National College of Ireland

Analyzing the relationship between asthma and air impurity in developed cities using machine learning approach

MSc Research Project
Data Analytics

Jyoti Chavda
Student ID: x18114831

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Jyoti Chavda
Student ID:	x18114831
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	12/07/2018
Project Title:	Analyzing the relationship between asthmaand air impurity in developed cities using machine learning approach
Word Count:	XXX
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analyzing the relationship between asthma and air impurity in developed cities using machine learning approach

Jyoti Chavda
x18114831

Abstract

Now a day, many metropolitan regions experience environment changes due to air pollution issues. Air pollution and allergens are the principle variables which causes a significant effect on human health. Many illnesses are caused by air pollution, in which the main disease is asthma. This research examines the relationship of human health and air impurity in which the pollutant is hazardous to health. This is evaluated by using various machine learning techniques. To investigate this, Various regression models and ensemble models such as LASSO regression, Ridge regression, ElasticNet regression, Gamboost, decision tree and Random forest were used. this investigation is very useful in area where high air pollution levels are present and it can also help to reduce the asthma mortality by maintaining high air quality, which improves the health of country.

1 Introduction

With the constantly increasing development of infrastructures, social and economic norms, respiratory emergencies were visited at a high level. Among the most polluted natural resources is air, which is one of the most fundamental necessities to live. The alarming tendency that necessarily goes hand in hand with modern conveniences is an uncontrollable and harmful contributor to atmospheric pollution(Paul S.T. (2018)). Asthma can be caused due to various factors. The harmful health results of air contamination are well recognized, mortality estimates, and disease association indicate that substantial health burdens can be imposed at comparatively small concentrations as well[1]. Asthma is a lifelong illness that affects thousands in the United States every year.

As any visible or invisible air composition or gasses, including nitrogen oxides, carbon dioxide, particulate matter or ozone, air pollution is determined by the United States Environmental Protection Agency(EPA)[2]. Urbanization makes significant contributions with regard to asthma and is partially due to increasing air pollution in outdoor environments, as well asfast development in population and increasing outdoor air pollution in many modern cities in developing countries. While air pollution is nearly always a mixture of its different elements, air quality is regulated by most jurisdictions and this has

[1]https://www.who.int/mediacentre/news/releases/2014/air-pollution/en
[2] https://www.aafa.org/air-pollution-smog-asthma/

an indirect impact on the health economy of the country(Guarnieri and Balmes (2014)).

The assessment of the influence of mixtures on single pollutants has numerous explanations. A pollutant may reasonably be suspected of being dangerous if one delicate region of the body has already been weakened by other chemical substances. The breathing and circulatory system differently depend on single chemicals. The material in the atmosphere is closely linked to seasonal weather patterns and other climate and environmental conditions such as precipitation. The government can discover and initiate preventive measures to control air pollution by imposing economic losses related to health, Such results can also make environmental protection-sensitive for the local administration and the citizens(Brusseau et al. (2019)).

Recent research showed more than one instance of the most common asthma risk factors being social demographic factors. Ongoing studies have been conducted on the causes of asthma, comprehensive medical and clinical trials have examined the factor of asthma morbidity in different situations like smoke, allergens, temperature, and domestic dust. However medical researchers traditionally conduct asthma research, public health has been influenced by asthma morbidity (Kim and Ahn (2018)) because of its apparent association to social and demographic changes. Numerous studies have identified children as one of the most susceptible asthma group. In some studies, children can correlate asthma exposure with traffic. Oversight of air quality and its effect on human health globally has been done through distinct parameters(Ahn and Kim (2019)). Different modeling methods, such as spatial regression models, random forests, RF-based partition models and many more data patterns were developed during several decades in the data mines.Research on the link between environment and human asthma morbidity is limited.

Modern modeling requires an accurate method of cases, which can not only determine the individual circumstances but also determine the impact of pollution exposure on the geographic information system. A variety of techniques must be assessed on different situations since each situation conflicts so that all other methods which lead to the selection of a specific set of techniques according to information can be chosen. This study also fulfills the study gap, so it is anticipated that the output of the model will increase to predict accurately, by using different machine education.

This research aims to examine the air contamination which contributes to human asthma morbidity, taking into consideration all the above-mentioned questions, with the hypothesis that the constructed environment pollutes the air by releasing hazardous pollutants. This paper also examines how the constructed environment influences asthma in older and younger people in a variety of conditions. To that extent., in the next section, the paper is organized: section 1 contains associated work which provides a review of all-present health systems literature and developed asthma surveillance services. Section 2 Methodology dedicated to the proposed algorithms and section 3 is accompanied by an implementation of this research section 4 covers the evaluation and results and section 5 gives the overall research conclusion and future work.

2 Related Work

An increase of the population of urban settlements, elevated demographic density, industrial operations and enhanced traffic essential to air pollution. These operations use elevated concentrations of energy which release a significant amount of pollutants into the atmosphere including PM10, CO, O3, PM2.5., NO2 Air contamination is the biggest environmental hazard in the last several decades with an influence on fitness, economy, welfare, and lives (Piyatilake and Perera (2018)). Previous research has been carried out to simulate air quality levels and identify the uncertainty of air contamination in developed cities and the relationship between air contamination and asthma (Cox Jr. (2018)).

2.1 Indoor/Outdoor Air pollution

The population exposure measures are generally based on the external element and rarely accountable for the majority of individuals spending their time indoors. The housing industry represents a significant difference in air appearance because of external contamination and indoor ventilation discharge. An especially distributed home inventory model in England and Wales is defined as an inner air pollution model by Taylor et al. (2019). First of all, the PM2.5, NO2 was estimated for both indoor and outdoor atmospheric pollution and total indoor atmospheric levels. Indoor and outdoor sources showed the ability to determine the concentration value of pollutant by illustrating the effective adaptation of CO concentration. The indoor atmosphere for England and Wales as internal sources, including warming, cooking, and smoking. Indoor air pollution was found by the statistical method while allowing estimates on a measure for single residences of indoor air pollution. This study estimates average indoor exhibitions of No2 and Pm2.5 in outdoor references at 0.14 for NO2 and 0.60 for PM2.5.

The primary variables causing death and diseases are the fine particles such as PM2.5, but indoor measurements are often not sustainable for a large population. This analysis Yuchi et al. (2019) established a randomized monitored experiment using Multiple linear regression and Random Forest Regression has been created to predict indoor PM2.5 for pregnant females involved in an air purification system for indoor air contamination. Also, the researcher created a 10 times cross-validation method of the heterogeneous models connecting MLR and RFR algorithms. The MLR and RFR cross-validation findings were uniforms. The best model was the blended MLR with RFR gives a good prediction. The intervention status was only 6% of the evaluated for indoor PM2.5. The predictive concentration of indoor air contamination is promised by machine learning and combining methods. The proposed technique is good for predicting one concentration but cannot predict many variables.

High air nitrogen dioxide levels affect several aspects of peoples health in the gradually urbanized area. This research (Kamiska (2019)) implements the technique which includes two distinct models (ML for low level and Mu for high-level values) to predict the atmospheric No2 concentration value. The maximum boundary at which the normal absolute error of the prediction was achieved by an iterative method. This model was constructed to traditional random model, where the traditional model asserts that traffic, followed by weather has the greatest effect on No2 levels and establishes the rationale for

constructing distinct models for moderate and great pollution levels. This study was not giving proper justification if there are comparable relationships with other air impurity.

Previous investigations proposed urbanization and industrialization are responsible to increase the air contamination. Air pollution is a major cause of breathing illnesses in contribution to serious financial, social and sanitary issues (KoşanI et al. (2019)). Greater knowledge of the number of PM2.5 concentration in the space-time area is essential to estimate the risk and epidemiological studies. For the prediction of PM2.5 contribution in environment bidirectional long short-Term memory and Recurrent Neural Network is done by researcher Tong et al. (2019). Overall, the above studies multiple effects of air pollution are not properly understood. Air pollution impacts human health in large measure. Different methods including analytical and machine learning techniques study the relationship between individual wellness and air contamination.

2.2 Statistical Approach

Mathematical approaches are used to understand the correlation between different determinants such as the depending variable and independent variable. To overcome the effect of overfitting and underfitting of data, LR and MLR are considered. Asthma has a significant global influence and is a very common fitness problem. The global responsibility of dysfunction investigations has shown, in recent comprehensive analysis, that nearly 334 million people have experienced from asthma. Asthma burden as estimated by inability and unexpected mortality rises among the children (age between 10 to 14) as well as among older people (aged between 75-79). In the previous investigation (Ahn and Kim (2019)), (Kim and Ahn (2018)), three spatial regression model were developed such as SL, SS, and SAC. To examine transport-related asthma with an in-depth focus on vehicle contributions to childhood asthma. The study did not bring into account ordinary least square recurrence due to spatial knowledge dependency, the independence of observation interferes with one of the main hypotheses about OLS regression. Typically, a spatial relationship such as spatial autocorrelation, spatial heterogeneity or both is evaluated by geographical units for further inquiry. Spatial autocorrelation relates to the place in the vicinity and spatial heterogeneity relates spatial differentiation, the average difference in composition and variance. The researcher suggests that SL model assumes dependencies are mainly considered to have annoyance to the spatial correlation between the dependent and SE models level. The overall spatial model involves both the word spatial lag and a spatial error structure, which is a mixed spatial strategy. The SE and SAC model shows that only three transport parameters with serious asthma are statistically important concerning infantile asthma. This involves an adverse impact on the severity of infant asthma caused by bus transits as well as the beneficial impact of active transport variables.
21

Chiang et al. (2016) Study of allergic rhinitis, bronchitis and asthma occurrences in kids living with SO2 contamination at air counselor station in a petrochemical complex. Child classification 11- 14 was based on the 10 km range of the network into elevated or low vulnerability groups. There were substantial differences between elevated exposure and low vulnerability groups between smoking rates, drinking rates, passive smoking exposure and permit for burning. Using t-test significant variations were discovered among allergic rhinitis groups. After this chi-square testing implemented, variable HE and LE groups

have been conducted to verify the correlation between category and disease incidence. In this case, the student t-test has been considered to compare the difference between variables. Incidence of allergic rhinitis and asthma was calculated with the Kaplan-Meier technique from moment to moment probabilities function. The log-rank k test was used to assess the distinction in feature between these HE and LE groups. The proportional risk regression model from Cox has been used for assessing the demographic, ambient and prevalence of allergic rhinitis, bronchitis, and asthma. Constantly higher-level bronchitis and asthma are present in the large exhibition group. The probability of allergic rhinitis, bronchitis, and asthma, HE children were smaller than LE groups. The likelihood was considerably different between 67% in HE and 89% in LE. The HE groups had a greater level of incidence of asthma (18.5%) than the LE group (11.0%) but were not statistically significant in HE. Allergic rhinitis and asthma were less likely in HE children than in LE groups. There were several constraints, firstly misclassification of the air exhibition by unusual subjects was also observed and this research used level of SO2 at the air condition control first of peculiar air quality data.

Study of minor modification in ambient PM2.5, PM10, NO2, O3 levels and fresh incidence of asthma in a historically high-risk ozone region. The impact of enhanced short-term pollutants on the timing of asthma occurs through a cross-over pattern and an unfriendly logical regression stated byWendt et al. (2014). Every metric exposure and pollutant were measured by conditional logistic regression. The bias result from compliant data analysis is reduced with this technique. The cumulative mean measure from the logistic regression model was calculated with the coefficient and p-value for all pollutants. This document also anticipates the uncertainty of asthma from air pollutant due to race. This research did not show clearly in what way the sampling methodology followed a particularly elevated stage of allergens against modifications.

To predict which environmental pollutants and accessibility factor affects the human health Wang et al. (2018) use Ordinary least square method, principal component analysis, and least absolute shrinkage and selection, operator. The suggested research hypothesized that asthma spatial distribution frequency has a comparable social and environmental impact and their primary object was to analyze the significance and impact of distinct variable quantitatively. The main restriction of this research was that this technique was difficult to distinguish environmentally indoor and outdoor pollutants. The effect from air pollutants and temperatures were wrongly evaluated on the assumption that individuals do not relocate and are subjected to natural, rather than unnatural environment.

The scientific risk assessment of air contamination health damages in the jing-jin-Ji are can be the base for determinations to develop and upgrade strategies for reducing environmental pollution. The suggested research assessed the negative health effects of particulate pollution in the jing-jin-Ji region using the linear exposure-response feature. The result of this study shows that air pollution poses unintelligible health and economic risk. This paper provides insight into the financial loss of health that is impacted by air impurities.

2.3 Machine Learning Approach

Studies in-country health showed that asthma has various connections to several variables. It is essential to create an extensive model to identify important components based upon distinct conditions of asthma and several people responded. Machine learning can be helpful for the management of multi-data (Wang et al. (2018)). Air pollutants and allergens are the major incentives that have a major health effect on asthmatics. Asthma can be improved through sequential pain and environmental control, limiting exposure to allergens and irritants, and reducing wider symptoms. By addressing this, Kaffash-Charandabi et al. (2019) created an omnipresent pharmaceutical system to fit all users with assistance everywhere and every moment through any device and network. For this scheme, a machine learning algorithm was implemented such as a support vector machine. This is the most suitable way to predict accurately for a high volume of data. the suggested technique was extremely efficient in the surveillance and management of asthma.

Asthma was criticized for the extremely problematic analytical efficacy and ineffectiveness of the alone analysis and in conjunction of the effects of various pollutants utilizing the association rule mining by the traditional research od external pollution effects (Toti et al. (2016)). This method is simple to comprehend and appropriate for various susceptibility assessments, but conventional exposure-related risk differential metrics were substituted by exposure. The methodology based on ARM produces rule concerning relevant odds that limit the number of a final rule to 0.5%, even at very little support levels. The effects of various air impurity on pediatric asthma were evaluated in a crossover study. Combined law stress policy constraints on air quality based on single contaminant limit and indicate that a higher proportion of odds is caused to harm chemical mixtures. Improvement of algorithms to define the best use limit on the grounds of the information investigated, where the threshold is used to guarantee and safeguard the quality of outdoor air. This algorithm showed no significant relationship between NO2 and asthma exacerbation probabilities.

The asthma system is the main element of these programs. Some attacks affect multiple varieties including environmental variables. Asthma is a chronic airway disease and the investigator believes that the self-management plans to monitor illnesses are essential. This study (Taylor et al. (2019)) aimed at developing a tool that will assist individuals to understand and regulate asthma more efficiently. The research included logistic regression, decision-making tree, boosting gradient and random forest. A decision tree was then accepted for a more reliable model to check these models performance. Evidence has accumulated regarding the impact on respiratory infections and morals of lung cancer. The researcher conducts out time-series studies to estimate the short-term impacts of air contamination in Hefei, China (Zhu et al. (2019)). For the investigation, they consider SO2, NO2, and PM10 as the factor increasing the mortality of respiratory diseases. In comparison, the lung cancer mortality rate was only considerably correlated with SO2. These results indicate that air pollution could enhance respiratory disease and lung cancer mortality.

The interpolation of illnesses such as asthma, which happened by air impurity when two to three pollutants were considered in any prior studies. When statistics such as LR,

MLR and several analytical methods are used, the constraints are often addressed. Many statistical techniques lead to an error in the classification of air contaminant exposures by individual subjects and to how much a particular high allergen is reflected in the sampling method. This restrictive machine learning strategy was overcome. Previous research uses machine learning methods to correlate between human health and air pollution, although the main focus of the study was one pollutant. It is not clear which pollutant impacts human health more.

In our research, we examine how human health is related to the contamination of air by multiple machine-learning algorithms which include Random Forest, ridge regression, lasso regression, elastic net regression, gamboost, and Decision tree and which pollutants have a greater effect on human health, It also concentrates on which machine learning offers a more suitable outcome of air pollution and asthma. These methods are applied to the original information that predicts exact results and offers total precision.

3 Methodology

To achieve a great association between air contamination and human health CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is chosen. The CRISP-DM deal with parts of these problems through a standard model that shows a structure for the implementation of data mining projects, independent both industry and of the technology used Huber et al. (2019). The CRISP-DM method offers an overview of the data mining projects life cycle. It also gives an overview of project stages, activities and results. This cycle is divided into six stages of the data mining project, which involve Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment.

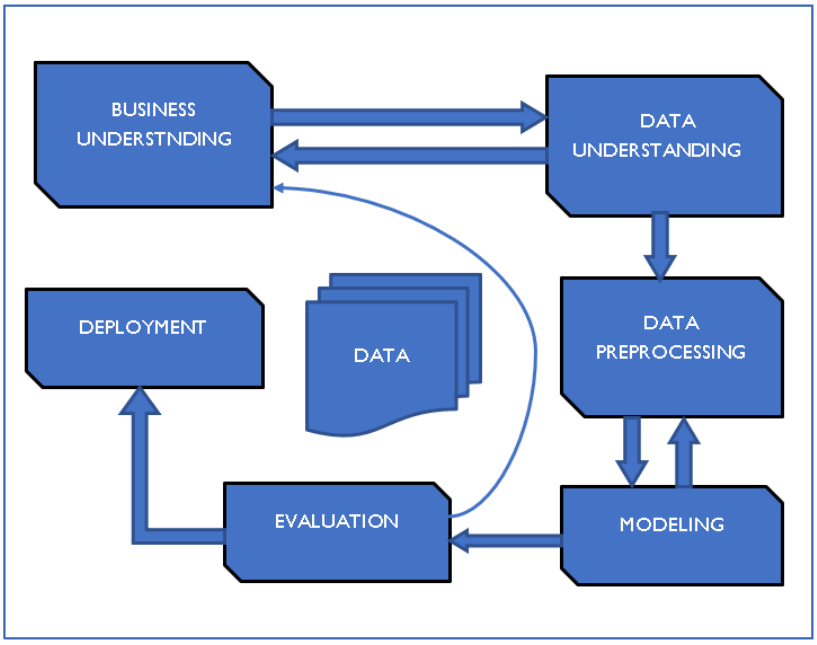

Figure 1: CRISPDM

3.1 Business Understanding

During the preliminary stage, the aim of the data analytics project were understood, in this research our main objective is to be finding the relationship between air impurity and asthma. these specifications were translated from the perspective of the target field into a definition of the concern problem. Which ultimately led to the formulation of a plan for data understanding.

3.2 Data Understanding

This stage begins with primary data gathering and data access. The data quality problems must be identified, and the initial assumption is established. There is a close association between business understanding and data understanding. At some knowledge of the accessible information is necessary for the implementation of the data mining problems.

3.2.1 Data Collection

Different data sources for this study are being applied. CHHS open data portal that presents an asthma patient mortality rate in California. The asthma burden is becoming more comprehensible by measuring the percentage of asthma amongst standard population due to the association between asthma and air condition. This dataset involves a count and rates of asthma patients (per 10,000 people) visit an emergency department by county and age group of 0-17 and 18+ inhabitants in California from 2011 to 2017. This information contains visits to the emergency department in all certified clinics in California[3]. Information about air pollution is collected from United State Environment Protection Agency-EPA. It covers air pollutants estimates that are released into the atmosphere from different sources, which provides knowledge about air contamination including CO, NO2, O3, PN10, PM2.5. This Meteorological data contain concentration value and AQI value of each pollutant measured by air monitors stations in a city, county or state of California[4]. 17 different counties of California were selected for each county five air pollutant were gathered together from the year 2011 to 2017. In total there are 400 CSV's which contain hourly based data for all 17 counties.

3.3 Data preparation:

The data pre-processing stage comprises all operations for building the ultimate dataset from the initial raw data. This task can be done several times. This phase includes data cleaning, data transformation, and feature selection.

3.3.1 Data Cleaning

Once all significant information is collected from various data sources, it is essential to prepare the data into proper form. The outcomes of the models could be poor and may affect the results when raw data is used, therefore data pre-processing is necessary for better outcomes. The different RStudio libraries like dplyr, tidyr, etc were used. That helps to eliminate special characters, missing values, and unwanted records. By using the

[3]https://data.chhs.ca.gov/dataset/asthma-ed-visit-rates-lghc-indicator-07/resource/781708cb-7b25-4967-b760-54b2a4b8cfed

[4]https://www.epa.gov/outdoor-air-quality-data/download-daily-data

gsub function of dplyr library is used to remove the special character like comma which is present in the raw data. There is no NA value present the original dataset.

3.3.2 Data Transformation

For the further process, raw data is converted into certain informative data. Primarily, asthma data contain race ethnicity and gender of the patients. according to this project requirements subset of asthma patients by gender is selected. There are many variables whose data types need to be changed according to the recodes present in the variables. In this case gender and age group are converted character to factor. Meteorological data provides the records of concentration value and AQI value of air pollutants on the hourly bases from year 2011 to 2017 for each country. This data aggregate and converted in to yearly data. date is present in dd-mm-yyyy form is separate using the parse_date_time function of lubridate library. All csvs are merged using rbinf and cbind function. Raw data is randomly distributed which generate the outliers in dataset to overcome, transformation of the variable is necessary. In this case log is taken to remove the outliers and improve the correlation with other variables which make the normally distributed data.

3.4 Modeling

Different modeling methods are chosen and implemented during this stage and their parameters are calibrated to optimum values. Typically, the same sort of data mining problems is subjected to several methods. This phase is thus often carried out in an iterative manner until the selected quality criteria of the models are met. In this research various machine learning algorithms were used such as LASSO Regression, Ridge Regression, Elastic net Regression, Decision Tree, Random Forest, Gam boost.

1. **LASSO Regression**
   LASSO regression is a type of shrinkage linear regression. Shrinkage means that data values shrink to a certain point, such as the average. This regression is suitable for models with high multicollinearity. Thus, LASSO shrinks the coefficient of the least important feature to zero(Shechter et al. (2019)). The estimate result from multivariate LASSO regression indicates how much air contamination affects the asthma patients.

2. **Ridge Regression**
   Ridge regression analyzes multiple regression and can manage multilinearity. Due to multicollinearity least squares are not biased but the variance of the data is large from the actual predictor. Ridge regression reduce the standard error.

3. **ElasticNet Regression**
   ElasticNet is a regularized regression technique combining LASSO and Ridge penalties on a linear basis.

4. **Gamboost**
   Generalized additive models offer a framework for a generalized additive model, enabling non-linear features of each variable to be extended by a conventional linear model while preserving additivity. As linear model generalized additive model used both quantitative and qualitative response. GAMs can fit non-linear data. so

automatically non-linear relationship is considered that fail in conventional linear regression. The non-linear fits could potentially make prediction for target variable more accurately.

5. **Decision Tree**
   Decision tree is one of the method to identify, which variable influencing more and relationship between two or more variables. A decision tree is built from a root node and requires the partitioning of the data into subsets obtaining instance of similar records. In order to calculate numeric sample homogeneity, standard deviation is used. The standard deviation is zero if he numerical sample is entirely uniform. There is no effect to performance on non-linear interaction between parameters.

6. **Random Forest**
   A random forest is an ensemble technique capable of performing both regression and classification tasks by making use of multiple decision trees. The fundamental concept behind it is that various decision tree should be combined to determine the final outcomes instated of using individual decision tree. Random forest select feature randomly and accordingly generate the predicted outcomes.

3.5 Evaluation

Various methods are used in this study to analyze the correlation between asthma patient and air impurity like NO2, CO, O3, PM2.5, PM10 in developed cities. Each of the machine learning models has a special set of features, which helps to achieve effective outcomes. To predict the model outcomes accurately, it is essential to compare model efficiency. Once the algorithm is built, outcomes of the models are evaluated using the appropriate factors by checking the root mean square error(RMSE), Mean absolute error(MAE), Mean percentage absolute error(MAPE), R-Square.

4 Implementation

Implementation is done by adopting the R programming language. For this investigation, various data sources are used. The record of an asthma patient who visits the emergency department is collected from CHHS open data portal and weather impurity data is download from united states Environmental Protection Agency. Next, to this, data is stored in RStudio using read.csv() function. To process the data all required packages including ggplot2, corrplot, caret, rpart, gamboost, glmnet, randomForest, etc are installed in R. All these packages are loaded to the memory using the library function once these packages are installed. Data are transformed into the necessary format before any model is implemented. According to the requirement of this research, the target variable is transformed by taking a log of all records. Important features are extracted from the original datasets which help to fulfill the research objective. The parameter which has taken into consideration to divine the association between asthma patients and air contamination are concentration values of all pollutants, AQI values and age group of patients and their gender. The following section gives a brief description of data exploratory analysis and model outcomes.

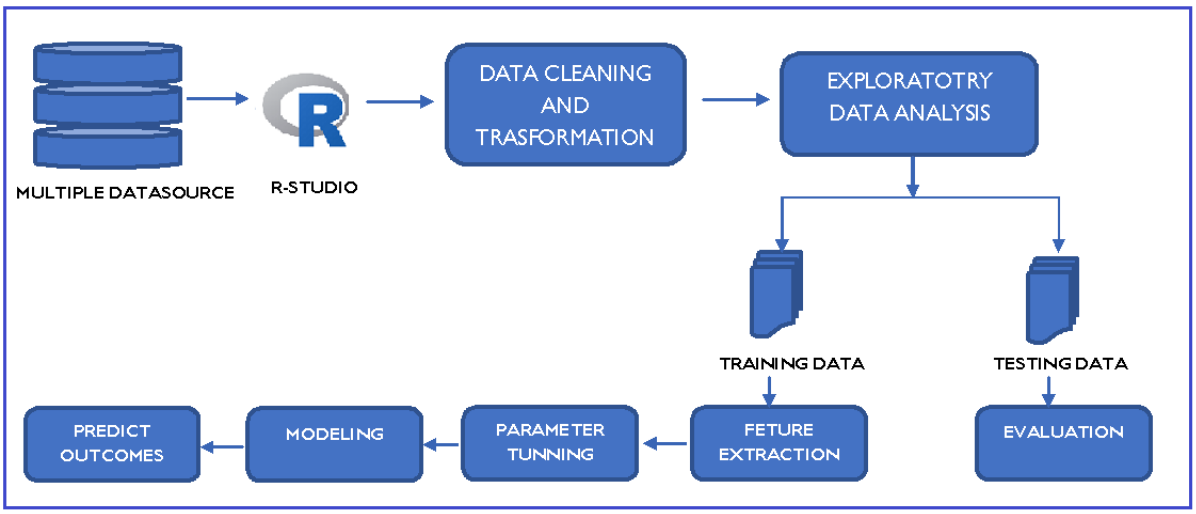

Figure 2: Work Flow Diagram

4.1 Data Exploratory Analysis

Introductory data interpretation was carried out with multivariate analysis and data normality, identification of outliers, the correlation between the features, distribution of the variables. Most of the data is in continuous form except some feature like gender, age group, and counties which are in categorical form.

Preliminary data analysis was carried out with multivariate analysis and data normality, identification of outliers, the correlation between the features, distribution of the variables. Most of the data is in continuous form except some feature like gender, age group, and counties which are in categorical form.

The outliers in the dataset reported by data exploration. For checking the outliers boxplot is used. Patients who visit the emergency department of asthma (Numerator) has these outliers which means some value is above the normal scale. This outlier is handling by taking a log of the records.

An examination on the correlation between the numerical attributes namely CO2 concentration value, NO2 concentration value, O3 concentration value, PM10 concentration value, PM2.5 concentration value. CO AQI, NO2 AQI, O3 AQI, PM10 AQI, PM2.5 AQI, Numerator and NumeratorTransformed are used for checking the correlation with the target variable. If the correlation is more than 0.7 indicate a strong correlation, below 0.3 indicate a weak correlation. If the value is in-between 0.3 and 0.7 means the moderate correlation is present. Figure 3 hows the correlation with the actual variable in which correlation between target variable i.e numerator is weak with other variable but after transforming the target variable correlation is increased.

According to the Figure 3 correlation between the independent variable is strong, this highlight the multicollinearity in the dataset. Therefore, the correlation between these variables are considered if the VIF score between them is below 10 otherwise, the finding can be difficult to predict.

Further investigation shows the pattern between the dependent and independent variables where the record of these variables is non linearly distributed, to predict the outcome of the research objective non -linear models are implemented.

According to the Figure 3 correlation between the an independent variable is strong,

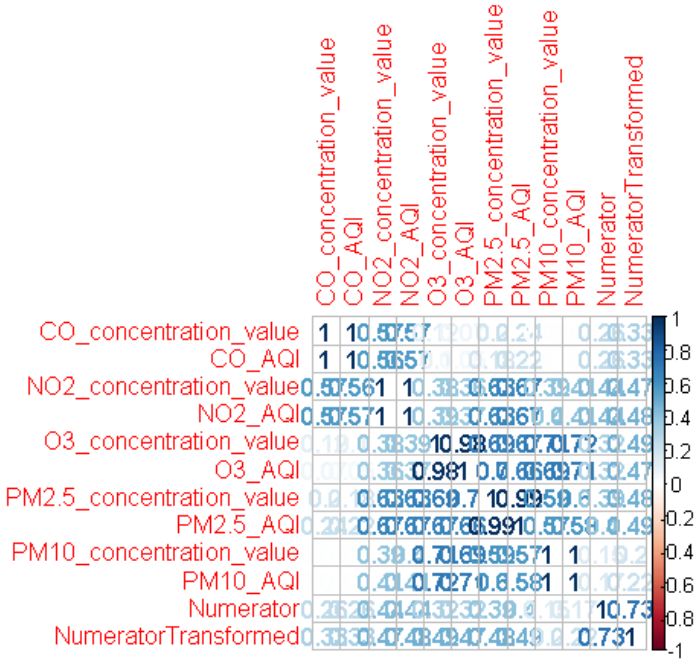

Figure 3: Correlation with variables

this highlight the multicollinearity in the dataset. Therefore, the correlation between these variables are considered if the VIF score between them is below 10 otherwise, the finding can be difficult to predict.

Further investigation shows the pattern between the dependent and independent variables where record of this variables is non linearly distributed, to predict the outcome of the research objective non -linear models are implemented.

4.2 Feature Selection

Once the data is converted into appropriate form, data required for further analysis have been extracted from data. For fracture selection Boruta package is used. Which compares the importance of attribute iteratively with the target attributes. Attributes which are considerably worse that variable are rejected and variable who show good relationship with target variable are confirmed. As our datasets contain 23 variables in which only 19 variables are confirmed other are rejected. This confirmed variable is used for the further analysis.

4.3 Statistical Test of Assumption

In this research project, statistical assumption test has been implemented. To Check the occurrence of multicolineariry among the independent variables vif function was used. vif value of all variables is less than 2 which indicate that there is no influence of multicolinearity. Multivariate Normality test was carried out to check whether residuals are normally distributed or not by using Shipro Wilk test. Relation between residuals and predicted value was tested to check the assumption of linearity and homoscedasticity

Table 1: Feature Description

Attributes	Description	Data type
county	Name of california county	Factor
year	year between 2011 to 2017	Num
CO_concentration_value	Concentration value of CO	Num
CO_AQI	AQI value of CO	Num
NO2_concentration_value	Concentration value of NO2	Num
NO2_AQI	AQI value of NO2	Num
O3_concentration_value	Concentration value of O3	Num
O3_AQI	AQI value of O3	Num
PM10_concentration_value	Concentration value of PM10	Num
PM10_AQI	AQI value of PM10	Num
PM2.5_concentration_value	Concentration value of PM2.5	Num
PM2.5_AQI	AQI value od PM2.5	Num
Age_group	Age group divided into 18+ and below 18	Factor
Strata_Name	Gender of patient	Factor
Numerator	Number of patient who visit emergency department	Num
NumeratorTransformed	logarithm of Numerator	Num

where both assumption are satisfied. Durbin Watson Test was performed to find Auto correlation of residuals. residuals and independent variables are uncorrelated.

4.4 Data Splitting into Testing and Training

These insights were acquired through the evaluation of exploratory data analysis facilitated by the preparing or preparation of information. Data are further distributed in to large portion of training data i.e 70% and small portion of testing data i.e 30%. This separation has also kept a balanced set of data for optimal testing.

4.5 Hyperparameter Tuning and Cross Validation

While training the regression model, 10- fold cross validation was used with 3 repeats due to the restricted observations and the overfitting. For the better computation of model hyperparameter tuning for done using random search and grid search in order to get the accurate result. Where random search is a method used to discover the best alternative for the built-in model with random combinations of hyperparameters. It attempts a number of random combinations. To optimize the function with a random search, several random configurations of the parameter are evaluated.

4.6 Model Performance

The construction of the various models was carried out with the use of the train function in the (caret) package. The method was initially to implement easier models and then use more complicated models. According to the exploratory data analysis, data utilized in this study project has some non-linear features. By considering this LASSO regression, Ridge regression, Elasticnet regression, Gamboost, Decision tree, and Random forest were

applied one after another. Once the data is normalized, the control parameter was set up using the train control function from the caret package were random search is defined.

At first LASSO regression was carried out using the grid search method which provides the optimum model parameter. This grid function has been defined by expand.grid function in R-studio before running the trained model. This grid search enables optimize parameters defining the value of alpha 1 and giving the range of lambda. Similarly, ridge regression was implemented where an alpha value was defined as 0 and the range of lambda was same as LASSO regression. For the ElasticNet regression, the alpha value is elected randomly between 0 to 1 and accordingly lambda value is suggested by the random search tuning function.

After performing regression models, Gamboost, Decision tree, Random Forest has been implemented. The parameter of this model is optimized by using the tuning function of random search using the caret package. On tuning of these parameters, the best-identified values obtained were mstop- 957, step_size- 0.1, Offset- 7.056, Number_of_baselearners- 5 using these parameters gamboost model was implemented to predict the number of patients who visits the emergency department of asthma. Similarly, random forest identified values obtained were ntree- 1, mtry- 1, forest- 11. And the decision tree has been identified by using random search, which performs the multiple iterations and gives the best possible outcomes. To analyze its efficiency and reliability, the model has been tested with test data.

5 Evaluation

The main objective of this research is to find the relationship between the human health and air contamination. To predict the number of patients who visit emergency department of asthma. Various machine learning algorithms was used for finding the result. Each algorithm has their own important functionality based on this, model gives the result. It is important to evaluate the model performance with testing data by using the some mathematical measurement like RMSE(root mean squared error) and MAE(mean absolute error)

1. **Root Mean Square Error**

   RMSE is the square root of mean squared error, which is used to measure average prediction error. The difference between the predicted value and actual value for each stage is to be computed. Lower value of RMSE indicate better fit while root mean square percentage is indicate the decline in model performance.

2. **Mean Absolute Error**

   The MAE is calculated as an average total difference of the observed value from the predicted value. The MAE is a linear result that the individual differences are weighted in the same way.

3. **R-Squared**
   It is difficult to understand, whether the model is good fit or not looking into the absolute error(MAE) and Root mean square error(RMSE). To check the how accurately our model fits to the base line R-Square is calculated. R-Square is always between -Inf to 1, where negative sign indicates the worst fit model and large positive value indicate good fit.

5.1 Result

This research shows the relationship between air contamination and asthma patients. For analyzing purpose five air pollutant was considered such as NO2, CO, O3, PM2.5, PM10 in which NO2, O3, PM2.5 show the strong relationship than other pollutants. Once we get the correlation between then we predict the number of who visits the emergency department of asthma by using various machine learning algorithms.

Table 2: Model Performance Outcomes

Models	**RMSE**	**R-Squared**	**MAE**
Ridge Regression	0.042	0.46	0.056
ElasticNet Regression	0.0413	0.500	0.0534
LASSO Regression	0.0412	0.503	0.0531
Gamboost	0.039	0.57	0.051
Decision Tree	0.026	0.72	0.032
Random Forest	0.017	0.92	0.021

In the Table 2, Random Forest gives better performance with the 10-fold cross-validation with 3 repeats. where RMSE, R-Squared, MAE was computed. the values of models are as follows RMSE= 0.017, R-Squared= 0.92 and MAE= 0.021. Random Forest gives the 92% of better fit when compared to others.

Figure 4 shows the various regression residual plot including residual vs fitted, QQ plot, coefficient plot and cumulative distribution plot. Generally, residual is calculated by taking the difference of observed value of dependent variable and predicted value. Main aim behind using the plot is to justify the assumptions like linearity and homoscedasticity. this plot has been implemented for all regression models. residual vs fitted plot does not generate any particular pattern and residuals are distributed near about zero line.The obtained outcomes satisfied the above both assumption.

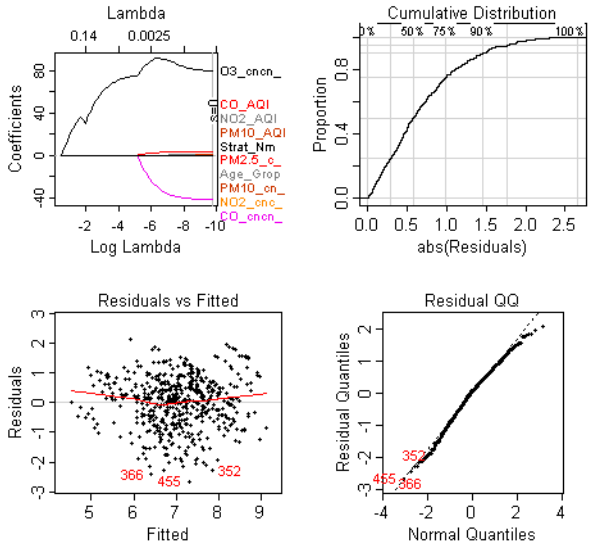

Figure 4: Residual Plot

6 Discussion

A significant challenge for developed cities and government health organizations, which create policies and strategies to guarantee the conservation of the living environment and mitigate inadequate air quality Kim and Ahn (2018). Environmental factors built in urban developed cities contributes to the aggravation of asthma symptoms. Figure 5 Adults are more prominently affected by asthma then child. It is also true that the dense metropolitan environment worsens air quality, causing an adverse impact on the morbidity of asthma in elderly local individuals. Therefore, various machine learning methods was introduced to establish a powerful relationship between asthma and air impurity.

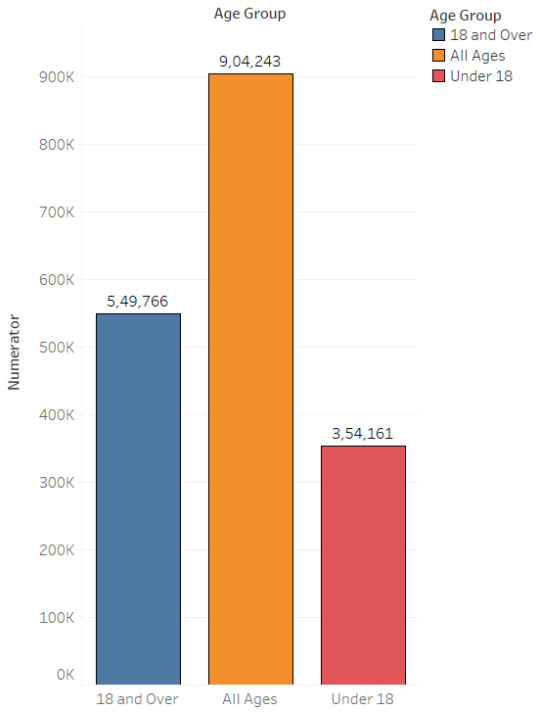

Figure 5: Count of patient based on Age group

Earlier research made use of IOT-enabled Hadoop-based analysis, the suggested framework can lead to some revolutionary modification in younger generations respiratory health. In this research we used the machine learning techniques to carry out the relation of human health and air impurity. Amongst them decision tree provides the variable importance. As shown in Figure 6 No2 concentration value, O3 concentration value has high priority. The output of the regression model suggests the air contamination affects the asthma patients. Our finding gives the insights of every pollutants with asthma patients including their age group. Various algorithms were compared to find the accurate outcomes between the dependent variable i.e numerator(count of asthma patient) and independent variables such as all air pollutants.

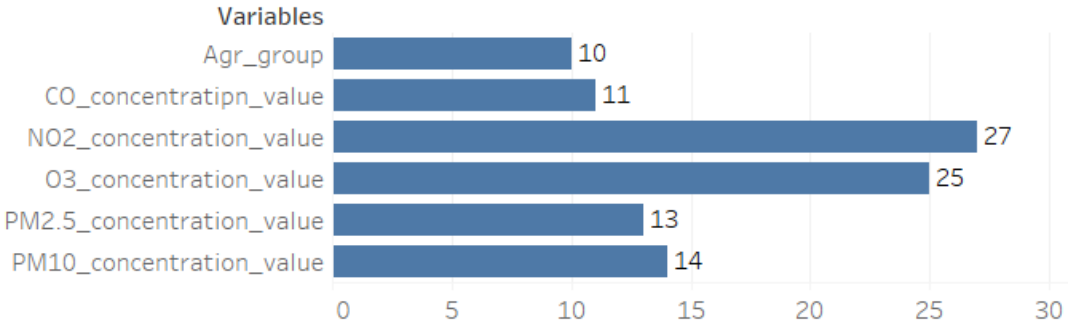

Figure 6: Feature Importance

7 Conclusion and Future Work

This research focuses on the relationship between air impurity and respiratory diseases. In our study, machine learning was rewarding for a simpler choice of variable, better outcomes and prediction. Various algorithms such as Ridge regression, ElasticNet regression, LASSO regression, Gamboost, Decision tree, Random forests were used. To identify the parameters that influence the model outcomes feature selection was performed by using boruta package which give the 11 conformed attributes and decision tree and random forest also gives the feature importance. Model outcome is evaluated using the statistical measurements like RMMSE, MAE and R-Squared. This research will help to improve air quality measurement and reduce the asthma mortality rate which indirectly enhancing the country health.

Our research had several restrictions, as the data contain information about air pollutant and count of asthma patients with their age group. Indoor and outdoor exposure of individuals was difficult to discriminate against the environment. Air contamination is not only the influencing factor, other then this genetic data could also influence. If relevant data is taken into consideration there is possibility to get better outcomes.

References

Ahn, Y. and Kim, D. (2019). The prevalence of asthma and severe asthma in children influenced by transportation factors: Evidence from spatial analysis in seoul, korea, *Cities* **85**: 30 – 37.
URL: *http://www.sciencedirect.com/science/article/pii/S0264275118307650*

Brusseau, M., Ramirez-Andreotta, M., Pepper, I. and Maximillian, J. (2019). Chapter 26 - environmental impacts on human health and well-being, pp. 477 – 499.
URL: *http://www.sciencedirect.com/science/article/pii/B9780128147191000264*

Chiang, T.-Y., Yuan, T.-H., Shie, R.-H., Chen, C.-F. and Chan, C.-C. (2016). Increased incidence of allergic rhinitis, bronchitis and asthma, in children living near a petrochemical complex with so2 pollution, *Environment International* **96**: 1 – 7.
URL: *http://www.sciencedirect.com/science/article/pii/S0160412016302999*

Cox Jr., Louis Anthony & Popken, D. A. . S. R. X. (2018). *Descriptive Analytics for Public Health: Socioeconomic and Air Pollution Correlates of Adult Asthma, Heart*

Attack, and Stroke Risks, Springer International Publishing.
URL: *https://doi.org/10.1007/978-3-319-78242-3_3*

Guarnieri, M. and Balmes, J. R. (2014). Outdoor air pollution and asthma, *The Lancet* **383**(9928): 1581 – 1592.
URL: *http://www.sciencedirect.com/science/article/pii/S0140673614606176*

Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications a holistic extension to the crisp-dm model, *Procedia CIRP* **79**: 403 – 408.
URL: *http://www.sciencedirect.com/science/article/pii/S2212827119302239*

Kaffash-Charandabi, N., Alesheikh, A. A. and Sharif, M. (2019). A ubiquitous asthma monitoring framework based on ambient air pollutants and individuals' contexts, *Environmental Science and Pollution Research* .
URL: *https://doi.org/10.1007/s11356-019-04185-3*

Kamiska, J. A. (2019). A random forest partition model for predicting no2 concentrations from traffic flow and meteorological conditions, *Science of The Total Environment* **651**: 475 – 483.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969718336416*

Kim, D. and Ahn, Y. (2018). Built environment factors contribute to asthma morbidity in older people: A case study of seoul, korea, *Journal of Transport & Health* **8**: 91 – 99.
URL: *http://www.sciencedirect.com/science/article/pii/S2214140517300671*

KoşanI, Z., Kavuncuoğlu, D., Yılmaz, S., Bilici, A. S. and Aras, A. (2019). An evaluation of the relation between air quality and hospital presentations due to respiratory tract diseases: A cross-sectional study.

Paul S.T., Raimond K., K. G. (2018). An iot-enabled hadoop-based data analytics and prediction framework for a pollution-free smart-township and an asthma-free generation, *Advances in Big Data and Cloud Computing. Advances in Intelligent Systems and Computing* **750**: 577–587.

Piyatilake, I. T. S. and Perera, S. (2018). Mathematical model to quantify air pollution in cities, pp. 147–178.
URL: *https://doi.org/10.1007/978-981-13-1153-6_7*

Shechter, J., Roy, A., Naureckas, S., Estabrook, C. and Mohanty, N. (2019). Variables associated with emergency department utilization by pediatric patients with asthma in a federally qualified health center, *Journal of Community Health* .
URL: *https://doi.org/10.1007/s10900-019-00653-6*

Taylor, J., Shrubsole, C., Symonds, P., Mackenzie, I. and Davies, M. (2019). Application of an indoor air pollution metamodel to a spatiallydistributed housing stock, *Science of The Total Environment* **667**: 390 – 399.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969719308484*

Tong, W., Li, L., Zhou, X., Hamilton, A. and Zhang, K. (2019). Deep learning pm2.5 concentrations with bidirectional lstm rnn, *Air Quality, Atmosphere & Health* .
URL: *https://doi.org/10.1007/s11869-018-0647-4*

Toti, G., Vilalta, R., Lindner, P., Lefer, B., Macias, C. and Price, D. (2016). Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining, *Artificial Intelligence in Medicine* **74**: 44 – 52.
URL: *http://www.sciencedirect.com/science/article/pii/S0933365715301032*

Wang, M., Qiu, J. and Ong, P. (2018). Machine learning in spatial study of asthma rate distribution in los angeles county, pp. 3696–3700.
URL: *https://ieeexplore.ieee.org/abstract/document/8623677*

Wendt, J. K., Symanski, E., Stock, T. H., Chan, W. and Du, X. L. (2014). Association of short-term increases in ambient air pollution and timing of initial asthma diagnosis among medicaid-enrolled children in a metropolitan area, *Environmental Research* **131**: 50 – 58.
URL: *http://www.sciencedirect.com/science/article/pii/S0013935114000401*

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S. B., Janes, C. R., Lanphear, B. P., McCandless, L. C., Takaro, T. K., Venners, S. A., Webster, G. M. and Allen, R. W. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city, *Environmental Pollution* **245**: 746 – 753.
URL: *http://www.sciencedirect.com/science/article/pii/S0269749118323017*

Zhu, F., Ding, R., Lei, R., Cheng, H., Liu, J., Shen, C., Zhang, C., Xu, Y., Xiao, C., Li, X., Zhang, J. and Cao, J. (2019). The short-term effects of air pollution on respiratory diseases and lung cancer mortality in hefei: A time-series analysis, *Respiratory Medicine* **146**: 57 – 65.
URL: *http://www.sciencedirect.com/science/article/pii/S0954611118303901*