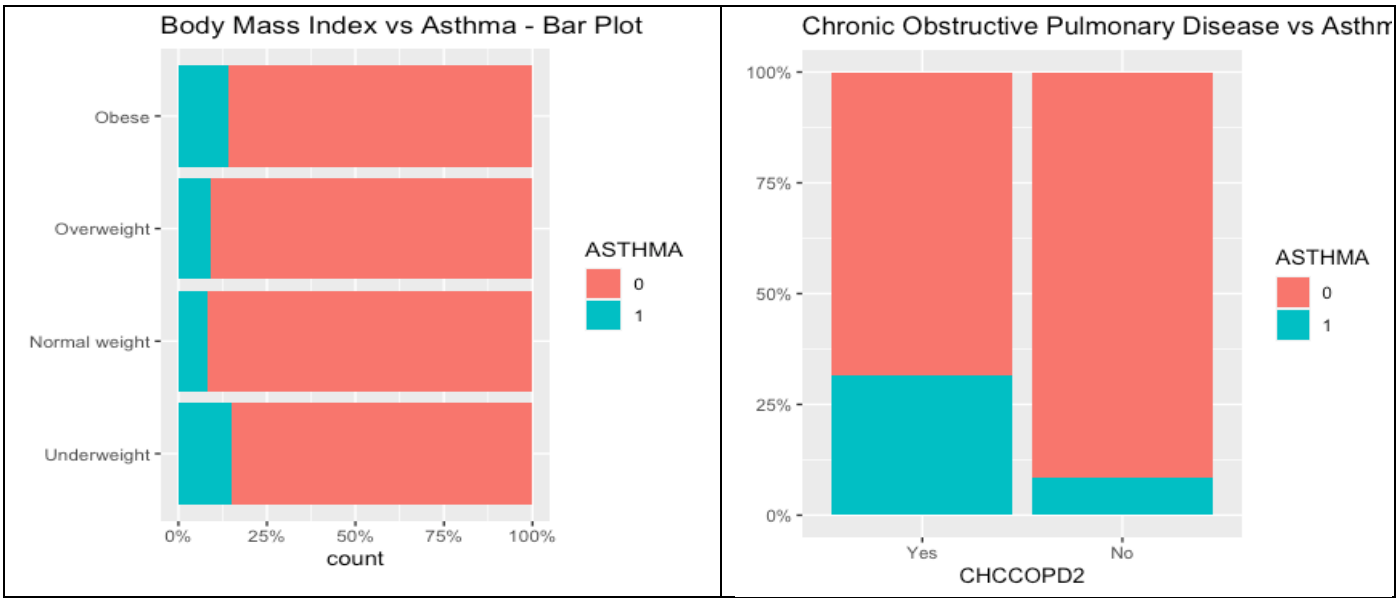



Those who attended below high school had the highest asthma rates while those who were graduated from college/technical school had the lowest. This somehow establishes that participants' education level had a bearing on asthma prevalence. Regarding income level, participants in the lowest income level had the highest asthma rates.




Participants residing in urban counties have higher asthma rates than in rural counties. Diabetes seemed to be having some association with asthma. People with pre-diabetes or in the borderline

diabetes had highest asthma rates. On the other hand, people with no diabetes reported lowest asthma rates.




As established by several other studies [6,9], obese people were the ones with highest rate of asthma prevalence. Chronic obstructive pulmonary disease (COPD) showed an association with asthma as people with COPD had higher rates of asthma than people without COPD.

## Model Building

We built eight predictive models for the prediction of response variable ASTHMA. In building these models, data were first partitioned into training and testing data for the purpose of building and testing the developed models, respectively. As figure 1 shows, the distribution of response variable is quite imbalanced, so ROSE (Random Over-Sampling Examples) technique [18] was used on training data to address the challenges posed by imbalanced data. Due to imbalanced classes, the classifiers were biased towards the majority class. When one of the class in binary classification task is rare, ROSE produces a synthetic, possibly balanced sample of data simulated according to a smoothed-bootstrap approach.

All models were built using **caret package** (short for Classification And REgression Training) [19] in R. The package contains functions to streamline the model training process for complex regression and classification problems. A 10-fold cross-validation was used and parameters were tuned to obtain optimal models from training data after addressing the issue of imbalanced classes.

## Model Development

Predictive model building techniques employed in this study have been briefly summarized below. A detailed discussion of these can be found in [20,21].

*Logistic Regression*: This is parametric modeling technique and has a major advantage of being easily interpreted. The technique is widely used for categorical target, especially for binary target. The logistic regression uses logistic function,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

With some manipulation,

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

The logarithm term in the left-hand side is called the log-odds or logit.

*Partial Least Squares*: This is a dimension reduction method in which the input variables are first transformed and combined into a smaller number of new variables $Z_1$, …, $Z_M$ called principal components. A linear model is then fitted via least squares using these M new variables. This technique uses the response Y to identify new features that not only approximate the old features well, but also are related to the response. The tuning parameter is the number of principal components. Setting tune length in R will test different values of the tuning parameter. The optimal value is selected so that the cross-validation error is minimized.

*Random Forest*: This involves building several decision trees on bootstrapped training samples. When these trees are built, in each split only a random sample of m predictors is chosen as split candidates from the full set of p predictors. By allowing only a subset of predictor at each split, the correlation between trees will be reduced which in turn makes the average of resulting trees less variable. When a large number of correlated predictors are present, a small value of m is desirable. This m, known as tuning parameter for random forest algorithm is tuned to get optimal model.

*Gradient Boosting*: Random Forest involves building several trees on bootstrapped data sets. Hence, each tree is independent of the other. With boosting, each tree is grown using information from previously grown trees. Unlike random forest, boosting uses a modified version of the original data set to fit each tree. This method has three tuning parameters: the number of trees,

shrinkage parameter that controls rate at which boosting learns, and number of splits in each tree. Different values of these parameters are tried to achieve the final best model.

*LASSO*: This is known as a shrinkage method that shrinks the coefficient estimates towards zero while fitting a model. While the least squares method minimizes the residual sum of squares (RSS) given by $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)$, LASSO works by minimizing the quantity $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) + \lambda\sum_{j=1}^{p}|\beta_j|$ where the second term is called a shrinkage penalty. The tuning parameter λ controls the relative impact of the two terms viz. RSS and shrinkage penalty on the regression coefficient estimates. The value of λ is often chosen by cross-validation method. We provided a grid of λ values to choose the best one from. LASSO produces simpler and more interpretable models that involve only a subset of predictors since some of the estimates shrink to zero.

Elastic Net: This method seeks to minimize a different quantity compared to LASSO which is given by $RSS + \lambda\sum_{j=1}^{p}\beta_j^2 + (1-\lambda)\sum_{j=1}^{p}|\beta_j|$. It is a weighted combination of LASSO and ridge regression (another shrinkage method). Elastic net controls the impact of correlated predictors. The two tuning parameters: λ for complexity and α for the compromise between LASSO and ridge are tuned to obtain the optimal model.

*K-Nearest Neighbors*: It is a simple, easy-to-implement supervised machine learning algorithm and assumes that similar things exist in close proximity. It is conducted by choosing the K (number of neighbors) that reduces the number of errors while making predictions. The choice of k, called the tuning parameter, has a drastic effect on the algorithm result.

*Support Vector Machine:* SVM, mostly used in classification problems can be used for regression as well. The idea behind it is each data value is plotted as a point in n-dimensional space and we seek for a hyper-plane that differentiates the class very well. The hyperplane can be linear or nonlinear. Parameter tuning is done for the tuning parameter C, known as cost, that essentially imposes a penalty to the model for making an error in classification.

### Model Comparison

To compare the performance of the models developed, confusion matrix was obtained to note down accuracies, sensitivity and specificity of each model. Similarly, area under the Receiver Operating Characteristics (ROC) curve was computed. Other important statistics such as F Measure, G Measure, Matthew's Correlation Coefficient, and Cohen's Kappa that are particularly useful while dealing with imbalanced classes [22] were also computed to decide about the best classifying algorithm. Due to huge proportion of non-event cases (ASTHMA = 0) in the population, we were more focused on the sensitivity of the predictive models.

## RESULTS

We start by studying the statistical significance of the association between ASTHMA and predictors considered in the study. Chi-square test has been used for this purpose and the corresponding p-values are noted. P-values less than 0.05 are considered statistically significant.

Table 3: Association between response and predictors

| **Variables** | **ASTHMA** | | **Chi-Square** | **P-Value** |
|---|---|---|---|---|
| | **No** | **Yes** | | |
| **SEX** | | | | |
| Female | 5043 (86.8) | 764 (13.2) | 73.396 | <0.0001 |
| Male | 4241 (92.1) | 363 (7.9) | | |
| **AGE_G** | | | | |
| 18-24 | 635(87.5) | 91(12.5) | 18.342 | 0.002 |
| 25-34 | 962(88.6) | 124(11.4) | | |
| 35-44 | 1009(88.7) | 129(11.3) | | |
| 45-54 | 1332(87.2) | 195(12.8) | | |
| 55-64 | 1808(88.9) | 225(11.1) | | |
| 65 or older | 3538(90.7) | 363(9.3) | | |
| **EDUCAG** | | | | |
| Below High School | 342(83.6) | 67(16.4) | 15.7 | <0.001 |
| High School | 2377(88.7) | 304(11.3) | | |
| Some college or more | 6565(89.7) | 756(10.3) | | |
| **INCOMG** | | | | |
| Below $25000 | 1597(83.04) | 326(16.96) | 100.33 | <0.0001 |
| $25000 - $50000 | 1892(89.0) | 234(11.0) | | |
| $50000 or more | 4172(91.8) | 394(8.2) | | |
| Don't know/Not Sure | 1623(90.7) | 173(9.3) | | |
| **IMPRACE** | | | | |

| | | | | |
|---|---|---|---|---|
| White, Non-Hispanic | 7804(89.6) | 903(11.4) | | |
| Black, Non-Hispanic | 743(85.5) | 126(14.5) | | |
| Asian, Non-Hispanic | 145(95.6) | 7(4.4) | 28.819 | <0.001 |
| Hispanic | 244(89.7) | 28(11.3) | | |
| Other | 348(84.7) | 63(15.3) | | |
| **URBSTAT** | | | | |
| Urban counties | 8539(89.8) | 1037(10.2) | <0.0001 | 1 |
| Rural counties | 745(89.2) | 90(10.8) | | |
| **VETERAN3** | | | | |
| Yes | 1056(92.6) | 85(7.4) | 14.735 | <0.001 |
| No | 8228(88.6) | 1042(11.4) | | |
| **EMPLOY1** | | | | |
| Employed | 4458(90.9) | 485(10.1) | | |
| Out of work | 391(87.8) | 57(12.2) | | |
| Homemaker | 450(88.9) | 56(11.1) | 136.84 | <0.0001 |
| Student | 288(88.1) | 39(11.9) | | |
| Retired | 3118(90.8) | 315(9.2) | | |
| Unable to work | 579(76.8) | 175(13.2) | | |
| **SMOKER3** | | | | |
| Current smoker | 1377(85.8) | 228(14.2) | | |
| Former smoker | 2770(89.8) | 315(10.2) | 22.464 | <0.0001 |
| Never smoked | 5137(89.8) | 584(10.2) | | |
| **USENOW3** | | | | |
| Yes | 267(89.3) | 32(10.7) | <0.0001 | 1 |
| No | 9017(89.2) | 1095(10.8) | | |
| **SMOKE100** | | | | |
| Yes | 4134(88.4) | 540(11.6) | 4.523 | 0.033 |
| No | 5150(89.7) | 587(10.3) | | |
| **DIABETE4** | | | | |
| Yes | 1256(85.7) | 217(14.3) | | |
| No | 7843(90.0) | 876(10.0) | 33.769 | <0.0001 |
| Pre-diabetes or borderline diabetes | 185(84.5) | 34(15.5) | | |
| **BMI5CAT** | | | | |
| Under weight | 135(84.9) | 24(15.1) | | |
| Normal Weight | 2674(91.8) | 238(8.2) | 79.77 | <0.0001 |
| Overweight | 3332(90.7) | 340(9.3) | | |
| Obese | 3143(85.7) | 525(14.3) | | |
| **CHCCOPD2** | | | | |
| Yes | 692(68.5) | 318(31.5) | 492.2 | <0.0001 |
| No | 8592(91.4) | 809(8.6) | | |

| | | | | |
|---|---|---|---|---|
| **MEDCOST** | | | | |
| Yes | 852(84.1) | 161(15.9) | 29.283 | <0.0001 |
| No | 8432(89.7) | 966(10.3) | | |
| **FLUSHOT7** | | | | |
| Yes | 4388(88.3) | 581(11.7) | 7.238 | 0.007 |
| No | 4896(90.0) | 546(10.0) | | |

As table 3 shows, among the predictors considered in the study, we found asthma to have a statistically significant association with demographic and socio-economic variables such as sex, age-group, education, income, race, veteran status, employment, and cost to see a doctor. Personal habits such as smoker status and smoked at least 100 cigarettes were also significantly associated with asthma. Regarding health characteristics, diabetes, body mass index, chronic obstructive pulmonary disease, and flu shot/spray during past 12 months were found to have a significant association with asthma. These predictors were used to build predictive models. Predictors that did not have significant associations were dropped.

The table below presents summary of the eight predictive models built and evaluated on the test data set.

Table 4: Model performance in the test data

| Technique | Sensitivity | Accuracy | Area under ROC curve | F measure | G measure | Matthew's Correlation Coefficient | Cohen's Kappa |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.6114 | 0.6828 | 0.6516 | 0.3016 | 0.6504 | 0.2016 | 0.1598 |
| Partial Least Squares | 0.6029 | 0.6869 | 0.6552 | 0.3014 | 0.6485 | 0.2006 | 0.1603 |
| Random Forest | 0.4029 | 0.7209 | 0.5819 | 0.2444 | 0.5537 | 0.1182 | 0.1046 |
| Gradient Boosting | 0.5857 | 0.6911 | 0.6506 | 0.2982 | 0.6423 | 0.1949 | 0.1574 |
| LASSO | 0.6171 | 0.6847 | 0.6552 | 0.3049 | 0.6541 | 0.2064 | 0.1638 |
| Elastic Net | 0.6114 | 0.6837 | 0.6523 | 0.3023 | 0.6509 | 0.2024 | 0.1607 |
| KNN | 0.5743 | 0.5896 | 0.5829 | 0.2387 | 0.5829 | 0.1057 | 0.0744 |
| SVM (Linear Kernel) | 0.5171 | 0.7455 | 0.6457 | 0.3129 | 0.6328 | 0.2100 | 0.1856 |

Regarding the selection of best model among the eight, based on the value all statistics computed in table 4 and ease of interpretability, logistic regression was selected as the best model. For this model, nine inputs were found to be statistically significant at 5% level of significance. The following table shows the order of variable importance for each classification algorithms.

Table 5: Variable Importance Order

| | | Models | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Factors Selected | LR | PLS | RF | GB | LASSO | EN | KNN | SVM |
| CHCCOPD2 | 1* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| SEX | 2 | 3 | 6 | 2 | 3 | 4 | 2 | 2 |
| FLUSHOT7 | 3 | 7 | 2 | 5 | 8 | 9 | 6 | 6 |
| AGE_G | 4 | 8 | 9 | 8 | 4 | 3 | 5 | 5 |
| DIABETE4 | 5 | 6 | 5 | 7 | 7 | 6 | 10 | 10 |
| EMPLOY1 | 6 | 5 | 12 | 4 | 5 | 5 | 8 | 8 |
| IMPRACE | 7 | | 10 | 12 | 2 | 2 | | |
| INCOMG | 8 | 4 | 3 | 10 | 10 | 8 | 4 | 4 |
| VETERAN3 | 9 | 11 | | 6 | 9 | 11 | 12 | |
| SMOKE100 | 10 | | | | | 7 | | 12 |
| MEDCOST | 11 | 9 | 8 | | 11 | | 7 | 7 |
| SMOKER3 | 12 | 10 | 4 | 11 | | 10 | 9 | 9 |

**denotes the most important variable.*

The best model i.e. logistic regression was further explored to obtain important information. We were basically focused on the odds ratios and their p-values.

Table 6: Parameter and Odds ratio estimates of logistic regression model

| Variables | Parameter Estimate | Standard Error | Odds ratio | P-value |
|---|---|---|---|---|
| **CHCCOPD2** | | | | |
| Yes (Ref.) | | | | |
| No | -1.514 | 0.083 | 0.220 | <0.001 |
| **SEX** | | | | |
| Female | 0.533 | 0.057 | 1.704 | <0.001 |
| Male (Ref.) | | | | |
| **FLUSHOT7** | | | | |
| Yes (Ref.) | | | | |
| No | -0.316 | 0.054 | 0.729 | <0.001 |
| **AGE_G** | | | | |

| | | | | |
|---|---|---|---|---|
| 18-24 (Ref.) | | | | |
| 25-34 | -0.157 | 0.125 | 0.855 | 0.208 |
| 35-44 | -0.284 | 0.129 | 0.753 | 0.028 |
| 45-54 | -0.481 | 0.125 | 0.618 | <0.001 |
| 55-64 | -0.685 | 0.123 | 0.504 | < 0.001 |
| 65 and above | -0.663 | 0.133 | 0.515 | < 0.001 |
| **DIABETE4** | | | | |
| Yes (Ref.) | | | | |
| No | -0.375 | 0.075 | 0.687 | <0.001 |
| Pre-diabetes or borderline diabetes | 0.238 | 0.181 | 1.269 | 0.189 |
| **EMPLOY1** | | | | |
| Employed for wages (Ref.) | | | | |
| Out of work | 0.101 | 0.119 | 1.106 | 0.4 |
| Homemaker | -0.203 | 0.123 | 0.816 | 0.101 |
| Student | 0.092 | 0.152 | 1.096 | 0.547 |
| Retired | -0.386 | 0.086 | 0.680 | <0.001 |
| Unable to work | 0.447 | 0.105 | 1.564 | <0.001 |
| **IMPRACE** | | | | |
| White, Non-Hispanic (Ref.) | | | | |
| Black, Non-Hispanic | 0.131 | 0.092 | 1.140 | 0.157 |
| Asian, Non-Hispanic | -1.237 | 0.293 | 0.290 | <0.001 |
| Hispanic | -0.599 | 0.17 | 0.549 | <0.001 |
| Other | 0.082 | 0.117 | 1.085 | 0.486 |
| **INCOMG** | | | | |
| Below $25000 (Ref.) | | | | |
| $25000 - $50000 | -0.137 | 0.082 | 0.872 | 0.094 |
| $50000 or more | -0.302 | 0.077 | 0.739 | <0.001 |
| Don't know/Not Sure | -0.339 | 0.085 | 0.712 | <0.001 |
| **VETERAN3** | | | | |
| Yes (Ref.) | | | | |
| No | 0.291 | 0.096 | 1.338 | 0.002 |
| **SMOKE100** | | | | |
| Yes (Ref.) | | | | |

| | | | | |
|---|---|---|---|---|
| No | 0.521 | 0.298 | 1.684 | 0.081 |
| **MEDCOST** | | | | |
| Yes (Ref.) | | | | |
| No | -0.149 | 0.086 | 0.862 | 0.084 |
| **SMOKER3** | | | | |
| Current smoker (Ref.) | | | | |
| Former smoker | -0.005 | 0.086 | 0.995 | 0.955 |
| Never smoked | -0.477 | 0.299 | 0.621 | 0.11 |

Based on the odds ratio, the effect of some selected inputs can be interpreted as follows:

- People without chronic obstructive pulmonary disease are 78% less likely to report asthma compared to people with it.
- Females are nearly 70% more likely to develop asthma than males. Using 2000 BRFSS data, Gwynn (2009) found females to be 91% more likely. Similarly, Greenblatt et.al. (2017) found females 80% more likely than men from 2007-2012 BRFSS datasets.
- Adults who have not taken flu shot were nearly 27% less likely to report current asthma.
- People in all other age-groups are less likely to get asthma than people in 18 – 24 years age-group. Gwynn (2009) found that adults aged 35-64 and ≥65 years were less likely to report current asthma than adults aged 18-34 years.
- People without diabetes had nearly one-third less chances of reporting current asthma than people with diabetes. Study by Ehrlich [11] found after adjusting effects of some variables, patients with diabetes had a hazard ratio of 1.08 [95% CI 1.03-1.12] to develop asthma.
- Regarding race, both Asian and Hispanic adults were less likely to report current asthma than White, Non-Hispanic. The American Lung Association [23] also reports the same regarding asthma disparities among races.
- People in the higher income categories were less likely to report current asthma than people in the lowest income category (below $25000). Greenblatt et al. [9] also found the lowest income category to be two-third more likely to report asthma.

# DISCUSSION, CONCLUSION AND FUTURE WORK

Using a secondary dataset from a national survey, we attempted to develop several machine learning models for the prediction of asthma and compared their performances on the test data set. Important inputs were also identified from each of the model.

The issues that come with using secondary dataset such as missing values were addressed during the data cleaning stage. The accuracies of all models (except KNN) developed were similar at around 70%. Logistic Regression (LR), LASSO, and elastic net (EN) had comparable performances among all algorithms. Logistic regression model was selected as the final best model based on accuracy measures, area under ROC curve and ease of interpretability. As table 5 shows Chronic obstructive pulmonary disease, Sex, Age group, Diabetes, Employment, Race, and Income were among the most important predictors particularly from LR, LASSO, and EN.

Our work discusses the issues related to missing data, imbalanced data and then predictive models are built. The best model can be used to make predictions of asthma development and implement early intervention for the treatment in higher risk groups.

Some studies such as the followings may be conducted in future to know about the disease even better:

- Similar study may be conducted in other states besides Michigan to see the model performances and explore important inputs.
- Predictive models may also be built for data from entire United States. However, while doing so, one should be aware about the differences caused due to geographical variations.
- Additional models such as Neural Networks may be developed to compare performance with models in this study.
- BRFSS surveys are conducted every year; therefore, using annual survey data, asthma trends over the years may be studied.

## REFERENCES


1. Center for Disease Control and Prevention (CDC). Asthma. https://www.cdc.gov/asthma/default.htm Accessed on July 15, 2021

2. Center for Disease Control and Prevention (CDC). BRFSS Asthma Prevalence Data. https://www.cdc.gov/asthma/brfss/default.htm Accessed on July 19, 2021

3. Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. Ann N Y Acad Sci. 2017 Jan;1387(1):153-165. doi: 10.1111/nyas.13218. Epub 2016 Sep 14. PMID: 27627195; PMCID: PMC5266630

4. Harvey, Julie & Kumar, Sathish. (2019). Machine Learning for Predicting Development of Asthma in Children. 596-603. 10.1109/SSCI44817.2019.9002692

5. Zein JG, Wu CP, Attaway AH, Zhang P, Nazha A. Novel Machine Learning Can Predict Acute Asthma Exacerbation. Chest. 2021 May;159(5):1747-1757. doi: 10.1016/j.chest.2020.12.051. Epub 2021 Jan 10. PMID: 33440184; PMCID: PMC8129731.

6. Gwynn RC. Risk factors for asthma in US adults: results from the 2000 Behavioral Risk Factor Surveillance System. J Asthma. 2004 Feb;41(1):91-8. doi: 10.1081/jas-120026066. PMID: 15046383.

7. Hsu, J., Chen, J., & Mirabelli, M. C. (2018). Asthma Morbidity, Comorbidities, and Modifiable Factors Among Older Adults. *The journal of allergy and clinical immunology. In practice*, *6*(1), 236–243.e7. https://doi.org/10.1016/j.jaip.2017.06.007

8. Hatice S. Zahran & Cathy Bailey (2013) Factors Associated with Asthma Prevalence among Racial and Ethnic Groups—United States, 2009–2010 Behavioral Risk Factor Surveillance System, Journal of Asthma, 50:6, 583-589, DOI: 10.3109/02770903.2013.794238

9. Greenblatt, R., Mansour, O., Zhao, E. *et al.* Gender-specific determinants of asthma among U.S. adults. *asthma res and pract* **3,** 2 (2017). https://doi.org/10.1186/s40733-017-0030-5

10. Rivera, A. C., Powell, T. M., Boyko, E. J., Lee, R. U., Faix, D. J., Luxton, D. D., Rull, R. P., & Millennium Cohort Study Team (2018). New-Onset Asthma and Combat Deployment: Findings From the Millennium Cohort Study. *American journal of epidemiology*, *187*(10), 2136–2144.

11. Ehrlich SF, Quesenberry CP Jr, Van Den Eeden SK, Shan J, Ferrara A. Patients diagnosed with diabetes are at increased risk for asthma, chronic obstructive pulmonary

disease, pulmonary fibrosis, and pneumonia but not lung cancer. *Diabetes Care*. 2010;33(1):55-60. doi:10.2337/dc09-0880

12. Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo Á, Barreto SM, Duncan BB. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. Sao Paulo Med J. 2017 May-Jun;135(3):234-246. doi: 10.1590/1516-3180.2016.0309010217. PMID: 28746659.

13. M. R. Ahmed, M. A. Ali, J. Roy, S. Ahmed and N. Ahmed, "Breast Cancer Risk Prediction based on Six Machine Learning Algorithms," *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1-5, doi: 10.1109/CSDE50874.2020.9411572.

14. Dahal, K. and Gautam, Y. (2020) Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. *Open Journal of Statistics*, **10**, 694-705. doi: 10.4236/ojs.2020.104043.

15. Center for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System. Survey Data & Documentation. https://www.cdc.gov/brfss/data_documentation/index.htm. Accessed July 18, 2021.

16. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Gra mmar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/packa ge=dplyr

17. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

18. Nicola Lunardon, Giovanna Menardi, and Nicola Torelli (2014). ROSE: a Package for Bi nary Imbalanced Learning. R Journal, 6(1), 82-92.

19. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28*(5), 1 - 26. doi:http://dx.doi.org/10.18637/jss.v028.i05

20. G. James, D. Witten, T. Hastie, R. Tibshirani. (2013). An Introduction to Statistical Learning. Springer Texts in Statistics.

21. Zou, Hui; Hastie, Trevor (2005). "Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society, Series B*. **67** (2): 301–320

22. Akosa, J. (2017). Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc. Available at https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf

23. American Lung Association. Current Asthma Demographics. https://www.lung.org/research/trends-in-lung-disease/asthma-trends-brief/current-demographics. Accessed July 18, 2021.