

**A NOVEL APPROACH ON AIR QUALITY
PREDICTION AND CLASSIFICATION USING
MACHINE LEARNING**

A PROJECT REPORT

Submitted by

HARINI.R.A 211420243019

SHREEMATHI.S 211420243051

in partial fulfilment for the award of degree

of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE



PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

MARCH 2024

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report “**A NOVEL APPROACH ON AIR QUALITY PREDICTION AND CLASSIFICATION USING MACHINE LEARNING**” is the bonafide work of **HARINI.R.A [REGISTER NO: 211420243019]**, **SHREEMATHI.S [REGISTER NO: 211420243051]** who carried out this project work under **Dr.P.KAVITHA** supervision.

SIGNATURE

DR. S. MALATHI M.E., Ph.D.,
HEAD OF THE DEPARTMENT
DEPARTMENT OF AI&DS,
PANIMALAR ENGINEERING
COLLEGE,
CHENNAI-600123.

SIGNATURE

Dr.P.KAVITHA M.Tech., Ph.D.,
ASSOCIATE PROFESSOR
DEPARTMENT OF AI&DS,
PANIMALAR ENGINEERING
COLLEGE,
CHENNAI-600123.

Certified that the above mentioned students were examined in End Semester project (AD8811) viva-voice held on 27.03.2024.

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION BY THE STUDENT

We **HARINI.R.A (211420243019), SHREEMATHI.S (211420243051)** hereby declare that this project report titled “**A NOVEL APPROACH ON AIR QUALITY PREDICTION AND CLASSIFICATION USING MACHINE LEARNING**” under the guidance of **Dr.P.KAVITHA** is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

ACKNOWLEDGMENT

Our profound gratitude is directed towards our esteemed Secretary and Correspondent, **Dr. P. CHINNADURAI, M.A., Ph.D.**, for his fervent encouragement. His inspirational support proved instrumental in galvanizing our efforts, ultimately contributing significantly to the successful completion of this project

We want to express our deep gratitude to our Directors, **Tmt. C. VIJAYARAJESWARI, Dr. C. SAKTHI KUMAR, M.E., Ph.D., and Dr. SARANYASREE SAKTHI KUMAR, B.E., M.B.A., Ph.D.**, for graciously affording us the essential resources and facilities for undertaking of this project.

Our gratitude is also extended to our Principal, **Dr. K. MANI, M.E., Ph.D.**, whose facilitation proved pivotal in the successful completion of this project.

We express our heartfelt thanks to **Dr. S. MALATHI, M.E., Ph.D.**, Head of the Department of artificial intelligence and data science, for granting the necessary facilities that contributed to the timely and successful completion of project.

We would like to thank our supervisor **Dr.P.KAVITHA**, coordinators **Dr. K.JAYASHREE**, and all the faculty members of the Department of AI&DS for their advice and encouragement for the successful completion of the project.

HARINI.R.A

SHREEMATHI.S

Abstract

Background: Air pollution has become a serious environmental issue. It is responsible for hundreds of fatalities each year and it poses a serious threat to human health and environment. It leads to global warming, greenhouse effect and it also causes respiratory problems like asthma, lung cancer etc. It is important to predict the quality of the air to regulate air pollution. Air quality index (AQI) is a measure of air quality which describes the level of air pollution. Machine learning algorithms can help in predicting the AQI. Linear regression, LASSO regression, ridge regression, and SVR algorithms were used to forecast the AQI.

Objectives: The main objective of the thesis is to build and train a models using machine learning algorithms and find out the most accurate model in predicting the AQI.

Methods: Literature Review and Experimentation were chosen as methods to answer the research questions. There are a number of research papers written on prediction of AQI and literature review helped us a lot in research and references. Experimentation is also used to find out the most accurate machine learning model in predicting the air quality. In the experimentation phase, four machine learning algorithms were trained with air quality data to create predictive models for forecasting AQI.

Results: Algorithms like Logistic Regression, Ridge Regression, LASSO Regression, and SVR are selected through literature review. Upon experimentation and training the algorithm with "Air Quality Data in India (2015-2020)" data set has showed that Ridge regression has the least MAE and RMSE and the highest R-square, which shows that it has the highest performance in predicting the AQI.

Conclusions: Four models are built by training with machine learning algorithms like Logistic Regression, Ridge Regression, LASSO Regression, and SVR and "Air Quality Data in India (2015-2020)" data set. After experimentation, it was found that Ridge Regression and LASSO regression has the better performance in the prediction of AQI.

Keywords: Air Quality Index, Air Quality Monitoring, Machine Learning, Regression.

Contents

Abstract

Acknowledgments

1	Introduction	1
1.1	Aim and objectives	2
1.1.1	Aim:	2
1.1.2	Objectives:	2
1.2	Research questions	2
1.3	Scope of the thesis	3
1.4	Overview	3
2	Background	4
2.1	Air Quality	4
2.2	Machine Learning	4
2.2.1	Supervised Learning Algorithms	4
2.2.2	Unsupervised Learning Algorithms	5
2.2.3	Semi-supervised Learning Algorithms	5
2.3	Regression	5
2.3.1	Linear Regression	5
2.3.2	Ridge Regression	6
2.3.3	LASSO Regression	6
2.3.4	Support Vector Regression	6
3	Related Work	7
3.1	Air Quality Index Predictive Models	7
3.1.1	Regression Models	7
4	Method	9
4.1	Literature review	9
4.2	Experimentation	10
4.2.1	Environment	10
4.2.2	Data	11
4.2.3	Data preprocessing	11
4.2.4	Training and testing the model	12
4.2.5	Performance metrics	13

5	Results and Analysis	14
5.1	Literature review outcome.....	14
5.2	Experimentation outcome	17
5.2.1	Linear Regression	17
5.2.2	Lasso Regression.....	19
5.2.3	Ridge Regression	21
5.2.4	Support vector regression	23
5.2.5	Comparison of results.....	25
6	Discussion	27
7	Conclusions and Future Work	28
7.1	Conclusion	28
7.2	Future work	28
8	References	29

List of Figures

4.1	Sample of air quality data in India (2015 - 2020)	11
5.1	Prediction of air quality index using linear regression	17
5.2	Scatter plot graph for linear regression model.....	17
5.3	Performance metrics (Linear regression)	18
5.4	Predicted vs observed (Linear regression).....	19
5.5	Prediction of air quality index using lasso regression.....	19
5.6	Performance metrics (Lasso regression).....	19
5.7	Predicted vs observed (Lasso regression)	20
5.8	Prediction of air quality index using ridge regression.....	21
5.9	Performance metrics (Ridge regression)	21
5.10	Predicted vs Observed (Ridge regression).....	21
5.11	Prediction of air quality index using support vector regression.....	23
5.12	Performance metrics (SVR)	23
5.13	Predicted values vs Test values (SVR).....	24
5.14	Mean absolute error bar chart	25
5.15	Root mean square error bar chart	25
5.16	R-square error bar chart.....	26

List of Tables

5.1	Literature review.....	14
5.2	Literature review.....	15
5.3	Literature review.....	16
5.4	Comparison of performance metrics for Linear regression model	18
5.5	Comparison of performance metrics for lasso regression	19
5.6	Comparison of performance metrics for ridge regression.....	21
5.7	Comparison of performance metrics for SVR model	23
5.8	Comparison of performance metrics for all models	26

List of Acronyms

AQI Air Quality Index.

LASSO Least Absolute Shrinkage and Selection Operator.

MAE Mean Absolute Error.

RMSE Root Mean Square Error.

SVR Support Vector Regression.

Chapter 1

Introduction

The oxygen we breathe is the basic element of life as it is very much essential for the human body cells to function. Air is one of the most important elements on the planet earth. We humans can stay up to days without water whereas, we cannot live for up to a few minutes without air. Air also maintains the temperature of the planet's surface by circulating hot and cool air. The water cycle also depends on the air. The air we breathe not only keeps us alive but also determines the quality of life we can live. Health changes can be highly triggered by the low quality of air.

Polluted air can be very harmful in many ways such as it can cause respiratory diseases like tuberculosis, pneumonia, bronchitis, asthma, and lung cancer [3]. It is found that air pollution kills about 7 million people every year worldwide. Air pollution can also cause global warming which is a process in which heat is trapped in the air that resulting in effects like temperature hikes, rising sea levels, diseases caused by heat, and transmissions that can cause infectious diseases.

The quality of air can be measured. AQI is a measure that determines the quality of the air. AQI is a random scale ranging from 0 to 500 [8]. Generally, the more the AQI, the more the level of air pollution and more the health concern. For example, an AQI of 50 or less is said to be good air quality and AQI of 300 or above resembles hazardous quality. AQI is basically classified into six categories of health concerns. Each category is given a color for easy interpretation. The six categories are Green (0-50), Yellow (51-100), Orange (101-150), Red (151-200), Purple (201-300), and Maroon (300 and higher). Calculating AQI is essential because it provides critical information about the state of the air. We can stop people getting effected due to pollution by warning about the air quality whenever AQI exceeds the maximum value.

The purpose of AQI is to warn people about the air they are living in and also letting them know which group of people might get affected by the air and also measures that an individual can take to prevent submission to air pollution. AQI focuses on those health effects that people might suffer from breathing the polluted air within a few hours to a few days.

Prediction of data can be achieved using many machine learning algorithms [7]. Machine Learning is a field of science where models are build by training algorithms to understand patterns in data and make possible decisions. Machine learning can

be supervised, unsupervised, and semi-supervised. Supervised learning approach is a process where each of the input data has a mapped output data. In this project, we decided to go with the supervised learning approach. We worked with supervised learning algorithms as we have access to labelled data and also we learned that the supervised algorithms are better at accuracy than unsupervised algorithms[6][4]. This project is about comparing various supervised machine learning algorithms and selecting the most accurate algorithm for the prediction of AQI with help of "Air quality index in India (2015-2020)" data set. PM 2.5, PM 10 and NO2 attributes present in the data set will be used in computation of AQI.

The authors of reference [5] have compared the machine learning algorithms for predicting the AQI in different areas and stated that neural network algorithms were superior than other machine learning algorithms in predicting AQI. But this lacks in determining hourly pollutant levels. The authors of reference [2] have stated that the regression machine learning algorithm was efficient in determining the AQI. In this research, we will try to compare these two supervised machine learning algorithms based on the previous research done and determine the efficient supervised machine learning algorithm.

1.1 Aim and objectives

1.1.1 Aim:

The primary aim of this study is to determine the most accurate supervised machine learning algorithm to build a model that can predict the AQI. We also aim to outline how machine learning helps in controlling air pollution.

1.1.2 Objectives:

The followings are the objectives set to achieve the aim:

- To review studies related to AQI.
- To determine the most usual supervised machine learning algorithms that were used to predict the AQI by performing a literature review.
- To build predictive machine learning models for forecasting AQI.
- To evaluate the performance and accuracy of created models and determine the most accurate supervised machine learning algorithm in the prediction of AQI.

1.2 Research questions

RQ1: How well can machine learning help in reducing air pollution by forecasting AQI?

Method opted: Literature review.

Motivation: There are many research papers that are done to predict the AQI. Therefore, literature review can be the most suitable method for the problem. The main aim of this research question is to summarize how supervised machine learning techniques help in reducing air pollution by predicting the air quality index. By performing literature review, the most usual supervised machine learning algorithms that were employed in predicting the AQI will be identified.

RQ2: Which is the most effective machine learning algorithm and has the most accuracy in predicting the AQI?

Method opted: Experimentation.

Motivation: Experimentation is chosen for this research question because we build a model by training an algorithm to analyse the performance of algorithms. Experimentation can give better results when compared to reviews because we build an efficient model. The main aim of the research is to find the most effective and accurate machine learning algorithm for the prediction of AQI.

1.3 Scope of the thesis

The main aim of the thesis is to develop a machine learning model to predict the AQI. The models are built by training machine learning algorithms like Linear regression, Ridge regression, LASSO regression, and SVR. The data set "Air quality data in India (2015-2020)" used in the research does not have large amount of data. However, if the data is taken on a world scale, Regression algorithms might take a long time to train and might lose accuracy.

1.4 Overview

The research study is divided into chapters and a brief description of each chapter is given below:

In chapter 2, the background of the study such as the information about machine learning and algorithms used are discussed. In chapter 3, a brief overview of the previous studies related to AQI is discussed. Chapter 4 discusses the methods used to answer the research questions. In chapter 5, results obtained by performing literature review and experimentation are discussed. Chapter 6 is about the discussion on the thesis. Chapter 7 discusses about the conclusion and future of the thesis.

2.1 Air Quality

Worldwide, many cities continuously assess the air quality using various monitoring techniques to record the concentrations of the pollutants in the air. Air quality can be defined as the measurement of quality of the air we breathe and the concentrations of the pollutants in the air that can cause various health issues. Air quality can be used for various purposes such as the communication of air quality with the public, to plan strategies that can be used to reduce air pollution, and to monitor short term and long term trends. [17]

Air quality can be measured using various machine learning algorithms. Many countries and their environmental agencies in the world use the AQI for the real time spreading of the information on the air quality. Although the basic concepts of air quality are similar, the practical implementations of each can differ. Applying AQIs on a common set of data can show large differences in the index values and concentration of pollutants.

2.2 Machine Learning

Machine learning is a field of study where models are built by training various learning algorithms. Machine learning algorithms are trained to build the models for identifying patterns in the data and make predictions. When trained, computer programs can take their own decisions and give outputs to the user. Machine learning is closely related to mathematical computations where algorithms perform various computations for predicting data. The efficiency of each algorithm is directly proportional to the amount of training data. Machine learning algorithms are basically categorised into supervised learning algorithms, unsupervised learning algorithms, and semi-supervised learning algorithms.

2.2.1 Supervised Learning Algorithms

Supervised learning algorithms are those algorithms where for every input data there is a output data mapped. The input data (basically a vector) always have a desired output value (signal). Supervised learning algorithms analyses the patterns between the data and develops a mapping function that can map any input to a output for new examples. There are various supervised learning algorithms such as Support

Vector Machine (SVM), Decision tree, K-means, Naive Bayes, Random forest, and Artificial neural networks (ANN). [22]

2.2.2 Unsupervised Learning Algorithms

Unsupervised machine learning algorithms are the learning algorithms where, like the supervised learning algorithms, there isn't any developing of a mapping function to map a input data to a output data. The main notion of a unsupervised algorithms is to build representations of input data that can be used to predict future output, take decisions, and efficiently communicating the input to other machines. Few examples for unsupervised learning algorithms are K-means clustering, Principal Component analysis, and Hierarchical clustering. [5]

2.2.3 Semi-supervised Learning Algorithms

As the name suggests, semi supervised learning lies between supervised and unsupervised. Most of the semi-supervised learning algorithms are basically extensions of either supervised or unsupervised algorithms to include additional information of the other paradigms. In most of the tasks, there is a very small amount of labelled data because acquiring labels can be very difficult as the process requires human annotators, special devices, and slow experiments. Semi-supervised learning can be more effective than supervised learning because semi-supervised uses both labeled and unlabeled data for learning. In other words, the performance of semi-supervised learning is on par with that of supervised learning but with fewer labeled data. [26]

2.3 Regression

Regression is a supervised machine learning approach that is generally used to predict continuous values. Regression is majorly used for two purposes. First, it is used in forecasting and prediction of data, which are the applications of machine learning. Second, regression analysis is used to determine the relation between the dependent and independent variables in the data set. Linear regression, Ridge regression, and LASSO regression are a few examples for regression models. [14]

2.3.1 Linear Regression

Linear regression is a machine learning approach which is used to establish a relation between independent variable and dependent variable. Here, the variable which is predicted is called the dependent variable and the variable that is used to predict the dependent variable is called the independent variable. Linear regression is said to give the simplest form of the regression function as a linear equation of variables. The interpretation of the parameters is easy due to the linear form of the regression function. [23]

2.3.2 Ridge Regression

Ridge regression model is one of the widely used parameter estimation method. Ridge regression is introduced to solve the co-linearity problem that occurs in multiple linear regression. In other models, to avoid co-linearity, few independent variables are removed to increase the final co-relation matrix of the other independent variables. In ridge regression, co-linearity is solved without removing the variables from the original data set of independent variables. [15]

2.3.3 LASSO Regression

LASSO regression is a regression approach that uses shrinkage. Shrinkage is where the values of the data are shrunk towards a common value like the mean of the values. It is a shrinkage and variable selection model. It's main aim is to identify those variables that minimise errors in prediction. [18]

2.3.4 Support Vector Regression

SVR is an analytical machine learning approach which is used to explore the relationship between one or more predictor variables and a real valued dependent variable. It is an approach in which the model learns the importance of distinguishing the co-relation between the input and output. It can applied to regression problems by introducing a loss funvtn. The basics of SVR is to map a input data to a random high dimensional feature space and then to obtain and solve the regression problem in that particular feature space. [25]

This section consists of a brief overview of previous studies related to the prediction of AQI. The techniques related to predicting AQI have been studied and reviewed. The machine learning approaches were examined to demonstrate why these approaches are capable to perform well in forecasting air quality.

3.1 Air Quality Index Predictive Models

An predictive air quality model is a computation tool which predicts the AQI. AQI can give complete description of the air quality. Regression algorithms are the most usual supervised machine learning algorithms that were used in predicting the AQI.

3.1.1 Regression Models

For predicting the AQI, both linear regression and non linear regression models have been used. Linear regression model attempts to establish a relation between the independent variables and dependent variable. The common usage of linear regression model is to know about the linear relationship between predictors and a predictand.

P Arulmozhivarman et al. (2017) presented multiple regression models for predicting AQI [4]. They implemented regression models like SVR and multiple linear regression model for forecasting AQI. Among those, SVR exhibited high level of performance. They considered statistical criteria like MAE, Mean absolute percentage error (MAPE), Correlation coefficient (R), RMSE, and Index of agreement (IA) for evaluating the performance of regression models. In our study we considered some of those statistical metrics for evaluating the performance.

Huixiang Liu et al. (2019) presented a research paper which used regression models for the prediction of air quality [11]. They implemented machine learning models such as SVR and Random forest regression (RFR) for the forecasting of AQI. Between the two models, RFR model performed better than SVR model because the time complexity of SVR has increased cubically with the increasing number of samples. They used performance metrics like RMSE, correlation coefficient (R), and coefficient of determination (R²). In our research, we have considered some the above mentioned metrics for performance evaluation.

In the research done by Soubhik Mahanta et al. (2019), AQI is predicted using

various regression models in the sklearn library [13]. They used models like Linear regression, Neural network regression, LASSO regression, ElasticNet, Decision forest, Extra trees, Boosted decision trees, XGBoost, KNN, and Ridge regression for the prediction of AQI. Out of all the regression models, Extra trees model had the highest accuracy and the least RMSE. They used statistical criteria like Accuracy and RMSE. In our research, we used RMSE as one of our performance metrics for evaluation.

Anikender Kumar and Pramila Goyal (2011) has presented a research paper in which Air quality of Delhi is forecasted [6]. They used regression models like Principal Component Regressor (PCR) and Multiple Linear Regressor (MLR) to predict the AQI. Their study have been made for four seasons namely, Summer, monsoon, post-monsoon, and winter. The statistical analysis of the model showed that it performed well in winter. Co-linearity, which has been found in MLR, has been eliminated with the use of Principal Components (PC). It was also found that the performance of PCR was better than MLR. They used Normalised Mean Square Error (NMSE), RMSE, and coefficient of determination (R^2) as the performance metrics for evaluation.

Mauro Castelli et al. (2020) presented a research paper in which they used SVR to predict the air quality in California [1]. The study is about the use of SVR to accurately calculate the AQI and the concentrations of the pollutants. The study has produced a highly suitable model for the hourly prediction of AQI, accurate concentrations of pollutants such as o_3 , co_2 , and so_2 , and the hourly atmospheric pollution in the area. They used MAE, Normalised Mean Square Error (NMSE), and RMSE as the performance evaluation metrics.

Bing-chun liu et al. (2017) presented a research paper in which they used multi-dimensional collaborative SVR model [10]. In this paper the data used was collaborative multiple city air quality data. As the data used was from more than one city, the training is complex that leads to increase in training time. The RMSE values for the training and testing data sets were less than 12, and they concluded that SVR is strong and is an efficient model for the prediction of AQI. They used the performance metrics such as RMSE and Mean absolute percentage error (MAPE).

Jasleen Kaur Sethi Mamta Mittal (2021) presented a research paper in which they used a feature selection method named Correlation based Adaptive LASSO regression method for air quality index prediction [20]. In this study the model evaluation depicts that the feature subset extracted by proposed model performs better than subset extracted by LASSO with an average classification accuracy of 70 percent.

Chenchen Li et al. (2021) presented a research paper in which they used ridge regression, k nearest neighbour regression, decision tree regression, random forest regression and gradient boosting regression and other supervised machine learning algorithms for predicting the air quality index [7]. They concluded that random forest regression and gradient boosting regression performed better in predicting the air quality index.

We've chosen Literature review and experimentation as research methods to answer the research questions. In the literature review, we've studied various supervised machine learning algorithms and identified predictive algorithms for forecasting AQI. By performing literature review, we analysed how well machine learning helps in forecasting AQI. We've collected data from central pollution control board which is the official portal of the government of India. Experimentation is done using the collected data and identified algorithms from the literature review.

4.1 Literature review

To analyze how machine learning helps in forecasting AQI and to identify the predictive algorithms for forecasting AQI we've used Literature review research method.

- Firstly, we've identified keywords like air quality index, forecasting, prediction, supervised machine learning algorithms, machine learning which are mainly related to our thesis.
- Using the identified keywords, we searched for the previously done researches related to prediction of AQI in google scholar, IEEE, etc.
- Some research works related to AQI prediction were gathered from the search.
- Important research works were filtered by considering inclusion criteria and exclusion criteria.

Inclusion criteria:

1. Include studies that are related to both prediction of AQI and machine learning.
2. Peer assessment of the articles linked to AQI prediction.
3. Include the articles written in english.
4. Only published papers should be included.

Exclusion criteria:

1. The gathered works which are not scientific are excluded.
2. The works which don't follow proper guidelines like without having abstract are excluded.

4.2 Experimentation

We've performed experimentation using "Air quality data in India (2015 - 2020)" data set and algorithms identified from literature review. The following stages were involved in experimentation.

- Collecting the data set from online resource kaggle.
- Preprocessing the data set.
- Build models using preprocessed data set and algorithms identified during literature review.
- Testing the models.
- Evaluating the performance of the models by considering performance metrics.

4.2.1 Environment

To implement experimentation we've used following technologies. A brief description of used technologies along with version is represented below.

- Python (Version 3.10.4) - Python is an interpreted, high level and object oriented programming language. It is an open source.
- Anaconda navigator (Version 2.1.1) - It is a graphical user interface (GUI) that allows to launch applications easily and manage packages.
- Jupyter notebook (Version 4.8) - It is a web based interactive computing platform. It helps in developing, documenting, and executing code.
- Pandas (Version 1.4.2) - It is a free open source python library. It is mainly used for data analysis. It helps to perform various data manipulation operations.
- Numpy (Version 1.22.3) - It is a python library which is used for scientific computing in python. It is used to perform wide mathematical operations on data.
- Matplotlib (Version 3.5.2) - It is a Python package used to create static, animated, and interactive visualizations.
- Seaborn (Version 0.11.2) - It is a python library which is used for data visualization. It is based on matplotlib library. It helps to make statistical graphics using python.
- Scikit-learn (sklearn) (Version 1.0.2) - It is a python library which is used for machine learning. It is largely written in python. It consists of many machine learning algorithms. It is best suitable for predictive analysis.

4.2.2 Data

We've collected data set from Central pollution control board (CPCB) official website (<https://cpcb.nic.in/>). It is government of India's official portal. The data set contains air quality data and AQI of various cities in India. The data set contains 29531 rows, 16 columns and the size of data set is 76 MB. The data set contains following variables particulate matter 2.5, particulate matter 10, nitric oxide, nitric dioxide, nitric x-oxide, ammonia, carbon monoxide, sulphur dioxide. The sample of "air quality data in India (2015 - 2020)" data set is shown below.

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN

Figure 4.1: Sample of air quality data in India (2015 - 2020)

4.2.3 Data preprocessing

Initially the data set contains noisy, inconsistent data and missing values. The data has to be preprocessed to remove the unwanted data and to make the data useful. Data preprocessing helps to transform data into useful format. The following steps were involved in data preprocessing.

- 1) Data cleaning
- 2) Data reduction
- 3) Data transformation

Data cleaning: Data cleaning is the process of removing unwanted data like incorrect data, duplicate data, unformatted data from the data set. By cleaning the data, we can improve the accuracy of the result. The following steps were involved while cleaning the data.

- **Remove duplicates:** When the data is collected from different sources, there will be a high possibility for data to have duplicated entries. These duplicates will create confusion in results. It is advisable to remove those duplicates to improve our results.
- **Remove irrelevant data:** Performing analysis on irrelevant data slows down the process as it wasn't useful. For example, if we only need particulate matter concentration for analysis then we've to exclude other components as they're irrelevant to our analysis to save time.
- **Handling missing values:** We can handle missing data by removing the entire tuple of data or else by filling the missing values in it. We can place

approximate value in the missing field. If the data is too large then we can remove the tuple data that has missing values.

- **Clear formatting:** If the data is formatted heavily, machine learning models can't process the information. If we have different formats in our data it will be confusing.
- **Convert data types:** Sometimes numbers will be inputted as text. Then the data type of those numbers will be string. As they were strings, we can't perform mathematical operations on them. So we have to convert the data types to perform operations on them.

Data transformation: Data is transformed to improve its structure. It enables better data driven decision making. It is the process of changing the format, structure (or) values of data. It involves normalization, attribute selection, discretization, concept hierarchy generation.

Data reduction: When working with huge amounts of data, analysis becomes more challenging. To handle this we use data reduction technique. Data reduction focuses on enhancing the storage efficiency while minimizing data storage and analysis cost. The various steps involved in data reduction were data cube aggregation, attribute subset selection, numerosity reduction, dimensionality reduction.

4.2.4 Training and testing the model

Firstly, we've collected the data set from kaggle. Then we performed data preprocessing techniques like data cleaning, data reduction on the data. At first the data related to the city Delhi is extracted. After that all the tuples which has missing values were ignored. Further data cleaning is performed for better and accurate results. Then data is split into two parameters which contains air particles composition in one parameter and AQI data in other parameter. We've considered PM 2.5, PM 10 and NO2 variables for predicting the AQI variable. We've split the data into 70% training data and 30% testing data. We've used matplotlib library to visualize how well the predicted values matches the actual values using scatter plot graphs and plot graphs.

The machine learning algorithms which were identified from the literature review such as linear regression, LASSO regression, ridge regression, and SVR algorithms were used to build models for prediction. The same data which was preprocessed before is used to train the model with all those identified algorithms. At first a model is built using linear regression algorithm followed by LASSO regression, ridge regression and SVR algorithm. Building machine learning model involve analyzing data to identify patterns and make predictions. Then the built models will be tested by calculating the performance metrics. These performance metrics will test how well the model can predict the AQI.

4.2.5 Performance metrics

To analyze the performance of a machine learning model we need some metrics. These metrics are statistical criteria that can be used to measure and monitor the performance of a model. As our thesis deals with prediction, we've considered MAE and RMSE as the performance metrics.

Mean absolute error (MAE): MAE is the arithmetic average of the difference between the ground truth and the predicted values. It can also be defined as measure of errors between paired observations expressing same phenomenon. It tells us how far the predictions differed from the actual result. Mathematical representation for MAE is given below.

$$MAE = \frac{1}{N} \sum_{j=1}^N |y_j - \bar{y}_j| \quad (4.1)$$

Where,

y_j = Prediction

\bar{y}_j = True value

N = Total number of data points

R squared (R^2): R square performance metric indicates how well predicted values matches actual values. To compute R squared value, we can use the `r2_score` function of `sklearn.metrics`.

$$R^2 = 1 - \frac{\text{sumsquaredregression}(SSR)}{\text{totalsumofsquares}(SST)} \quad (4.2)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4.3)$$

Root mean square error (RMSE): RMSE is the square root of the average of the squared difference between the target value and the value predicted by the model. It is square root of mean square error (MSE). The implementation is very much similar to MSE.

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (y_j - \bar{y}_j)^2} \quad (4.4)$$

Where,

\bar{y}_j = True value

N = Total number of data points

The machine learning models are validated by comparing the performance metrics. The lower the MAE, RMSE and higher the r-squared, the machine learning model performs better.

Chapter 5

Results and Analysis

This section contains the outcomes obtained by performing literature review and experimentation.

5.1 Literature review outcome

To understand and summarize how well machine learning helps in reducing air pollution and forecasting the AQI. To identify the best supervised machine learning algorithms to predict the AQI. We performed literature review and collected some previous research papers that helped us to achieve the above. The research papers were presented below.

No.	Title	Observation
1	Machine Learning - Based Prediction of Air Quality	In this research the machine learning algorithms like SVM, Regression, ANN were employed to predict the AQI. R-square, RMSE, MAE were the performance metrics considered to evaluate the performance of the model [9].
2	AQI and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms	In this research the regression models were built to forecast the AQI. In regression models, they used SVR and random forest regression algorithms. They considered R-squared and RMSE for performance evaluation [11].
3	AQI Prediction using Machine Learning - A Review	In this research paper, importance of machine learning in calculating AQI is explained and machine learning algorithms like linear regression, decision tree, random forest, artificial neural network, support vector machine were implemented to forecast AQI. Accuracy is considered to find the best fit model for the data [12].

Table 5.1: Literature review

No.	Title	Observation
4	Forecasting AQI using regression models: A case study on Delhi and Houston	In this research, SVR and linear models like multiple linear regression consisting of gradient descent, stochastic gradient descent, mini-batch gradient descent were implemented to predict the AQI. Among these models, SVR exhibited high performance in terms of investigated measures of quality [4].
5	Urban Air Quality Prediction Using Regression Analysis	Based on pollution and meteorological data in New Delhi, India, this paper explored how successful several available prediction models are in predicting AQI values. They performed regression analysis on the data, and the results revealed which meteorological parameters had the most impact on AQI levels and how useful predictive models are for air quality forecasting [13].
6	An efficient correlation based adaptive LASSO regression method for AQI prediction	In this research, LASSO regression model was implemented to predict the AQI. In this research the model assessment demonstrates that the suggested method's feature subset performs better than the entire dataset and the subset extracted by LASSO Regression, with an average classification accuracy of 78% [21].
7	Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine	In this research, linear regression and support vector machine algorithms were used to train a predictive model for forecasting AQI. They concluded that SVR has higher accuracy in predicting the AQI [2].
8	AQI Forecasting using Auto-regression Models	This research focused to develop a daily AQI forecasting model that can be applied to local and regional air quality management. For developing these models they used auto regression techniques [24].
9	Air Quality Prediction Of Data Log By Machine Learning	In this research, a machine learning model is developed which helps to predict the future data of pollutants using machine learning techniques. By predicting the future data the model can predict the future pollution levels [16].

Table 5.2: Literature review

No.	Title	Observation
10	Estimating the Impact of Urbanization on Air Quality in China Using Spatial Regression Models	In this research, regression models were used to find out the impact caused by urbanization on air quality. They were inversely proportional to each other because urbanization leads to air pollution [3].
11	Analyzing the impact of heating emissions on AQI based on principal component regression	In this paper a new method was proposed for investigating the quantitative impact of the heating emissions on the AQI and its assumptions and accuracy were verified as well [8].
12	Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review	In this research, Regression, Neural networks and Support vector machine algorithms were reviewed which were employed to monitor the outdoor air quality [19].

Table 5.3: Literature review

By conducting the literature review and reading the above research papers we understood how machine learning helps in reducing the air pollution and calculating the AQI. By performing literature study on the above research papers, we identified that regression algorithms were the most usual supervised machine learning algorithms that are used to predict the AQI.

5.2 Experimentation outcome

This section shows the analysis of the experiment with the preprocessed data set as well as the identified machine learning algorithms from the literature review. For the same collection of air quality data, Linear regression, LASSO regression, Ridge regression, and SVR are used to build the model individually. For experimentation purpose, we considered PM 2.5, PM 10, and NO2 attributes as our input in order to obtain the AQI attribute which is the final output.

5.2.1 Linear Regression

The data is preprocessed and trained with linear regression algorithm to predict the AQI. The figure 5.1 shows how the linear regression model is configured.

```
: from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=101)
from sklearn.linear_model import LinearRegression
linreg = LinearRegression()
linreg.fit(X_train,y_train)

: LinearRegression()

: ypred = linreg.predict(X_test)
linreg.intercept_

: 23.72962438322378

: linreg.coef_

: array([1.12885225, 0.46842454, 0.28577763])

: plt.scatter(y_test,ypred)
```

Figure 5.1: Prediction of air quality index using linear regression

The figure 5.2 is a scatter plot graph. Here X-axis and Y-axis are observed AQI value and predicted AQI value respectively.

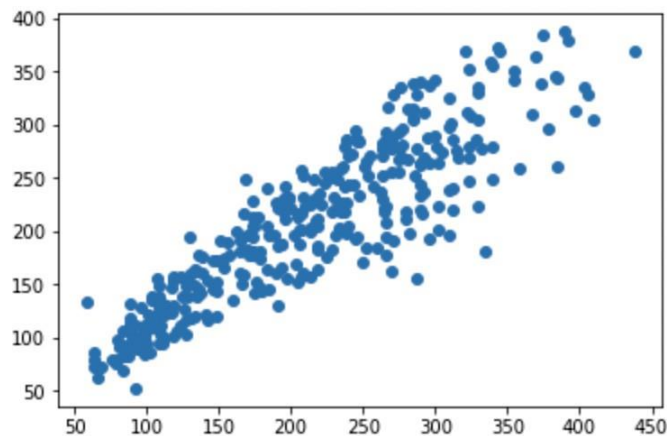


Figure 5.2: Scatter plot graph for linear regression model

The figure 5.3 shows the calculated performance metrics.

```
: from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

: print('Mean absolute error : {}'.format(mean_absolute_error(ypred,y_test)))
: print('Root Mean square error : {}'.format(np.sqrt(mean_squared_error(ypred,y_test))))
: print('R-squared : {}'.format(r2_score(ypred,y_test)))

Mean absolute error : 29.227562900959548
Root Mean square error : 38.485034202614074
R-squared : 0.7441403375729978
```

Figure 5.3: Performance metrics (Linear regression)

MAE	RMSE	R-square
29.227	38.485	0.7441

Table 5.4: Comparison of performance metrics for Linear regression model

The 5.13 is a graph which shows how predicted values using linear regression algorithm were much similar to observed values. Here X-axis and Y-axis are Time (days) and AQI value respectively.

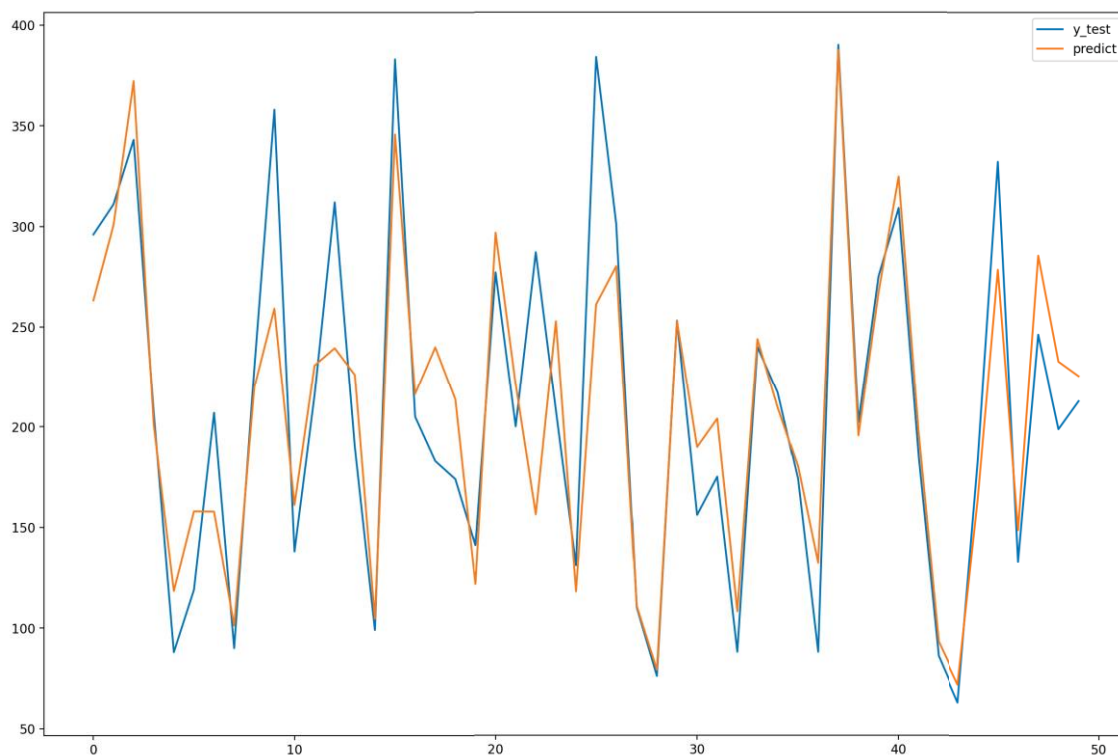


Figure 5.4: Predicted vs observed (Linear regression)

5.2.2 Lasso Regression

The data is preprocessed and trained with LASSO regression algorithm to predict the AQI. The 5.5 shows how the LASSO regression model was configured.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=101)
from sklearn.linear_model import Lasso

model = Lasso(alpha=0.1)
model.fit(X_train, y_train)

Lasso(alpha=0.1)
```

Figure 5.5: Prediction of air quality index using lasso regression

The figure 5.6 shows the calculated performance metrics.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

pred = model.predict(X_train)

print('Mean absolute error: {}'.format(
    mean_absolute_error(y_train, (pred))))
print('Root mean square error: {}'.format(
    np.sqrt(mean_squared_error(y_train, (pred)))))
print('R-squared: {}'.format(
    r2_score(y_train, (pred))))

Mean absolute error: 27.908180955437096
Root mean square error: 36.79150156134114
R-squared: 0.8089727838403192
```

Figure 5.6: Performance metrics (Lasso regression)

MAE	RMSE	R-squared
27.908	36.791	0.8089

Table 5.5: Comparison of performance metrics for lasso regression

The 5.7 is a graph which shows how predicted values using LASSO regression algorithm were much similar to observed values. Here X-axis and Y-axis are Time (days) and AQI value respectively.

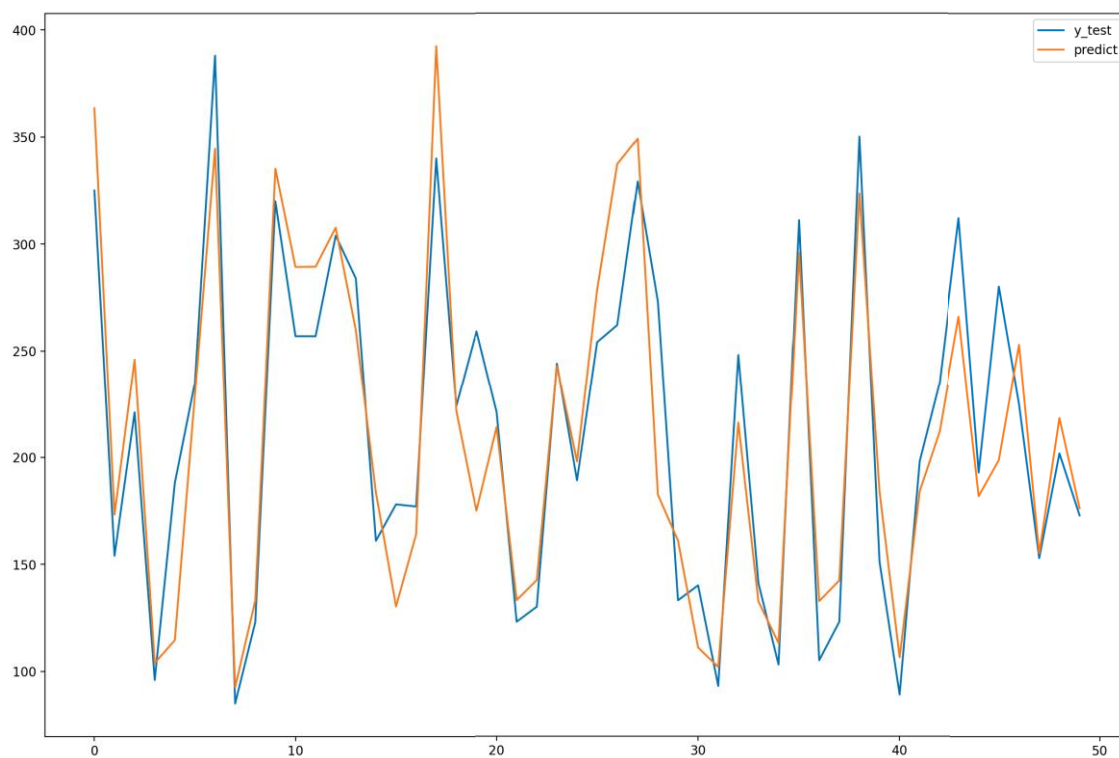


Figure 5.7: Predicted vs observed (Lasso regression)

5.2.3 Ridge Regression

The data is preprocessed and trained with ridge regression algorithm to predict the AQI. The figure 5.8 shows how the ridge regression model was configured.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=101)
from sklearn.linear_model import Ridge

model = Ridge(alpha = 1)
model.fit(X_train, y_train)

Ridge(alpha=1)
```

Figure 5.8: Prediction of air quality index using ridge regression

The figure 5.9 shows the calculated performance metrics.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

pred = model.predict(X_train)

print('Mean absolute error: {}'.format(
    mean_absolute_error(y_train, (pred))))
print('Root mean square error: {}'.format(
    np.sqrt(mean_squared_error(y_train, (pred)))))
print('R-squared: {}'.format(
    r2_score(y_train, (pred))))

Mean absolute error: 27.907866715355404
Root mean square error: 36.79150046674403
R-squared: 0.808972795206957
```

Figure 5.9: Performance metrics (Ridge regression)

MAE	RMSE	R-squared
27.907	36.791	0.8089

Table 5.6: Comparison of performance metrics for ridge regression

The 5.10 is a graph which shows how predicted values using ridge regression algorithm were much similar to observed values. Here X-axis and Y-axis are Time (days) and AQI value respectively.

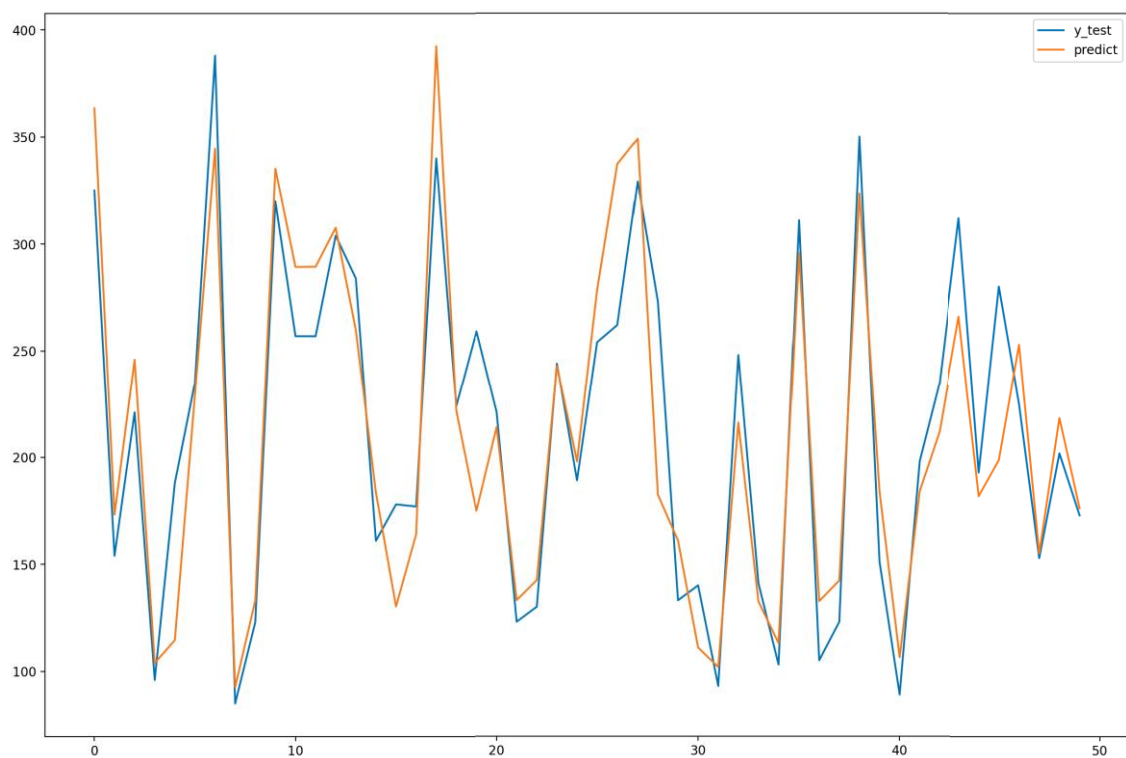


Figure 5.10: Predicted vs Observed (Ridge regression)

5.2.4 Support vector regression

The data is preprocessed and trained with SVR algorithm to predict the AQI. The 5.11 shows how the SVR model was configured.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=101)
from sklearn.preprocessing import StandardScaler
Scaler = StandardScaler()
scaled_X_train = Scaler.fit_transform(X_train)
scaled_X_test = Scaler.transform(X_test)
from sklearn.svm import SVR
param_grid = {'C':[0.001,0.01,0.1,0.5,1],
              'kernel':['linear','rbf','poly'],
              'gamma':['scale','auto'],
              'degree':[2,3,4],
              'epsilon':[0,0.01,0.1,0.5,1,2]}
from sklearn.model_selection import GridSearchCV
svr = SVR()
grid = GridSearchCV(svr,param_grid=param_grid)
grid.fit(scaled_X_train,y_train)

GridSearchCV(estimator=SVR(),
              param_grid={'C': [0.001, 0.01, 0.1, 0.5, 1], 'degree': [2, 3, 4],
                           'epsilon': [0, 0.01, 0.1, 0.5, 1, 2],
                           'gamma': ['scale', 'auto'],
                           'kernel': ['linear', 'rbf', 'poly']})

grid.best_params_

{'C': 1, 'degree': 2, 'epsilon': 2, 'gamma': 'scale', 'kernel': 'linear'}

model_predict=grid.predict(scaled_X_test)
```

Figure 5.11: Prediction of air quality index using support vector regression

The 5.12 shows the calculated performance metrics. MAE, RMSE, R-squared were calculated.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

print('Mean absolute error : {}'.format(mean_absolute_error(model_predict,y_test)))
print('Root Mean square error : {}'.format(np.sqrt(mean_squared_error(model_predict,y_test))))
print('R-squared : {}'.format(r2_score(model_predict,y_test)))

Mean absolute error : 29.82830898853934
Root Mean square error : 41.33061093517168
R-squared : 0.6814468969375063
```

Figure 5.12: Performance metrics (SVR)

MAE	RMSE	R-square
29.828	41.330	0.6814

Table 5.7: Comparison of performance metrics for SVR model

The 5.13 is a graph which shows how predicted values using SVR algorithm were much similar to observed values. Here X-axis and Y-axis are Time (days) and AQI value respectively.

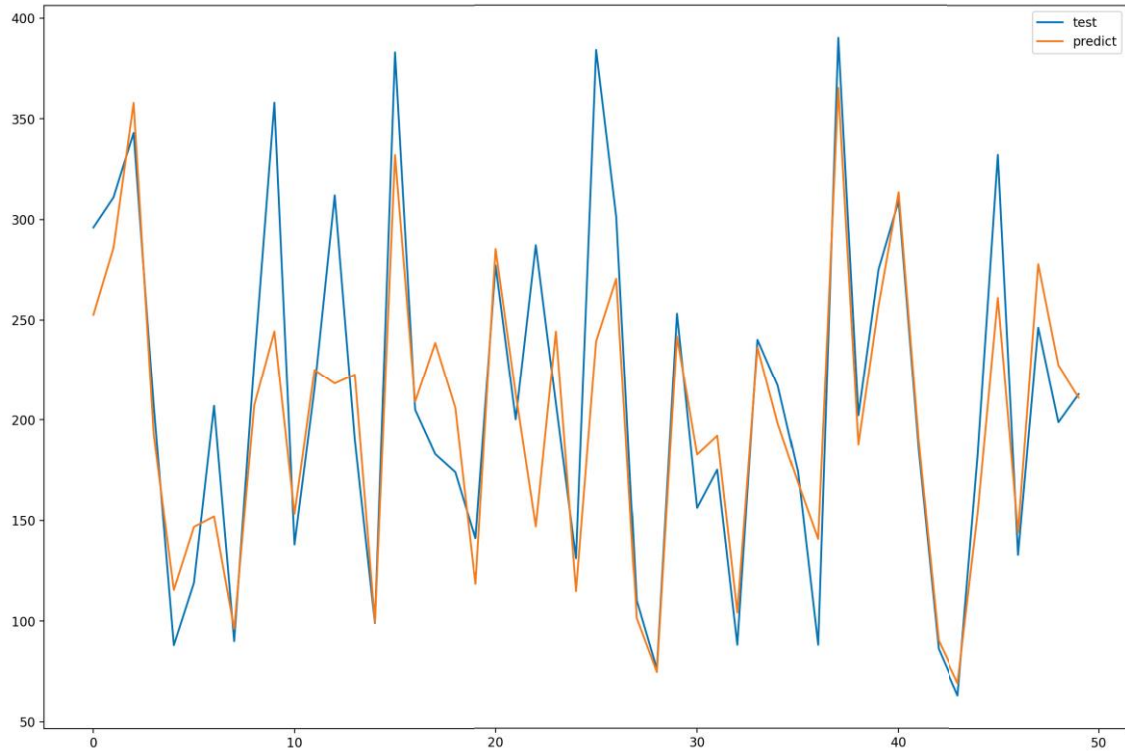


Figure 5.13: Predicted values vs Test values (SVR)

5.2.5 Comparison of results

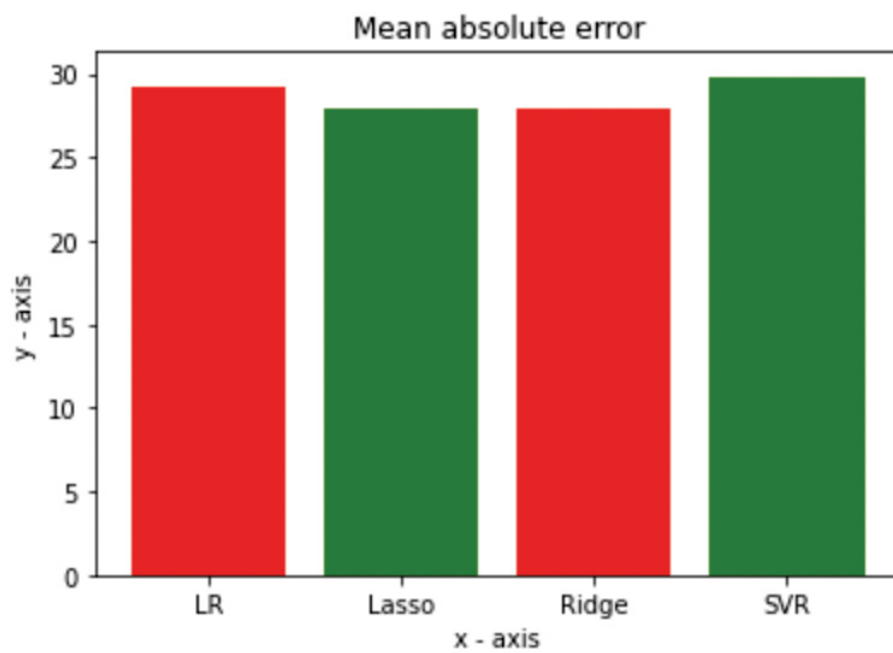


Figure 5.14: Mean absolute error bar chart

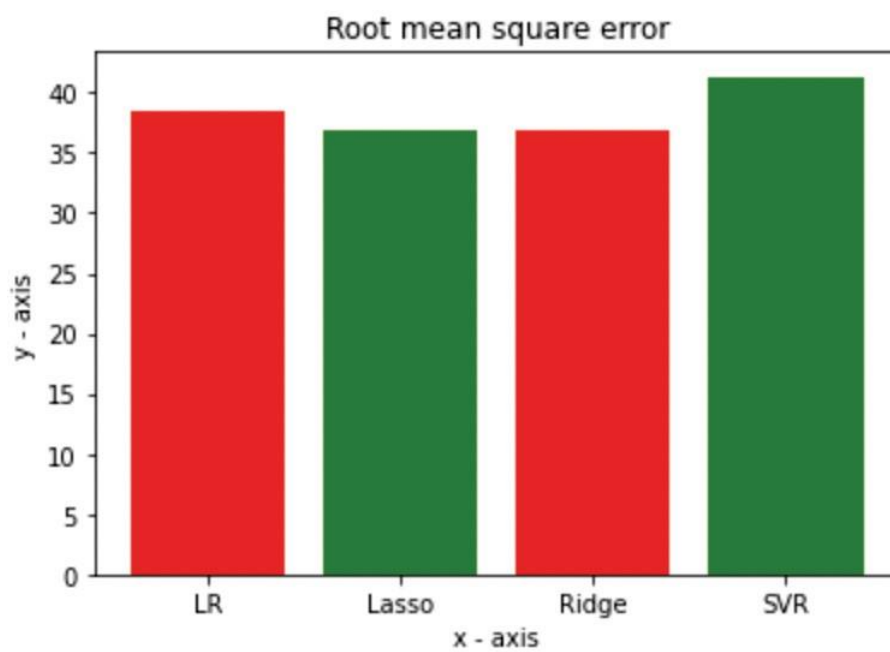


Figure 5.15: Root mean square error bar chart

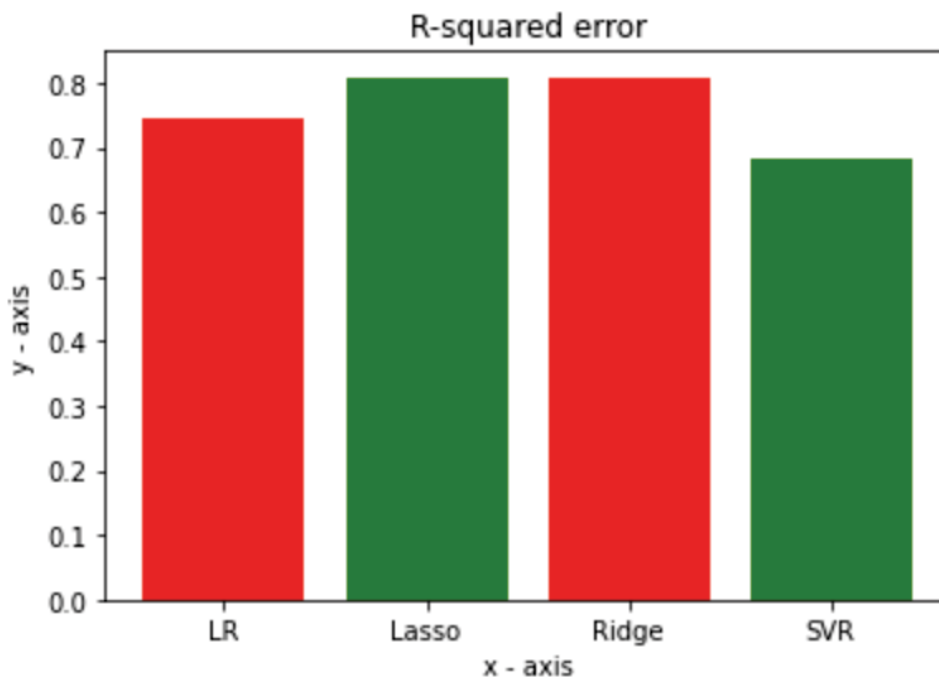


Figure 5.16: R-square error bar chart

In the above bar graphs X-axis represents the algorithms and Y-axis represents performance metrics.

Model	MAE	RMSE	R-square
Linear regression	29.227	38.485	0.7441
Lasso regression	27.908	36.791	0.8089
Ridge regression	27.907	36.791	0.8089
Support vector regression	29.828	41.330	0.6814

Table 5.8: Comparison of performance metrics for all models

By observing the table 5.8, when compared to all algorithms the model has lower MAE, lower RMSE and higher r-squared error when built using ridge regression algorithm. This indicates that **ridge regression model** is good for forecasting the AQI. However ridge regression and LASSO regression has same RMSE, r-squared error and almost same MAE indicating that these two models performs well in predicting the AQI in this case. Finally, by observing performance metrics for all the models, we can say that among all regression models ridge regression model and LASSO regression model performs well in forecasting AQI for given data.

Chapter 6

Discussion

Table 5.1 represents the outcomes of the literature review of our research. Many papers and articles have used machine learning algorithms for the prediction of AQI. The articles were found by searching Google Scholar with some keywords such as Machine learning, Air Quality Index, Prediction of AQI using regression, Regression analysis, SVR and so on. All the important highlights related to our research are noted down that consists of the use of classification in the prediction. In our literature review, we have found that there are various algorithms that are used for the prediction of AQI such as Linear Regression, logistic regression, decision tree, random forest, XGBoost, KNN, Ridge regression, LASSO regression, SVR and Artificial neural networks (ANN). Of all the machine algorithms we chose to use Linear regression, ridge regression, LASSO regression and SVR because these four have the higher accuracy than others. Therefore, we wanted to find out the most accurate among the accurate.

Table 5.8 shows the results of experimentation. We found that the model built with Ridge regression algorithm has the highest R-square and least MAE and RMSE. SVR got the least performance with the least R-square and highest MAE and RMSE. Although LASSO regression and ridge regression got the exact RMSE and R-square, ridge regression has lower Ridge regression, therefore it achieves the highest performance among the four. Therefore, based on the results of experimentation, ridge regression can be called as the most efficient algorithm in the prediction of AQI whereas the SVR has the least performance and is not much suitable for the prediction of AQI.

SVR has lower performance when compared with LASSO regression, ridge regression and linear regression. The lower performance of SVR model was possibly due to lack of enough training data. Too little training data leads to poor performance.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

In this study, we've conducted literature review and identified some machine learning algorithms to predict the AQI. We identified linear regression, LASSO regression, ridge regression and SVR algorithm from the literature review. We've preprocessed the data and successfully trained the linear regression, LASSO regression, ridge regression, SVR algorithms. Same set of data is used to build every model. In this study we've considered MAE, RMSE and r-square error to evaluate the performance of the models. We can conclude that the models ridge regression and LASSO regression have shown better performance with lower MAE, root mean squared error and higher r-squared.

7.2 Future work

In this study the data used was static that means the data will be fixed and it remains the same after it's collected. However the government updates the data hourly. Further we can use real-time data analysis using cloud to obtain better outcomes for greater performance as the data updates for every particular interval of time. We can further ensemble two or more machine learning algorithms and process large data to get more accurate results.

References

- [1] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, “A machine learning approach to predict air quality in california,” *Complexity*, vol. 2020, 2020.
- [2] M. Dun, Z. Xu, Y. Chen, and L. Wu, “Short-term air quality prediction based on fractional grey linear regression and support vector machine,” *Mathematical problems in engineering*, vol. 2020, 2020.
- [3] C. Fang, H. Liu, G. Li, D. Sun, and Z. Miao, “Estimating the impact of urbanization on air quality in china using spatial regression models,” *Sustainability*, vol. 7, no. 11, pp. 15 570–15 592, 2015.
- [4] S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, “Forecasting air quality index using regression models: A case study on delhi and houston,” in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017, pp. 248–254.
- [5] Z. Ghahramani, “Unsupervised learning,” in *Summer school on machine learning*. Springer, 2003, pp. 72–112.
- [6] A. Kumar and P. Goyal, “Forecasting of air quality in delhi using principal component regression technique,” *Atmospheric Pollution Research*, vol. 2, no. 4, pp. 436–444, 2011.
- [7] C. Li, Y. Li, and Y. Bao, “Research on air quality prediction based on machine learning,” in *2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, 2021, pp. 77–81.
- [8] H. Li, S. You, H. Zhang, W. Zheng, W.-I. Lee, T. Ye, and L. Zou, “Analyzing the impact of heating emissions on air quality index based on principal component regression,” *Journal of cleaner production*, vol. 171, pp. 1577–1592, 2018.
- [9] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, “Machine learning-based prediction of air quality,” *Applied Sciences*, vol. 10, no. 24, p. 9151, 2020.
- [10] B.-C. Liu, A. Binaykia, P.-C. Chang, M. K. Tiwari, and C.-C. Tsao, “Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang,” *PloS one*, vol. 12, no. 7, p. e0179763, 2017.
- [11] H. Liu, Q. Li, D. Yu, and Y. Gu, “Air quality index and air pollutant concentration prediction based on machine learning algorithms,” *Applied Sciences*, vol. 9, no. 19, p. 4069, 2019.

- [12] T. Madan, S. Sagar, and D. Virmani, "Air quality prediction using machine learning algorithms –a review," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 140–145.
- [13] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor, "Urban air quality prediction using regression analysis," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 1118–1123.
- [14] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, 2020.
- [15] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.
- [16] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, "Air quality prediction of data log by machine learning," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 1395–1399.
- [17] A. Plaia and M. Ruggieri, "Air quality indices: a review," *Reviews in Environmental Science and Bio/Technology*, vol. 10, no. 2, pp. 165–179, 2011.
- [18] J. Ranstam and J. Cook, "Lasso regression," *Journal of British Surgery*, vol. 105, no. 10, pp. 1348–1348, 2018.
- [19] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approaches for outdoor air quality modelling: A systematic review," *Applied Sciences*, vol. 8, no. 12, p. 2570, 2018.
- [20] J. K. Sethi and M. Mittal, "An efficient correlation based adaptive LASSO regression method for air quality index prediction," *Earth Science Informatics*, vol. 14, no. 4, pp. 1777–1786, Dec. 2021. [Online]. Available: <https://doi.org/10.1007/s12145-021-00618-1>
- [21] —, "An efficient correlation based adaptive lasso regression method for air quality index prediction," *Earth Science Informatics*, vol. 14, no. 4, pp. 1777–1786, 2021.
- [22] M. Somvanshi, P. Chavan, S. Tambade, and S. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *2016 international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2016, pp. 1–7.
- [23] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.
- [24] N. Tomar, D. Patel, and A. Jain, "Air quality index forecasting using auto-regression models," in *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2020, pp. 1–5.
- [25] W. Wang and Z. Xu, "A heuristic training for support vector regression," *Neurocomputing*, vol. 61, pp. 259–275, 2004.

- [26] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

ICAAMT 2023

International Conference on Advances in ADDITIVE MANUFACTURING TECHNOLOGIES

CERTIFICATE

This is to certify that

HARINI. R.A


from PANIMALAR ENGINEERING COLLEGE


has participated / presented a paper entitled COMPARATIVE


ANALYSIS OF BREAST CANCER USING

MACHINE LEARNING

in the International Conference on ADVANCES IN ADDITIVE
MANUFACTURING TECHNOLOGIES (ICAAMT 2023) organised
by the Department of Mechanical Engineering, Chennai Institute of
Technology from 27th to 29th November, 2023.


Dr. P. Gurusamy
Convener


Dr. M. D. Vijayakumar
Co-Convener


Dr. A. Ramesh
Principal

INTERNATIONAL CONFERENCE ON
ADVANCES IN ADDITIVE MANUFACTURING TECHNOLOGIES
(ICAAMT 2023)

ICAAMT 2023 International Conference on Advances in
**ADDITIVE MANUFACTURING
TECHNOLOGIES**

CERTIFICATE

This is to certify that

SHREEMATHI.S

from PANIMALAR ENGINEERING COLLEGE

has participated / presented a paper entitled COMPARATIVE

ANALYSIS OF BREAST CANCER USING

MACHINE LEARNING

in the International Conference on **ADVANCES IN ADDITIVE
MANUFACTURING TECHNOLOGIES (ICAAMT 2023)** organised

by the Department of Mechanical Engineering, Chennai Institute of
Technology from 27th to 29th November, 2023.

Dr.P.Gurusamy
Convener

Dr.M.D.Vijayakumar
Co-Convener

Dr.A.Ramesh
Principal

Comparative Analysis of Breast Cancer using Machine Learning

Harini Ramachandran Aruna^{1,a*}, Shreemathi Santharaman^{2,b},

Dr.Kavitha Prithiviraj^{3,c}, Dr.Chitra Devarajalu^{4,d}

Department of Artificial Intelligence and Data Science, Panimalar Engineering College,
Chennai, Tamil Nadu, India^{1,2,3}

Department of Master of Business Administration, Panimalar Engineering College, Chennai,
Tamil Nadu, India⁴

harinirams212@gmail.com^a, shree0207mathi@gmail.com^b,
drkavitha.ads2021@gmail.com³, chitrambapec@gmail.com^d

Keywords: Breast cancer, Analysis, Machine learning

Abstract: This paper gives us an idea about breast cancer detection. The global battle against cancer, especially in developing nations, necessitates innovative approaches for early detection and treatment. This study conducts a rigorous comparative analysis of breast cancer detection methodologies employing machine learning techniques including KNN, Naive Bayes, SVM, Decision Tree, Logistic Regression, Random Forest, SVM Kernel, XGBoost, and AdaBoost. The accuracy in addition to the efficiency of the trained models are evaluated using performance metrics such the confusion matrix, exactitude, order back, F- score, substructure and preciseness. These performance measures show that Logistic Regression has the greatest accuracy rate (97%). We find distinctive biomarkers and diseases characteristics by combining several algorithms and analyzing various data sources, including genetic factors and clinical records. Our cross-population comparison reveals customized insights necessary for individualized therapies that promise better patient results.

1. Introduction

Millions of people all over the world are afflicted by the widespread and potentially fatal disease of breast cancer. The likelihood of surviving this, the most prevalent cancer in women, is greatly increased by early identification long-term survival and a successful treatment. With the introduction of cutting-edge large volumes of readily accessible medical data, technological advancements, and machine learning. The field of healthcare has seen the emergence of learning approaches as effective instruments for illnesses, such as breast cancer, can be predicted and detected. Breast cancer is possible when the cancer cells are able to spread through the lymphatic or blood systems and the body parts were then transferred. In breast cancer screening, mammograms are the primary means of detection. If an abnormality is detected, other tests, such as a breast ultrasound, MRI, or

biopsy, may be performed to help confirm the diagnosis. Breast cancer is impacted by the type, stage, and individual characteristics treated. The most common types of therapy include immunotherapy, targeted therapy, and hormone therapy. Customized care is provided based on the patient's particular circumstances. Depending on the diagnosis's stage and its type, Breast cancer has a wide range of prognoses. Generally speaking, early-stage cancer is preferable to late-stage cancer. The main goal of the research is to investigate how machine learning algorithms might be applied to the field of breast oncology and how successful medical therapies can be. By leveraging the inherent patterns within a comprehensive dataset, we delve into the realm of predictive analytics to develop a reliable and efficient system for identifying potential cases of breast cancer. This project is not merely a technical endeavour. It represents a significant stride toward more personalized and accurate healthcare solutions.

2. Literature Survey

Using machine learning techniques, Siyabend Turgut et al. published "Microarray Breast Cancer Data Classification" [IEEE 2018]. The classification of the patients in the research is done using machine learning techniques employing microarray breast cancer data. In the first instance, the dataset is subjected to the application of eight various machine learning algorithms, and the classification outcomes are recorded. Then, in the second scenario, some fifty features were chosen as the standard reason after applying two distinct intellection methods, including a popular selection model, Recursive Feature Elimination (RFE) and Randomized Logistic Regression (RLR), to the data file for microarray breast cancer. All over again, the adjusted dataset was subjected to the similar eight expert system algorithms. Comparisons between the classifications' outcomes and those of the first example are made. Techniques including MLP, SVM, KNN, Random Forest, Decision Tree, Logistic Regression, Ad Boost, and Gradient Boosting Machines were employed. The two feature selection techniques were applied, and SVM produced the best outcomes. To study how the count of layers and neurons affects categorization accuracy, several layers and neurons are used when applying MLP. [3].

Varalatchoumy M et al., "Four Novel Approaches for Detection of Region of Interest in Mammograms - A Comparative Study" [ICISS 2017]. This study looks at four novel approaches based on real-time and database images for identifying regions of interest in mammography images. The preprocessing of Approach I used techniques such as dynamic thresholding and histogram equalization. Particle swarm optimization and k-means clustering were used to extract the Region of Interest (ROI) from the preprocessed image. In Approach II, preprocessing was carried out employing distinct morphological processes, such dilation and erosion. ROI was calculated by modifying the 10 watershed segmentation approach. In Approach III, data is preprocessed using an upgraded level set method for data segmentation and histogram equalization. Adaptive histogram equalization, contrast-limited, and other morphological approaches are used to preprocess images in approach IV, which is thought to be the most successful strategy. A highly inventive method was developed in order to discover Regions of Interest. Only the

images from the Mammographic Image Analysis Society (MIAS) database could be utilized to implement Approaches I and II. Real-time hospital pictures and MIAS both benefited from Methods III and IV. The comparison study's multiple graphs clearly demonstrate that radiologists should adopt the novel strategy—which used a novel ROI detection algorithm—to detect cancers in MRI images since it is the most reliable, accurate, and efficient approach. [4].

Ammu P. K. et al., "Review on Feature Selection Techniques of DNA Microarray Data," IJCA 2013, This study analyzes a few key feature selection methods used with microarray data and outlines the benefits and drawbacks of each method. One of the most critical steps in bioinformatics is feature selection from DNA microarray data. Based on the concepts that mutation is a natural process and that species migrate between habitats during their evolutionary journey, biogeography-based optimization, or BBO, is an optimization technique. A search space's movement of particles serves as the foundation for the Particle Swarm Optimization (PSO) algorithm. The selected genes can be made more representative of the population by removing redundant genes using redundancy-based feature selection techniques. An information gain filtering criterion-based two-stage hybrid filter wrapper strategy wherein the first stage generates a subset of the original feature set. Second, the set of filtered genes is processed by the genetic algorithm. Gene selection based on the dependencies between features, where features are divided into 11 dependent, half-dependent, and independent features. Any feature that is independent of any other feature is referred to as such. The significance of half-dependent characteristics increases when combined with other features, whereas dependent features rely only on other features. [5]

Integrating level set approaches with spatial fuzzy clustering for automatic segmentation of medical images, Bing Lan Li et al. Elsevier (2010) In order to facilitate automated medical image segmentation, a new Fuzzy Level Set technique is suggested in this study. It is possible for it to directly develop from the initial segmentation based on spatial fuzzy clustering, where a specified cost function is minimized by adaptively calculating each subclass's centroid and range. Considering the outcomes of fuzzy clustering, governing factors for Level Set evolution are also estimated. Dynamic variational boundaries are used in the level set approaches to segment images. Initializing level set segmentation is automated using the novel fuzzy level set method and parameter setup using spatial fuzzy clustering. To identify the approximate Fuzzy-C mean (FCM) with spatial constraints is used to identify contours of relevance in a medical image. Further, the technique for adding locally regularized evolution improves the Fuzzy Level Set. Improvements like these reinforce segmentation and allow level sets to be adjusted. The suggested strategy was assessed on a variety of medical images from different modalities. Results demonstrate how effective it is for segmenting images. [6].

3. Proposed System



Fig. 1. Architecture Diagram

3.1 Data Collection

The most typical malignancy among women worldwide is breast cancer. Over 2.1 million people were impacted by it in just 2015, accounting for 25% of all cancer cases. When breast cells start to proliferate out of control, the condition begins. These cells typically develop tumors that can be perceived as breast lumps or seen on an X-ray. How to categorize tumors as malignant (cancerous) or benign (non-cancerous) is one of the main obstacles to its diagnosis. Please finish the analysis of identifying these tumors using machine learning (with SVMs) and the Breast Cancer Wisconsin (Diagnostic) Dataset.

3.2 Data Analysis and Preprocessing

The data analysis and pre-processing step involve the exploration and visualization of the data. The first step is to clean the data which includes the process of repairing or eliminating data that is missing, duplicated, corrupted, improperly formatted, or erroneous from a dataset. It covers feature scaling, selecting relevant features, and handling missing and categorical data. In the dataset, "NaN" is used to represent missing values. The next step is to visualize the data. The visualization of the data is done using histogram which is an approach to bar charts that divide continuous measures into distinct bins for distribution analysis and heat maps, which are a kind of geographical visualization represented as a map that show distinct data values as distinct colors. The histogram is represented in the form of count plot which is used to represent the occurrence(counts) of the observation present in the categorical variable.

3.3 Training of the Model

For model training, we use various machine-learning algorithms for the Breast Cancer analysis. Since the output of the model is a class, we are taking the algorithm as a classifier as it helps in categorizing data into different classes. The dataset is divided into training and test sets in order to train this model. The training set is used to assist train the model, and the test set is used to help evaluate the model. The machine learning algorithms used for this analysis are Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, SVM Kernel, Support Vector Machine, XGBoost, AdaBoost, KNN(K-Nearest Neighbour).

3.4 Evaluation and Comparison of the Model

After training the models with machine learning algorithms, we must assess the models' performance using a range of metrics that are detailed in the categorization report, one of a grouping-based machine learning model's enactment estimation measurements. It exhibits your model's exactness, recollection, a function of precision score, and provision. It delivers a appreciation of the complete enactment of our skilled prototypes.

4. Results and Discussion

In this proposed system, we utilised 9 machine learning algorithms for training the model. Also we used six performance measures for the evaluation of the model. Based on these parameters, we will compare the performance of each model and higher accuracy of the analysis is considered as a primary metric for comparison which has the highest accuracy. Based on the comparison, Logistic Regression has the accuracy of 97% and considered as the best model for analysis of breast cancer.

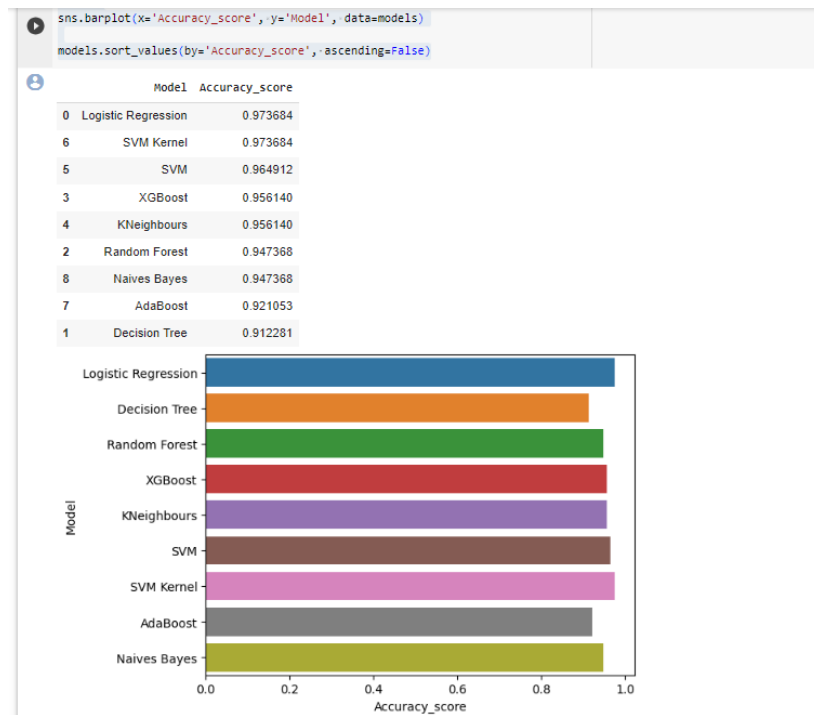


Fig. 2. Comparison of the Models

5. Conclusion

The use of expert system techniques in the comparative analysis of breast cancer has shown to be a useful and promising method in the realm of medical research and healthcare. Machine learning's aptitude for working with a variety of data sources, including genetic, imaging, and clinical data, is one of its main advantages in the analysis of breast cancer. Machine learning algorithms can provide a thorough picture of the condition by combining these different data sources, which will enhance patient outcomes through more effective diagnosis and individualized treatment plans. On the clinical data of patients identified with breast cancer based on symptoms connected to the disease, we employed well known machine learning techniques as KNN, Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, SVM Kernel, XGBoost and AdaBoost. The Logistic Regression shows the best performance in the analysis, with accuracy rate of 97%, according to the performance validation criteria recall, accuracy, precision, and support. The study's future focus may involve the diagnosis of the condition using numerous or sizable data sets.

6. References

- [1] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians. 2005 Jan 1; 55(1):10-30.
- [2] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centres for Disease Control.
- [3] Siyabend Turgut; Mustafa Dağtekin; Tolga Ensari, “Microarray breast cancer data classification using machine learning methods”, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 10.1109/EBBT.2018.8391468.
- [4] M.Ravishankar, Varalatchoumy M, “Four Novel Approaches for Detection of Region of Interest in Mammograms - A Comparative Study”, 2017 International Conference on Intelligent Sustainable Systems (ICISS), 10.1109/ISS1.2017.8389410.
- [5] Ammu P K, Preeja V, “Review on Feature Selection Techniques of DNA Microarray Data”, International Journal of Computer Applications (0975 – 8887) Volume 61– No.12, January 2013.
- [6] Bing Nan Li, Chee Kong Chui, Stephen Chang, S.H. Ong, “Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation”, Computers in Biology and Medicine, Volume 41, Issue 1, January 2011, Pages 1-10, <https://doi.org/10.1016/j.compbimed.2010.10.007>.
- [7] B. N. Dontchos, A. Yala, R. Barzilay, J. Xiang, C. D. Lehman, “External validation of a deep learning model for predicting mammographic breast density in routine clinical practice”, Acad. Radiol., 28 (2020), 475-480, <https://doi.org/10.1016/j.acra.2019.12.012>.
- [8] Siddhant Rao ,”MITOS-RCNN:A novel approach to mitotic figure detection in breast cancer histopathology images using region based convolutional neural networks”, arXiv preprint arXiv:1807.01788 (2018), <https://doi.org/10.48550/arXiv.1807.01788>.
- [9] Péter Bánci, Oscar Geessink, Quirine Manson, Marcory Van Dijk, et al; “Detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge”, IEEE Transactions on Med. Imaging (2018), 10.1109/TMI.2018.2867350 .
- [10] B. N. Dontchos, A. Yala, R. Barzilay, J. Xiang, C. D. Lehman, “External validation of a deep learning model for predicting mammographic breast density in routine clinical practice”, Academic Radiology, Volume 28, Issue 4, April 2021, Pages 475-480, <https://doi.org/10.1016/j.acra.2019.12.012>.



HARINI RAMACHANDRAN <harinirams212@gmail.com>

Your paper has been accepted for publication in International Conference on Advances in Additive Manufacturing Technologies

International Conference on Advances ... <9783036404493@scientific.net>

Mon, Feb 5, 2024 at 11:05 AM

Reply-To: "International Conference on Advances ..." <9783036404493@scientific.net>

To: "Miss Harini R.A" <harinirams212@gmail.com>

Dear Miss Harini R.A,

Your article «Comparative Analysis of Breast Cancer Using Machine Learning» has been preliminary accepted for publication in the «International Conference on Advances in Additive Manufacturing Technologies». The final verification of it will be done by the Publisher - as soon this is done, you will get an additional confirmation for the final acceptance. Although no further action is required, you can verify the status of your article by logging in to the publisher's website :

Please go to <https://www.scientific.net> and log in using the credentials below.

Username : harinirams212@gmail.com

Password : Google1234\$\$

After you log in please select « Author » role near the top of the screen.

If any further changes in your article become necessary you must notify and obtain a permission from an Editor via E-mail PRIOR to uploading a new version.

The final acceptance notice you will receive after publisher's internal check. Thank you very much.

Best regards,
Gurusamy Pathinettampadian
gurusamp@citchennai.net