

**Sub. Code : 21BIO111**

**Sub Name : Introduction to Biological Systems**

**Name of the Project : Asthma Prediction and comparison using various algorithms**

**Team members**

**1.Sabarinath (21141)**

**2.Shyam Ganesh (21149)**

**3.Sidesh Sundar (21150)**

**4.Sasank Sami(21160)**

**5.Bharadwaj (21165)**

**ACKNOWLEDGEMENT**

We would like to express our gratitude to our professor Dr S.Dhivya mam who gave us this goldenopportunity to do the project on the topic **“Asthma Prediction and comparison using various algorithms”.**

We would like to extend our gratitude to the AIE department of Amrita Vishwa Vidhyapeetham, Chennai who assigned this project to us and helped us to improve our knowledge.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Page No** |
| 1 | Abstract | 3 |
| 2 | Introduction | 4 |
| 3 | Literature Review | 5 |
| 4 | Algorithms Used | 6 |
| 5 | Methodology | 14 |
| 6 | Output | 15 |
| 7 | Conclusion | 15 |
| 8 | Future works | 16 |
| 9 | Reference | 16 |

**ABSTRACT:**

Person who have asthma need to be treated as soon as possible, which requires early diagnosis. It has proven challenging to develop a model that reliably predicts asthma. As of now, research has produced asthma prediction models with low accuracy and tiny, focused sample sizes. There is a dearth of research that examines a sizable population of Persons using a set of criteria in order to create a model that can be applied in a clinical environment. In this study, we created predictive models to examine a dataset on the health of Person with asthma. These prediction models are created using machine learning classifiers, such as Linear Regression, Decision Tree, Random Forest, KNN, and Naive Bayes approach. The Random Forest Classifier produced the highest prediction accuracy (90.9%) of all the deployed classifiers.The following are the factors: difficulty and sex The three factors most closely associated with asthma are breathing, allergies, and medication. Medical professionals can utilise the analysis of recent research and the model's findings to make predictions about the onset of asthma in Person and to carry out early intervention for the treatment of asthma development in a clinical environment.

**Keywords—**allergies; asthma; bronchiolitis; Person; health;

machine learning; National Survey of Person’s Health;

predictive models.

**INTRODUCTION:**

**Asthma:**

The prevalent chronic lung condition known as asthma is brought on by airway inflammation. Recurrent episodes are brought on by the respiratory tract's narrowed airways. Breathlessness, chest tightness, coughing, and wheezing are typical symptoms that can be fatal. Asthmatic reactions can be brought on by smoke and other air pollutants, allergens, respiratory diseases, environmental factors, physical activity, and gastric reflux. 8,3 percent of American Person had asthma as of 2016, according to the Centers for Disease Control and Prevention. On the basis of data gathered in the clinical context, a precise model that can be applied to determine whether a Person would acquire asthma is required. The dataset used in this article.

**Machine learning:**

Many different fields have used machine learning and data science. Data mining has also been used extensively in studies on person's heath. The research has mostly concentrated on old data and tiny sample numbers. The models created by other researchers have not demonstrated good accuracy in asthma prediction. The purpose of this study is to construct an efficient model for finding variables that can be used to forecast the onset of asthma in an individual. For medical practitioners, the models given in this research are simple to utilise. Data mining in a person's health records takes time and varies from doctor to doctor as well as from practise to practise.Medical personnel want to be able to recognise when a patient needs a particular type of care and then deliver that care in a timely manner. Medical professionals can concentrate on early identification of risk factors for the development of asthma in Persons and early intervention with potential preventative care by identifying the characteristics that are most influential to the development of asthma in Persons and using that data to determine the appropriate medical care.

**LITERATURE REVIEW:**

The majority of machine learning studies in healthcare are based on low-accuracy logistic regression models and evaluate limited sample sizes. The majority of the study is out of date, and machine learning was not used in the methodologies used to forecast the onset of asthma. Using a variety of machine learning predictive models to accurately predict asthma using massive datasets is the subject of a small number of research. Recent research looks at how a small number of categorization algorithms can be used to foretell the onset of asthma. 248 people's asthma was assessed by Andersson et al.In Sweden, 3430 parents of children aged 7 to 8 were asked to complete a questionnaire [8]. 248 individuals with asthma were monitored until they were 19 and 205 of them were still taking part in the trial. The data gathered from annual surveys was examined using logistic regression. The analysis revealed a relationship between 10 factors and asthma in these people. The findings indicate that female gender, sensitivity to allergens in furred animals, and high BMI are associated with asthma. According to earlier research, children whose moms smoked while they were pregnant had a higher risk of developing asthma . According to this study, people in that cohort were more likely to have nonallergic asthma that went into remission during the course of the investigation. Ram, Zhang, Williams, and Pengetnze looked at combining different data sources to estimate the proportion of visits to emergency rooms for symptoms associated with asthma in 2015. The model's forecasts of ER visits utilising near-real-time environmental data, internet searches, and social media data showed a 70% accuracy rate. Using a dataset of 4500 tweets, artificial neural networks (ANN) were utilised to find tweets from Twitter that were pertinent to the study. To ascertain the trends of asthma-related phrases, Google Trends was used. The environmental data was collected using EPA databases.The first tool utilised was the Pearson correlation coefficient, and then four different classification models—decision tree, Naive Bayes, SVM, and ANN—were applied. Then, the accuracy of these models was contrasted. Asthma tweets and three pollutant indexes were elements incorporated into the prediction algorithms (CO, NO2, and PM2.5). The authors recommended that future research concentrate on fusing traditional data with real-time data.

**ALGORITHM USED:**

**Randomforest:**

A classification system made up of several decision trees is called the random forest. It attempts to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree by using bagging and feature randomness when generating each individual tree.



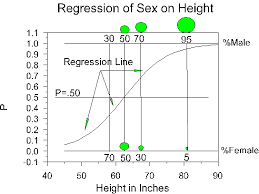
a decision tree might be beneficial to begin by briefly outlining the decision tree algorithm because the random forest model is made up of numerous decision trees. Should I surf? is a common starter question for decision trees. A sequence of queries, such as "Is it a long period swell?" and "Is the wind blowing offshore?," can then be used to arrive at an answer. These inquiries serve as the decision nodes in the tree, which divide the data. Each query aids a person in coming to a conclusion, which is indicated by the leaf node. The "Yes" branch will be followed by observations that meet the requirements, while the opposite path will be taken by observations that don't.Decision trees seek to find the best split to subset the data, and they are typically trained through the Classification and Regression Tree (CART) algorithm. Metrics, such as Gini impurity, information gain, or mean square error (MSE), can be used to evaluate the quality of the split.

**Logistic Regression:**

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. The following formulas are used to represent this logistic function, which is sometimes referred to as the log odds or the natural logarithm of odds:

Logit(pi) = 1/(1+ exp(-pi)) ln(pi/(1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k

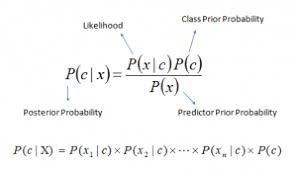
Logit(pi) is the dependent or response variable in this logistic regression equation while x is the independent variable. either the beta parameter



**NaiveBayes:**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in text classification that includes a high-dimensional training dataset.Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.



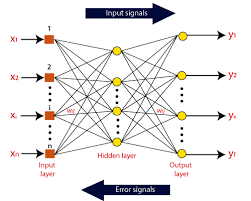
Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

**Naïve:** It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

**Bayes:** It is called Bayes because it depends on the principle of [Bayes' Theorem](https://www.javatpoint.com/bayes-theorem-in-artifical-intelligence" \t "https://www.javatpoint.com/_blank).

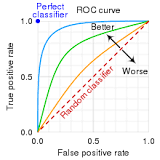
**ANN:**The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes.



Artificial neural network tutorial covers all the aspects related to the artificial neural network. In this tutorial, we will discuss ANNs, Adaptive resonance theory, Kohonen self-organizing map, Building blocks, unsupervised learning, Genetic algorithm, etc.

**ROC-curve:**

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.



This curve plots two parameters:

True Positive Rate

False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

TPR=TPTP+FN

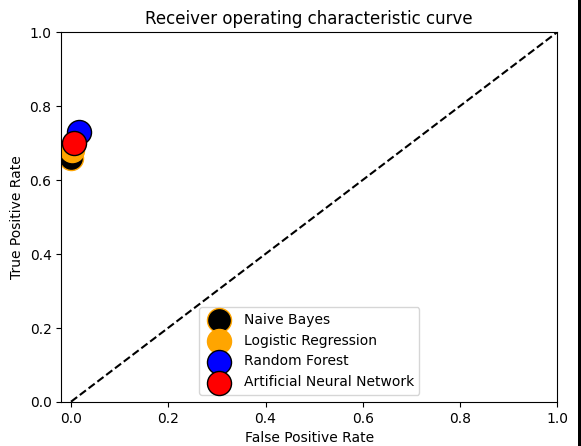
False Positive Rate (FPR) is defined as follows:

FPR=FPFP+TN

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

**METHODOLOGY:**First we extract the required data set from National Survey Of Human Health which is data based on symptoms and reports on Asthma patients, then we classify and extract the features which are common and are medically much related to the specified disease via manual knowledge and survey based analytical data. These data is then analysed and the model is buiilt based on the parameters and tested by spllitting up the data in the ratio of 7:3 as train and test data to analyse its accuracy (corrections are made in general to increase computational speed and are not related to data).The trained model is then subjected to accuracy test and is then cross validated with the other models with similar implementation where the data is mapped with respective parameter resulting in like structurally organised data.The cross validation is then displayed via graphical representation and the output of each graph is differntiated in a graph to compare the accuracy with respect to dataset in each model

**OUTPUT:**



**CONCLUSION:**

When independently validated, existing prediction models showed only a moderate degree of generalizability and mediocre predictive accuracy. The shortcomings of conventional approaches have been found to reduce prediction accuracy and resolution. By looking into new methodology, modifying or cross-validating current procedures, or by applying cutting-edge techniques like machine learning techniques and prediction and optimization algorithms, these constraints for future asthma prediction may be overcome.

**FUTURE SCOPE:**

The categorization algorithms employed in this study can be applied to other datasets or to the same data set to predict additional health metrics. medical care

The models can be used by providers to select features that they want to examine for correlations and see whether they can predict a child's development of asthma. Then, preventive care, diagnosis, and treatment for kids can be implemented using these prediction models. When asthma is identified, prompt diagnosis and early intervention are crucial for enhancing patients' quality of life.

We will add few more algorithms to get precise output and we can create an app which will be more effect as it always keeps a track of our regular activities.

**CODE:**

Github link:<https://github.com/Sideshsundar06/IBS_project>

**REFERENCE:**

[1] https://www.cdc.gov/asthma/asthmadata.htm

[2] Messinger, A., Nam, B., Vu, T., & Deterding, R. (2017). “Predicting

Asthma Severity Using Machine Learning Algorithms: A Pilot Study. In

A66.” THE MANY FACES OF ASTHMA IN CHILDHOOD (pp. A2231-

A2231). American Thoracic Society.

[3] Luo, G., Nkoy, F. L., Stone, B. L., Schmick, D., & Johnson, M. D.

(2015). “A systematic review of predictive models for asthma development

in children.” BMC medical informatics and decision making, 15(1), 99.

[4] Patel, S. J., Chamberlain, D., & Chamberlain, J. M. (2018). “A

Machine-Learning Approach to Predicting Need for Hospitalization for

Pediatric Asthma Exacerbation at the Time of Emergency Department

Triage.”

[5] Nam, E., Sarkar, I.N., Stey, P.C., Chen, E. S. (2016). “Predicting

Asthma Risk Using Social, Behavioral, and Environmental

Factors.Retrieved from https://github.com/elliotnam/PredictAsthma

[6] Andersson, M., Hedman, L., Bjerg, A., Forsberg, B., Lundbäck, B., &

Rönmark, E. (2013). “Remission and persistence of asthma followed from

7 to 19 years of age.” Pediatrics, peds-2013.

[7] Bjerg, A., Hedman, L., Perzanowski, M., Lundback, B., Ronmark, E.

(2011). “A strong synergism of low birth weight and prenatal smoking on

asthma in schoolchildren.” Pediatrics. 127(4).

[8] Princy, M. J. C., & Sivaranjani, M. K. (2016). “Survey on Asthma

Prediction Using Classification Technique.”

[9] Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. (2015).

“Predicting asthma-related emergency department visits using big data.”

IEEE J. Biomedical and Health Informatics, 19(4), 1216-1223.

[10] Sears, M. R. (2015). “Predicting asthma outcomes.” Journal of Allergy

and Clinical Immunology, 136(4), 829-836.

[11] Belsky, D. W., & Sears, M. R. (2014). “The potential to predict the

course of childhood asthma.” Expert review of respiratory medicine, 8(2),

137-141.

1. Bannier, M. A. G. E., van de Kant, K. G., Jöbsis, Q., & Dompeling, E.

[13]https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiRw4ry7Yj8AhWF4DgGHTpTBHIQFnoECA4QAw&url=https%3A%2F%2Fdevelopers.google.com%2Fmachine-learning%2Fcrash-course%2Fclassification%2Froc-and-auc%23%3A~%3Atext%3DAn%2520ROC%2520curve%2520(receiver%2520operating%2CTrue%2520Positive%2520Rate&usg=AOvVaw3Nkh2piez\_MjbOF5mdRyY9