Predicting Asthma using Imbalanced Data Modeling Technique: Evidence from 2019 Michigan BRFSS Data

Nirajan Budhathoki[1*], Ramesh Bhandari[2], Suraj Bashyal[3] and Carl Lee[1]

[1]Department of Statistics, Actuarial & Data Sciences, Central Michigan University, USA
[2]Department of Physics, Central Michigan University, USA
[3]Department of Geography & Environmental Studies, Central Michigan University, USA

*budha1n@cmich.edu

Abstract

Studies in the past have explored asthma prevalence in the United States using national survey data. The findings may not be applicable for specific states due to different environmental and socioeconomic characteristics. Using data from the 2019 Behavioral Risk Factor Surveillance System (BRFSS), several modern machine learning techniques are applied to predict asthma and identify risk factors of asthma among adults in Michigan. A total of 10,518 participants were surveyed in Michigan for 2019 BRFSS. A sample of 10,411 individuals, resulted after data cleaning, is analyzed. Of those, 1127 (10.8%) reported having asthma during the survey time. The typical machine learning techniques perform poorly due to the problem of imbalanced data. In this regard, random over-sampling examples (ROSE) method is applied to generate synthetic data. The performances of logistic regression, partial least squares, LASSO, and elastic net are somewhat similar with sensitivity at around 61% and area under the curve (AUC) at around 65%. Due to ease of interpretability, logistic regression is chosen for further exploration. Presence of chronic obstructive pulmonary disease, female sex, taken flu shot/spray in the past twelve months, 18-24 age group, inability to work, being white, non-hispanic, and lower income level are identified as important predictors.

Introduction

In general terms, the Centers 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. In the survey, 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].

In this study, we attempt to develop predictive models and evaluate their performance in predicting asthma among Michigan adult population. Although studies in the past have identified risk factors of asthma in the US population [3,4], no studies have focused specifically on Michigan population. Knowing important modifiable factors can help to design appropriate interventions for asthma control. Since the data was imbalanced, i.e., observations in one of the categories in response variable was significantly fewer than the other category, a popular synthetic data generation technique known as Random Over-Sampling Examples (ROSE) is referred. Machine learning models namely logistic regression, partial least squares, random forest, gradient boosting, least absolute shrinkage and selection operator (LASSO), elastic net, K-nearest neighbors, and support vector machine are built on the balanced training set. Model performances are compared to find the "best" model based on the predictive ability on the test dataset. Variable importance order as identified by each model are presented. Further exploration of the selected model is made to study association of asthma with the predictors considered.

Review of Literature

Machine learning algorithms have been extensively used for predicting health outcomes such as diabetes [5], breast cancer [6], coronary artery disease [7], among others. Regarding asthma prediction, using data from the 2016 National Survey of Children's health, Harvey and Kumar [8] developed linear regression, decision trees, random forest, K-nearest neighbors and naïve bayes models for prediction of asthma development in children. The study included demographic and health related variables such as information on allergies, sleeping habits, doctor visits, etc. Among all the classifiers considered, random forest resulted in highest prediction accuracy of 90.9%.

Zein et.al.[9] 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.

Finkelstein and Jeong [10] 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.

Previous studies have also explored risk factors of asthma among US population using BRFSS data. Gwynn [3] 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. Zahran and Bailey [4] 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.

Using BRFSS 2006 – 2010 data, Hsu et.al. [11] 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.

Greenblatt et.al. [12] 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. Similarly, Rivera et. al. [13] 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. [14] revealed that respondents with diabetes had higher asthma rates than those without diabetes.

Regarding the use of imbalanced data modeling techniques in machine learning, Alghamdi et.al. [15] developed models to predict diabetes mellitus from an imbalanced data that was treated using synthetic minority oversampling technique (SMOTE). Liang and Martell [16] developed models for sleep/awake classifications from Fitbit data. They compared the performances of four resampling strategies namely random up sampling, random down sampling, ROSE and SMOTE on the models developed. The study found ROSE to have a consistently better performance than the others.

Methodology

The process of selecting and comparing the optimal models was conducted in the following steps. Step 1 was to perform preliminary predictor variable selection using the Chi-Squared association test between ASTHMA and each predictor variable. Step 2 was to perform data cleansing including missing data imputation. Step 3 was to apply the ROSE method using the ROSE Package [17] in R [18] to generate synthetics to create a more balanced data during the pre-modeling process. Step 4 was to build and select the optimal model for each modeling technique using 10-fold cross-

validation. Step 5 was to apply each optimal model to the original data and compute several model performances measures for comparing the model performances.

Data Source and Variable Description

Data for this study was 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 [19]. This study utilizes the 2019 BRFSS data collected only from Michigan. BRFSS data are publicly available on CDC's website https://www.cdc.gov/brfss/annual_data/annual_2019.html. Of the 354 variables measured for different purposes, only those variables which were found to be relevant based on the past literature reviews were selected. Table 1 presents a description of the variables used in the study.

Table 1: Variable Description

Variable Type	Variable Names (Labels)
Demographic/Socioeconomic Characteristics (DS)	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 (PH)	SMOKER3 (Smoker Status), USENOW3 (Smokeless tobacco products), SMOKE100 (Smoked at least 100 cigarettes)
Health Characteristics (HC)	DIABETE4 (Diabetes), BMI5CAT (Body Mass Index), CHCCOPD2 (Chronic Obstructive Pulmonary Disease), FLUSHOT7 (Flu shot/spray past 12 months)

Missing values for the response variable ASTHMA were removed and rows of other variables with respect to those missing value were also removed. Table 2 summarizes the distribution of missing values for each explanatory variables 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

Missing value imputation for explanatory variables was carried out using *mice* package [20] in R. For the imputation of missing values in continuous variables, the package uses predictive mean matching (PMM) and for categorical variables with binary class, it uses logistic regression. Regarding categorical variables with multiple classes, it uses multinominal logistic regression. We ran three imputations, with different random seeds, each with five iterations to compare the imputed values and selected the run giving the best imputed values. Among the three imputed datasets, since all variables in the dataset were categorical, the one that had frequency distribution closest to the original dataset was selected. We chose a few variables such as 'USENOW3' and 'SMOKE100' to compare these frequency distributions in the original and imputed datasets.

Distribution of the participants by response variable ASTHMA is shown in Figure 1. Of the total participants in the study, 10.8% were told by doctor (nurse or other health professional) that they had asthma (ASTHMA = "YES") at some period before the survey and were still having it. Majority (89.2%) of the survey participants reported not having asthma (ASTHMA = "NO"). The distribution of explanatory variables for each level of ASTHMA and the association between ASTHMA and each explanatory variable was observed by Chi-Square test, which are presented in Table 4 in the Result section.

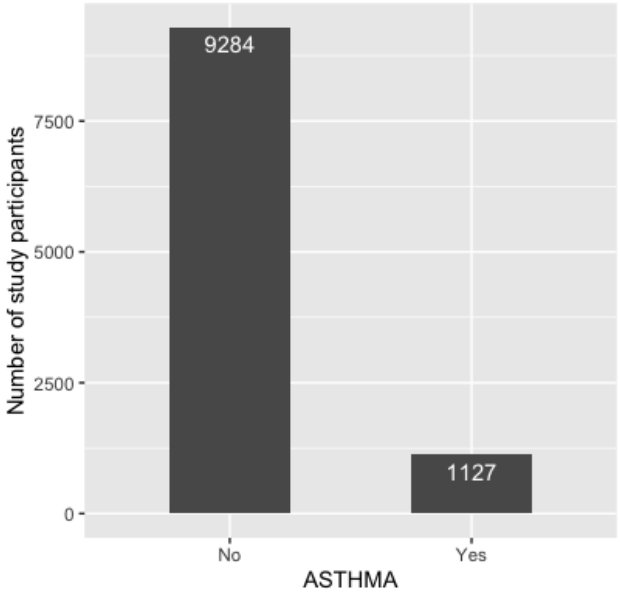

Fig1: Distribution of participants by Asthma Status

Brief summary of techniques for modeling Asthma data

The data were partitioned into training data (70%) and testing data (30%). The training data were used to build and select the best model for each modeling technique and test data were applied for model comparisons among different modeling techniques. All models were built using **caret package** (short for Classification And REgression Training) [21] 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. Predictive modeling techniques applied in this study are briefly summarized below. A detailed discussion of these can be found in [22,23].

Logistic Regression: This is a parametric modeling technique and has a major advantage of being easily interpreted. The technique is widely used for categorical target, especially for binary target (Y). Denoting set of predictors by X, the logistic regression uses logistic function,

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

With some manipulation,

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

The logarithm term in the left-hand side is called the log-odds or logit. The exponentiation of log-odds or logit gives odds ratio which indicate the odds of an event occurring in one group compared to the odds of it occurring in another group.

Partial Least Squares: Partial least squares (PLS) regression is a technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components, instead of on the original data. PLS regression is especially useful when predictors are highly collinear, or when you have more predictors than observations and ordinary least-squares regression either produces coefficients with high standard errors or fails completely. The technique reduces the number of predictors using a technique similar to principal components analysis to extract a set of components that describes maximum correlation between the predictors and response variables. Let Y, n x q matrix, be the response variables and X, n x p matrix, be a set of predictors. If Y is a univariate responsible variable, then, q=1. Assuming X and Y are standardized, then PLS method attempts to determine a linear decomposition of X and Y as X = TP' and Y = UQ', where T and U are both n x r matrix. The T matrix is the X-principal components. Each X-principal component is a linear combination of X, and U is the Y-principal components. Then, perform regression between T, as the predictors and U as the response variables.

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 sequentially using the information from previously grown trees. The technique starts with fitting a small tree, often with two or four terminal nodes. Given the current model, another small decision tree is fitted to the residual from the model. The new tree is added to the fitted function, and residual is recomputed. The fitted function is slowly improved until the improvement is ignorable. This method has three tuning parameters: the number of trees, shrinkage parameter that controls the rate at which boosting learns, and number of splits in each tree. Grid method and cross-validation are often applied 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. 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 nonparametric supervised machine learning algorithm. Each predicted response at a given predictor is the average of the responses at the K nearest predictors. The choice of K decides the smoothness of the predicted response surface.

Support Vector Machine (SVM): SVM was originally developed for classification problems. The idea behind SVM is to seek for a hyperplane that differentiates the class as accurate as possible. Consider a binary classification problem. If the two classes are separatable by the predictors, there exists many possible hyperplanes to separate the two classes. SVM seeks to identify the hyperplane that has maximum margin for separating the two classes as the classifier. If the two classes are not separatable by the predictors, there are points that will be wrongly classified. A slack variable is introduced and SVM seeks to identify the hyperplane that has the maximum margin as the

classifier. The hyperplane may be linear or non-linear. For problems involving non-linear hyperplane, a kernel function is introduced to capture the non-linear characteristics. For details, one may refer to [22,23].

3.3: The issue of imbalanced data classification for Asthma data

The modeling techniques summarized in section 3.2 were applied to classify the asthma data. Table 3 summarizes the performance of the fitted models using Sensitivity, Specificity, Accuracy and ROC value.

Table 3: Performance of machine learning models on original dataset

Technique	Sensitivity	Specificity	Accuracy	Area under ROC curve
Logistic Regression	0.040	0.996	0.889	0.518
Partial Least Squares	0.000	1.000	0.888	0.500
Random Forest	0.000	1.000	0.888	0.500
Gradient Boosting	0.046	0.996	0.890	0.521
LASSO	0.029	0.997	0.889	0.513
Elastic Net	0.040	0.996	0.889	0.518
KNN	0.014	0.999	0.889	0.507
SVM (Linear Kernel)	0.000	1.000	0.888	0.500

Sensitivity measures the True Positive (TP) rate within the event class (that is, ASTHMA = "YES"); while specificity measures the True Negative (TN) within the non-event class (that is, ASTHMA = "NO"). Similarly, accuracy is the ratio of correct predictions (that includes true positive and true negative) to total predictions made. Area under ROC curve measures overall performance of a specific classifier and has value in the range 0.5 to 1.0. A value of 0.5 represents performance of a random classifier while value of 1.0 represents a perfect classifier. As shown in Table 3, sensitivity measures are very small, specificity measures are very high, and the accuracy measures are also quite high for all modeling techniques applied. The results indicate all modeling techniques are biased towards the majority class which in our case means that the models were predicting cases of ASTHMA = "NO" very well but had miserable performance predicting ASTHMA = "YES". The accuracy is strongly biased towards the majority class due to large proportion of majority class (ASTHMA = "NO"), which does not effectively provide useful statistic for classification of minority class (ASTHMA = "YES"). However, the ROC values are barely over 0.5, which suggests these modeling techniques are not adequate for modeling the asthma data.

Potential problems that may cause the weak performances of these eight modeling techniques may include:

(a) The predictor variables surveyed in the national survey may not be appropriate for predicting asthma. Some additional predictor variables may need to be surveyed.
(b) Additional data cleaning and manipulations may provide better prediction of asthma.
(c) Other modeling techniques may be more appropriate.
(d) The asthma data only contains 10.8% of ASTHMA = "YES". This is an imbalanced data. The typical modeling techniques require the assumption that the data is approximately balanced. One cannot apply these modern machine learning techniques to classify highly imbalanced data.

Regarding (a), since this is a National Survey, which was created by many experts and has been conducted over many years, and the data have been analyzed in many literatures. This is beyond the discussion of this article. We revisited the data and review the variables used in similar studies in the literatures and concluded the predictor variables used are solid and appropriate for our study. However, there still might be several variables that can contribute to asthma prediction such as family history of asthma, which was not collected in the survey. Regarding (b), possible additional data manipulation and variable transformations may be performed by using different numerical transformations. We reviewed various related literatures and chose similar approach by keeping the data in their original scales, so that the resulting selected predictors can be meaningfully interpreted. Regarding (c), the eight modeling techniques applied are the common modern machine learning techniques, which have been shown evidence of their successes in many applications. We did not include the modern deep learning techniques for the reason that the observational survey data has weak signal to noise ratio, which is the major weakness of deep learning techniques. In addition, we would like to be able to observe predictors that show high association relationship with asthma. Deep learning cannot provide such information. We, therefore, focus on how to deal with the imbalanced data classifications.

Several techniques have been developed for dealing with the imbalanced data classifications. Among them, the approach of generating synthetic data to increase the minority class during the pre-modeling stage have been shown more successful, in particular, the SMOTE and the ROSE techniques. These techniques also are known as over-sampling techniques for generating synthetic minority cases. SMOTE generates synthetic data for the minority data using K-Nearest Neighbor technique so that the sizes of minority and majority classes are balanced. A detailed discussion of SMOTE can be found in [24]. ROSE generates synthetic data based on a smoothed bootstrap approach. A detailed description of the ROSE method can be found in [17]. The comparison study of different synthetic data generating techniques for imbalanced data classification by Liang and Martell [16] indicated ROSE performed the best in different applications. Therefore, ROSE is chosen for generating synthetic data in this study during the pre-modeling data processing.

As shown in Table 3, the accuracy appears to be quite high, while the fitted model is very poor, if not totally useless. This is an indication that accuracy measure often is misleading for imbalanced data classifications, and the ROC curve appears to be more robust. Other statistics such as F Measure, G Measure, Matthew's Correlation Coefficient, and Cohen's Kappa have been shown to be more useful measures of model performances for imbalanced data classifications [25].

Results

Before building machine learning models, it is essential to perform a preliminary screen of variable selection determine what predictors should be considered. Accordingly, the statistical significance of the association between ASTHMA and various categorical predictors were studied using Chi-square test. The results are presented in Table 4. Predictors with p-value < 0.05 were selected.

Table 4: Association between asthma and predictor variables in the study

	ASTHMA			
Predictors	**No**	**Yes**	**Chi-Square**	**P-value**
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)	28.819	<0.001
Black, Non-Hispanic	743(85.5)	126(14.5)		
Asian, Non-Hispanic	145(95.6)	7(4.4)		

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	1.000
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)	136.84	<0.0001
Out of work	391(87.8)	57(12.2)		
Homemaker	450(88.9)	56(11.1)		
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)	22.464	<0.0001
Former smoker	2770(89.8)	315(10.2)		
Never smoked	5137(89.8)	584(10.2)		
USENOW3				
Yes	267(89.3)	32(10.7)	<0.0001	1.000
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)	33.769	<0.0001
No	7843(90.0)	876(10.0)		
Pre-diabetes or borderline diabetes	185(84.5)	34(15.5)		
BMI5CAT				
Underweight	135(84.9)	24(15.1)	79.77	<0.0001
Normal Weight	2674(91.8)	238(8.2)		
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 4 shows, asthma is more prevalent among females than males (13.2% vs 7.9%). This association between sex and asthma is found statistically significant. The age-group of 45-54 years had the highest prevalence while 65 years or older had the lowest prevalence. Education is also found significantly associated with asthma as those below high school was the most prevalent group. Regarding income, participants in the lowest income group (below $25000) were the ones who reported asthma more than other groups. Black, Non-Hispanic communities reported higher asthma rates than White, Non-Hispanic, Asian, and Hispanic group. However, the highest rates were reported by 'other' races besides these. No significant association was found between urban-rural residence status and asthma. Regarding veteran status, those who did not report of being a veteran had higher asthma rates (11.4% vs 7.4%). Higher asthma rates were also observed among those who were unable to work, current smokers, and smoked at least 100 cigarettes in their life. Diabetes status was also found to be significantly associated with asthma as more people having diabetes or borderline diabetes had asthma. Underweight and obese categories had higher asthma rates than reported by people in normal weight categories. There seems to exist a strong association between chronic obstructive pulmonary disease and asthma since nearly one-third of the people with the disease had asthma while only 8.6% of the people without disease had asthma. Also, higher asthma rates were reported by people who could not see a doctor due to high medical cost. The study also showed that people who received a flu shot/spray in the past 12 months had slightly higher asthma rates than those who did not receive the shot.

To summarize, asthma was found 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 current smoking and smoked at least 100 cigarettes throughout the life 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. Predictors that were found to be statistically significant were used to build predictive models using training data set. Predictors that did not have significant associations were dropped.

After imputing some missing data, ROSE technique was applied to the original training data (N= 7287) to create the ROSE training data, which is approximately balanced new training data for modeling. The ROSE algorithm applies over-sampling techniques to generate synthetic data for minority class (ASTHMA = "YES") and under-sampling techniques to select majority class (ASHMA = "NO") so that both classes is approximately 50%. Table 5 summarizes the distributions of original training data and ROSE training data of two ASTHMA levels.

Table 5: Distribution of ASTHMA response variable in original and after-ROSE training data

	Total N	ASTHMA = "YES"	ASTHMA = "NO"	Proportion of "YES"
Original training data	7287	777	6510	10.7%
After ROSE training data	7287	3614	3673	49.6%

The ROSE training data was used to build the models using 10-fold cross-validation and the obtained optimal model based on each machine learning techniques are applied to the original test data (N = 3124, N (ASTHMA = "YES") = 338 and N (ASTHMA = "NO") = 2768) for comparison. The results are summarized in Table 6.

Table 6: 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

Among the models developed, four models had similar performances in terms of sensitivity and area under the ROC curve. These were logistic regression, partial least squares, LASSO, and elastic net. Other performance metrics were also somewhat close to each other for these models.

Models namely random forest, gradient boosting, KNN and SVM with a linear kernel had lower sensitivity values. When comparing the ROC values between the results in Table 6 and the result in Table 3, it is noticed that the performance of the models using the synthetic data is better than those using the original data. The ROC values are above 0.65 using synthetic data for most modeling techniques, while the ROC values are near 0.5 based on the original data.

The order of variable importance for each modeling technique is shown in Table 7. The results suggest that the predictor *Chronic Obstructive Pulmonary Disease* (CHCCOPD2) appears to be the most important chosen in each model. The next is the predictor, *Sex*, followed by two variables having similar average level of importance are *AGE* and *Flu shot/Spray in the past 12 months (FLUSHOT7).* The next in the group are *Employment* and *Income.* However, the degrees of importance of predictors are not consistent, except CHCCOPD2 and SEX. Both smoking related predictors are selected by some models, but not all models. In general, our study identified similar risk factors shown in the literatures. In addition, our study identified that *Chronic Obstructive Pulmonary Disease* appears in every model. This is not surprising. Although these two diseases are different, they have various similar symptoms, thus are highly associated with each other. Somewhat surprising is that none of the Smoking related predictors was selected by all models. These may be due the fact that these predictors are highly correlated. Thus, the effects of smoking related variables were explained by other predictors.

Table 7: Variable Importance Order

		Models								
Factors Selected	Var. Type	LR	PLS	RF	GB	LASSO	EN	KNN	SVM	Average rate
CHCCOPD2	HC	1*	1	1	1	1	1	1	1	1
SEX	DS	2	3	6	2	3	4	2	2	3.00
FLUSHOT7	HC	3	7	2	5	8	9	6	6	5.75
AGE_G	DS	4	8	9	8	4	3	5	5	5.75
DIABETE4	HC	5	6	5	7	7	6	10	10	7.00
EMPLOY1	DS	6	5	12	4	5	5	8	8	6.63
IMPRACE	DS	7		10	12	2	2			
INCOMG	DS	8	4	3	10	10	8	4	4	6.38
VETERAN3	DS	9	11		6	9	11	12		
SMOKE100	PH	10					7		12	
MEDCOST	DS	11	9	8		11		7	7	
SMOKER3	PH	12	10	4	11		10	9	9	

**denotes the level of importance with 1 meaning the most important variable.*

Among the eight models we analyzed, logistic regression, partial least squares, LASSO and elastic net performed equally well. We choose logistic regression model as the best model and apply it to the entire original data for further analysis of each predictor for the reason of meaningful interpretation in terms of odds ratio of each level for each predictor. Table 8 presents the model information using the selected logistic regression model on the entire dataset.

Table 8: Summary information from logistic regression model

Variables	**Parameter Estimate**	**Standard Error**	**Odds ratio**	**P-value**
CHCCOPD2				
Yes (Ref.)				
No	-1.605	0.090	0.201	<0.001
SEX				
Female	0.522	0.076	1.685	<0.001
Male (Ref.)				
FLUSHOT7				
Yes (Ref.)				
No	-0.253	0.069	0.776	<0.001
AGE_G				
18-24 (Ref.)				
25-34	-0.227	0.166	0.797	0.172
35-44	-0.325	0.169	0.723	0.055
45-54	-0.383	0.163	0.682	0.019
55-64	-0.733	0.165	0.480	< 0.001
65 and above	-0.886	0.177	0.412	< 0.001
DIABETE4				
Yes (Ref.)				
No	-0.163	0.093	0.850	0.080
Pre-diabetes or borderline diabetes	0.138	0.214	1.148	0.519
EMPLOY1				
Employed for wages (Ref.)				
Out of work	0.052	0.160	1.053	0.745
Homemaker	-0.018	0.160	0.982	0.908
Student	0.003	0.210	1.003	0.989
Retired	-0.018	0.115	0.982	0.876
Unable to work	0.432	0.122	1.540	<0.001
IMPRACE				
White, Non-Hispanic (Ref.)				
Black, Non-Hispanic	0.169	0.111	1.184	0.126

Asian, Non-Hispanic	-0.893	0.396	0.409	0.024
Hispanic	-0.223	0.213	0.800	0.293
Other	0.297	0.151	1.346	0.050
INCOMG				
Below $25000 (Ref.)				
$25000 - $50000	-0.162	0.102	0.850	0.111
$50000 or more	-0.304	0.099	0.738	0.002
Don't know/Not Sure	-0.409	0.107	0.664	<0.001
VETERAN3				
Yes (Ref.)				
No	0.167	0.132	1.182	0.207
SMOKE100				
Yes (Ref.)				
No	0.303	0.393	1.354	0.442
MEDCOST				
Yes (Ref.)				
No	-0.186	0.101	0.830	0.067
SMOKER3				
Current smoker (Ref.)				
Former smoker	-0.009	0.104	0.991	0.928
Never smoked	-0.143	0.392	0.867	0.716

Based on the odds ratios, the effects of some selected inputs can be interpreted as follows:

- People without chronic obstructive pulmonary disease are 80% less likely to report asthma compared to people with it.
- Females are nearly 69% more likely to develop asthma than males. Using 2000 BRFSS data, Gwynn [3] found females to be 91% more likely. Similarly, Greenblatt et.al. [12] found females were 80% more likely than men to develop asthma from 2007-2012 BRFSS datasets.
- Adults who have not taken flu shot were about 22% less likely to report current asthma. This association could be the other way too: adults who had asthma could be more likely to show up for the flu shot. Since the study is observational, we cannot infer causal effects.
- People in all other age-groups are less likely to get asthma than people in 18 – 24 years age-group. Again, Gwynn [3] found that adults aged 35-64 and ≥65 years were less likely to report current asthma than adults aged 18-34 years.
- Regarding race, both Asian and Hispanic adults were less likely to report current asthma than White, Non-Hispanic. The American Lung Association [26] 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. [12] also found the lowest income category to be two-third more likely to report asthma.

Summary and Conclusion

In this study, our objective was to analyze and build predictive models to predict asthma for the population in the State of Michigan using data from the 2019 Behavioral Risk Factor Surveillance System (BRFSS). Eight modern machine learning techniques are applied and compared. The straightforward application of the modeling techniques based on the original data indicated the obtained models are practically useless. As illustrated, machine learning models perform poorly when the distribution of response variable is highly uneven among its classes. The performances of such models are highly biased towards the majority class. A thorough analysis was conducted using five-steps strategies as described in Section 3. The critical problem of asthma data classification is the distribution of two levels of asthma are highly skewed with only 10.8% of minority class (ASTHMA = "YES"). The modern machine learning techniques require the assumption of approximately balanced majority and minority classes. By applying ROSE techniques to generate synthetic data, the newly formed training data is more balanced in the two levels of ASTHMA. The obtained eight optimal models were applied to the original data as a comparison with the models solely based on the original data. Improved performances of the models were observed: accuracies of all models (except KNN) were similar at around 70%. Logistic regression (LR), partial least squares, LASSO, and elastic net (EN) had similar level of performances. Logistic regression model was selected as the best model based on sensitivity, area under ROC curve and ease of interpretability. Presence of chronic obstructive pulmonary disease, female sex, taken flu shot/spray in the past twelve months, 18-24 age group, inability to work, being white, non-hispanic, and lower income level are identified as important predictors. The best model can be used to make predictions of asthma development and implement early intervention for the treatment in higher risk groups.

Machine learning models based on synthetic data may overfit the minority class. One should always be careful about evaluating the model performance in the context of imbalanced learning. It should also be noted that the issues of class overlapping in which data samples appear as valid instances of more than one class and data fracture in which there is a change in data distribution between train/test splits, often in the minority class, may affect bias and variability of the accuracy estimator [27]. Another popular method for generating synthetic data is the synthetic minority oversampling technique (SMOTE). Although not one method will always dominate the other, studies such as [16,27] have demonstrated that ROSE produces consistently better performance than SMOTE. For future work, additional models for asthma prediction may be developed such as those involving neural nets. Also, various other techniques for handling imbalanced data may be used for performance comparison.

References

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

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

3. Gwynn RC. Risk factors for asthma in US adults: results from the 2000 Behavioral Risk Factor Surveillance System. Journal of Asthma. 2004 Jan 1;41(1):91-8.

4. Zahran HS, Bailey C. Factors associated with asthma prevalence among racial and ethnic groups—United States, 2009–2010 behavioral risk factor surveillance system. Journal of asthma. 2013 Aug 1;50(6):583-9.

5. 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 Medical Journal. 2017 May;135:234-46.

6. Ahmed MR, Ali MA, Roy J, Ahmed S, Ahmed N. Breast Cancer Risk Prediction based on Six Machine Learning Algorithms. In2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) 2020 Dec 16 (pp. 1-5). IEEE.

7. KR, Gautam Y. Argumentative comparative analysis of machine learning on coronary artery disease. Open Journal of Statistics. 2020 Jul 10;10(4):694-705.

8. Harvey JL, Kumar SA. Machine learning for predicting development of asthma in children. In2019 IEEE Symposium Series on Computational Intelligence (SSCI) 2019 Dec 6 (pp. 596-603). IEEE.

9. Zein JG, Wu CP, Attaway AH, Zhang P, Nazha A. Novel machine learning can predict acute asthma exacerbation. Chest. 2021 May 1;159(5):1747-57.
10. Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. Annals of the New York Academy of Sciences. 2017 Jan;1387(1):153-65.

11. Hsu J, Chen J, Mirabelli MC. Asthma morbidity, comorbidities, and modifiable factors among older adults. The Journal of Allergy and Clinical Immunology: In Practice. 2018 Jan 1;6(1):236-43.

12. Greenblatt R, Mansour O, Zhao E, Ross M, Himes BE. Gender-specific determinants of asthma among US adults. Asthma research and practice. 2017 Dec;3(1):1-1.

13. Rivera AC, Powell TM, Boyko EJ, Lee RU, Faix DJ, Luxton DD, Rull RP, Millennium Cohort Study Team. New-onset asthma and combat deployment: findings from the Millennium Cohort Study. American Journal of Epidemiology. 2018 Oct 1;187(10):2136-44.
14. Ehrlich SF, Quesenberry Jr CP, 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 Jan 1;33(1):55-60.

15. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one. 2017 Jul 24;12(7):e0179805.

16. Liang Z, Chapa-Martell MA. Combining resampling and machine learning to improve sleep-wake detection of Fitbit wristbands. In2019 IEEE International Conference on Healthcare Informatics (ICHI) 2019 Jun 10 (pp. 1-3). IEEE.

17. Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. R journal. 2014 Jun 1;6(1).)

18. R Core Team (2020). R: A language and environment for statistical computing. R Foundat ion for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

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

20. Van Buuren S, Groothuis-Oudshoorn K, Robitzsch A. Package 'mice': multivariate imputation by chained equations. CRAN Repos. 2019.

21. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008 Nov 10;28:1-26.

22. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013 Jun.

23. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: springer; 2009 Aug.

24. Chawla B. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research. 2002;16:321-57.

25. Akosa J. Predictive accuracy: A misleading performance measure for highly imbalanced data. InProceedings of the SAS global forum 2017 Apr 2 (Vol. 12, pp. 1-4).

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

27. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery. 2014 Jan;28(1):92-122.