Predicting Asthma Risk Using Social, Behavioral, and Environmental Factors

Elliot Nam[1], Indra Neil Sarkar, PhD, MLIS[2], Paul C. Stey, PhD[2], Elizabeth S. Chen, PhD[2]
[1]Seoul International School, Seoul, Korea
[2]Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

The aim of this study was to build an asthma-predicting model based on the Behavioral Risk Factor Surveillance System (BRFSS)[1] to identify features (ranging from social to environmental factors) that may affect vulnerability for asthma and develop predictive models using machine learning algorithms implemented in the Julia implementation of scikit-learn[2,3]. Five-fold cross validation was used to compare the accuracy of the tested models.

Introduction

Known for its universality, asthma today is the most ubiquitous chronic condition that affects the entire demographic. Its increasing prevalence has led to numerous studies for identifying and determining risk factors encompassing environmental and genetic factors. Yet, the causes of asthma remain obscure, as the correlation with social, behavioral, and environmental factors has not been fully studied[4].

Methods

This study consisted of four major processes: (1) Pre-Processing of BRFSS Data, (2) Running Machine Learning Algorithms/Building Predictive Models, (3) Optimizing Algorithms, and (4) Validating Models. From the 464,664 respondents in 2014 BRFSS, a sample of 59,749 respondents was extracted that included "Yes" or "No" answers to whether one was diagnosed with asthma (responses such as "Don't know/Not Sure" or "Refused to answer" were excluded). The extracted dataset was then processed to eliminate and replace all blank spaces with "NA" values in the columns for 14 features (representing gender, race, residing state, physical activity, sleep, education, employment, income level, tobacco use, and alcohol use). Missing values were imputed using mean imputation for continuous variables; categorical variables were imputed probabilistically according to the variables' original distributions[5]. Next, machine learning algorithms implemented in the Julia *ScikitLearn.jl* package were used to analyze the dataset. Each column in the dataset was analyzed by each algorithm as separate features, and in all, a total of nine different algorithms were used. Important features of each algorithm were extracted to eliminate those that do not contribute to (or may hinder) performance. Algorithm optimization was done through Random Search Parameter Tuning and Grid Search Parameter Tuning, where either the algorithm or the user generates arrays of possible parameter values or weights for each algorithm in order to compare and contrast the performance of each combination of parameter values. Finally, the optimized models were validated using five-fold cross validation in *ScikitLearn.jl* and Receiver Operating Characteristic (ROC) curves that represent the specificity and sensitivity of each model were generated.

Results

Figure 1 depicts the comparison of the performance for the nine algorithms. The top performing algorithms were found to be Logistic Regression, Linear Discriminant Analysis, Naïve Bayes, Gradient Boosting Classifier (GBC), and Ada Boost Classifier (ABC), which performed up to 71.9% in accuracy. Figure 2 shows the ROC curves and important features for ABC and GBC. The most important features of the ABC model were: state, employment status, income level, sleep duration, race, and interval since last smoked. The most important features of the GBC model, on the other hand, were: employment status, exercise in past 30 days, income level, race, state, and sleep duration. While the algorithms had different sets of most important features, state, sleep duration, income level, and employment status were common features in both models. The code is publicly available in GitHub[6].

Discussion

The association between various environmental and societal factors and asthma has been identified through this research. Yet, further research into the relationship between the important features and asthma is needed to further ascertain the correlation. Next steps include studying how different geographic locations influence the risk of asthma by identifying clusters of similar geographic features present in states with high asthma occurrence. Future efforts also involve further studying the relationship between particular features and asthma risk (e.g., income level that was found to be one of the most frequent indicators).

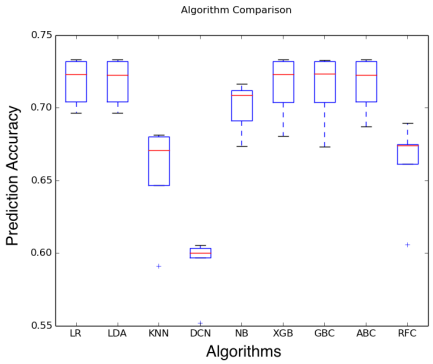
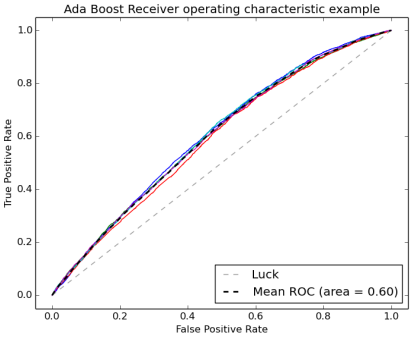
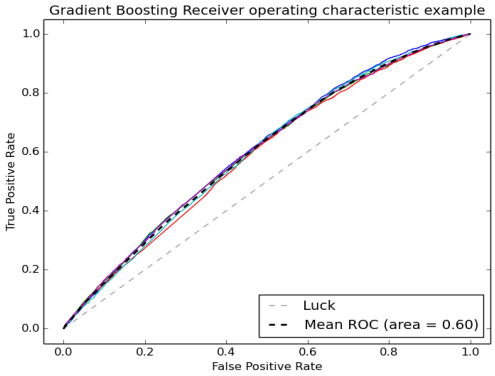


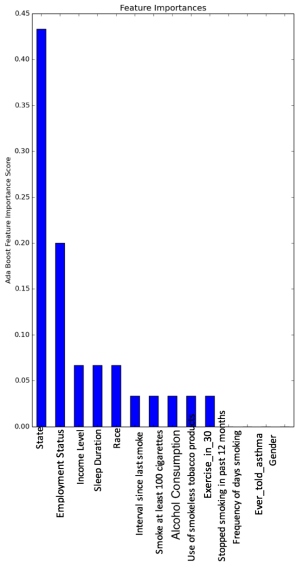
Figure 1. Comparison of Algorithms. (LR=Logistic Regression, LDA=Linear Discriminant Analysis, KNN=k Nearest Neighbor, DCN=Decision Tree (Classification and Regression Trees [CART]), NB=Naïve Bayes, XGB=X Gradient Boosting classifier, GBC=Gradient Boosting Classifier, ABC=Ada Boost Classifier, RFC=Random Forest Classifier)



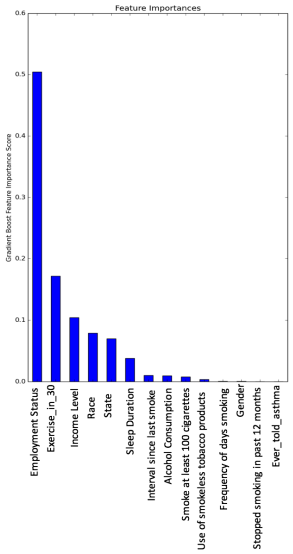
(A)



(C)



(B)



(D)

Figure 2. Comparison of ROC Curves and Important Features for ABC (A and B) and GBC (C and D).

Acknowledgment: This work was supported in part by National Library of Medicine grant R01LM011364.

References

1. http://www.cdc.gov/brfss/
2. http://scikit-learn.org
3. https://github.com/cstjean/ScikitLearn.jl
4. Eder W, Ege MJ, von Mutius E. The asthma epidemic. N Engl J Med. 2006 Nov 23;355(21):2226-35.
5. https://github.com/JuliaStats/Distributions.jl
6. https://github.com/elliotnam/PredictAsthma