Some improved Forecasting Models using Data Mining

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

In

MATHEMATICS

By

MINAKSHI SHARMA

Registration Number: 41900720

Supervised By

Prof. (Dr.) A. K. Awasthi (25155)

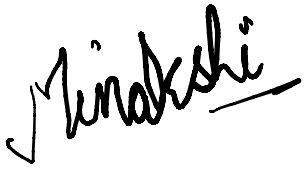
Professor, Department of Mathematics

School of Chemical Engineering and Physical Sciences

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB
2023

DECLARATION

I, hereby declared that the presented work in the thesis entitled **"Some improved Forecasting Models using Data mining"** in fulfilment of degree of **Doctor of Philosophy (Ph.D)** is outcome of research work carried out by me under the supervision **Prof (Dr.) A.K. Awasthi**, working as a **Professor**, in the **School of Chemical Engineering and Physical Sciences** of Lovely Professional University, Punjab, India. In keeping with general practice of reporting scientific observations, due acknowledgements have been made whenever work described here has been based on findings of other investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.



(Signature of Scholar)

Name of the scholar: Minakshi Sharma

Registration No.: 41900720

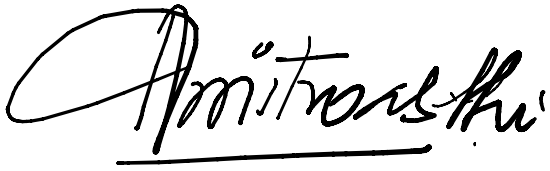
Department/school: Department of Chemical Engineering and Physical Sciences

Lovely Professional University,

Punjab, India

Certificate

This is to certify that the work reported in the Ph.D. thesis entitled **"Some improved Forecasting Models using Data mining"** submitted in fulfilment of the requirement for the award of degree of Doctor of Philosophy (Ph.D.) in the Department of Mathematics, School of Chemical Engineering and Physical Sciences, is a research work carried out by Minakshi Sharma, 41900720 , is bonafide record of her original work carried out under my supervision and that no part of thesis has been submitted for any other degree, diploma or equivalent course.



Name of Supervisor: Prof. A.K. Awasthi

Designation: Professor

Department/school: Department of Mathematics, School of Chemical Engineering and Physical Sciences

University: Lovely Professional University, Punjab, India

Abstract

The research reported in this thesis deals with the forecasting models by using data mining along with their applications in various fields. To predict or estimate future trends based on historical data or observations forecasting is an important concept. In this the technical methods which is based on mathematics and artificial intelligence dealt with the big data for prediction / forecasting and there are many non- technical methods which works like the backbone for machine learning or technical methods based on short term or long term forecasting like Arima, Sarima, Autoregressive methods. The present study provides the widen scope to researcher to take better decision in the most difficult time as it is important and benefitted in many fields like businesses because it allows them to better plan for the future. It helps them to anticipate customer needs, identify potential opportunities and risks, and allocate resources more effectively. It helps companies to make more informed decisions about pricing, product development, marketing strategies and inventory management to maximize profitability. The various forecasting models plays an important role for the prediction. It doesn't show perfectness basically it is always an improved model comparable to the existing model.

This study is dealing with the forecasting using data mining and data mining is to get the knowledge or to extract out the hidden information which could be of any type. Thus, the improved models can be implemented in various fields like agriculture, healthcare, education and many more. In this the main focus is to improve the models to reach towards perfectness in the field of forecasting. Based on the various types of data used for forecasting in field of healthcare, electricity generation, rainfall count of a state, production of crops and price of the crops etc. For forecasting, various machine learning methods and time series methods has been used like regional frequency analysis, regression methods, classification models, and Multifarious grey models along with the combination of linear and nonlinear models. The hybrid model improves the accuracy of forecasting by reducing the error created by individual model. In addition, thesis explores the potential of new machine-learning techniques and their role in improving forecasting models. Some basic forecasting models such as Naïve bayes, moving average, auto-regression, auto-regressive integrated moving average model, etc., are discussed and some of them are applied to different types of time series data set by machine learning. For the analysis of the forecasting used some software's as python 3 in jupyter notebook, SPSS, Weka, Minitab etc., for analysis. As a hybrid model i.e., RF-DT used for the forecasting of

the covid-19 data, Weka software used for data analysis and prediction techniques in healthcare, for the forecasting of crop yield and price of crop Minitab software along with artificial intelligence has applied.

This thesis helps to know which type of knowledge you required or want to draw out by using data mining. There are many techniques and abundant of data but it supports and make researcher's to choose best fit algorithm and methods on various types of data like big data, time series etc.

The various approaches and methods used here can enhance the knowledge by providing a platform to other researchers to make use of these proposed methods for future implications.

Acknowledgment

At the outset, I'd like to express my sincere gratitude to the almighty God and my supervisor, **Prof. (Dr.) A.K. Awasthi**, Professor, School of Chemical Engineering and Physical Sciences, Lovely Professional University, Punjab, for his interest, excellent rapport, untiring cooperation, invaluable advice and encouragement throughout my Ph.D. Degree. Without his unfailing support and belief in me, this thesis would not have been possible. I indeed feel short of words to express my sense of gratitude towards him. Besides being a scholar par eminence, he is a person par excellence a perfect embodiment of dedication, intelligence, and humanity.

I acknowledge with pleasure the cooperation from the esteemed faculty of the University for their Enthusiastic Support and encouragement throughout the research work.

A special thanks to my family. Words cannot express how grateful I am to my in-laws and my parents. I would also like to thank to my husband Neeraj Sharma and my children Sohalia & Reyansh for all of the sacrifices that they've made on behalf of me.

I am indebted to the participants who contributed to this study, as well as the numerous researchers and authors whose work served as a foundation for my research. Their collective efforts have enriched the knowledge based upon which this thesis is built.

I am also grateful to the anonymous reviewers who provided valuable feedback during the thesis review process. Their constructive criticism has undoubtedly improved the quality and rigor of this work.

Finally, I would like to extend my gratitude to all those, whose names may not be mentioned here but have played a part in shaping my academic journey. Each of you has left an indelible mark on my growth as a researcher and as an individual.

To everyone mentioned above and to those unmentioned but not forgotten, thank you for your support, guidance, and belief in me. Your contributions have been pivotal in making this thesis a reality.

Minakshi Sharma

CONTENT

List of Figures

List of Tables

Chapter I

GENERAL INTRODUCTION

"Forecasting is not about predicting the future perfectly, but about minimizing uncertainty and making informed decisions."

1.1 Introduction

Introduction of the forecasting is the main purpose of this chapter. Since the forecast word itself comprises of two words “fore” and “cast” which means before and to tell respectively. Thus, Forecasting is to tell something before it actually happens either to control the damage or to get benefit most of it by proper planning so, there has been so much work done on ways for prediction and hence is the first important motivation for forecasting. People are existing just based on forecasting either by logical forecasting or by making guesses. As people by seeing the clouds can assume it will rain or not also, they get weather data from API providers like Open weather, Aeris Weather, Climacell, Hurricane Mapping etc. to understand how much rainfall they are going to receive and adjust the irrigation accordingly. During the winter, they indulge in other kinds of farming activities, such as raising animals, winter crops, etc. Thus, weather forecast helps farmers in many ways. It helps them to plan things so that crops get enough nutrition with proper watering and fertilizers. The various organizations can also decide on the basis of forecasting that which actions which will obtain the greatest benefits and avoid those actions which don’t provide benefits. There are many ways of forecasting based on the type of situation in which it is to be used and the type of data. All the forecasting methods are good because it depends on the various factors moreover there are factors in the factors on which the accuracy depends. This is not a new concept, even three decades ago the concept of forecasting started in the field of economics and now it left no field uncovered. There are many ways of forecasting based on the problem of prediction as

Forecasting can be used in many areas, including weather forecasting, economics, demand planning, and finance with the process of Data science.

1.2 Aim and Objectives of research in forecasting using data mining:

The main focus of this on the development of various forecasting models for the improvement of exactness in prediction. Based on quadratic modelling, the models combined to form the most suitable model for forecasting.

1. To develop accurate and reliable forecasting models based on data mining techniques.
2. To identify key factors that influences the forecasting accuracy and develops strategies to improve the forecasting accuracy.
3. To identify and analyse the relationships between different variables in the data set and use this information to improve the forecast accuracy.
4. To develop novel data mining techniques for forecasting and evaluate their performance.
5. To analyse the effect of different types of data on forecasting accuracy
6. To develop automated methods for selecting the most appropriate forecasting model for a given set of data.
7. To find out the trends in data that may be applied to predict future outcomes.
8. To develop predictive models that can be used to make accurate forecasts.
9. To develop methods for improving the accuracy of forecasts.
10. To identify potential areas of improvement in data collection and data processing.
11. To develop strategies for more efficient data mining and analysis.
12. To identify underlying relationships between variables that can be used to improve forecasting accuracy.
13. To develop strategies for incorporating new data sources into existing forecasting models.

1.3 Data science and types of data

As the name “Data science” itself comprises of two words data and science. It is the systematic study of data. In this it make experiments with raw or structured data to get the information or knowledge. Today, we can see that everything is depends on data like business, education, entertainment etc. so data is known as the driving component. Expert says that every second generates the quintillion of data from various sources like face book, Whatsapp, YouTube, Flip kart, Amazon and many more websites. Data plays an important role in life through which we can take many decisions and forecasting can be done from this type of data. In this it is observed that to draw out the best knowledge and information, the type of data plays a significant role. So it is mandatory to pre-process the data and make it error free to get desired result. Thus, it can be said that “The interdisciplinary field of machine learning for data visualisation, and predictive analytics by using scientific methods and various, algorithms to extract knowledge from raw data is well known as Data science. It also involves the application of these techniques and algorithms to solve real world problems. Data science employs techniques and theories drawn from a range of disciplines such as mathematics, statistics, machine learning, information sciences, and computer science. The ultimate purpose of data science is to gain insights and knowledge from data.”

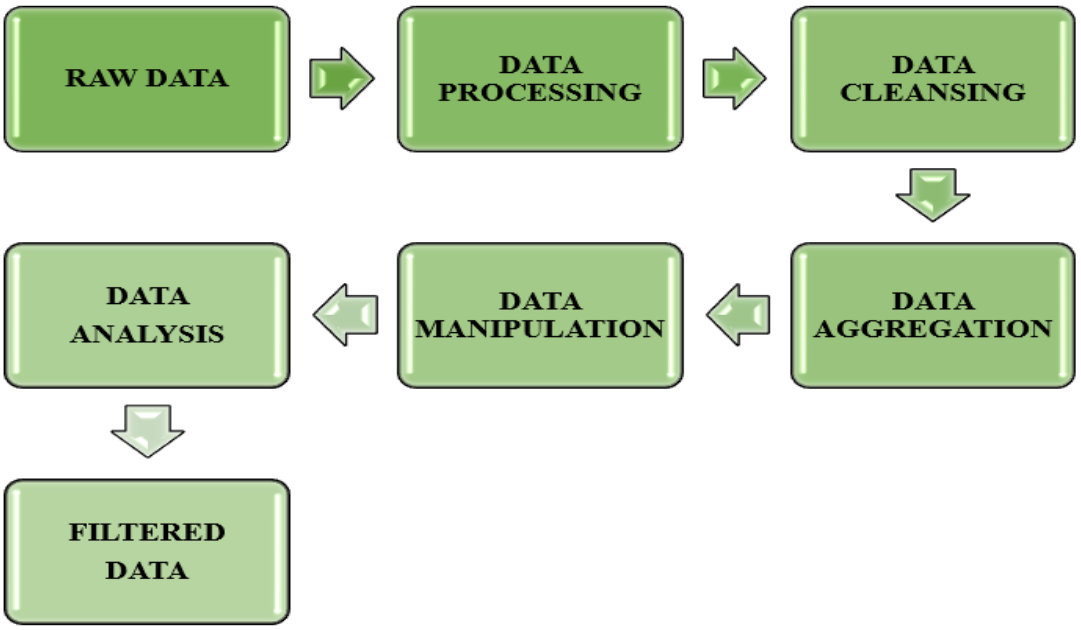


Figure 1.1 Workflow process of data science

The data can be defined in many ways according to the requirement but for the purpose of AI the suitable definition may consider as “The data is the information gathered in digital form which is used for reasoning, calculation that can be transmitted or processed”. The data is defined on the basis of their storage capacities also. Nowadays, the data is stored in abundant forms and a new unit is found for this i.e., brontobyte which means 10 raises to the 27th power of bytes. The data which we get from the primary sources is raw data. In computers the data stored is binary data which is in the form of either 1 or 0. The record of a specific quantity is Data. It includes set of rows, columns, facts and figures for the purpose of survey and analysis. The data become information when arranged in an organized form. Moreover, for analysis, the authenticity of data depends on the source of data i.e., Primary source or Secondary source of data. The data can be arranged according to the need of researcher the analysis can be done either manually or by using data mining.

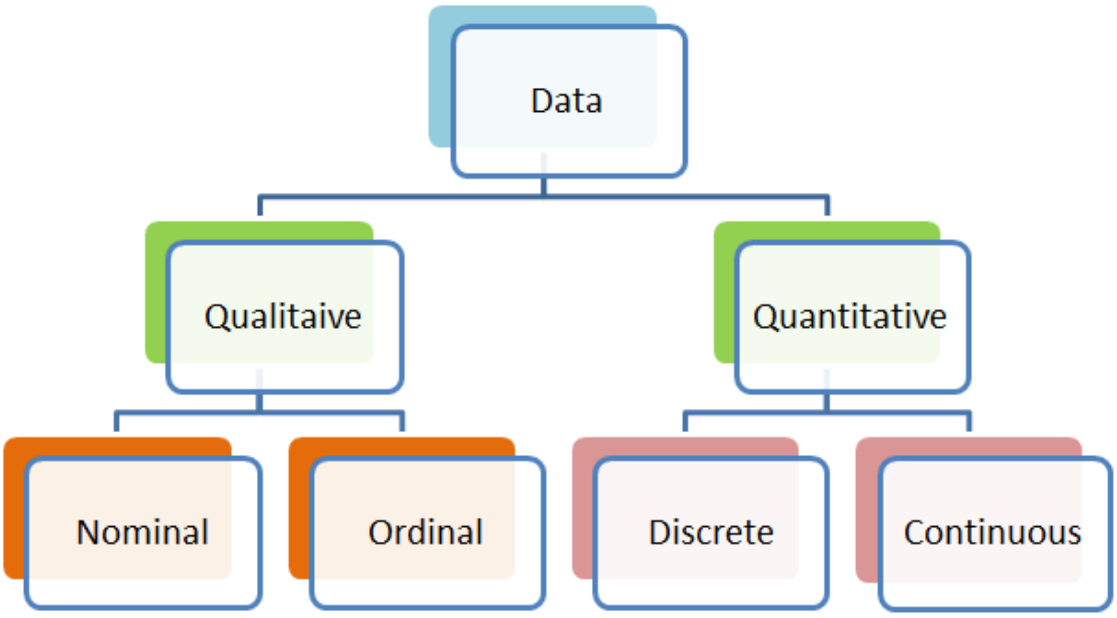


Figure 1.2 Data and Its types

Qualitative and Quantitative are two types of data.

(i) Qualitative Data

Characteristics or attributes are shown in this type of data. It is for the descriptive type of data. The data cannot be computed or calculated but only can be described.

The data of the student samples has been taken and attributes are of qualitative nature. It is also known as categorical data. In smart phones Category, colour and current rating of the phone is of qualitative nature and categorized as qualitative data.

It further subdivided into two parts: Nominal and Ordinal

Nominal: In this the natural ordering is not possessed by the nominal data. The colour of laptops is incomparable as it is not compared with another so it is considered as nominal data type. Nominal data is not of quantifiable type and hence it can't be measured through numerical units. Genders like Male, Female are intangible which cannot be measured in any unit and are therefore regarded as Nominal Data Type.

Ordinal: This type of data maintains their class and also holds natural ordering. In case of brands of laptops or smart phones we can easily sort it in the term of their size i.e Large & medium &small. The student's grading system is also an example of Ordinal Data where Grading of A++ is better than Grade A+. To handle these types of data which is not in numerical form will be encoded into binary form so that the machine learning models can easily handle and perform various operations. There is one-hot encoding and label encoding which is similar to binary coding and integer encoding respectively.

(ii) Quantitative Data Type: This data gives the numerical value as it quantifies things that make it countable. The price of anything, discount upon them, ratings of the product all these are of quantitative type. This type of data is not only for description or observation. It is for various calculations like forecasting, analysis, comparison and many more. E.g., in schools, if we have taken the data of students who are interested in sports and it shows that how many students like which game according to the data collected and this kind of information can be classified as quantitative and of numerical type. The price of any product varies from m amount to n amount and it can be further divided in fractional values so this type of data is also further subdivided or well explained in two categories:

Discrete: This type of data has fixed value in the form of integers or whole numbers. The number of research scholars, number of students and films release in this year all are examples of these types of data. Thus, from these examples it is to say that these values can't be measured but only can be counted as objects which have fixed value. In this the values

counted are represented by hoe numbers and it can also be identified through charts, bar graphs, tally charts and pie charts.

Continuous: In discrete values there are no continuous values. Due to the whole values, it discontinued the values which is the fractional values as there are infinite number that lies between two whole numbers hence the fractions that exists between whole numbers are actually considered as continuous values. There are many examples of this type of continuous values like values of ECG, band width of Wi-Fi. The variation between the high value and low value is known as its range. As the temperature, weight of human break down the continuous value and can take any value. This type of Continuous data is represented using line graph, frequency distribution and histogram. It can be well used for analysis and interpretation.

(iii) Big Data

Big data, the word itself well description for very large datasets that are too complex to be analyzed using trending methods. Big data can include complex social media data, sensor data, transactional data, and more. Big data can be used to uncover insights, trends, correlations, and anomalies in complex datasets to unlock better business decisions and outcomes.

In terms of data mining, it refers to the application of large datasets and advanced analytic techniques, such as artificial intelligence, machine learning, and natural language processing, to find insights and correlations that may not be readily apparent from small data sets. By using bigger datasets and powerful analytics, organizations are able to glean information from many sources and make more accurate decisions than ever before. This is particularly useful in industries such as medicine, finance, and retail, which benefit from having a larger pool of data. Specifically, data mining can be used to find trends in customer purchasing behaviour, market patterns, data distributions, and other metrics that may not be easily accessible through traditional methods.

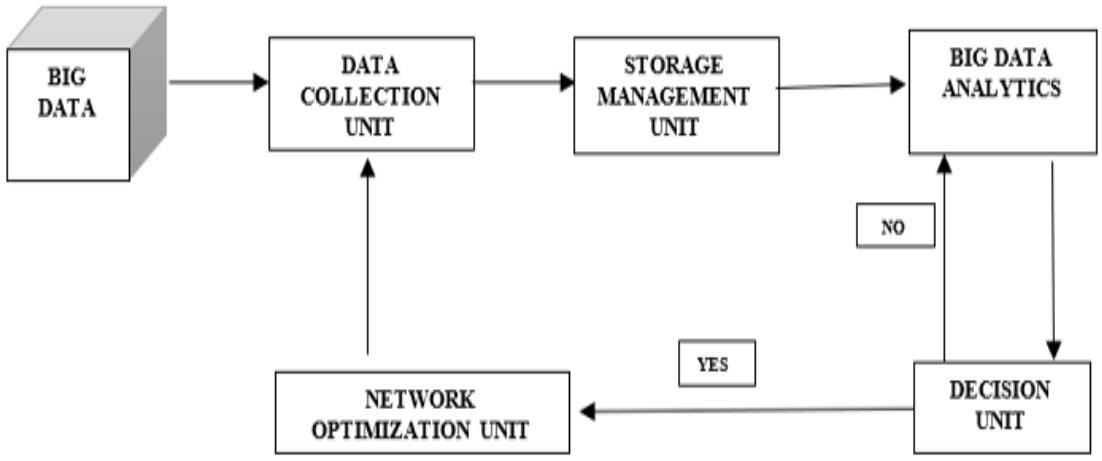


Figure 1.3 Big data phenomenal storage units with process

The following are the three categories of big data:

- **Structured Data**

  The data which is in well organized and as the name shows in structured form is known as structured data. The data which follows any pattern and must be in the form of tables or in other form which can be easily processed. Data stored in spreadsheets, Relational Databases with multiple rows and columns are the best examples of structured data type.

- **Unstructured Data**

  Data which is neither structured nor organized and as it is in unorganized manner so cannot be easily accessed or processed to draw out the information and knowledge. Images and Graphics types of data along with many audio, video, emails, social media's data are all the unstructured types of data. Apart from this the locations send by the people or their updation of exact location, orders on flipkart, amazon, Zomato all are examples of unstructured type of data.

- **Semi-Structured Data**

This type of data format does not conform to the traditional data model. The data which is in structured form but not in organized form. Types of files are the example of semi-structured

data files. To work on this type of data special type of software like Apache, Hadoop is required. Examples of semi-structured data include XML, JSON, and HTML.

1.4 Data Collection

Data can be classified as Primary data or secondary Data, based on its source of collection which is explained below:

- **Primary Data**

The data collected directly without any intermediate stage is known as Primary data. This type of data is collected and then immediately used by a person for a specific purpose. Primary data is an immaculate and original data without any loss of information. Census of India is an example of primary data.

- **Secondary Data**

The data which is collected by someone and passed one or many intermediate stages. This type of data is collected by a person and the passed to some other person or organizations for some specific reason so, it is not original or direct data. This also means that this kind of data is adulterated or has some information loss. There are many changes in the data as per the requirement of analyst or researcher so there must be addition, deletion or modification of data. This type of data is available on data repositories site like WHO, Kaggle and many government data websites.

1.5 Data mining and Confluxes in data mining

Data mining phrase came into existence in 1990 originally while since seventies the concept of Data mining have been utilised. The knowledge discovery can be alternatively used with the concept of KDD and since 1990s this concept is using by authors interchangeably. DM as the 'detection, innovation, in the huge volumes of data, of intriguing and valuable patterns and connections and describes them as sub-fields of prophetic demonstration.

These are the combination of algorithms, processes and techniques to identify patterns and relationships in large datasets. The goal of the confluxes is to enable data analysts to uncover meaningful insights and ultimately make predictions and decisions that are more accurate than conventional techniques. These confluxes can involve a combination of data exploration, data

analytics, machine learning, and artificial intelligence to develop a better understanding of the data at hand. Examples of confluxes in data mining include sentiment analysis, predictive models, and natural language processing.

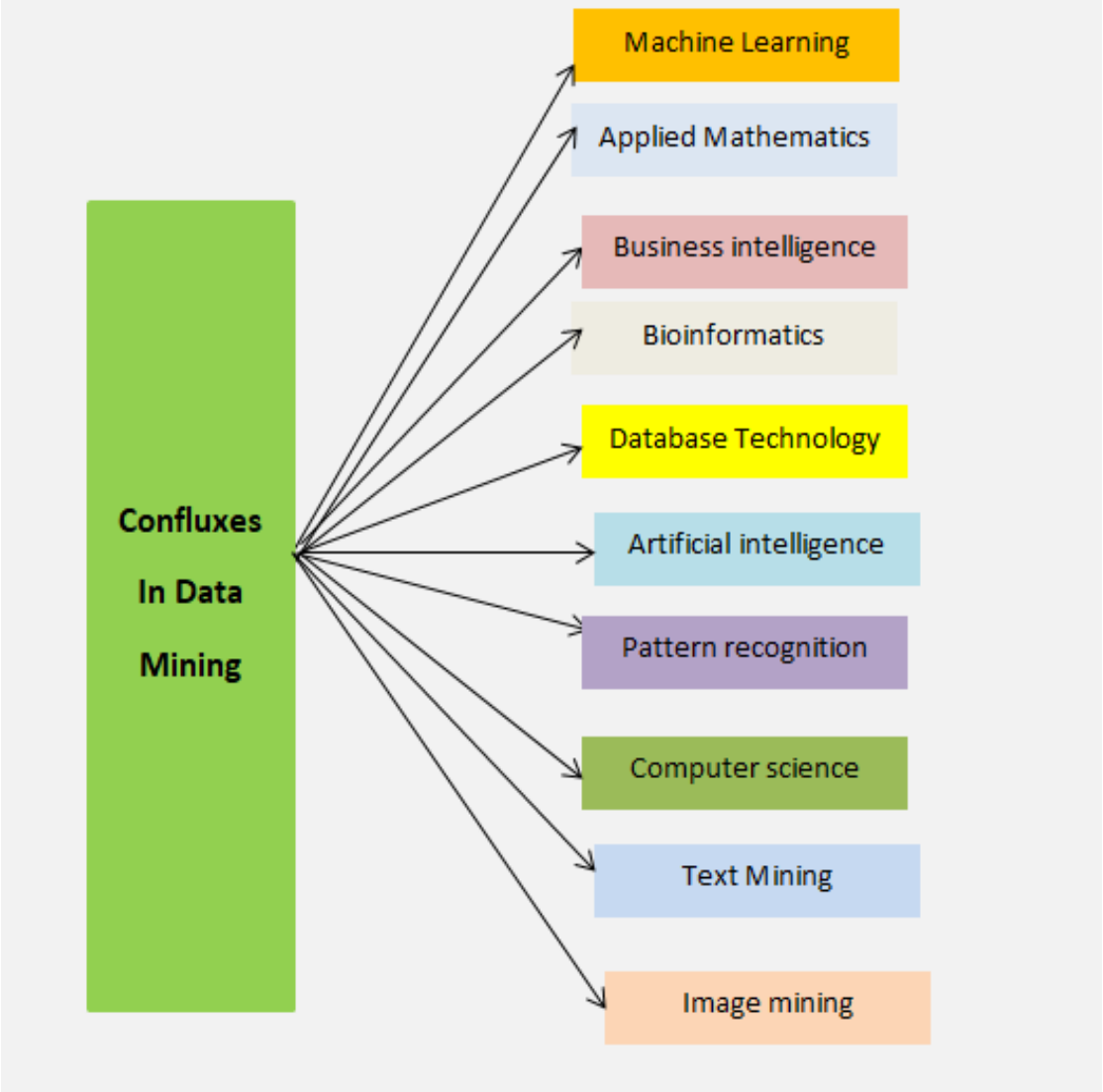


Figure1.4 Conflux of multidiscipline in Data Mining

1.5.1 Classification of Data Mining System: The concept of Data mining is part of KDD which is employed in this thesis part by using various machine learning, statistical and data analysis approaches to gain new knowledge, non – trivial like hidden patterns, prediction etc. for decision making from datasets especially from large datasets. To extract knowledge various techniques can be used like A.I, statistical approaches and data mining. In this the statistical approaches basically used to analyse primary data and data mining to analyse

secondary data. The statistical method usually employed on a sample of dataset. In statistical analysis, a formulation of hypothesis made while The induction procedures allows in data mining. Statisticians employed only formal mathematical methodologies to avoid heuristic techniques while data mining methods are also based on mathematics but these methods employ approximation methods and heuristic methods to find out the solution from large amount of data of both categories data.

1.5.2 Process of Data mining

The way of extracting useful insights and patterns from big data is data mining. It involves a variety of techniques used to analyse, model, transform, and extract insights from data. Data mining typically involves three important steps: Data Pre-Processing is to cleaning the dataset and prepare it for analysis then Data Modelling Uses statistical or machine learning techniques to identify patterns in the data and Data Analysis to interpreting the data to find insights. In addition to these steps, data mining may also include data visualization strategies to better understand the results of the analysis.

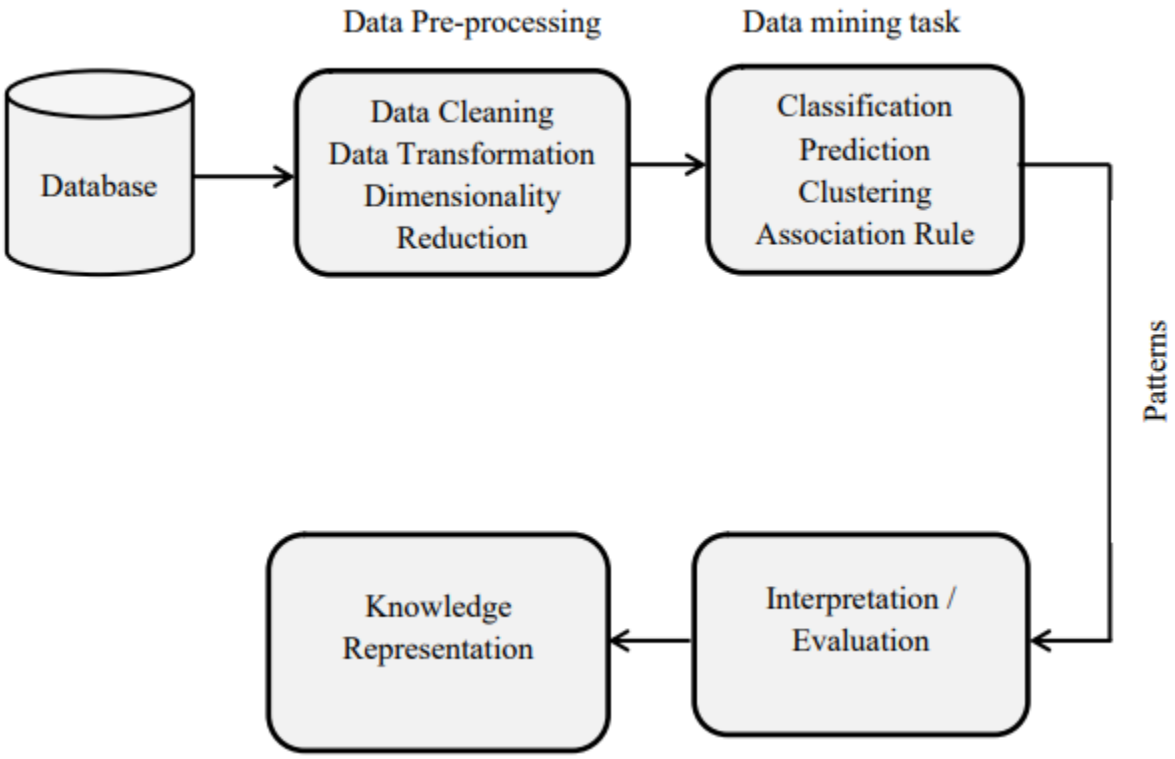


Figure1.5 Architecture of Data mining Phenomenon

I. **Database**: A database is a collection of data that is organized to provide efficient retrieval of information. It is typically composed of multiple tables or forms that allow users to store and manage their data. Databases can be used in a variety of fields such as tracking nutritional information, providing an address book/contact list, or storing customers' purchase records. Furthermore, databases are often used to analyze the collected data and help in decision making.

II. **Data pre-processing**: It is a crucial step in data mining which is used to prepare the data for subsequent analysis. The motive of this process is to ensure that the data provided to the machine learning model is of high quality, clean, and compatible with the learning algorithm. Data pre-processing involves cleaning, transforming, and normalizing the data. It also includes scaling, discretization, correcting irregularities in the data, such as outliers etc. The transforming the data from one format to another, such as from text to numerical is known as Transformation. Normalization includes standardizing the values of the data, or rescaling them to a common range. Scaling is another form of normalization which is used to ensure that data points are comparable. Discretization involves binning or grouping the data into categories. Feature selection is the task of selecting relevant features or attributes in the data set.

III. **Data Mining task:** After Pre-processing the next step is data mining tasks in which on the basis of deterministic results either the data is to be classify or to make predictions or to make clusters among data or to associate the data.

IV. **Classification:** It is the process of organizing data into categories based on similarity or common attributes. It is commonly used in machine learning and data mining to better understand and analyse data. Classification can be used to determine which group, class or category an observation belongs to. The categories used may be qualitative (unordered) or quantitative (ordered). Examples of classification include categorizing animals according to species, sorting books by genre, or labelling customers according to their purchasing behaviour.

V. **Prediction**:

1. Describe or diagnose the problem: At the beginning of prediction, diagnose the problem of forecasting, data available (structured, unstructured, etc.), and the potential applications of the data.

2. Clean and pre-process data: After, identifying the data required for prediction, pre-process the data. This includes cleaning the data (e.g. removing outliers, normalizing values) and transforming it into a format that is suitable for the predictive modelling process.

3. Select and train a model: Once the data is in the desired format, select and train a model. This involves selecting a model type, such as a decision tree, neural network, or support vector machine; and then tweaking the parameters of the models to optimize its performance.

4. Test and evaluate the model: Once the model is trained, then check how it performs on unseen data. This is typically done by comparing the model's predictions against known values in the data set and can also checked by evaluation parameters like recall precision etc.

5. Deploy the model: Once the model performed satisfactorily, deploy the model for prediction and get the required results.

VI. **Clustering and association**: These both Clustering and association are known as rule mining and both two are closely related concepts in data mining, that is, to extract the knowledge from data. Clustering is the process of finding groups and patterns in a dataset, while association rule mining aims to uncover relationships between items in the data.

1. Clustering involves analysing a dataset to find groupings of similar items, which often require an understanding of the underlying structure of the dataset. Clustering algorithms can be unsupervised or supervised, depending on the type of data in question. Unsupervised methods use only the data while supervised methods use prior knowledge to create a model of the data. Cluster analysis aims to identify underlying structure within the data, including any patterns or trends that are useful in the analysis.

2. Association rule mining is a method of discovering relationships between items in a large database. A rule is defined as a set of conditions that, when satisfied, constitute a relationship between items in the dataset. Association rule mining finds all rules that have a certain level of confidence, making it possible to uncover relationships that

were previously unknown. It is also useful for finding correlations between items that were not explicitly stated in the data. For example, in a transaction dataset.

VII. **Interpretation or evaluation** in data mining is the process of analyzing and interpreting the patterns and trends discovered in large datasets using data mining techniques. This is done through visualizing the data in charts, graphs, and other visual aids, as well as through using statistical methods such as hypothesis testing to interpret the results. It is an important part of the data mining process, as it allows for the generation of useful insights based on the observations made. These insights can then be used to make better decisions within the organization.

VIII. **Knowledge and visualization** are two important components in data mining.
Knowledge discovery involves the identification of relationships and patterns in data. Through knowledge discovery, the patterns of data can be understood so that effective decision making can be made. This is especially important in fields such as healthcare, financial services, and marketing while Data visualization is a technique that shows relationships among variables in a dataset by displaying them graphically. It allows the user to see patterns and trends that may not be obvious when looking at the data in tables or lists. Data visualization can be used to identify potential correlations between variables that would indicate the presence of a specific pattern. The data mining process may shows many errors caused by data type, software for analysis etc. which is known as the outliers and well explained below.

1.5.3 Outliers and its approaches in data mining:

The observations that are considered to be significantly unique from other data. They can be either extreme values or values that deviate from the expected pattern of the data. Outlier's occurrence can be by data entry errors, or even the presence of rare events. Outliers can be identified by using various methods such as visual examination, statistical tests, clustering, and data transformation. Data mining tasks significantly affected by outliers thus it is important to identify and remove them. Thus, outliers can be defined as the points which are different from rest of others and various forms of outliers are defined as follows:

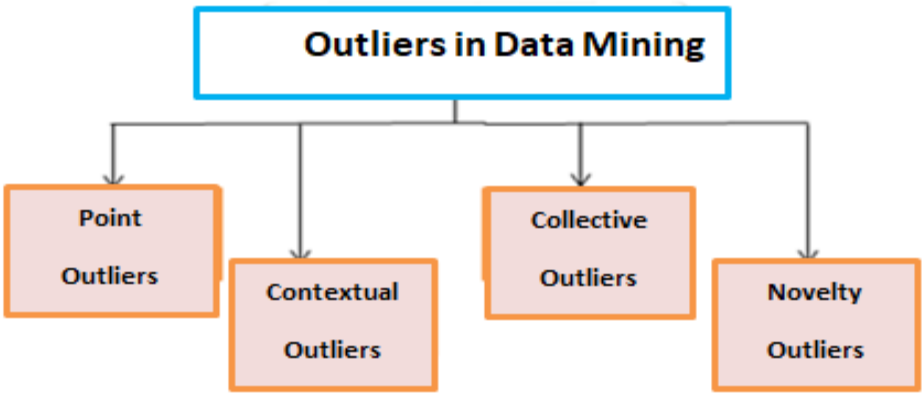


Figure1.6 Outliers in data mining

Types of Outliers in data mining

I. **Point outliers:** These are single data points which are totally different from the rest of the dataset. Data entry errors or extreme values that occur rarely can be the causes of point outliers.

II. **Contextual outliers:** These are data points that are considered outliers when compared to their neighbours. For example, a house with a much higher price than the surrounding houses.

III. **Collective outliers:** These are groups of data points that are remarkably different from the rest of the dataset. They can be caused by a shift in the data distribution or by a completely different population.

IV. **Novelty outliers:** These are data points that belong to a completely different population than the rest of the dataset. They can be caused by new or unknown phenomena.

a) Statistical approaches of outliers for data mining

I. Boxplot Analysis: Boxplots are effective in detecting outliers in a dataset. Boxplots are visual summaries of data that show where most of the data lies and how much variability exists. Boxplots also show the presence of outliers or suspicious values.

II. Z-Score Analysis: It is a numerical measurement used in statistics to indicate to mean distance from the center of points. Z-score can be used to identify outliers in a dataset.

III. **Interquartile Range (IQR):** It is used to measure the dispersion of the data, and is especially useful in identifying outliers. It can be used to detect outliers in a dataset by finding any number that falls outside of the range of theLet Q_1, Q_2 and Q_3 be the first quartile, 2nd quartile and third quartile respectively then the interquartile which is Q_2 $=\left(\frac{\eta+1}{2}\right) th$ term

$$\mathbf{Q_1} =\left(\frac{\eta+1}{4}\right) th \text{ term and } \mathbf{Q_3} =3\left(\frac{\eta+1}{4}\right) th \text{ term}$$

IV. **Modified Z-score:** The Modified Z-Score is a statistical measure developed by Edward Altman in the 1960s to assess the creditworthiness of a company. It is calculated by taking the ratios of financial statement items such as profits, total assets and liabilities and then "standardizing". It is expressed as a Z-score, which so that values below a certain threshold generally indicate a higher risk of default. The Z-score has been adapted over the years to incorporate additional factors, such as the size of the business.

V. **DBSCAN:** It is a clustering algorithm that is used to identify outliers in a dataset. It works by finding groups of similar data points and assigning outliers to different clusters.

VI. **The distance-based** approach is a method of detecting outliers in a dataset by measuring the distance between the data points and the mean or median of the dataset. This approach uses the theory that data points that are far away from the mean or median are more likely to be outliers than those that are closer. The distance can be calculated using any distance metric, such as the Euclidean distance, Manhattan distance, or Cosine similarity. Once the distance between each data point and the mean or median is calculated, data points with a large distance are considered to be outliers.

b) Index based algorithm for distance based approach of outliers

The index-based algorithm for distance-based outlier detection is a simple and efficient approach. It uses a distance function to measure the similarity between data points and then uses a threshold on the distance to identify outliers. The main steps of the algorithm are as follows:

1. Calculate the distance between all pairs of data points using a distance function.

2. Construct an index based on the distances.

3. Use a threshold to identify outliers.

4. Refine the threshold if necessary.

5. Repeat step 2-4 until a satisfactory result is obtained.

The index-based algorithm is suitable for small datasets, as it does not require the computation of all pairwise distances for large datasets. This makes it an efficient solution for detecting outliers in large datasets.

- **Nested loop algorithm for distance based approach**

```
//Loop through all points
for (int i=0; i<points. length; i++) {
    //If the distance is less than the specified threshold, add them to the result
    if (dist <= threshold) {
        result.add(points[i], points[j]);
    }
  }
}
```

for each point A in dataset and for each point B in dataset. Calculate the distance between A and B.

If the distance is less than the specified threshold; Add A and B to the list of similar points.

End.

d) Deviation based approach

The deviation based approach to determining the price of a product or service is one that takes into account the cost of production and other factors such as market conditions, customer

demand, and competitor prices. This approach involves analyzing the cost of production and then estimating the price of the product or service based on the deviation from the average cost. The deviation approach allows a business to adjust the price of a product or service to reflect changes in market conditions, customer demand, and cost of production. This approach is useful for businesses that want to remain competitive and maximize their profits by pricing their products or services at the most beneficial price point.

e) Density based local outlier approach

Density-based local outlier approach (LOF) is an unsupervised anomaly detection method that uses local density estimates to identify outliers in a dataset. The approach is based on the concept that outliers are data points that have significantly lower local density than their neighbours. It measures the local density of an object by calculating the number of objects in its neighbourhood and the distance of each object from its neighbours. Objects that have a significantly lower local density than their neighbours are considered outliers. The LOF algorithm can be used to detect outliers in both single and multi-dimensional datasets.

f) The cell based algorithm: It is used for the computation of the very straightway route between two points on graph which involves the use of a grid of cells. Each cell in the grid represents a node in the graph and contains information about its adjacent nodes. Starting from the starting point, the algorithm follows the shortest path by traversing the grid. At each step, the algorithm will evaluate the cost of moving to an adjacent node and will choose the node with the lowest cost. It is the iteration process and the cost of moving from one node to another is determined by the weights associated with the edges connecting the two nodes. The algorithm will save the cost of each node visited in order to prevent visiting the same node more than once. Once the goal node is reached, the algorithm will trace back its steps in order to find the shortest path.

1.6 Scepticism in Data mining

Uncertainty in data mining refers to the lack of clarity surrounding data sets or the analysis of those datasets. Uncertainty can arise in the form of missing data, conflicting data, errors in data collection, or insufficient data. Other sources of uncertainty include imprecise or incomplete models, and lack of knowledge about the context of the data. These techniques are

used to uncover patterns in data, and the presence of uncertainty can lead to inaccurate or incomplete results. To address this, data mining algorithms must be adjusted to account for uncertainty, or the data must be pre-processed to reduce its uncertainty.

I. **Missing Data**: Missing data can occur for a variety of reasons, such as when data is not recorded or recorded incorrectly. This can lead to imprecise or incomplete results.
II. **Conflicting Data**: Data from different sources can be conflicting, leading to uncertainty in data mining.
III. **Errors in Data Collection**: Errors in data collection may cause incorrect results.
IV. **Insufficient Data**: Insufficient data is also one of the reason for inaccurate prediction.
V. **Imprecise or Incomplete Models**: Models used to analyze data can be imprecise or incomplete, leading to uncertainty in data mining.
VI. **Lack of Knowledge about the Context of the Data:** The context of the data can be unknown, leading to uncertainty in data mining.
VII. **Data Quality**: Inaccurate or incomplete data can lead to incorrect results or misleading conclusions.
VIII. **Selection of data set:** Data mining problems require large and suitable datasets. Choosing the right dataset is a critical step in the data mining process.
IX. **Data pre-processing**: It is essential to pre-process the data before it is fed into data mining algorithms.
X. **Outliers:** Outliers can significantly affect the results of data mining, as they can skew the results in unexpected ways if not identified and eliminated.
XI. **Data Sampling**: Data sampling can also introduce uncertainty into the results. If the sample size is not big or not represents population then the results may be inaccurate.
XII. **Algorithm Selection**: The choice of algorithm used for data mining can significantly affect the results. The apropriate algorithms for the chosen problem can lead to correct and best results.
XIII. **Data Availability**: Data mining requires access to large amounts of data. If data is not available, it is not possible to perform data mining.
XIV. **Data Security**: Data mining can reveal sensitive information about customers or other individuals. Data security and privacy must be ensured by organisation.

XV. **Data Interpretation**: Data mining can provide valuable insights, but it is up to the analyst to interpret the data correctly and carefully. On the other hand negligence can give disastrous results.

XVI. **Model Over fitting**: Too complex models create over fitting and produces results that are more specific to the training data than to the general population. This can lead to inaccurate predictions when used with new data.

1.7 Building the data mining model

The process of building a data mining model involves several steps as it works to extract the knowledge from big data. For this, the data must be collected and cleaned up to remove any outliers or errors. Next, the data needs to be prepared for analysis by partitioning it into training and test datasets. Once the data is ready, the model can be made by using hybrid techniques like decision trees, cluster analysis, or neural networks. Finally, the model can be evaluated to determine its accuracy and predictive power.

1.8 Data mining tasks

Data mining tasks are processes and techniques used to identify patterns, correlations, and trends in large datasets. These techniques are used to identify meaningful information or to solve complex problems. Data mining tasks include clustering, classification, regression, association rule mining, outlier detection, and visualization. Data mining tasks are often used to uncover customer purchase history and customer demographic information and likewise the same in various fields.

I. **Identify and select the data:** Identify the data and by pre-processing make it ready for the analysis and select it for the model.

II. **Clean and pre-process the data:** In this remove irrelevant or incomplete data points, remove noise and fix errors to clean the data.

III. **Split the data:** There are two phases of data on which the model deployed and implemented known as training data and testing data.

IV. **Choose the type of model:** Decide on the type of data mining model, such as a decision tree, neural network, or clustering algorithm, that best suits the problem.

V. **Train the model:** Train the model using the training set and optimize its parameters.

VI. **Evalution of model:** Various evaluation metrics can check the accuracy on test dataset.

VII. **Fine-tune the model:** Iteratively fine-tune the model based on the evaluation results, until a satisfactory performance is achieved.

VIII. **Deployment of model:** The trained model will be deployed for the prediction in various fields.

1.9 Predictive data mining methods: These are used to harness the power of complex data sets and use them to predict future trends and behaviour. These methods use the data collected from past activities, trends and events to identify patterns and relationships. By analyzing these patterns and relationships, future events can be predicted with a high degree of accuracy. Predictive data mining methods involve Machine Learning techniques such as neural networks, decision trees, regression and clustering which can be used to make better and more accurate predictions. Additionally, predictive data mining, the methods can be implemented in many fields for the enhancement of performances like in business by providing important insights during the decision-making process.

1.10 Correlation between Data science, artificial intelligence and Machine learning

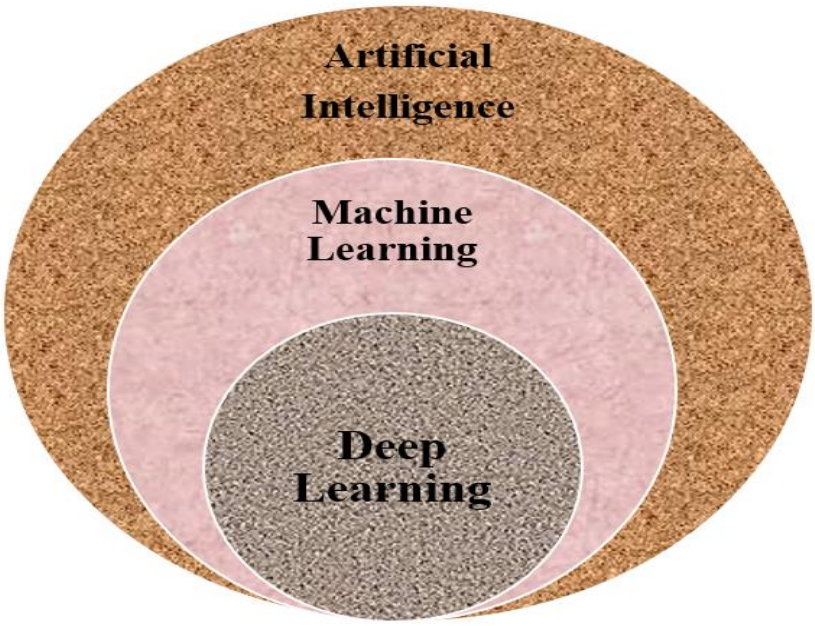


Figure1.7 Correlation between fields of data science

Artificial intelligence is the science of creating intelligent machines that can work and behave like human beings." A lot has been said about Artificial intelligence like there uses in daily life, work, mobility, healthcare, economy in communication and also about the dangers of this technology poses to human existence. Artificial Intelligence tries to mimic the human mind by putting logics, reasoning, and problem solving capabilities into machines or computers. Alan Turing described Artificial intelligence as the systems or machines which can replace human acts. It combines robust and computer sciences to solve the problems like forecasting in various fields. These fields are interrelated to each other as the artificial science is the super science of all sciences like machine learning, data science, deep learning etc. With the use of artificial science all hurdles gets reduced as it is similar to the forecasting i.e., as time flies, it may not equal to the forecasting but it approaches to the exactness in forecasting. Artificial intelligence is beneficial in every sphere of life when it is under human control. If it starts controlling humans it will become a vice for society as a whole.

"It is useful when it is mastered by man; when it masters man, it becomes disastrous."

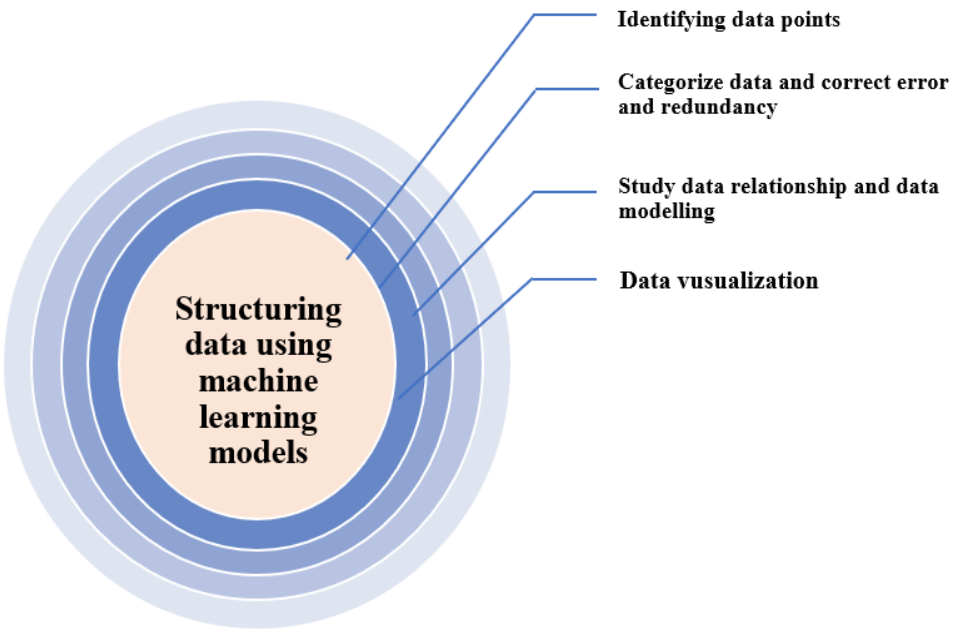


Figure1.8 Structuring the unstructured data using artificial intelligence or machine learning models

Artificial intelligence and machine learning models have revolutionized the way we structure unstructured data, providing invaluable insights and efficiency. Using AI and machine learning, we can analyze and structure unstructured data in ways that were previously impossible, unlocking its true potential. By leveraging artificial intelligence and machine learning, we can automatically extract and categorize information from unstructured data, making it more useful and accessible. With the help of AI and machine learning, raw unstructured data can be transformed into structured data, enabling faster and more accurate decision-making. In this the data is structured

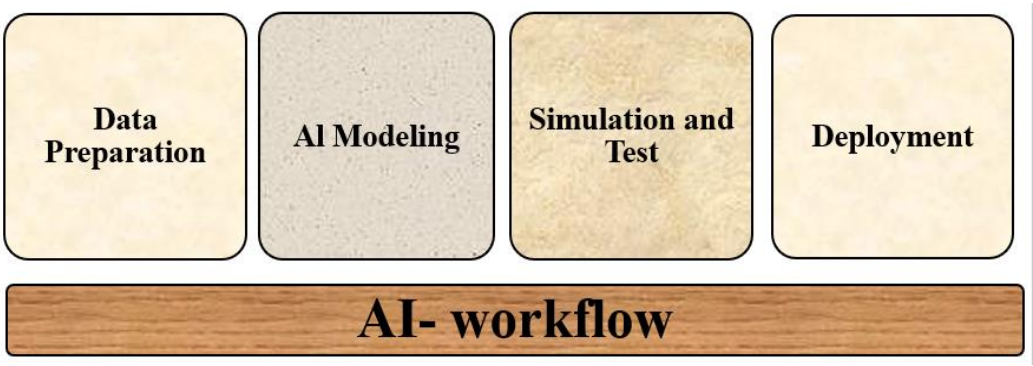


Figure 1.9 Artificial intelligence workflow

The typically consists of several key steps: Data preparation, AI modeling, simulation, testing, and deployment.

Data preparation involves collecting and cleaning the relevant data that will be used to train the AI model. This may involve gathering data from various sources, cleaning and pre-processing the data to remove noise or duplicates, and organizing it in a format suitable for training the AI model. The next step is AI modeling, where the data is used to train the AI model. The goal is to train the model to accurately predict or classify the desired outcomes based on the input data.

Once the AI model is trained, it is important to simulate and test its performance. This involves evaluating the model's accuracy, precision, recall, and other performance metrics using a separate test dataset. This step helps to identify any potential issues or weaknesses in the model and allows for adjustments to be made if necessary. Finally, once the AI model has been validated and optimized, it can be deployed to perform the intended tasks. This may

involve integrating the model into a larger system or application, setting up appropriate data pipelines, and ensuring that the model is capable of handling real-time data.

1.11 FORECASTING IN VARIOUS FIELDS

Forecasting starts from economics in early 1990s but now no field remained untouched from forecasting as it is the dire need of hour so, Forecasting is used in many areas, such as economics, weather, and the stock market. Forecasting can be used to plan business processes, anticipate customer demand, and make decisions on inventory control. In economics, forecasting is used to predict future economic trends, such as interest rates, inflation, and GDP growth. Forecasting also helps to understand the overall health of the economy, which helps policymakers make decisions on fiscal and monetary policy. In weather forecasting, meteorologists use data from a variety of sources, such as satellite imagery, radar, and atmospheric models, to predict future weather conditions. This helps to inform decision-making regarding disaster preparedness, agricultural activities, and other weather-related issues. In the stock market, forecasting is used to make predictions about future stock prices and market movements. This helps investors make informed decisions about which stocks to buy and sell. Forecasting is an important tool for decision-making in many different fields. By using past data and trends, forecasting can help to provide insight into the future and inform decisions on how to best prepare for it.

Forecasting can be used in all the area that is based on thinking/ techniques use for future demand, future requirements. Forecast is used in a different area for different situation, some areas are given below, where forecasting can be used like earthquake, Egain, technology, Political, land use, Transport, weather, Flood, Telecommunication, Agricultural, Land use, Sales, Player and team performance and economic forecasting.

Firstly, there were computers of very large size which can be only used for computing. As there is much evolution in computers from punch cards to smart cards by which it can do all the works. In 1960s computers work from professional to personal use also. Computer plays an important role in every field like forecasting by reducing the errors of prediction, by improving speed of calculations. Today every era depends on computers for their small to large business growth hence everyone does forecast by computers as it builds the models on old data and works on existing data thus it uses both old and new methods. Moreover, there

are various traditional methods but the data mining technique is fast, accurate, new and time saving technique hence we have chosen data mining techniques for forecasting.

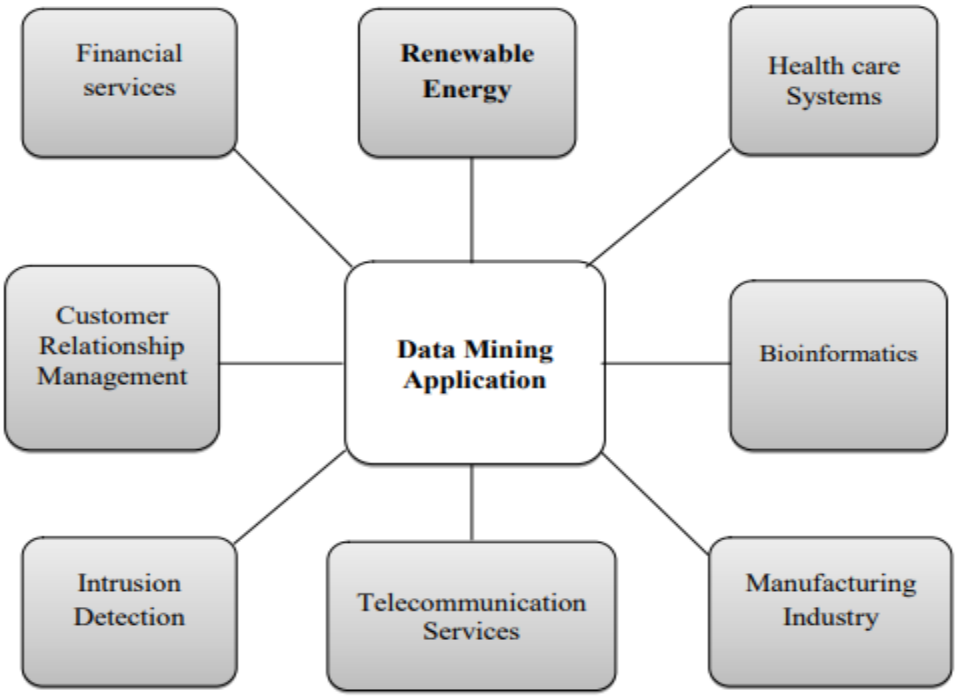


Figure1.10 Applications of Data mining

1.Customer Relationship Management (CRM): It can be used to analyze customer data such as demographic information, purchase histories, product preferences and website visits. This helps businesses better understand their customers and determine what products and services will best meet their needs.

2. Fraud Detection: Suspicious activities can be the one of best application of data mining. The data mining algorithms can spot patterns that might indicate fraudulent transactions and alert financial institutions to investigate further.

3. Healthcare: Healthcare organizations use data mining to better understand disease trends, identify patients that are at risk for certain conditions, predict how diseases will spread and more.

4. Network Intrusion Detection: Data mining algorithms can also be used to detect malicious activity in computer networks. The algorithms can detect patterns that indicate when a computer network is under attack and alert administrators to investigate further.

5. Predictive Maintenance: Predictive maintenance uses data mining to predict when machines or other equipment might need repairs or maintenance. This helps businesses cut costs by avoiding unnecessary repairs and scheduling maintenance when it's most needed.

1.11.1 Types of Forecasting:

1. Qualitative Forecasting: This type of forecasting uses subjective information such as opinions and intuition to predict the future. Examples include expert opinions and Delphi technique.

2. **Quantitative Forecasting:** This type of forecasting uses quantitative information such as historical data and trends to predict the future. Examples include time series analysis, regression analysis and econometric forecasting.

3. **Judgmental Forecasting:** This type of forecasting uses both qualitative and quantitative information to make predictions. Examples include market research and scenario planning.

4. **Causal Forecasting:** This type of forecasting uses causal relationships between variables to predict the future. Examples include ARIMA and exponential smoothing.

5. **Time Series forecasting:** It involves the data over a period of time it could be of any type and it can be of hourly, yearly or on the basis of days.

Basic time series analysis involves plotting the data over time, and examining the resulting graph for patterns. More advanced techniques involve smoothing the data to reduce noise, or using more sophisticated statistical tools to identify trends and patterns. Time series can also be limited to certain periods of time. For example, a company may wish to analyze their sales over the last five years, or a meteorologist may wish to look at weather data over the past decade. Limiting a time series to certain periods of time can help to identify changes in patterns and trends over the course of a longer period of time. The role of quantity in a time series analysis is important. The more data points that are included, the more accurately trends

and patterns can be identified. However, it is important to ensure that the data is reliable and accurate, as any errors or inaccuracies can lead to unreliable results.

Time series again are of two types Stationary and non-Stationary

In time series models, Stationary time series is an important assumption, as it allows for more robust predictions. Time series that are non-stationary can often be made stationary by applying a transformation such as differencing or de trending. This can also allow for more accurate predictions. It can help to provide more accurate predictions for future values of the series, and it can help to find out the series which can be used for forecasting and decision-making. Non-stationary can make it difficult to accurately forecast future values of the series, and can also make it difficult to identify patterns and trends. The most common way to check stationarity is to plot the time series graph and look for any obvious patterns or trends. weak, strong, and strict are three types of stationary time series.

1.11.2 Stationary Time Series are of three types

i) Autoregressive (AR) Model: In this model the past observations or data used to predict future outcomes. This type of model is based on the assumption that future values are dependent on past values and can be used to forecast outcomes such as stock market prices, economic indicators, and other trends. Autoregressive models are typically used in time-series analysis and consist of linear equations that incorporate a variable from the past and error terms. These models are useful for predicting short-term trends in data, but their accuracy may become less reliable when predicting longer-term trends.

ii) Moving Average (MA) Model: A Moving Average model is a statistical technique used to determine the averaged value of a data set over a period of time. This type of model is commonly used in finance to understand the average price of a stock or security over a certain time frame. It is also used to measure trends and predict future values given past values. The model is based on the premise that the average of a set of values will more accurately reflect the underlying trend than individual values.

iii) Autoregressive Integrated Moving Average (ARIMA) Model

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_1$$

$$Y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q}$$

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \varepsilon_t$$

$$Y_{(t-1)} = \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + .. + \beta_0 Y_0 + \varepsilon_{t-1}$$

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + .. + \phi_q \varepsilon_{t-q}$$

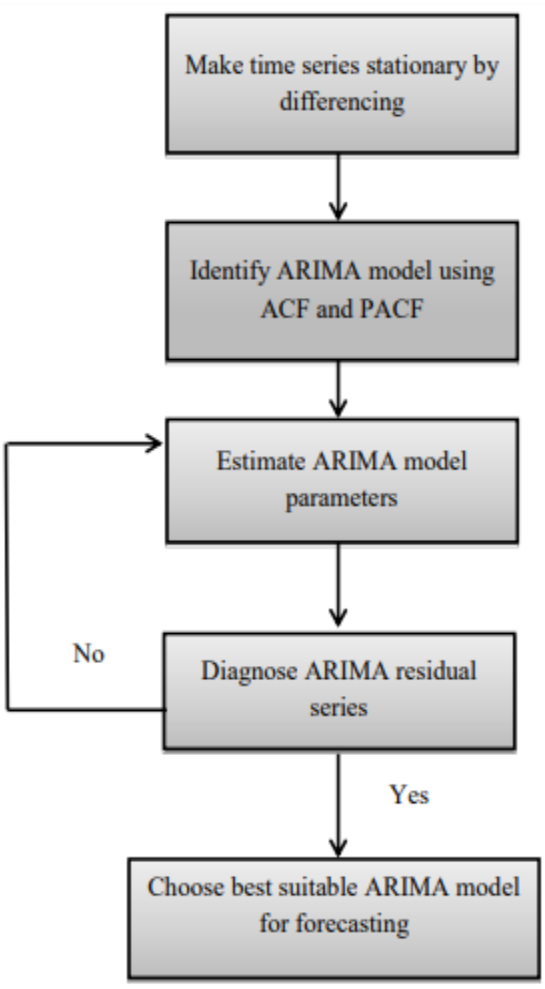


Figure1.11 Working of ARIMA Model

1.11.3 Non-Stationary Time Series

Constant mean or variance cannot be hold in Non stationary time series. These types of time series are often characterized by trends, seasonal fluctuations, or other patterns that can make forecasting and analysis more difficult. These types of time series are often the result of changes in the underlying system or environment that generates the data. In order to effectively analyse and forecast non-stationary time series, it is often necessary to use specialized methods such as decomposition, detrending, or differencing.

Non-stationary time series can be further classified into two types:

1. Trend-stationary series: These series have a mean and variance those changes over time due to a trend. These trends could be linear, exponential, or some other form.

2. Difference-stationary series: These series have a mean and variance that change over time due to seasonality or other factors. These series need to be differenced to make them stationary. This means subtracting the current value from the previous value, which removes the trend and seasonality. The components of the non-stationary series are constantly changing so it may become challenging due to its nature. To make them easier to work with, they often need to be transformed with techniques such as differencing or smoothing.

1.12 Multi-equation and Time series Methods of Forecasting

Multi-equation forecasting methods are techniques that involve the simultaneous estimation of multiple equations to make a forecast. These methods are used to capture the complex relationships between different variables, and to identify and measure the impact of different factors on the forecast. Examples of multi-equation forecasting methods include:

1. Vector Autoregressive (VAR) models: VAR models are a type of multi-equation forecasting model that use multiple time series to measure the impact of past events on a given target variable. They are typically used to understand the forecasting of macroeconomic trends.

2. Structural Equation Modelling (SEM): SEM is a type of multi-equation forecasting method that uses a system of equations to explain the relationships between variables. It is typically used to measure the factor which affects the results.

3. Bayesian Structural Time Series (BSTS): BSTS is a type of multi-equation forecasting method that uses Bayesian hierarchical models to estimate multiple equations simultaneously. It is also used to measure the factor which affects the results.

1.12.1 Time series and Composite methods of forecasting

This type of forecasting is based on past observation and past data. It is widely used for forecasting demand for a product or service, forecasting financial markets, predicting weather, and more. Time series forecasting methods uses historical data to evolve mathematical models for the prediction of future values.

1.13 Various test for time series forecasting

1. Augmented Dickey-Fuller Test: This test is used to check the stationary of time series types of data.

2. Autocorrelation Test: This test is used to find out the correlation between a given time series and its lagged values.

3. Box-Jenkins Method: This method is used to identify trends and seasonality in a time series, and also to identify parameters for an appropriate forecasting model.

4. Granger Causality Test: This test is used to identify whether one time series is caused by another.

5. ARIMA Modelling: This is an integrated model of autoregressive and moving average.

1. Econometric Modelling: Econometric modelling is a multi-equation forecasting method.

2. Exponential Smoothing: This is a technique used to smooth out noise in a time series. It is used to make forecasts more reliable.

Composite methods of forecasting combine the outputs of multiple forecasting models or techniques to create a single prediction. This type of forecasting uses multiple models and/or techniques to analyze historical data, identify patterns, and generate a more accurate forecast. The models and/or techniques used may include a mix of qualitative and quantitative approaches, such as market research, linear regression, time series analysis, and machine learning. Unlike traditional methods of forecasting, such as extrapolation or time series

analysis, composite methods of forecasting leverage the strengths of multiple models to create a more accurate and reliable forecast.

1. Delphi Method: A method of forecasting which involves the collection of opinions from experts in the field. The experts are asked to provide their predictions on the future of the given topic, which are then collected and analysed to come up with a consensus forecast.

2. Time Series Method: A method of forecasting which uses historical data to make predictions about the future. This forecasting method is based on the assumption that past patterns will continue into the future.

3. Regression Analysis: A method of forecasting which uses statistical techniques to identify the relationships between different factors and the outcomes of interest. It can be used to predict future results based on the correlations that are identified.

4. Neural Networks: A method of forecasting which uses artificial intelligence to simulate the behavior of the neurons in the brain. This forecasting method is based on the assumption that complex relationships between various factors can be identified and used to make predictions.

5. Monte Carlo Simulation: A method of forecasting which uses probability theory and random sampling to make predictions about the future. This method is often used in finance to estimate the probability of different outcomes.

Time series forecasting methods involve using the past data of a particular variable to predict future values of the same variable.

C**ross impact matrix method**

The Cross Impact Matrix Method is a tool used to analyse the potential impact of changes in one area of an organization or system on another area. It is a qualitative assessment tool that can be used to identify and assess the relationships between various factors, such as internal and external organizational processes, products, services, and systems. This method is useful for decision makers in helping to identify and assess the potential implications of proposed changes or scenarios. The Cross Impact Matrix Method is based on the premise that changes in one factor can have a direct or indirect impact on other factors. The matrix is composed of a

set of questions related to the factors under consideration and the relationships between them. Each factor is rated on a scale of probability of occurrence and magnitude of impact. The ratings are then used to construct a matrix which shows the likelihood and magnitude of the potential impacts. This method is useful in helping decision makers to identify and assess the potential impacts of proposed changes or scenarios and it gives the potential impacts of changes in one factor on other factors.

1**. Naive Method**: This is a basic method of forecasting where the most recent data point is used to predict the next data point.

2. Moving Average: This method takes the average of the most recent data points to predict the next data point.

3. Linear Regression: This method uses linear relationships between variables to predict future values.

4. Seasonal Autoregressive Integrated Moving Average (SARIMA): This method uses seasonal patterns in the data to forecast future values.

5. ARIMA: This method uses autoregressive and moving average models to forecast future values.

6. Exponential Smoothing: This method uses weighted averages of past data points to predict future values.

7. Neural Networks: This method uses artificial neural networks to learn from past data and make predictions.

8. Time Series Decomposition: This method uses the trend, seasonality, and residual components of a time series to make predictions.

1.13 Data mining Models

Forecasting models are mathematical models used to predict future events, such as economic trends, customer demand, or the performance of a company. Forecasting models can be used to identify trends, assess risk, and make decisions about resource allocation. Generally, forecasting models are based on historical data and can involve a variety of techniques,

including linear regression, time-series analysis, and simulation. Characteristics of forecasting models include:

1. Accuracy: Forecasting models must be able to accurately predict future trends and identify patterns in order to be useful.

2. Scalability: Forecasting models must be able to scale to accommodate changes in data or to handle larger datasets.

3. Flexibility: Forecasting models must be able to adjust to changes in the environment or in the data in order to remain accurate.

4. Reliability: Forecasting models must provide consistent results in order to be useful.

5. Robustness: Forecasting models must be able to handle unexpected inputs and outliers in order to remain accurate.

6. Time-efficiency: Forecasting models must be able to produce results in a timely manner in order to be useful.

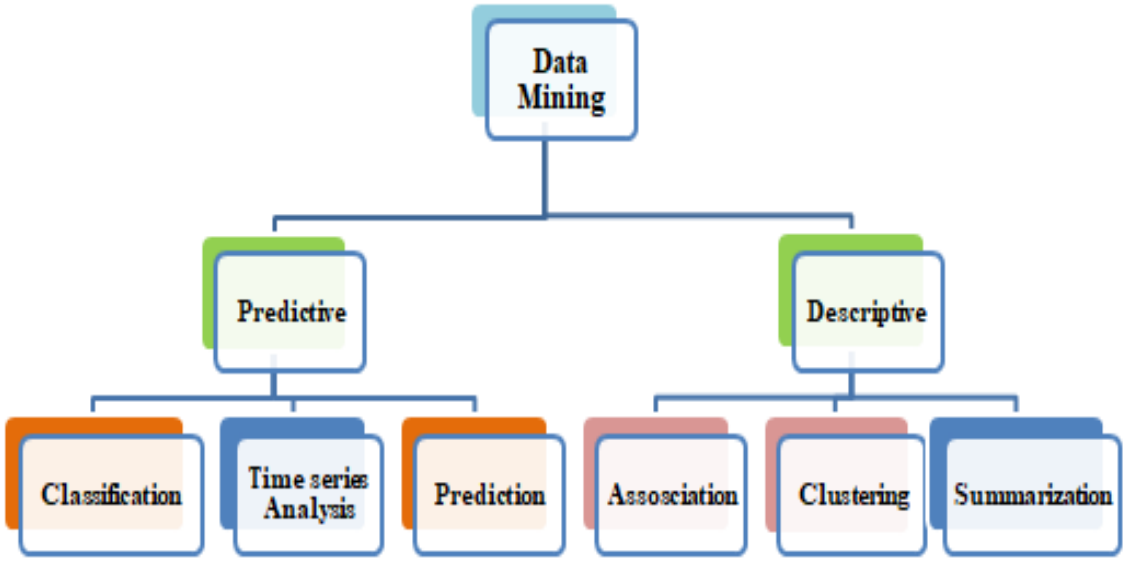


Figure1.12 Data mining models based on past observations

Predictive model and Descriptive model are two broad categories

A predictive model in data mining is a model that can predict the chances of the occurrence of an event in the future. It is built by extracting patterns from data that has already occurred, analyzing them, and using the results to create a prediction model. It can be used in events such as customer behaviour, purchasing patterns, financial fluctuations, and health outcomes. Predictive models can be used to group data, identify trends, create forecasts, and classify data.

Descriptive models in data mining are methods used to describe and summarize data for further analysis and can be applied to summarize large data sets and to find out correlations and patterns that can be used to better understand the data. Examples of descriptive models include decision trees, clustering, and regression. These models provide a basic representation of the data and can be used to develop more advanced models for prediction or further analysis.

1.14 Time series analysis: Time series data mining is a fairly broad field that deals with the investigation and modeling of data series containing measurements of some variable over time. Time series data mining often involves techniques from the fields of signal processing, machine learning, and statistics to extract meaningful information from the data.

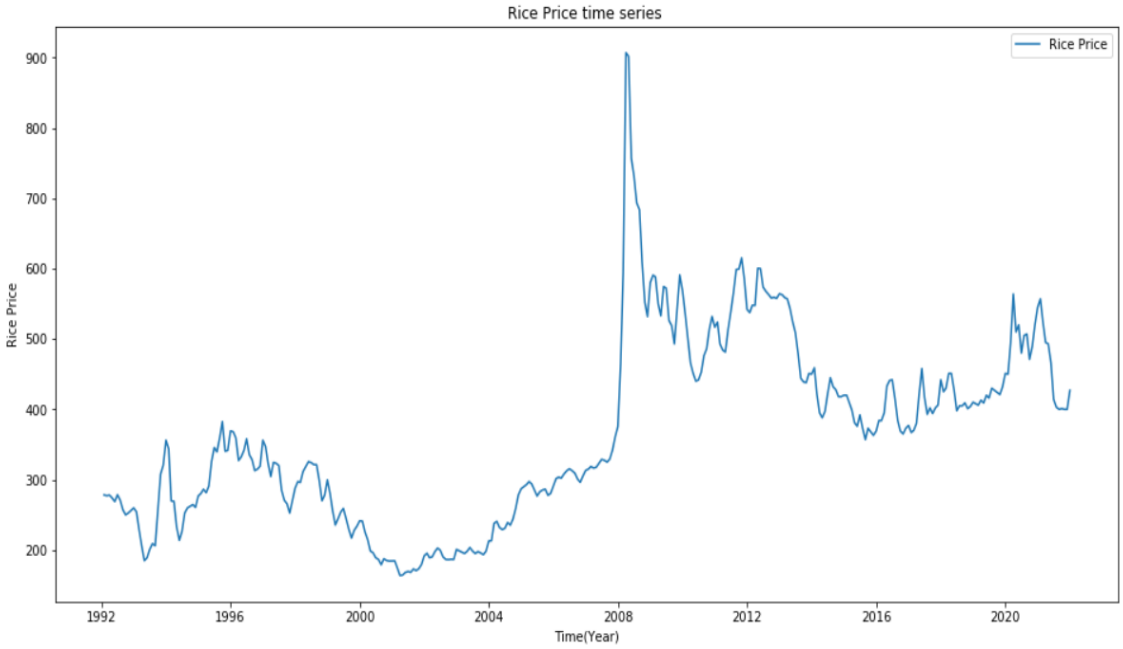


Figure 1.13 Rice Price Data from Feb 1992 to Jan 2022

It can be used to develop predictive models about future values of time series based on patterns and trends present in it, estimate parameters of underlying models, or identify the relationships among different series. It also can be used to detect outliers and anomalies, detect cyclical or seasonal patterns, summarize time series with few representative patterns or clusters, or provide timely alerts to changes. Stationarity, Seasonality and Trend are the three most common types of time series which is shown as below:

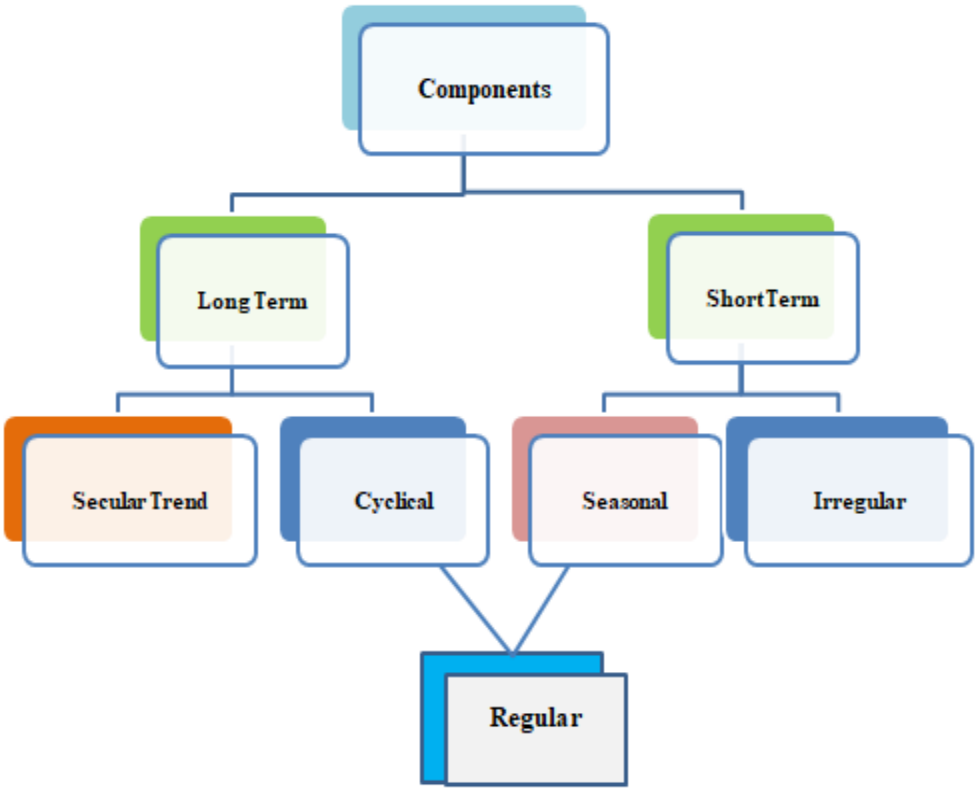


Figure1.14 Components of time series

1. Seasonal Components: Seasonal components involve repeating patterns that occur over preset time intervals, commonly known as seasonal cycles. This could be monthly, quarterly, or yearly.

2. Long-term Trend: A long-term trend is the overall direction of the data points over a period of several years. It is also known as the "big picture" of the data, and is used to provide an overall indication of whether the values are increasing or decreasing over time.

3. Cyclical Components: Cyclical components are patterns that can be observed in data that are longer than the seasonal cycle, such as economic cycles which can last five to ten years. They are linked to overall changes in the economy, and can alter the direction of the trend.

4. Irregular Components: Irregular components represent random, unpredictable changes in the data that don't follow a particular pattern. These are sometimes referred to as unpredictable, or noise, and can lead to inaccurate forecasts if not taken into consideration.

5. Trend: A long-term direction of the time series, such as an increase or decrease over time.

6. Seasonality: Seasonal patterns or cycles in the data, such as higher sales in the summer months or lower enrollment in summer semesters.

7. Cyclical Behavior: Recurrent up and down movements of varying length and magnitude; these vary between industries and can often be affected by external factors.

8. Irregular Components: Noise in the data that is unpredictable and cannot be explained by the other components.

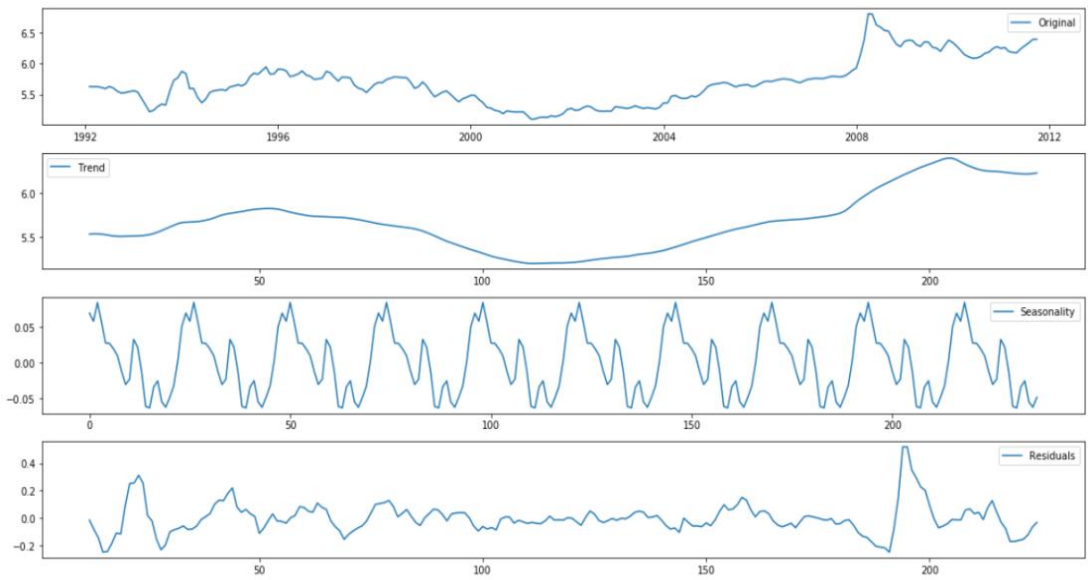


Figure 1.15 Decomposition of rice price time series into three components, i.e., Trend, Seasonality, Residuals.

1.14.1 Data mining algorithms

1. **Association Rule Learning:** This algorithm is used to find relationships between different items in a healthcare dataset. It can help identify correlations between different diseases, treatments, and other variables. There are many types of association rule mining

(I) Multi relational association rule: In this type of association it finds the relationship between more than two variables in the following manner

$$\ddot{\xi} - 2\ddot{\eta} = \frac{\partial \mathfrak{U}}{\partial \xi} - \xi - \sum \frac{m_i(\xi - \xi_i)}{\rho_i^3}$$

$$\ddot{\eta} + 2\ddot{\xi} = \frac{\partial \mathfrak{U}}{\partial \eta} - \eta - \sum \frac{m_i(\eta - \eta_i)}{\rho_i^3}$$

$$\ddot{\zeta} = \frac{\partial \mathfrak{U}}{\partial \zeta} = \sum \frac{m_i(\zeta - \zeta_i)}{\rho_i^3}$$

(II) Generalized association rule: It is to get rough idea of patterns followed to extract Knowledge and this can be find in the following manner:

$$\mathfrak{U} = \frac{\xi^2+\eta^2+\zeta^2}{2} + \sum \frac{m_i}{\rho_i^2}$$

$$\rho_i^2 = (\xi - \xi_i)^2 + (\eta - \eta_i)^2 + (\zeta - \zeta_i)^2, \quad i = 1,2,3$$

$$\dot{\xi}^2 + \dot{\eta}^2 + \zeta^2 + C = 2\mathfrak{U}$$

2. **Clustering:** Clustering algorithms are used to group together similar data points and provide insight into how different healthcare data sets are related. It can be used to identify subgroups of patients based on their medical condition, treatment history, and other factors.

It can be used for any type of grouping like to form clusters of some class. Let us suppose $(\chi_1, \chi_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \chi_p)$ are input data points and the clusters needed are k then as the initial centroids pick k points

By using Euclidean distance identify K points known as cluster centroids.

d(P, Q)= $\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$

In the same way, assign each data pint to closest centroid then so on the new centroid formed by taking the average of the points in each cluster group.

Arg min $c_i \varepsilon\ C\ dist\ (c_i, x)^2$ is way to assign each point to the nearest cluster.

3. **Decision Trees:** Decision trees are used to build predictive models that can be used to predict outcomes from healthcare data sets

$$\xi - \frac{m_1(\xi-\xi_1)}{\left[\,[\,\xi-\xi_1)^2+\,(\eta-\eta_1)^2+(\zeta-\zeta_1)^2]^{\frac{3}{2}}\right]} - \frac{m_2(\xi-\xi_2)}{\left[\,[\,\xi-\xi_2)^2+\,(\eta-\eta_2)^2+(\zeta-\zeta_2)^2]^{\frac{3}{2}}\right]} - \frac{m_3(\xi-\xi_3)}{\left[\,[\,\xi-\xi_3)^2+\,(\eta-\eta_3)^2+(\zeta-\zeta_3)^2]^{\frac{3}{2}}\right]}$$

$$\eta - \frac{m_1(\eta-\eta_1)}{\left[\,[\,\xi-\xi_1)^2+\,(\eta-\eta_1)^2+(\zeta-\zeta_1)^2]^{\frac{3}{2}}\right]} - \frac{m_2(\eta-\eta_2)}{\left[\,[\,\xi-\xi_2)^2+\,(\eta-\eta_2)^2+(\zeta-\zeta_2)^2]^{\frac{3}{2}}\right]} - \frac{m_3(\eta-\eta_3)}{\left[\,[\,\xi-\xi_3)^2+\,(\eta-\eta_3)^2+(\zeta-\zeta_3)^2]^{\frac{3}{2}}\right]}$$

$$\zeta - \frac{m_1(\zeta-\zeta_1)}{\left[\,[\,\xi-\xi_1)^2+\,(\eta-\eta_1)^2+(\zeta-\zeta_1)^2]^{\frac{3}{2}}\right]} - \frac{m_2(\zeta-\zeta_2)}{\left[\,[\,\xi-\xi_2)^2+\,(\eta-\eta_2)^2+(\zeta-\zeta_2)^2]^{\frac{3}{2}}\right]} - \frac{m_3(\zeta-\zeta_3)}{\left[\,[\,\xi-\xi_3)^2+\,(\eta-\eta_3)^2+(\zeta-\zeta_3)^2]^{\frac{3}{2}}\right]}$$

So on, in this way up to n number of decision trees find out the results and on the basis of maximum number of votes final output will be taken.

4. **Neural Networks:** It is used to analyze complex datasets and identify patterns. In healthcare, neural networks are often used to identify patterns for medical images.

5.**Naive Bayes Classifier:** Classification of data points based on the probability of certain characteristics is done by this algorithm. It can be used to classify patients based on their historical data like symptoms etc.

For a vector $\chi = \chi_1, \chi_2, \dots\dots\dots\dots\dots\dots\dots\dots\dots\chi_p$ and a class variable ς_κ

It states that $P(\varsigma_\kappa / \chi) = \frac{P(\chi/\varsigma_\kappa)\,P(\varsigma_\kappa)}{P(\chi)}$, for $\kappa = 1,2, \dots\dots\dots\dots\dots\dots\dots, \kappa$

$P(\varsigma_\kappa / \chi)$ is known for Posterior probability, $P(\chi/\varsigma_\kappa)$ is the likelihood, $P(\varsigma_\kappa)$ and $P(\chi)$ the prior probability of class and predictor respectively

Then by chain rule it can be decomposed as

$P(\chi/\varsigma_\kappa) = P(\chi_1, \chi_2, \dots\dots\dots\dots\dots\dots\dots\dots\dots\chi_p/\varsigma_\kappa) =$

$P((\chi_1/\chi_2, \dots\dots\dots\dots\dots\dots\dots\dots\dots, \chi_{p,}\varsigma_\kappa)P((\chi_2/, \dots\dots\dots\dots\dots\dots\dots\dots\dots, \chi_p, \varsigma_\kappa)P((\chi_{p-1}/\chi_p\varsigma_\kappa)$

6. **Linear Regression:** It is the problem of finding a relationship of the form

$$\Upsilon = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \cdots\dots\dots\dots\dots\dots\dots\dots\dots\dots\beta_p\chi_p$$

Where the $\chi_1, \chi_2, \dots\dots\dots\dots\dots\dots\dots\dots\dots\chi_p$ are covariates that describes certain system and the response in shown by Υ.In this the set of input output pairs (χ^i, Υ^j)arranged in the form of

matrix χ and γ then the correct $\beta = (\beta_0 + \beta_1 + \beta_2 + \cdots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . \beta_p)^T$ can be forecasted by solving the least squares optimization problem.

7. Gradient Boosting Decision Trees (GBDT) algorithms:

In the given dataset $\wp$ and a loss function $\mathcal{L} : \mathbb{R}^2 \longrightarrow \mathbb{R}$ the gradient boosting algorithm iteratively constructs a model F : $\varkappa \longrightarrow \mathbb{R}$ to minimize empirical risk $\mathfrak{R}_{\wp}[\mathcal{L}\ (f(x), y)]$

At each iteration t, the model is updated as

$F^{(t)}$ (x) = $F^{(t-1)}$ (x) + $\mathcal{E}$ $h^{(t)}$ (x) where $h^{(t)}$ (x) is a weak learner chosen and

$h^{(t)}$= arg min h$\mathcal{E}$ $\mathcal{H}$ $\mathfrak{R}_{\wp}$[$(-g^t\ (x, y) - h(x)^2$

1.15 Advantages and Disadvantages of data mining

- It draws out the best from big data so it helps companies to gather assured information.
- It is an efficient and time saving techniques compared to another techniques as it is the speedy process by which user can easily access huge amount of data in less time.
- It provides various business opportunities and helps to make profitable production.
- Data mining helps us to create data models by which one can easily identify the customers who are interested in their products.
- Data mining techniques can used to build risk model and hence it can improve safety.
- It builds models on existing platforms and can be implemented on new systems.
- It helps the overall growth of a company.
- It helps to compare with the competitors on the basis of information gathered, hence it enables us to understand better our customers and strengthens our brand.
- It maintains accuracy as it draws out the information from large amount of data without any problem.
- The collected information from many sources like market analysis, governments and other institutions arranged properly and analyze the past transactions of customer by which fraud and credit risk can be easily detected.
- It is an easy technique for forecasting of various fields.

Data mining also have few weaknesses as nothing is perfect so there are few challenges of implementation of data mining

- It is little bit complex so need skilled experts to formulate data mining queries.
- There may be the problem of over fitting.
- It is the process to discover knowledge from large data hence sometimes difficult to manage large data.
- Databases have heterogeneous type of data so sometimes it may become more complex in process of data integration.
- There are many data mining tools for different type of data therefore selection of appropriate tool is a very difficult task.
- Sometimes these techniques do not give correct result so one must not fully reliant on it.

1.16 Tools for Data Mining:

Data mining consists of big data which is not possible through pen paper works so there are various software which make easier task of forecasting. There are various software for forecasting, In general there are two types of tools Open source tools and Paid tools which are explained as below:

- **Open Source Tools for Data Mining:**

1. **Jupyter Notebook:** It is a web-based open-source platform that provides a wide range of tools for data mining and analysis. It can be used to explore data, create visualizations, and communicate insights. With the help of Jupyter Notebook, data scientists can quickly access and analyze large datasets, develop and test machine learning models, and share insights with colleagues. Additionally, Jupyter Notebook supports multiple programming languages which make it an ideal platform for data mining and analysis.

2. Rapid Miner: It is a powerful data mining platform that enables users to build predictive models and analyze data quickly and easily. It is an easy-to-use, comprehensive platform that provides a wide range of data mining functionality, including data pre-processing, machine

learning, statistical analysis, text mining, and visual analytics. Rapid Miner also includes advanced features such as automated modelling, model comparison, and model selection. With its intuitive graphical user interface, users can quickly create models, explore and visualize their datasets, and gain meaningful insights from their data. Rapid Miner is a great tool for both data scientists and business analysts alike, and is used by thousands of organizations worldwide.

3. Weka: It is a software suite for data mining and machine learning. It provides a graphical user interface for data mining, allowing users to develop models using a variety of data mining algorithms. It also includes attribute selection, pre-processing and visualization capabilities. Weka supports many tasks of association various programs.

4. Scikit-learn: It is one of the best Python libraries and designed for the easy use and efficient result and it is built on top of the popular NumPy, SciPy, and matplotlib libraries

5. PySpark: PySpark is an open source library for Python that provides a powerful set of tools for data mining and analysis. It is designed to be easy to use and integrate with the popular Python libraries such as NumPy, SciPy, and matplotlib.

6. **KNIME:** KNIME is powerful software that supports a wide range of algorithms and techniques. It is designed to be easy to use, with an intuitive graphical user interface, and can be extended with plugins.

7. Orange: Orange is open-source software for data mining and machine learning. It contains a range of machine learning algorithms, visualisation and data exploration tools, and is easy to use for predictive modelling and other tasks. It is available for Windows, Linux and Mac OS X. Along with the above Apache Mahout, Rattle, ML-Flex, Scikit-learn, ELKI, Apache Cassandra are also the open source tools for data mining.

- **Paid source tools for Data Mining**

1. SAS: SAS offers a comprehensive suite of software for data mining, predictive analytics and data exploration. The software includes powerful tools for data preparation, visualization and analysis, as well as comprehensive statistical, data mining and machine learning algorithms.

2. IBM SPSS: SPSS (Statistical Package for the Social Sciences) is a data mining software package that can be used for a wide range of data mining tasks. It is a powerful tool for predictive analytics and can be used for descriptive, diagnostic, and predictive analytics. It can be used to perform a wide range of analytics tasks including data exploration, predictive modelling, cluster analysis, forecasting, and more. Data mining with SPSS requires knowledge of the software and statistical analysis to gain insight from data. The software can also be used to create descriptive and predictive models, which can then be used to make decisions based on the results.

3. Minitab: Minitab is a software package used for data analysis, statistical process control, and graphical analysis. It is one of the most popular software packages for data mining, and it offers a variety of features that make it an ideal tool for uncovering patterns and relationships in large datasets. It's easy-to-use interface and powerful analytical capabilities make it an excellent choice for data mining and predictive analytics. Minitab is also an excellent choice for modeling and optimization, providing powerful algorithms and interactive tools that can help users quickly identify the most promising solutions.

4. Matlab: Matlab offers a wide range of data mining tools and functions that can be used to identify patterns and relationships in data and to create predictive models. It includes algorithms for clustering, classification, regression, feature selection, and time series analysis. It can also be used to visualize data and to explore data interactively. Additionally, Matlab's open source toolboxes offer users the ability to access and customize various data mining algorithms.

Microsoft SQL Server Analysis Services, Tableau, Google Cloud Data Mining, BigML, Oracle Data Mining are also paid software for data mining.

1.17 Composition of Thesis

The report in this thesis has analysed the improved forecasting models using data mining in healthcare, agriculture and for electricity demand with the Python software and jupyter notebook works as user interface for the best interpretation of results. The composition of thesis is as follows:

The first chapter provides the basic knowledge and preliminary concepts of forecasting, its models and techniques. The second chapter deals with a literature review of the existing studies related to the study of forecasting models using data mining in various fields. The third chapter focuses on role of data mining in healthcare as the third chapter has taken two case studies, first is of malaria and second of conformed corona cases while the fourth chapter deals with Regional frequency analysis approach for rainfall analysis while the fourth chapter studies the regression models in the agricultural field and the subsequent chapter focused on development of stochastic mathematical model for forecasting. Moreover, the summary and conclusion are given in chapter seven.

1.17.1 Objectives of research work:

1. To forecast COVID-19 cases by using improved random forest algorithm.
2. To study and analyse various data mining techniques in healthcare.
3. Application of classification and regression techniques in different fields.

1.17.2 Schematic diagram for composition of thesis:

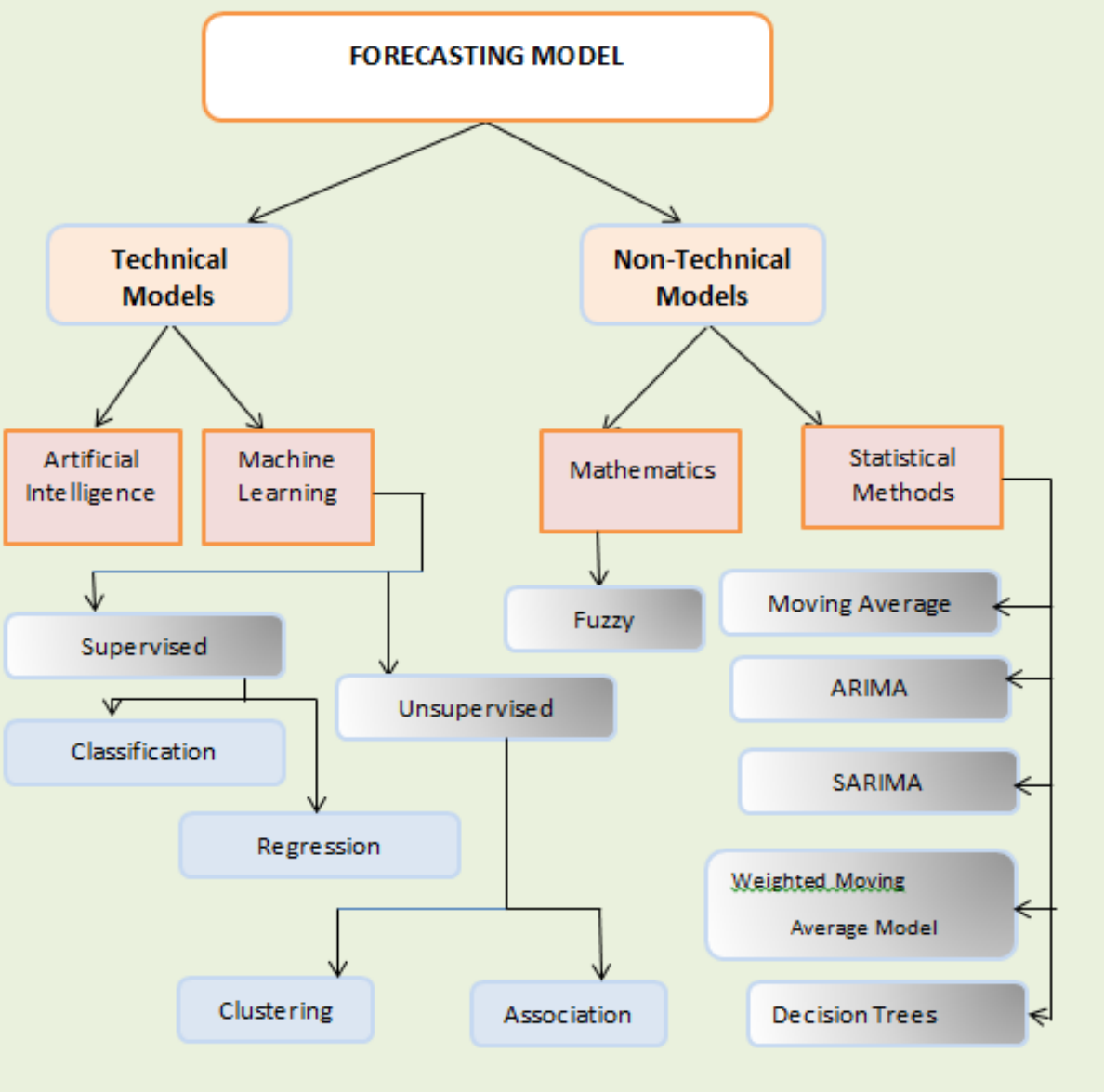


Figure 1.16 Forecasting models

Chapter II

LITERATURE REVIEW

In estimation and forecasting, there have been many developments over the past decades which show direct relation and applicability for forecasting of various fields. Literature review is the study of all concerned computational and theoretical study of authors which could be implemented in the proposed study. This chapter deals with the related studies of forecasting in various fields which enhances the ideas and helps to find out the research gap.

2.1 Data mining for the forecasting of malarial disease:

Lafferty, K. D. (2017) [1] discussed various marine disease which plays major role as the infectious agent and the author used machine learning algorithms and techniques to know which parasite will increase or decrease and thus causes transmission of various disease. The real case study of California is taken and a cor-relation is finding out by the author in between parasite and the food eaten by them which is known as the marine food. Moreover the infectious disease theory is also analysed and the results also find out by using various machine learning algorithms. Tai & Dhaliwal (2022) [2] tried to Know fully about malaria risk so to find out the genetic reason for individuals malaria risk take genetic variation data. In this the genetic algorithm applied on the dataset which consists of 20817 individuals data and to check the models accuracy mean absolute error evaluation metric has been applied and the individuals were at higher risk who have the MAE score in between 8.00E-06. Garrido C J A et al., (2019) [3] try to find out the global trend followed in terms of transmission of malaria. During the last so many years it is observed that P. vivax is the major causing agent of malaria in the Latin America, south and south east Asia. In this the authors has taken the research papers from Pub med, Scopus, web of science and Google scholar. Firstly the author searched the papers on Elsevier Scopus database by putting keywords malaria and vivax then from 1916 to 2018, 11166 documents came out but as per need and after proper analysis the author tried to find out the trend analysis during these years. The authors observed that in 1946 due to a virus chloroquine found which a causing agent for P. vivax and in 1973 there was intensification in the disease so it leads the enhancement in number of publication during

these years which was trend breaking. The author conclude that to get eradicate these types of disease the support of healthy countries like India, USA, UK is required and the role of anti-malarial were indispensable and the drugs recommended by WHO also shows the promising results for the removal of P.vivax and malaria. The effect of climatic conditions on malaria is studied by Kumar V et al. (2014) [4] With the various ARIMA model. On the 8 year's dataset of 12 months which consists of 14 rows and 10 columns. It shows the monthly malarial cases of Delhi from 2006 to 2013. In this by applying various values of P,d and q to know that which ARIMA is suitable for the disease then the author applied ARIMA (0,1,0) model to predict the malarial disease and its shows better result than others. Apart from, statistical analysis it also find out the impact of climate on transmission of disease and it is seen that when humidity is atleast 60% only that this disease can occur and as the temperature rises it is also expected to increase. All the methods and types of Arima shows good results. Patil and Gawande (2022) [5] used hybrid model under ROC curve on big data. The data was taken from Boston Medical center. The author used the alteration method of clustering and classification then by re-clustering for classifiers, classifies a new samples and checked its accuracy by using detection rate which is also known as sensitivity. Talapko, J et al., (2019) [6] focuses on the diagnostic testing which can diagnose the malaria this diseases many times may not be diagnosed due to which many people lost their lives. The tests discussed in this study were Light microscopy method which is also known as gold standard method used by Giemsa. For the detection of antigens Rapid Diagnostic test and even this test becomes the first choice of all and also recommend by WHO all across the world.To diagnose the presence of Parasite antigens immune chromatographic tests used. Immunofluorescence antibody testing used serologic testing for the diagnosis of malaria and polymerase chain reaction method is also very common for this. Malaria in Europe, Croatia, slovina and many more countries has been taken in consideration and it is concluded that all the antimalarial drugs are effective and the number of resources available are quite sufficient to eradicate this malarial disease.

2.2 Forecasting model to predict the Diabetes Mellitus

Sometimes the body doesn't make insulin then it is known as TYPE 1 diabetes and when body does not use the insulin produced is known as TYPE 2 diabetes. It may cause many problems such as heart disease, eye problems, kidney problems and many more. Increase in

body fat and indolent is the main cause for TYPE 2 diabetes which is the most common type of diabetes in most of the people. Various researchers are contributing in this field for the mankind. In this the analysis and forecasting are carried out in various fields by using various techniques. Eswari T et al. [7] uses Hadoop and map reduces method for the diabetes forecasting. In this they describe the types of diabetes amd methods for forecasting of type of diabetes, along with the symptoms and risk factor associated to each type of diabetes. Butwall M., and Kumar S.,[8] used various classifier to predict diabetes behaviour and to improve the accuracy it implements various methods even used the combination of methods also which is known as hybrid method. The method which shows the highest accuracy is random forest classifier and in this the number of trees has been increased to gain highest accuracy. Faruque and sarker [9] studied the disorders and risk factors by exploring various techniques of machine learning. It extracted the knowledge from the collected dataset and this research mainly focusses on the risk factors associated to these diseases. The author also employs NB, SVM, C4.5 decision tree, K-nearest neighbour on adult population diabetes dataset. Woldemichael F and Menaria S [10] discussed Pima Indian diabetic patient's forecasting by using back propagation method by implementing R Studio under R programming language. This research work also forecasted the diabetes by using J48, Naïve bayes, Support vector machine and neural network. In neural network, the author uses an input layer with 8 parameters and a hidden layer having 6 neurons which produces one output. In this research paper, the Hina S et al [11] forecasted the diabetes by using various data mining techniques like J48, ZeroR, Naïve bayes but in this research paper the author more focusses on the type of data. The researcher analysed and modified the data for better forecasting. Mujumdar, A., & Vaidehi, V. [12] imposed a new pipelining method on improved dataset. The forecasting has various factors on which it depends like dataset which consist of various attributes and applied methods. In this research paper the author shows the comparison between the prediction of diabetic patients by using the common dataset and in between the prediction of diabetic patients by using modified dataset.

2.3 Forecasting Models for the study of Electricity demand

Electricity demand is increasing. To meet the challenges which can occur in this line is the demand of this research that is why many researchers are focusing to forecast the power demand. Researchers forecasted power generation and power demand by using various

predictive models also checked the models adaptability whether it can help to forecast demand in other energy sectors. Tisdell C et al. [13] tested many machine learning models for same problem and in this the kernel ridge and multilayer perceptron shows the best results which was 13.9% improved model than the basic regression model [14]. Markovics D [15] used improved model to forecast electricity price and compared 27 models then it shows that the statistical methods outperformed. Saravanan S et al. [16] used various machine learning techniques to forecast India's annual energy consumption for 9 years starting from 2011 and much focused on Artificial neural network. Furthermore Kumar R [17] used only one method for forecasting as well as for accuracy which is MAPE which is also used by was used by Bhowte Y W et al. [18] to develop time series models of third category. S. Sravanan [19] used various models like RA, ANNs, PCR and PC-ANNs then in this study PC-ANN was concluded as the efficient method for forecasting. Dushant P et al. [20] used basic grey model with Bernoulli differential equation which further known as NGBM. Jatin Bedi et al.[21] in 2018 used non – linear complex data to predict electricity demand with power demand model. It is predicted by Inglesi R [22] in South Africa by Engle- Granger method and cuckoo search algorithm by Jiang P [23]. NMGM [1,1,β] developed by Wang and song [24] to predict oil demand in china.LS-SVM was used by Xie and Wang [25] in 2015 to predict electricity demand and leap then Emodi et al [26] used this model to predict electricity consumption of Iran. IRGM (1, 1) improved model used by Ning et al [27] for the prediction of electricity demand to compare with basic GM (1,1) model and optimized initial condition. Sifeng Liu et al [28] explained the problems of robust stability to explain the grey system theory.

2.4 Forecasting for crop analysis

Dharma raja et al. [29]. However with the rise in human population and decrease in fertile and cultivable land, time has come for government and other stakeholders to go for smart agriculture. The best farming practices of "Crop Yield Production" shows the good results as well as its impact on country's economy. Salpekar [30] discussed that forecasting of crops yield depends on various factors. This type of forecasting helps farmers and agricultural experts to plan crop production and make decisions regarding crop management. It also helps governments and other organizations to plan food security policies and develop sustainable agriculture practices. Various techniques such as remote sensing, data mining, machine

learning, and predictive analytics are used to make accurate crop yield forecasts. The effect of humidity on crop yield is complex and depends on many factors, including the crop species, soil type, water availability, and temperature is shown by Bang et al. [31]. Generally, higher humidity leads to higher rates of transpiration, which can be beneficial for some crops, but can also reduce yields for other crops. Hu et al. [32], Choudhury and Jones [33] explained that a higher relative humidity may increase water stress in crops such as corn and leads to reduce yield. On the other hand, higher levels of humidity may benefit crops such as tomatoes which thrive in warm, humid conditions. Therefore, the effect of humidity on crop yield will depend on the particular crop and environmental conditions. Li et al. [34], Chen et al. [35] used various remote sensing ways like multi temporal techniques to forecast crop yield. This study developed the Arima model which is also known as Box Jenkins model discussed by Hamjah [36] to forecast the crop yield of Jammu Kashmir. A potential ARIMA model for crop yield could be ARIMA (2,1,1). The two autoregressive terms would look at past values of the crop yield to help predict future values, while the differencing term would look at the change in crop yield over time to make more accurate predictions. The moving average term would help to smooth any random fluctuations in the data. Next is the multiple linear regression model which is used to measure the effect of a change in one or more independent variables on the dependent variable discussed by Conradt [37] also while controlling the effects of other independent variables. By using multiple regressions, it is possible to determine which of these factors have the most significant effect on yield and how much of an effect each factor has. This can then be used to inform decisions about crop management and to predict future crop yields. The model is then tested against actual yield data to determine its accuracy. Liu *et al*. [38] also explains that this method can be used to help farmers to optimize their crop production in order to maximize yield and profitability . For empirical analysis, the data of 40 years has been taken from Jammu and Kashmir (Agricultural Statistics at a Glance).

Spiertz and Ewert [39] discussed the crops production, resources for food and ways to avoid negative impacts on food security and energy .This study further focuses on the changes required in technology for the positive impact on agricultural sector. . Rise in Crop yield is the positive approach to meet food and fuel demand .In this study the focus is on implementation of new technologies by the government in the field of bio-energy.Balakrishna et al., [40] proposed single shot detection algorithms to detect animals and objects through images. It is

based on the machine learning and artificial intelligence in which it helps to the farmers to take decision for the protection of their fields. Sadenova et al., [41] used various methods for crop production and its yield. This study shows the reasons behind the less productivity. It focuses on digital technology as this proposed methodology is to forecast the remote crops in agro-ecosystems so these remote areas require the advancement in the form of digital tool implications in field of agriculture. Mishra et al., [42] reassess the studies based on machine learning models for agricultural crop production and list out the factors responsible for high yield and low yield. This study also shows the various methods for various crops. Chen et al., [43] developed the rice crop monitoring system. In this in between two seasons many type of cultivation can be done. The accuracy was achieved by neural network which was the proposed a system and this study compares the accuracy with govt. statistics. This study also measures various factors like environmental factors which are unpredictable also forecasts rice exports from Thailand by comparing study between the machine learning models by Co Ch et al., [44]. Garov, A. K., & Awasthi, A. K,,[45] discussed the main agricultural problem which can be seen mostly in the indian crops. It is seen that the crops gets affected by various disease and in this study, the main focus was on cotton crop and used machine learning algorithms along with few image processing methods for the detection of crop disease by Nischitha, K et al [46] also. The authors studied about the prediction of various diseases and then consider it as the prime duty to get rid of them. Van Klompenburg T et al., [47] deeply surveyed the 50 machine learning based papers and 30 papers based on deep learning approach. Initially 567 documents studied for the crop yield prediction along with environmental factors affecting the crop yield. Many of the authors used LSTM, Neural network, artificial intelligence, Hybrid networks, Multi task learning, deep recurrent q-networks for the study and during the review various evaluation parameters used frequently and in which the most common is root mean square error which was used 29 times in the study then the R squared error 19 times and the least estimated parameters used one which is four in number. The author than concluded that CNN and LSTM are most effective models for the crop yield prediction. Nigam A et al., [48] used very simple two models long short term memory and recurrent neural network based on four major factors. The author used confusion matrix to forecast and compare it with other models like regression methods and various classifiers for forecasting. In this the parameters are studied in isolation also and in

hybridisation also then the LSTM model outperformed for temperature prediction while random forest classifier shown highest yield accuracy. This study can help the farmers to choose which type of crop needs what type of climate or farmers can easily choose the crops based on climatic factors to get the maximum yield. As the climatic factors affects the yield so the regional distribution for rainfall is applicable everywhere the region after scaling (Gabriele & Arnell [49]). Regional frequency analysis involves; Formation of homogeneous groups/regions, data screening, choice of regional distribution, and quantiles estimation. Selection and parameter estimation of regional distribution are the most important components in regional frequency analysis. The method of moments is a traditional method for parameter estimation and it is easy as compared to other methods, but all moments do not exist every time. The maximum likelihood method is often considered the most effective method for parameter estimation. However, the maximum likelihood estimators may be difficult to compute and require numerical algorithms, particularly when the number of parameters is large. Introduction of L-moments estimators for parameter estimation of many probability distribution functions which are an exact analog of L-moment estimators and that are often superior to estimators obtained by the method of moment and maximum likelihood method for regional studies (Hosking, [50]; Hosking & Wallis [51]). The L-moments are easy to calculate and it is seen that they provide efficient estimators of parameters of distributions (Stedinger *et al*., [52]). In Haryana past studies related to rainfall and its probability distribution include Hooda [53]; Kumar *et. al* [54]; Babu & Hooda [55]; Nain & Hooda [56]. Recently, Nain and Hooda [57] successfully applied this approach and estimated rainfall quantiles for the maximum monthly rainfall data of Haryana and recommend it for seasonal as well as annual total rainfall in the state by Greenwood L et al [58] also. Keeping this fact in mind and the increasing popularity of L-moments among hydrologists and its wide applicability in hydrology and meteorology. To find the best-fit regional frequency distributions of annual total rainfall in Haryana, to estimate regional parameters of selected regional frequency distributions via L-moments [59], estimate the return period of the rainfalls[60], and Validate the result using Monte Carlo Simulation[61].

2.5 Data mining for Covid-19

Coronavirus is not a new virus even it come across in 1960's and there are various forms of virus as due to mutation it changes from one form to another form. In the field of research, various researchers are working on it for a different type of issues. It is observed that this virus is not only bound to human beings but the infection passes to birds and animals [66]. The coronavirus was declared as pandemic [67]. The rise in cases is extremely high, it was an exponential growth in the number of cases. Then after a high rise, it starts sigmoid growth. The first and second wave period was between 15th march and 30th June,1st July and 15th October respectively. And the period when corona estimated and again when it after saturation it fall down is the critical period [68] . Thus, this is the period where the health industries have to manage the system. Bambrick et al. [69] discussed various data mining techniques and analyzed the uniqueness of medical data mining, selection of the data and algorithms for the comparison between different algorithms on the dataset of near about three months. The author shows that the accuracy of forecasting much depends on type of data and the validity of data. Thus, in this study the main focus was on the data validation and then the effectively data analysis has been taken by using various machine learning algorithms. There is an incubation period which is considered here as the lag period and so the author effectively evaluated lag effect with cross correlation functions by using time series generalized linear model (GLM) and generalized additive model (GAM). Temperature effects on COVID-19 in transmissibility it shows that as the temperature decreases the transmission increases. Ahmad et al. [70] uses four stages formula to predict confirmed cases. This being a review paper can help others for prediction of confirmed cases as thus study suggests that there should be more focus on data collection so it is necessary to collect data from exact and assure sources. If The data can be quantified then the forecasting results are near to perfect but in this the data is from various health centres as well as from the society and the social data can't be quantified which was a limitation for whole during forecasting on COVID-19. So to overcome this, models trained on various types of data and then hybrid models chosen as the best models for forecasting with different types of data. Siddiqui et al. [71] shows that along with temperature there are many factors affecting the spread of COVID-19 pandemic. The author tries to find out the correlation between temperature and Covid-19 for all the three cases i.e., confirmed, suspected and death case. In this study, the temperature is added of various cities in dataset for

the analysis and to explore various trends. In this for analysis the K- means clustering method has been employed which shows the better results and along with this it was seen that there was diverse nature of trends in each city and only the solution was lockdown. Atef M et al.,[72] used three classification methods KNN, SVM, MLP for the covid-19 patients. In this study the authors used the dataset consists of 300 rows based on seven symptoms and this is to find the correlation between these features. The author did analysis by building SKlearn in Python 3 under pycharm. Due to not very large amount of data, the forecasting accuracy is very high and it enables to predict those people who were under very high risk and having critical conditions. COVID-19 forecasting by using neural network and combination of two models to make hybrid model was the proposed study of Namasudra et al. [73]. In this the author taken the dataset of various countries and discussed the incubation period along with predicted the confirmed and recovered cases. The author used various evaluation parameters to check the accuracy of individual model as well as hybrid model then rather CHAID, CART outperformed. Chimmula, V. K., and Zhang, L. [74] used hybrid model of artificial intelligence and deep learning for the forecasting of covid-19 transmission. Based on LSTM model, the author compared the covod-19 transmission with other countries as in this study the transmission rate of Canada is compared with USA and Italy .Moreover along with this the author also forecasted the number of patients on 2^{nd}, 4^{th}, 6^{th}, 8^{th}, 10^{th}, 12^{th} and 14^{th} day for two successive days which can help to the whole society to prevent themselves from it. Fanelli D and Piazza F. [75] did forecasting analysis only for three countries Italy, France and China by using various machine learning algorithms. In this the dataset of John Hopkins University from 22/02/20202 to 15/03/2020 has been taken and I is observed that on the basis of various factors like cultural, geographical by using first approximation is considered. By using SIRD (susceptible infected recovered deaths) model, the Italy and France shown same recovery rate and it is estimated that in the very peak time of transmission of coronavirus, the 2500 ventilation units are required. In this the author used the developed SIRD model, in which the infection rate is denoted by I and it is assumed that it may varies with time T*

$$I(t) = \begin{cases} I_0 & for\ T \leq T^* \\ I_0(1-\alpha)e^{-\frac{T-T^*}{\Delta T}} + \alpha I_0 & for\ T \geq T^* \end{cases}$$ and $I_0 = 7.90 \times 10^{-6}$ per day is the rate of transmission estimated.

Maleki M et al [76] used classical time series which requires the symmetry of errors distribution and this distribution and this distribution error consists of two –piece scale normal mixture. So, the author proposed simple time series model which is well suitable in all the conditions. In this the author used proposed model to forecast the number of confirmed cases of Covid-19 in the world. Thus , the author from this study shows that the proposed model is better than Gaussian time series model.Kumar M et al., [77] studied various methodologies which effectively detects the person suffering with COVID-19 based on various symptoms. In this dataset of images has been used like the chest x-rays and CT scan images. The author even improved the model by using A.I to enhance the use of this to know more effectively about this dreadful disease. In this study, the author even tries to find the cost-effectiveness during the treatment for the people who claims insurance and for their settlement to know financial implication were also one of the aims of this study. Rahimi I et al [78] verbosely explained the influential tool for bibliometric analyse from Scopus and web of science and in second section the authors addressed the models along with estimation of models . In this the author taken the papers of covid-19 from various subjects based on the interest and then in detail explained the methodology used in the papers. To extract out the knowledge from big data the author sets the parameters and then again verbosely explained the SEIR model i.e., susceptible, exposed, infected, and recovered. In this the susceptible gets exposed and then it becomes infective and in the last step the infective person sometime may get recovered or leads to death or it may quarantined to prevent from the transmission of infection. In this the author shows and compares various models for forecasting of covid-19 cases and it is suggested that rather than one model hybrid approaches must be inculcated for the investigation.

Chapter III

Role of data mining in healthcare

3.1 Introduction: Data mining in healthcare is becoming increasingly important as medical data continues to grow and become more complex. Data mining can be used to uncover relationships and patterns in medical data that would otherwise be too complex to identify. Thus, Forecasting in healthcare is a key tool for managing resources, predicting patient demand, and ensuring a safe and efficient system of care. Accurate forecasting can help hospitals and health systems anticipate the need for additional staff, supplies, equipment, and services. It can also enable them to better identify and address potential issues that may affect patients, such as staffing shortages, wait times, and overcrowding. This can help medical professionals make more informed decisions, predict outcomes, and improve patient care. Data mining can also be used to improve the accuracy of diagnoses and create more effective treatments. Additionally, data mining can be used to improve operational efficiency, reduce costs, and uncover fraud and waste in healthcare. Forecasting can also help healthcare organizations identify areas where additional resources may be needed, such as new technology, policy changes, and/or training. Ultimately, forecasting in healthcare can reduce costs, improve quality of care, and ensure patient safety

In healthcare, these models can help to predict a variety of outcomes, from patient volumes and readmission rates to cost and utilization.

I. These models can be used to forecast expected patient volumes for a given period of time.

II. This can be done by taking into account historical patient volumes, seasonal trends, and other factors that may affect patient volumes.

III. It can be used to forecast patient’s volumes and help healthcare organizations plan for staffing, resource allocation, and other related needs.

IV. This model can be used to forecast readmission rates. This can be done by taking into account factors such as patient demographics, health conditions, and treatments.

V. The model can then be used to predict future readmission rates and help healthcare organizations plan for interventions that can help reduce readmission rates.

VI. Forecasting models can also be used to forecast cost and utilization. This can be done by taking into account factors such as patient diagnoses, treatments, and procedures.

VII. Various models can then be used to predict future cost and utilization and help healthcare organizations plan for budgeting, resource allocation, and other related needs.

VIII. Overall models can be used to forecast a variety of healthcare outcomes.

IX. This can help healthcare organizations plan for staffing, budgeting, and resource allocation needs, as well as develop interventions to reduce readmission rates.

X. Forecasting models in healthcare are very promising. It can provide a powerful tool for healthcare professionals to identify, predict and manage trends in healthcare data.

XI. Medical practitioners can use these models to identify and predict the impact of new medical treatments, the progression of medical conditions, and the expected outcomes of patient treatments.

XII. It can also be used to analyze the cost of healthcare services and provide insights into cost-effectiveness.

XIII. This can help healthcare providers to optimize their resources and improve the cost-effectiveness of their services.

The uses of these models are also beneficial to the healthcare industry in terms of identifying and predicting the impact of emerging health trends. With the availability of large amounts of data, Machine learning models can be used to identify new trends in healthcare and provide insight into how these trends may affect the industry. Various models can also be used to improve the accuracy of predictive analytics in healthcare. With the development of new technologies, predictive analytics can be used to identify and predict the outcomes of medical

treatments and patient care. Forecasting using data mining can help healthcare providers to make more accurate predictions and improve patient outcomes. Lastly, these models can be used to monitor patient health over time and identify any changes in health status. This can help healthcare providers to provide more effective and personalized care to their patients.

Table 3.1 Comparison of data mining technique with respect to traditional technique

S.No	**Types of disease**	**Data mining tool**	**Technique**	**Algorithm**	**Traditional method**
1	Tuberculosis	WEKA	Naïve Bayes Classifier	KNN	Probability Statistics
2	Heart Disease	ODND,NCC2	Classification	Naïve Bayes	Probability
3	Kidney Dialysis	RST	Classification	Naïve bayes	statistics
4	Diabetes Mellitus	ANN	Decision tree	C4.5 Algorithm	Regression
5	Dengue	Dengue	Classification	C5.0	Exponential
6	Blood Bank Sector	WEKA	Decision tree	J48	Moving average
7	Hepatitis C	SNP	Decision tree	Gain	Delphi method

In this chapter we are dealing with two cases i.e., Data mining based neural network model for Forecasting Malarial disease and Rf-DT model for COVID-19 patient's Recovery.

Case 3.1.1 As, Malaria is a major public health issue in India. According to the World Health Organization (WHO), India accounted for the highest number of global malaria cases and deaths in 2017. In India, malaria is mostly concentrated in the states of Odisha, Jharkhand, Chhattisgarh, Madhya Pradesh, and Maharashtra. In 2017, over 6.6 million cases of malaria were reported in India, out of which 1.3 million were confirmed by laboratory tests. The number of deaths due to malaria in India was estimated to be at least 2,000 in the same year. The government of India has taken measures to reduce the impact of malaria in the country. The National Vector Borne Disease Control Programme (NVBDCP) is a multi-sectoral programme implemented by the government to control vector-borne diseases such as malaria. This programme includes vector control measures, preventive and curative measures, and health education. The government also provides free diagnosis and treatment for malaria in all public health facilities. In addition, the government has launched a national campaign called 'Eliminate Malaria' to raise awareness about the disease and its prevention.

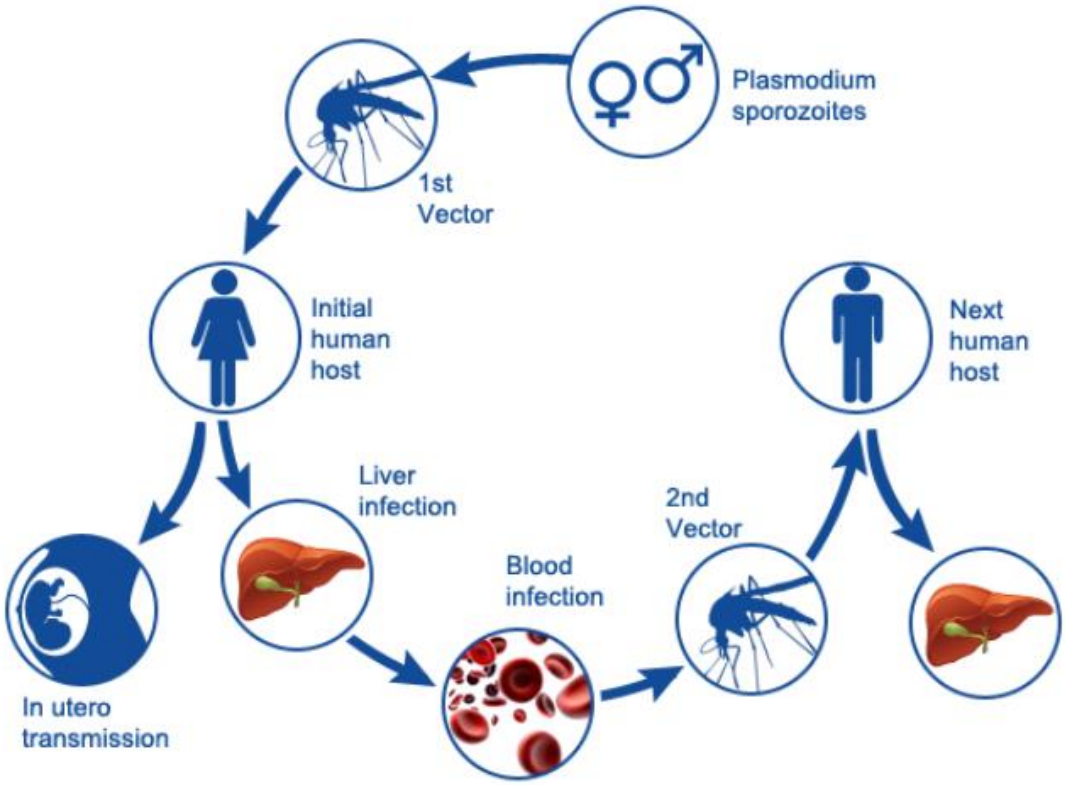


Figure3.1 Phenomenon for transmission of malarial parasites

3.1.2 Data mining and Malaria: A correlation

Data mining can be used for malaria forecasting by analysing large amounts of historical data and then predicting the prevalence of the disease in certain areas. Data mining can be used to identify areas that are prone to malaria outbreaks and to determine the risk factors associated with the disease. Data mining can also help to identify temporal patterns in the spread of malaria, as well as spatial patterns in the distribution of the disease. This information can be used to develop strategies for the prevention and control of malaria. In this the various types of models used for analysis which are:

1. Time series analysis: Time series analysis is a statistical method that is used to analyse data points collected over a period of time. It is used to explore and understand trends, patterns, and seasonal variations in data, and to make predictions. The goal of time series analysis is to identify hidden relationships and patterns in the data that can help inform decision-making. Time series analysis can be used in a variety of applications, such as predicting future stock prices, forecasting sales, and analyzing customer behaviour.

2. Stationary Time series analysis model:.Stationary models are typically expressed as a linear combination of two or more random variables. The equation for a stationary model is expressed as: $Y(t) = \alpha + \Sigma\ \beta_i X_i(t) + \varepsilon(t)$Where:$Y(t)$ is the output of the system at time t, α is a constant, $X_i(t)$ is a random variable at time t, β_i is a coefficient, $\varepsilon(t)$ is a random error term at time t.

The coefficients β_i are estimated using a process called regression analysis. Regression analysis is used to estimate the coefficients βi by minimizing. Once the coefficients have been estimated, the model can be used to make predictions about future behavior.

3. Moving Average: The moving average is one of the most widely used technical indicators because it is simple to calculate and interpret.

For a simple moving average, the equation is: $MA = (P_1 + P_2 + P_3 + \ldots + P_n) /n$ Where: MA = Moving Average P_1 = Price of the first period P_2 = Price of the second period P_3 = Price of the third period … P_n= Price of the nth period n = Number of periods

4. Sarima model: A SARIMA (Seasonal AutoRegressive Integrated Moving Average) model is a type of time series forecasting technique used to model and forecast seasonal effects in data. It combines the autoregressive (AR) and moving average (MA) models which are widely used in time series forecasting. It is often used to model seasonal trends and long-term cycles in data. Sarima model can be represented mathematically as:

ARIMA(p,d,q) x (P,D,Q)mwhere p, d, and q are the parameters of the ARIMA model; P, D, and Q are the parameters of the seasonal component and m is the seasonal period.

3.1.3 Feed forward neural network

It is called "feed-forward" as the connections are not connected to each other and donot form a cycle. In a feed-forward neural network, each node in a layer is connected to every node in the next layer. Information is fed into the input layer, which simply passes it to the next layer. Each subsequent layer performs computations on the inputs it receives and passes the results to the next layer until it reaches the output layer. Each connection between nodes in a feed-forward neural network has a weight associated with it, which determines the strength of the connection. The output of each node is computed by applying an activation function to the weighted sum of the inputs. Common activation functions include sigmoid, ReLU, and tanh.

Training a feed-forward neural network typically involves adjusting the weights of the connections between nodes based on a desired output. This is done through a process called back-propagation. In this the connections between nodes do not form a cycle. It is called "feed-forward".

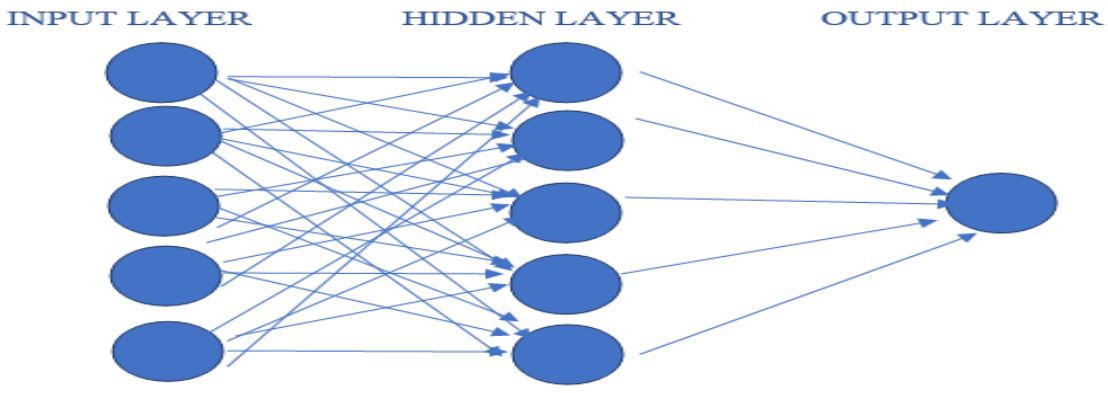


Figure3.2 Feed forward neural network

The above feed forward neural network consists of three layers in which Information is fed into the input layer, which simply passes it to the next layer. Each subsequent layer performs computations on the inputs it receives and passes the results to the next layer until it reaches the output layer. Each connection between nodes in a feed forward neural network has a weight associated with it, which determines the strength of the connection and there are few processes like adjusting the weight which is done in hidden layer and thus, the output of each node is computed by applying an activation function to the weighted sum of the inputs. Common activation functions include sigmoid, ReLU, and tanh. Training a feed forward neural network typically involves adjusting the weights of the connections between nodes based on a desired output. This is done through a process called back propagation, which involves calculating the error between the network's output and the desired output and then adjusting the weights to minimize this error network's output and the desired output and then adjusting the weights to minimize this error.

3.1.4 Methodology: The machine learning models has been applied and the given flow chart is the instruction to follow for the proposed prediction model. For this study, the dataset is taken from Delhi state along with weather details of that particular time to analyse the climatic effects on malaria. In this the humidity and rainfall is considered as the weather parameters which may affect the transmission of malarial disease. The analysis of data and prediction of malarial disease is in Python 3 in jupyter notebook. It follows the seven step process while prediction using neural network.

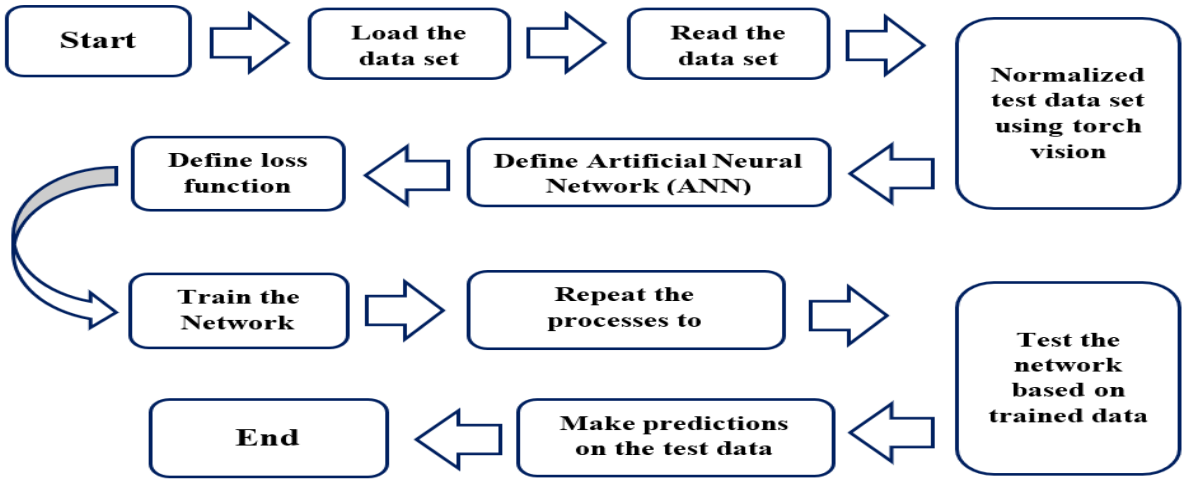


Figure3.3 Steps of prediction with Neural Network

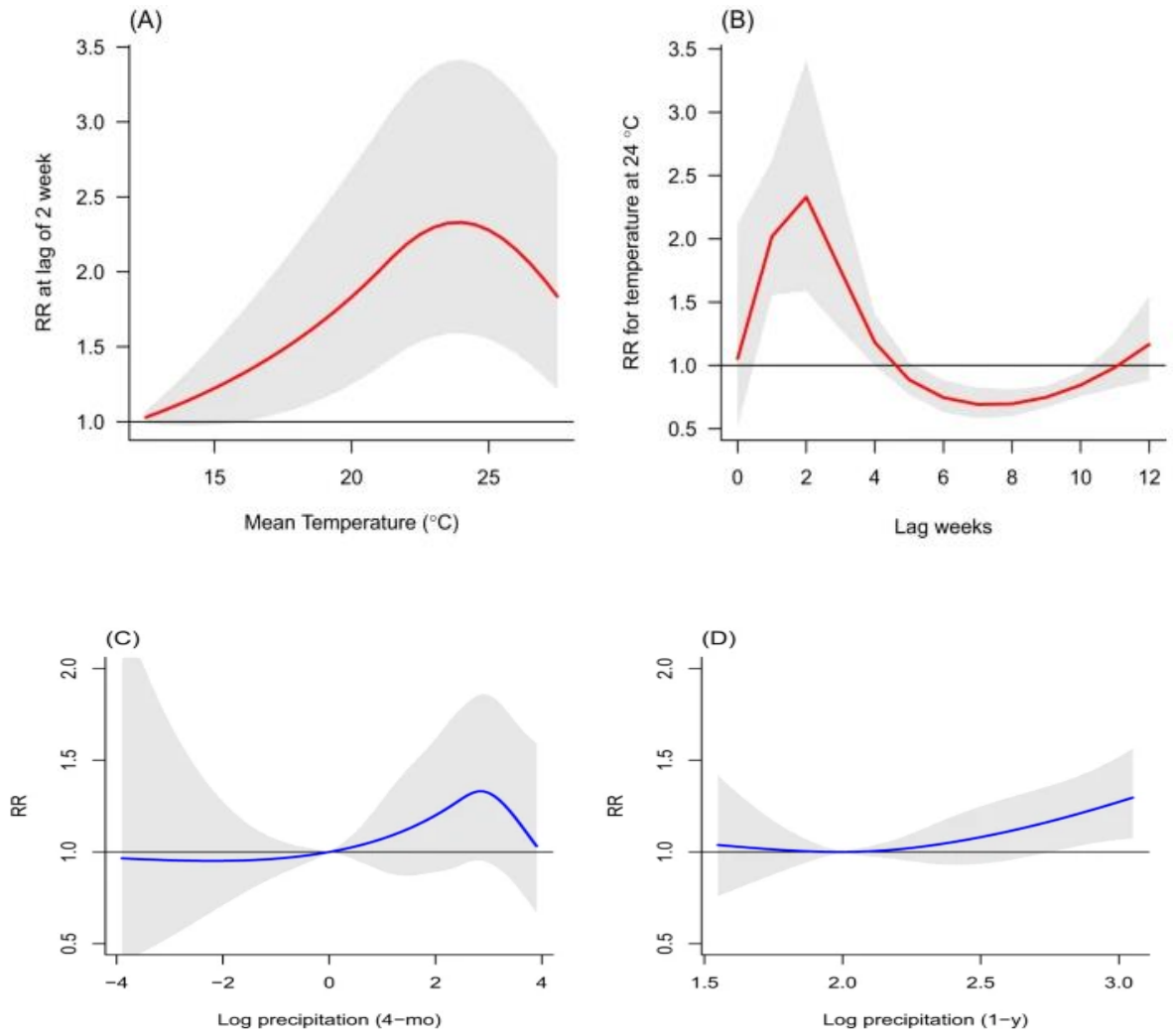


Figure3.4 Relative risk with respect to various factors

3.1.5 LSTM model for forecasting An LSTM model can be used for forecasting by using the historical data of a given time series and using it to train the model. The LSTM model will use the data from the past to make predictions about the future. The model can then be used to forecast future values of the time series data. This can be done by providing newly input data and letting the model infer the future values based on the past data. LSTM model for predicting the death cases in India due To Malaria.

In the field of health care Long short-term memory (LSTM) works on feedback connections in which it not only works on single data points but takes the entire sequence of data. It is a type of recurrent neural network. In this method, bagging and boosting techniques are most popular which provide composite prediction. In this the data set is divided into training and

testing data in which each training data (η) represented as (v, γ) v matrix consists of P variables and Q sequences which can be written as

$$[X_{1,}\, X_{2,} - - - - X_{q,} - - - -, X_Q] \tag{3.1}$$

X_q is a vector and x_q^p represents the p^{th} value at q^{th} step.

$$[X_q = x_q^1, x_q^2, - - - - - - - - - x_q^p - - - - x_q^P] \tag{3.2}$$

The LSTM model basically consists of three gates which are Input, Output and Forget gate. The input for LSTM block is X_q and in this the hidden or forget layer is H_q and for these layers the output can be computed as follows

$$\Gamma_q = \Omega(\omega_\Gamma\, [\Gamma_{q-1}, X_q\,] + \beta_\Gamma), \tag{3.3}$$

$$!_q = \Omega(\omega_!\, [\Gamma_{q-1}, X_q\,] + \beta_! \tag{3.4}$$

$$\theta_q = \Omega(\omega_\theta\, [\Gamma_{q-1}, X_q\,] + \beta_\theta), \tag{3.5}$$

$$C_q = \Gamma_q * C_{q-1} + !_q * \tanh(\omega_C\, [\Gamma_{q-1}, X_q\,] + \beta_C) \tag{3.6}$$

$!_q$, θ_q, Γ_q are the input, output and forget gates respectively. The hidden previous state is H_{q-1}

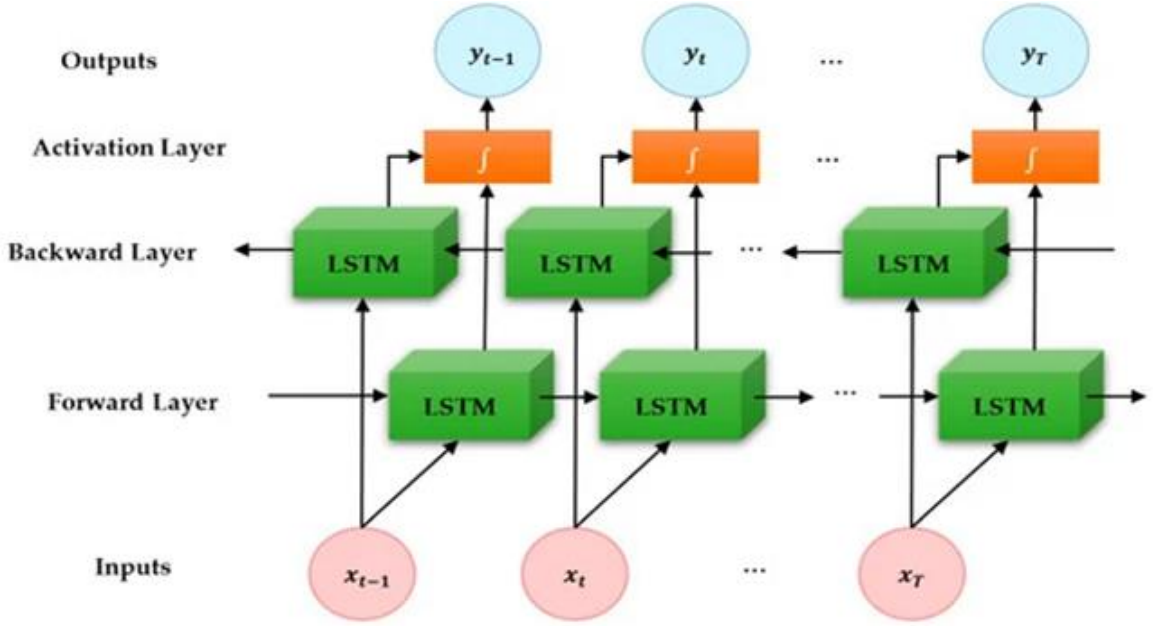


Figure3.5 LSTM Model and its layers

LSTM model consists of five layers. In this it inputs the data and with the hidden layers it analysed the data by using activation layer, backward layer and forward layer. It is the modification of recurrent neural network. The recurrent neural network works on vanishing gradient due to which it was unable to remember long term dependencies so to get rid of that problem LSTM made and hence have the potential to remember both type of dependencies that is long term and short term. In LSTM model, there are three stages, which are known as gates named as first gate, second gate and third gate. Further, it is named on the basis of allotted work, the first gate is forget gate, it forgets the data which is not as per the requirement. The selected data comes through input gate as input the data and after proper analysis in hidden gate it gives the final result as output.

3.1.5.1 Training and Validation of model: In order to train and validate a model, the first step is to obtain a dataset that contains training data as well as a set of validation data. This data should be randomly split into two sets. After splitting the dataset, the model can be created using either supervised or unsupervised learning techniques. Once the model is created, it should be tested on the validation set in order to measure its performance. The model may then need to be tweaked to improve its performance. Finally, the model can be tested on the training set to ensure that it can accurately predict labels or values for new data.

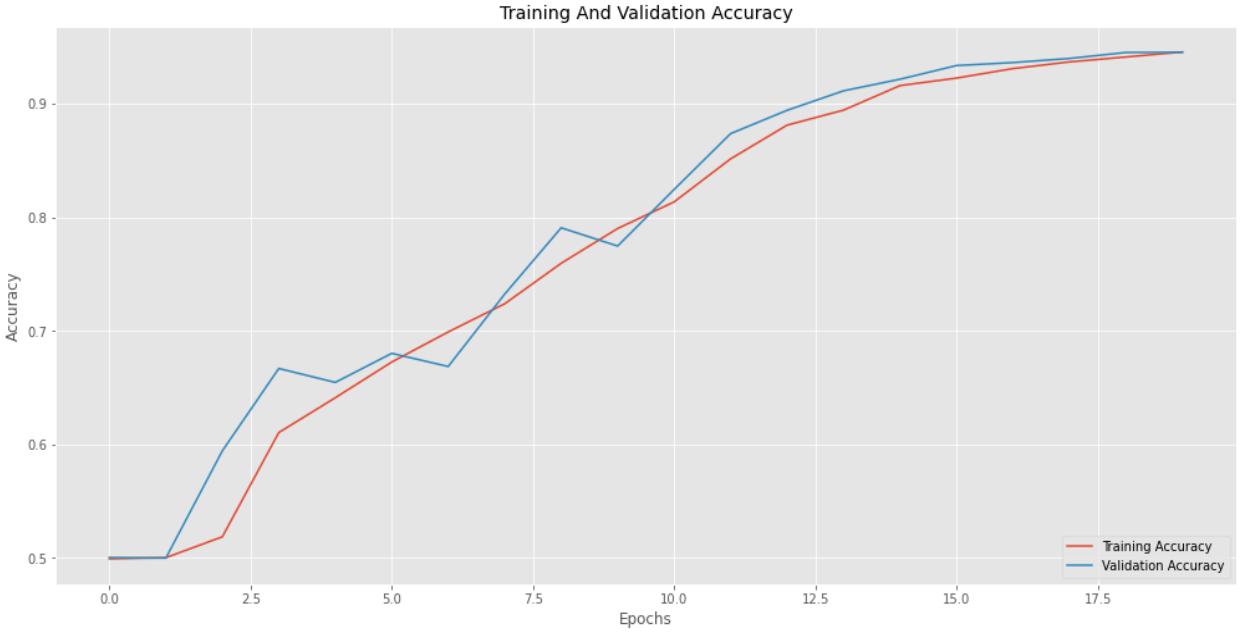


Figure3.6 Training and validation accuracy on trained model

The training and validation accuracy of a model is the percentage of correctly classified (or correctly anticipated) examples from the entire dataset. It measures how well the model fits the data which it was trained on, and can thus serve as a metric to measure the model's performance. Training accuracy is based on the training dataset, while validation accuracy is based on the validation dataset. Validation accuracy can give a better indication of how well a model will perform when given data that it was not trained with.

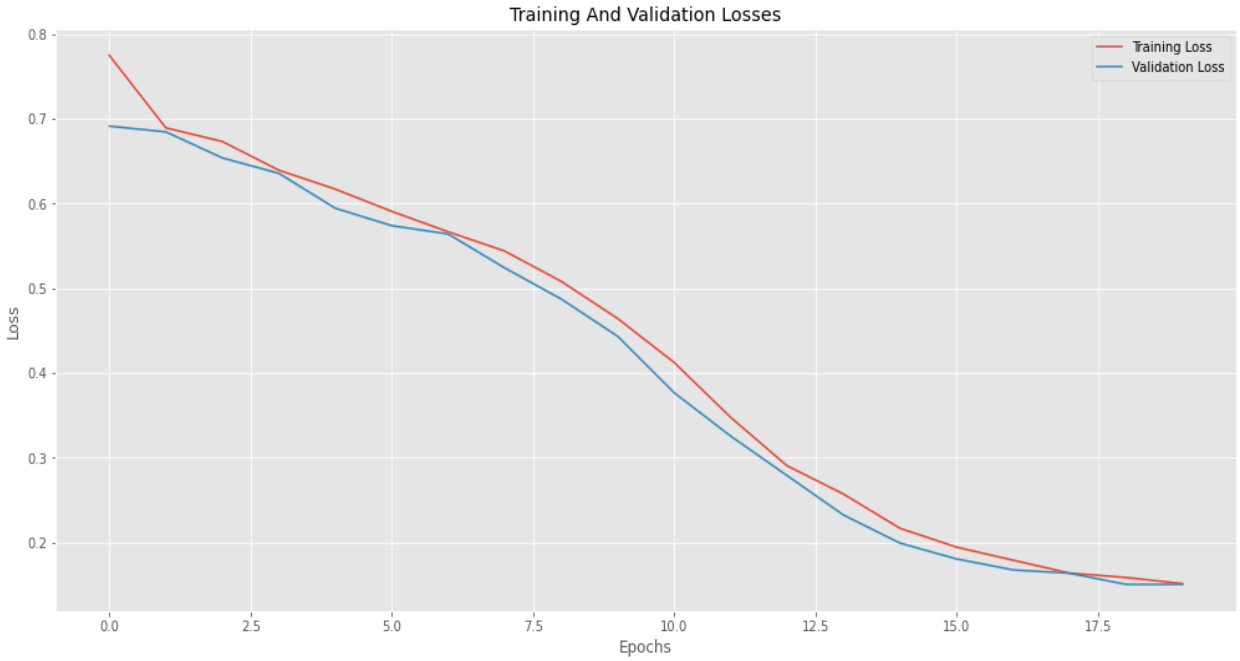


Figure3.7 Training and validation losses of trained model.

The training and validation losses of a model with respect to epochs are typically plotted in a graph format, so as to visually track the progress of the training and validation process. This graph provides a comparison between the training and validation losses as the number of epochs increase, and can be used to identify potential points of under or over fitting of the model when discrepancies between the train and validation losses arise. As the model is trained, the train and validation loss should generally tend to decrease, up until the point where the model no longer exhibits improved performance. This point of no longer improvement is often referred to as the model's convergence.

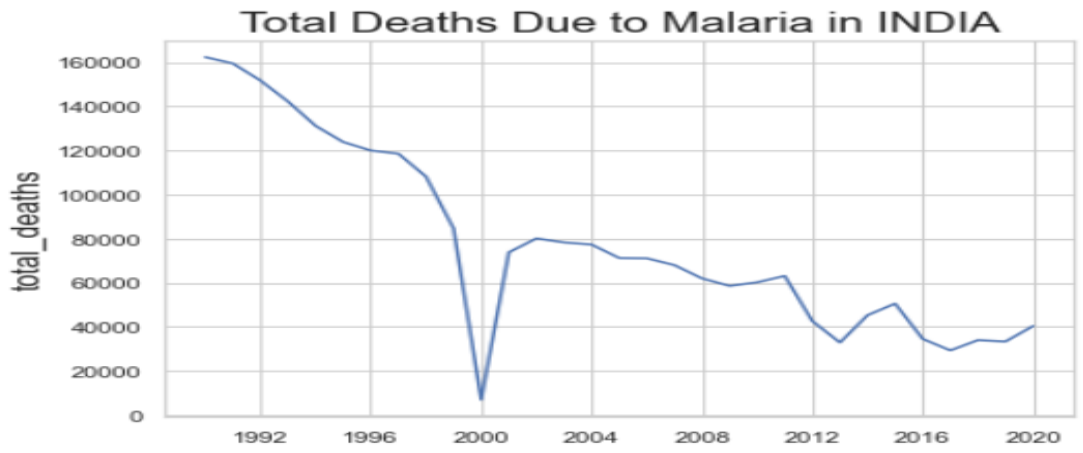


Figure3.8 Total death from 1990 to 2020 in India due to Malaria

The above figure is the representation of malarial cases in India. It is the dataset from 1992 to 2020 of country India. The special case of Delhi has been taken into consideration along with the climatic effects on it then in the following figure the Actual and predicted values has been shown. All the results like actual data, Predicted while deployment of model on training data set and prediction on testing dataset is shown by different colors and the following table shows the root mean square error values by which it is concluded that the prediction is almost following the actual trend and hence this study is helpful to aware the people that during which season Malaria becomes prominent by predicting the malarial cases.

```
1/1 [==============================] - 0s 467ms/step
1/1 [==============================] - 0s 32ms/step
Train Score: 22438.25 RMSE
Test Score: 19346.39 RMSE
```

The values of both scores were very close to each other, then the LSTM model was plotted.

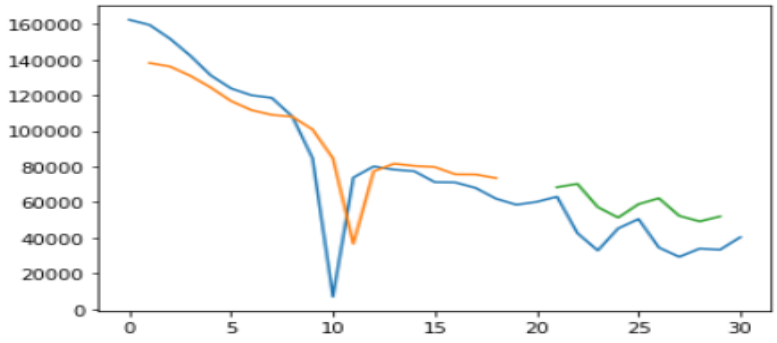


Figure3.9 LSTM model for Death Cases of India

For this, First of all read the CSV file, import required libraries and perform some basic code and visualize the data.

3.1.5.2 Ad fuller test: Ad Fullers are tests used to measure the completeness of a data set. It is used to detect whether the data is missing values or not. The test works by determining the ratio of the number of records with complete information to the total number of records. A score of 1 indicates that all records are complete, while a score of 0 indicates that no records are complete. Ad Fullers can be used to identify data quality issues, identify trends in data, and assess the completeness of data. They can also be used to identify data errors and ensure data is ready for analysis. Ad Fullers are especially useful when dealing with large datasets.

```
ADF Test Statstics:-2.0329380863821482
p-value:0.27232019168329513
#Lags used:0
Number of Observations Used:30
Weak evidence against the null hypothesis, time series has a unit root and is non-stationary
```

Figure3.10 Results of ADF first test

3.1.5.3 Order of differencing

Differencing is a technique used in time series analysis to remove the effects of trends and seasonality in a data set. It is the process of subtracting the value of a variable at one time point from its value at the previous time point. The order of differencing is the number of times the differencing is applied to a time series. The order of differencing is determined by how much the data needs to be de-trended to make it stationary. A stationary time series is one that has no trend or seasonality, and is the prerequisite for forecasting. The higher the order of differencing, the more de-trended the data becomes.

```
ADF Test Statstics:-4.438863508216068
Strong evidence against the null hypothesis(Ho),reject the null hypothesis,The series has no unit root and is stationary
p-value:0.0002531920220117824
Strong evidence against the null hypothesis(Ho),reject the null hypothesis,The series has no unit root and is stationary
#Lags used:9
Strong evidence against the null hypothesis(Ho),reject the null hypothesis,The series has no unit root and is stationary
Number of Observations Used:19
Strong evidence against the null hypothesis(Ho),reject the null hypothesis,The series has no unit root and is stationary
```

Figure3.11 Results of ADF second test after Series differencing

ADF test statistics is -4.438863508216068, p-value of 0.00025,

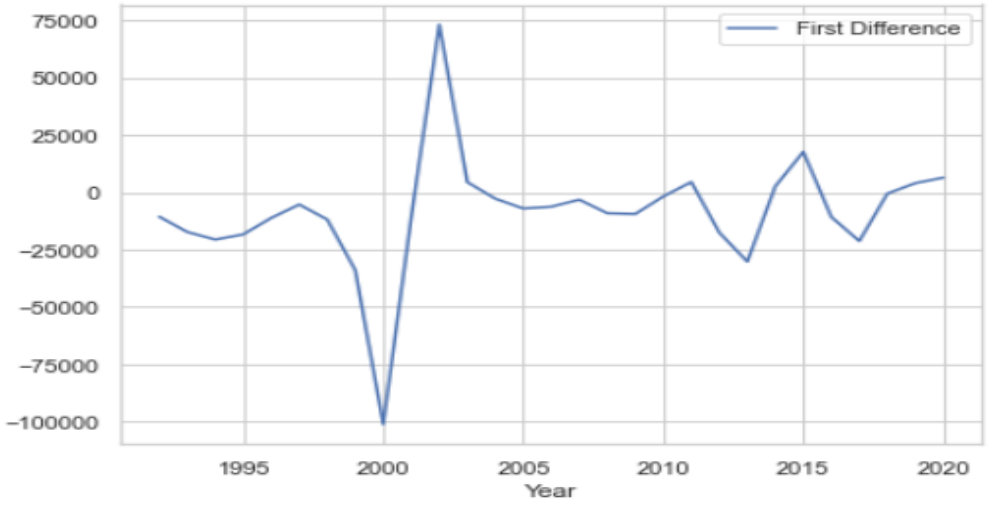


Figure3.12 Representation of the Stationarity series of data after the first difference

3.1.6 Autocorrelation Function (ACF) and Partial Autocorrelation Function (ACF)

The relationship between malaria and weather is complex and varies from region to region. Autocorrelations provide a way to measure the influence of weather on the spread and prevalence of malaria. Autocorrelations measure how strongly the current values of a variable are related to past values. Specifically, a strong autocorrelation of malaria and weather would indicate that past weather conditions affect future malaria occurrences, often in the form of lags (for example, rainy conditions in the preceding weeks could lead to an increase in malaria

cases in the coming weeks). It is important to note that autocorrelation does not necessarily indicate causality or a linear relationship, only that weather has an influence on malaria.

Autocorrelations between malaria and weather are generally positive, indicating a link between precipitation, temperature, and humidity and malaria cases. Studies have shown that high humidity, along with warm and wet temperatures, are the most conducive for mosquito breeding, which increases the risk of malaria transmission. Therefore, autocorrelations help to identify areas where prevention methods, such as insecticide-treated bed nets or insecticide spraying, could be most useful.

The partial autocorrelation between malaria and weather depends on the location of where the two variables are studied. Generally speaking, there is likely to be a correlation between malaria and weather in tropical regions with high humidity and rainfall, as these are the conditions that support the development of the disease-carrying mosquito. In addition, there are certain weather patterns, such as the El Niño phenomenon, that can establish environmental conditions that make the spread of malaria more likely. By examining the autocorrelation (or lack thereof) of weather over time, scientists can get an indication of how weather patterns affect the spread of malaria in an area.

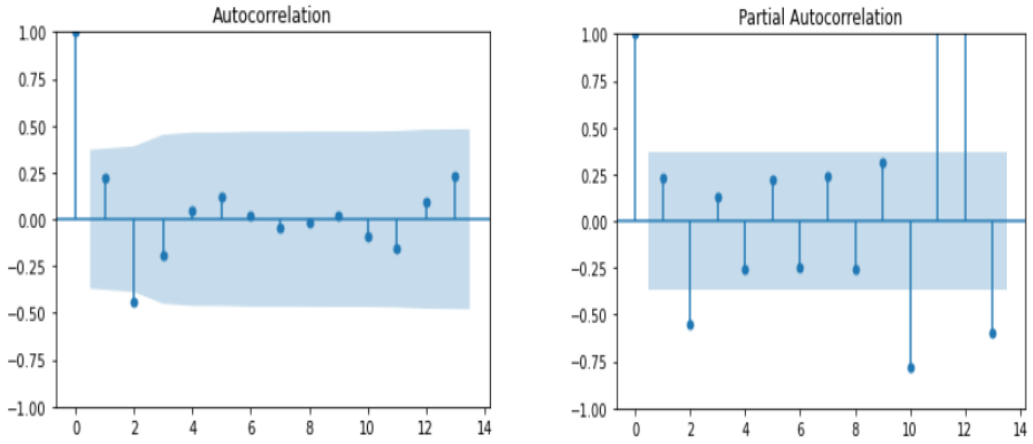


Figure3.13 Correlation between weather and Malaria

```
ARIMA(order=(0, 1, 0), scoring_args={}, suppress_warnings=True,
      with_intercept=False)
```

Figure3.14 Representing p, d, and q value

From the above figure p, d, and q is given by 0, 1, and 0 respectively.

Now moving on to prediction.

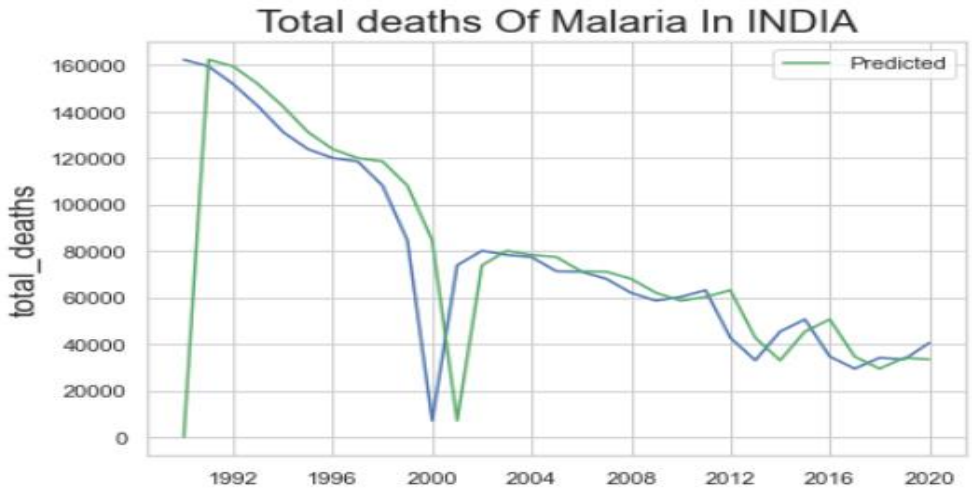


Figure3.15 Predicted malarial deaths using ARIMA model

Arima models are able to provide reasonably accurate forecasts for a range of data, including malaria data because malaria is an infectious disease that tends to follow seasonal patterns. An arima model that incorporates seasonal components can provide more accurate forecasts than one that does not incorporate such components. Depending on the malaria data, a researcher may want to use a model that includes exogenous variables, such as environmental factors or population size, to further improve accuracy of the forecast. While Arima models can be relatively accurate predictors of malaria data, they cannot provide insight into the underlying causes of transmission, which must be determined through other methods such as epidemiological or statistical inference.

3.1.7 SARIMA MODEL (Seasonal Autoregressive Integrated Moving Average) is a type of time series model used to analyze data with seasonal components. It combines both autoregressive and moving average components to capture both short-term and long-term

patterns in the data. SARIMA models are often used to forecast sales, inventory levels, and other time-series data.

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	31
Model:	SARIMAX(0, 1, 0)	**Log Likelihood**	-340.536
Date:	Thu, 03 Nov 2022	**AIC**	683.073
Time:	18:47:26	**BIC**	684.474
Sample:	0	**HQIC**	683.521
	- 31		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	4.098e+08	4.68e+07	8.754	0.000	3.18e+08	5.02e+08

Ljung-Box (L1) (Q):	1.75	**Jarque-Bera (JB):**	81.08
Prob(Q):	0.19	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.15	**Skew:**	-0.19
Prob(H) (two-sided):	0.01	**Kurtosis:**	11.04

Figure3.16 Sarima model summary

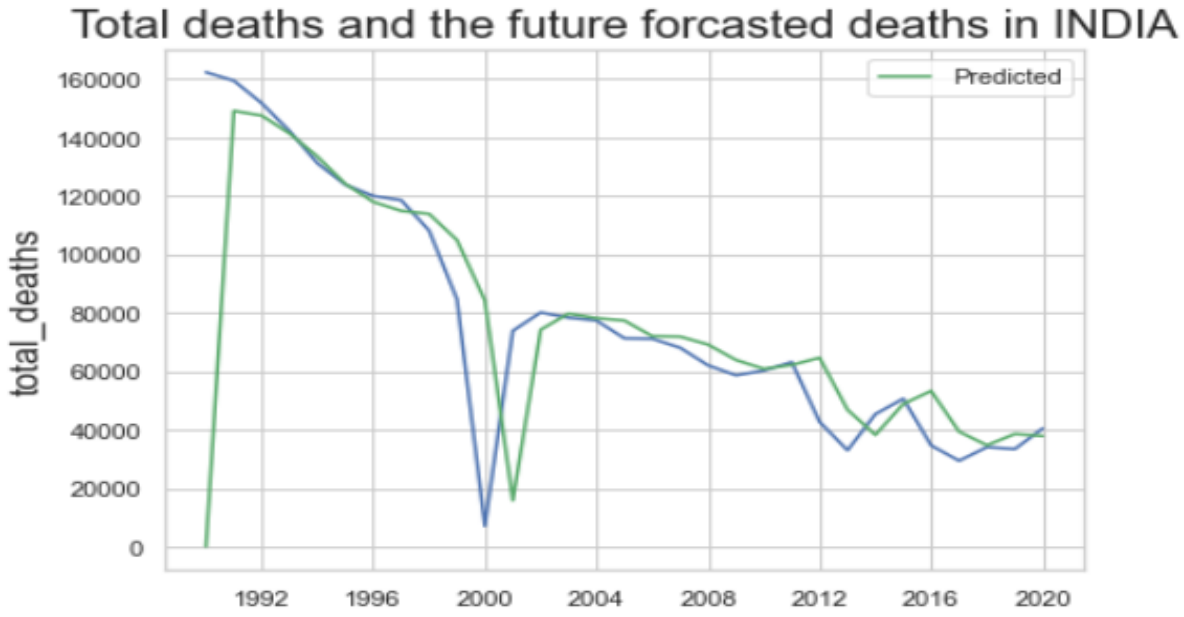


Figure3.17 Predicted malarial deaths using SARIMA model

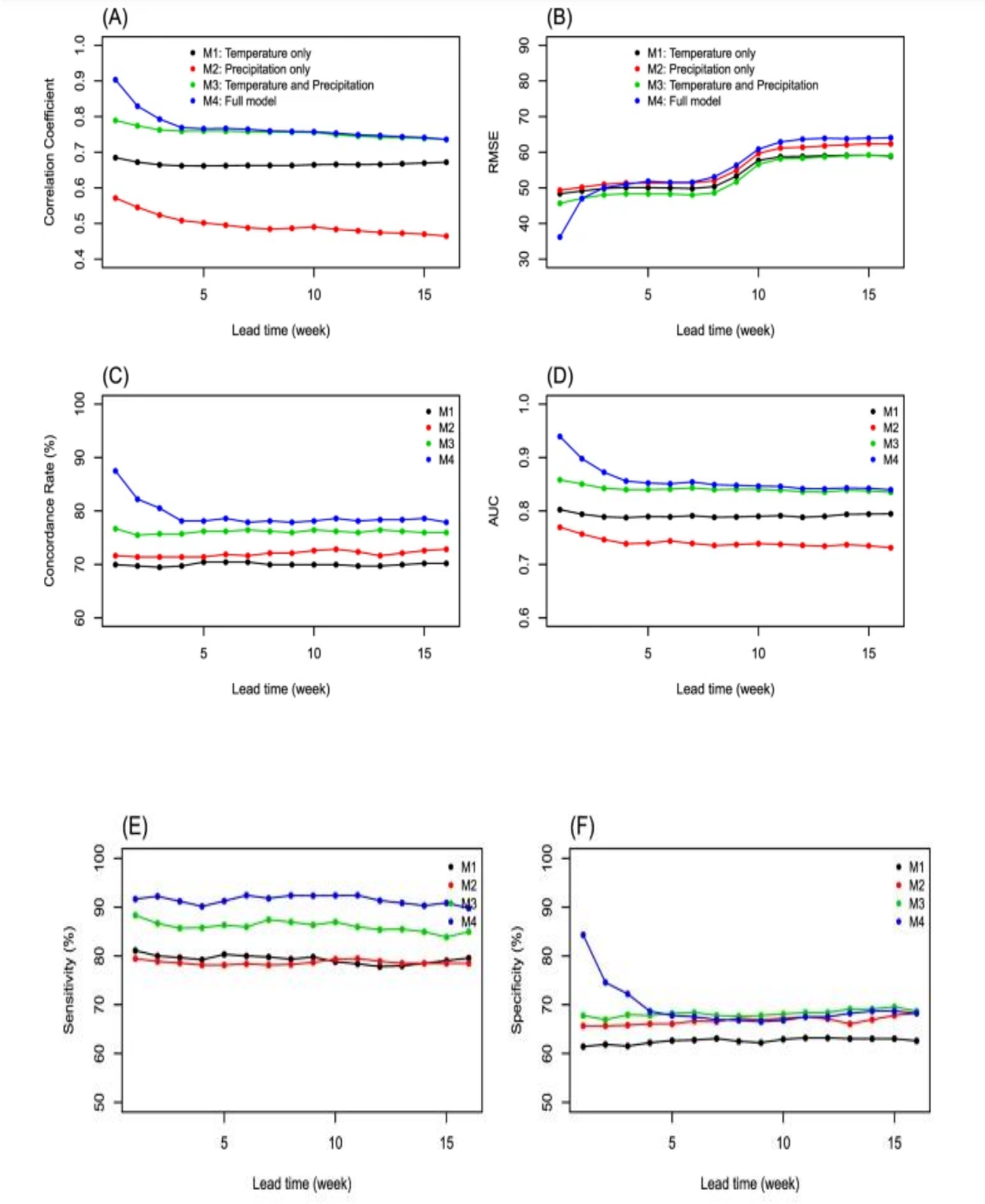


Figure3.18 Concordance rate, correlation coefficient and sensitivity with respect to time

The correlation coefficient between temperature and incidence of malaria and between precipitation and incidence of malaria changes over time. This is because temperature and precipitation affect the lifecycle and activity of the mosquito vector of malaria differently in different parts of the year. To calculate the correlation coefficients, researchers typically consider data over a period of several weeks or months in order to account for seasonal variations in temperature and precipitation.

The concordance rate of malaria prevalence under the temperature and precipitation factor varies over time and can be estimated each week. Studies have shown that as temperature increases and precipitation decreases, malaria risk increases. A study conducted in Kenya showed that between weeks 16 and 20 of 2005, the concordance rate between temperature and precipitation for malaria prevalence was 0.85, indicating that climate factors do have an impact.

The sensitivity of malaria with respect to time in weeks under temperature and precipitation factors depends on the region. In warmer regions, where temperatures are higher and precipitation is more abundant, malaria can remain active throughout the year. In cooler regions with fluctuating temperatures and lower precipitation, malaria can be seasonally active for shorter periods of time. Thus, the sensitivity of malaria with respect to time in weeks under temperature and precipitation factors is highly variable and region specific.

3.1.8 Conclusion

Climate is known to affect the transmission of malaria. Changes in temperature directly influence the rate of the mosquito life cycle, which is a major driver of malaria transmission. Warmer temperatures can increase the rate of parasite development in the mosquito, reducing the length of time it takes for the disease to be transmitted from a human host to a mosquito vector. Similarly, higher temperatures tend to increase the rate of mosquito reproduction, increasing the number of vectors. Changes in precipitation can also alter the transmission of malarial disease. Wetter conditions create more standing pools of water, which provide a suitable environment for mosquito larval development. As a result, increased precipitation can lead to a higher number of mosquito vectors and an increased risk of the disease spreading. Thus, climate change has been associated with an increased risk of malaria transmission.

Warmer temperatures and higher precipitation can all provide an environment conducive to the spread of the disease.

Case 3.2 COVID-19 patient's recovery using RF-DT model: The data mining proves the best results in field of healthcare like during corona as well as in many pandemic. It helps to forecast that what situation will happen so prior arrangements to tackle the situation leads to save many lives. As the proposed work related to COVID-19 pandemic is of various countries. In this the methods which we used is the combination of two models i.e., it is of the random forest and decision trees which leads to the more accurate results.

The use of random forest and decision tree models is common in the modeling phase.

First, in the data preparation stage, the dataset is preprocessed and cleaned to ensure data quality. This may involve handling missing values, removing outliers, and transforming variables if necessary. Next, the random forest model, which is an ensemble technique, is trained on the prepared data. Random forest combines the predictions of multiple decision trees to make more accurate predictions. Each decision tree is constructed by splitting the data based on different features and evaluating the best split using metrics such as information gain or Gini impurity.

After training of the model, accuracy is checked and then this evaluation helps to measure the performance of the model and provides insights into its strengths and weaknesses. Simulation testing can be performed by applying the trained model to a test dataset or by using cross-validation techniques to assess its generalization capabilities. Finally, if the model performs well in the simulation test, it can be deployed for further predictions. The deployment can be done through an API, a web application, or integrated into existing systems, depending on the specific use case.

	ObservationDate	Province/State	Country/Region	Confirmed	Deaths	Recovered
0	01/22/2020	Anhui	Mainland China	1	0	0
1	01/22/2020	Beijing	Mainland China	14	0	0
2	01/22/2020	Chongqing	Mainland China	6	0	0
3	01/22/2020	Fujian	Mainland China	1	0	0
4	01/22/2020	Gansu	Mainland China	0	0	0

Figure 3.2.1 Actual data of COVID-19

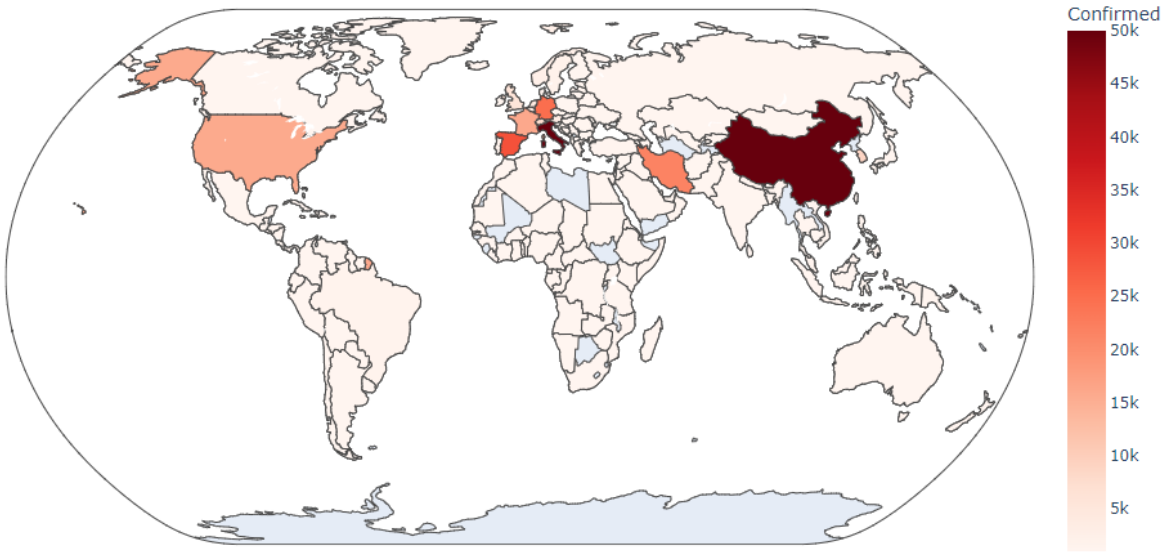


Figure 3.2.2 World map depicting confirmed cases

In the above figures, the confirmed cases of the country and world map where these affected countries lies has been shown by using Jupyter notebook with interface Python Language.

3.2.1 RF-DT Model: It is the combination of two models i.e., random forest and decision tree and in this by increasing the number of decision trees it increases the accuracy. Each decision tree gives their result and by voting the final output will be considered as that value which have maximum number of votes and for this the following algorithm is defined:

Step1. Selection of random samples from the dataset.

Step 2. For each selected sample, create a decision tree, then get a forecasting result from it.

Creation of decision tree

Step 2.1 Calculate the entropy (S) of the training set as follows:

$$\text{Entropy (S)} = -\sum \left[\frac{freq[\frac{freq\ (c_i,s)}{|s|}}{|s|} \right] log_2 \left[\frac{freq\ (c_i,s)}{|s|} \right]$$

$$\text{Entropy (S)} = -\sum_{i-1}^{k} \left\{ \left[\frac{freq[\frac{freq\ (c_i,s)}{|s|}}{|s|} \right] log_2 \left[\frac{freq\ (c_i,s)}{|s|} \right] \right\}$$

Where $|s|$ is the number of samples in the training set c_i is a dependent variable, i = 1,2,……k. k is the number of classes of the dependent variable and $freq(c_i, s)$ is the number of samples included in class c_i.

Step 2.2 Calculate the information Gain X(S) for test attribute X to partition:

$$\text{Information Gain X(S)} = \text{Entropy (S)} - \sum_{i=1}^{l} \left[\left(\frac{s_i}{s} \right) entropy\ (s_i) \right]$$

Where L is the number of test outputs X.

s_i is a subset of S corresponding to i[th] output and $|s_i|$ is the number of dependent variables of subset s_i.

Step 2.3 Calculate the partition information value Split info (X) acquiring for S partitioned into L subsets.

$$\text{Split info (X)} = \sum_{i=1}^{l} \left[\left(\frac{s_i}{s} \right) log_2 \left(\frac{s_i}{s} \right) + \left(1 - \left(\frac{s_i}{s} \right) \right) log_2 \left(1 - \left(\frac{s_i}{s} \right) \right) \right]$$

Step 2.4 Calculate the Gain ratio (X):

$$\text{Gain ratio (X)} = \frac{Information\ Gain_x\ (S)}{Split\ info\ (X)}$$

Step 2.5 The attribute having the highest gain ratio will be designates as the root node and the same computation from step 2.1 to step 2.4 is done for every intermediate node until all the instances re-exhausted and it reaches the leaf node according to step 2.2.

Step 3 voting of the forecasted result.

Step 4 Final forecasting from the most voted forecasted result.

	Country/Region	Confirmed	Deaths	Recovered	Active
0	Mainland China	81060	3261	72252	5547
1	Italy	59138	5476	7024	46638
2	US	33276	417	178	32681
3	Spain	28768	1772	2575	24421
4	Germany	24873	94	266	24513
5	Iran	21638	1685	7931	12022
6	France	16044	674	2200	13170
7	South Korea	8897	104	2909	5884
8	Switzerland	7245	98	131	7016
9	UK	5741	282	67	5392
10	Netherlands	4216	180	2	4034

Figure 3.2.3 Summary of the data Country/Region wise with various parameters

	Confirmed	Deaths	Recovered
count	7926.000000	7926.000000	7926.000000
mean	655.640172	22.993187	237.621625
std	4978.076030	229.440512	2704.176057
min	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000
50%	16.000000	0.000000	0.000000
75%	130.000000	1.000000	10.000000
max	67800.000000	5476.000000	59433.000000

Figure3.2.4 Statistical summary of covid-19 data

The evaluation of statistical summary shows the overall analysis in a single step of code which makes the analysis quite clear and helpful to take well decision for the proper arrangements as well as to know the confirmed cases of corona.

```
1  table['Country/Region'].value_counts().head()
```

```
Mainland China    1889
US                1617
Australia          323
Canada             254
France             127
Name: Country/Region, dtype: int64
```

```
1  table.isnull().sum()
```

```
ObservationDate       0
Province/State     3433
Country/Region        0
Confirmed             0
Deaths                0
Recovered             0
dtype: int64
```

Figure3.2.5 Values counts in the form of Country/region and null values of data

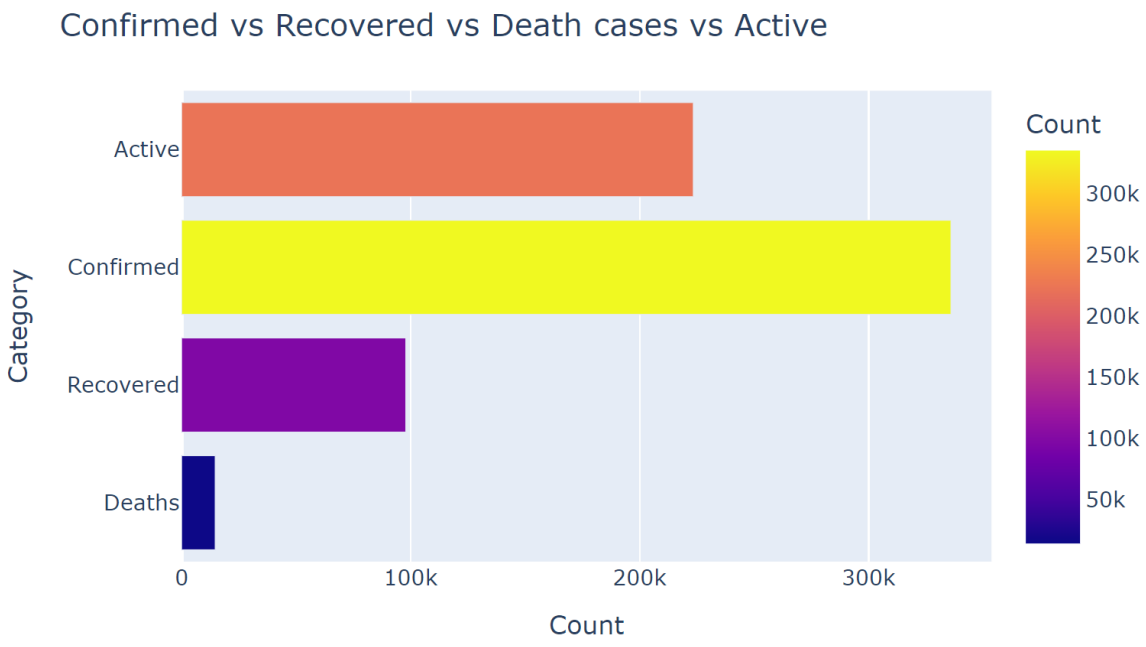


FIGURE3.2.6 Counts the numbers parameters wise

Used data mining algorithms such as the Random Forest model for Analysis and forecasting of the COVID-19 data. Used random forest model result various evaluation parameters were measured in terms of precision, accuracy, F-measure, recall, time taken to build a model, correctly classified instances percent and incorrectly classified instances percent.However,

accuracy is usually measured by the percentage of correct predictions made by the algorithm, and is calculated by

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ Negative + False\ Positive + False\ Negative}$$

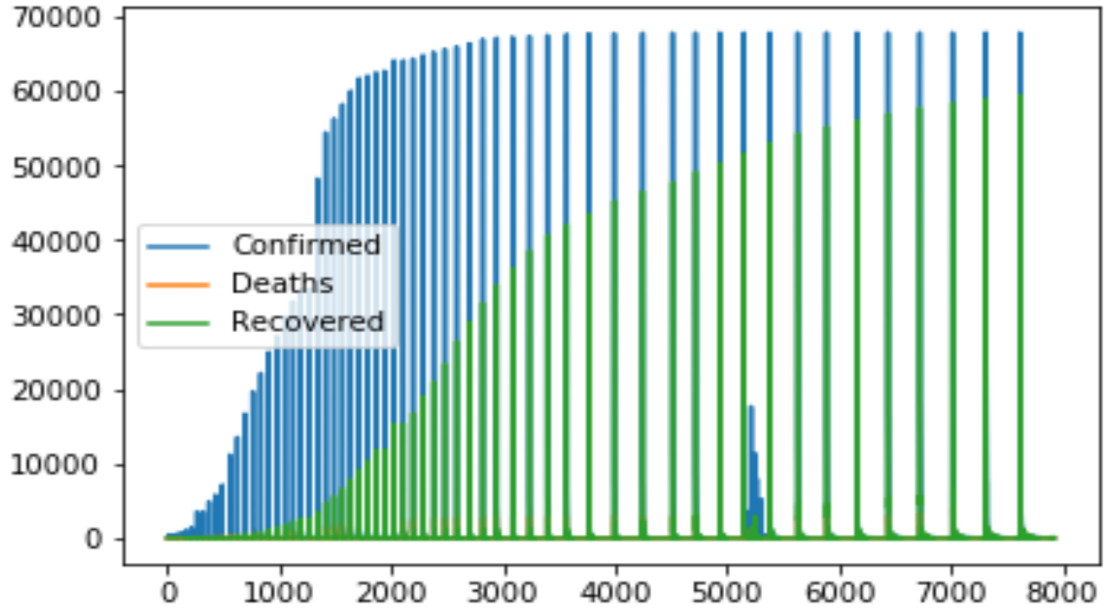


FIGURE 3.2.7 Graphical representation of COVID-19 dataset with different measures

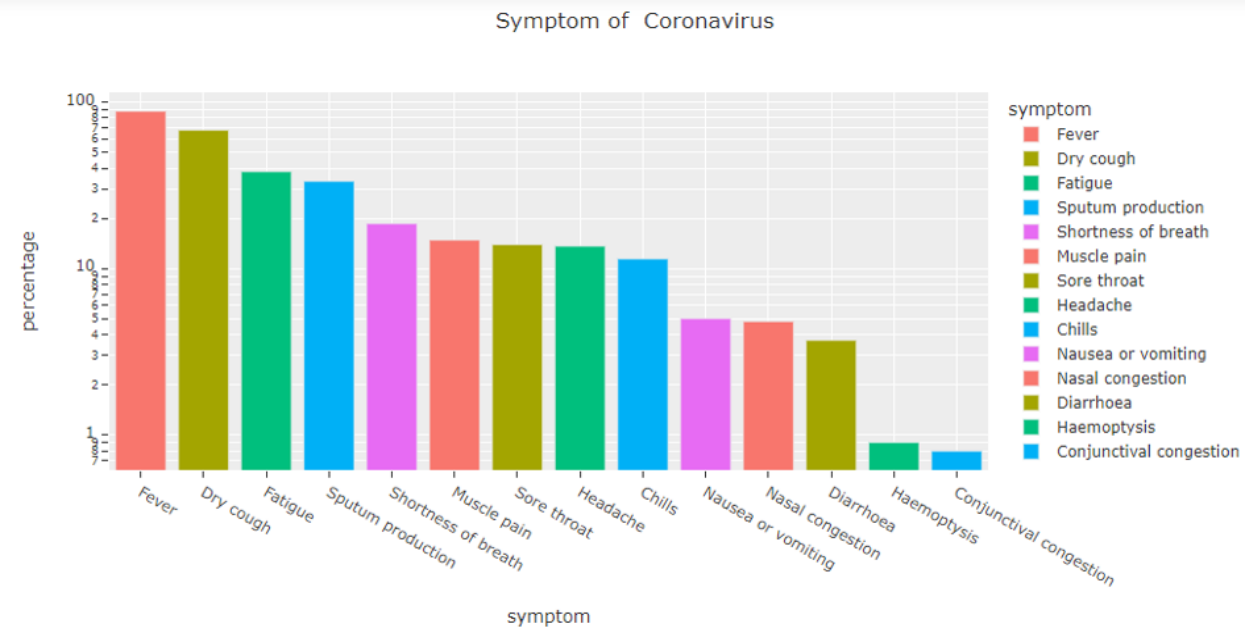


Figure 3.2.8 Graphical representations of the symptoms of COVID-19

According to the World Health Organization and other health Agencies coronavirus is defined as the collection of various viruses whose symptoms can be from a mild cold to severe diseases. It is seen that it is the respiratory disease that increases the cases of COVID-19 rapidly. It has many symptoms but the most common symptoms of COVID-19 are sneezing, coughing, fever, respiratory problems, loss of sense, smell chest pain. Moreover, the given chart shows the details of symptoms in Fig. 3.2.8 and it is observed that fever, dry cough and fatigue are the most common symptoms.

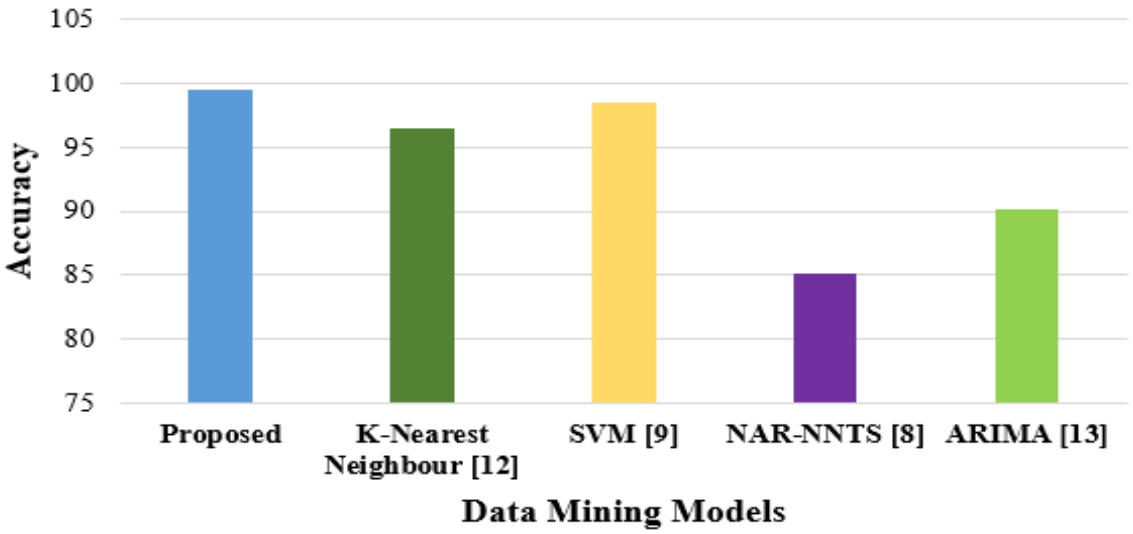


Figure 3.2.9 Comparison of Proposed model with other existing models

In the proposed model, the percentage of accuracy will be 99.5%, which is the highest compared to other models such as K-Nearest Neighbour, SVM, NAR-NNTS, ARIMA models as 98.5%, 98.45%, 85.10%, 90.1%, respectively. Then we can conclude that the proposed model would be very useful and helpful in the healthcare sector such as COVID-19.

3.2.2 Conclusion: Climate is known to affect the transmission. Changes in temperature directly influence the rate of the transmission. Warmer temperatures can increase the rate of parasite reducing the length of time it takes for the disease to be transmitted. Similarly, higher temperatures tend to increase the rate of virus reproduction, increasing the number of vectors. Thus, climate change has been associated with an increased risk of transmission. Warmer temperatures and higher precipitation can provide an environment conducive to the spread of the disease like malaria and corona and thus various factors are studied for the transmission on

different models such as K-Nearest Neighbour, SVM, NAR-NNTS, ARIMA models which shows various accuracy on the basis of considered factors also. It is concluded that the proposed model in case of COVID-19 would be very useful and helpful in the healthcare sector.

Chapter –IV

Regional frequency analysis approach for rainfall analysis

4.1 Introduction

Machine learning and data mining both are interrelated fields usually deal with the problem of prediction based on many techniques. In this, for the categorical division of data, classification techniques has been implemented so that data can be categorized into various labels on the basis of various parameters given to it and the labels can also be predicted from the given dataset. From all over Haryana 29 rain gauge stations were selected based on data availability as shown in figure 4.2. In the regional frequency analysis approach, data are gathered from a variety of sites/stations (with similar statistical properties) across the region, and a single regional distribution is fitted to this region.

These stations were clustered into three groups namely, G1, G2, and G3 with the help of Ward's method. L-moments-based Heterogeneity test (H) was used to test the region's homogeneity and for each station, a discordancy measure (D_i) was calculated. Further, quantiles were estimated at each station using regional quantiles by taking the station's mean as a scaling factor. The results showed that estimated quantiles using the best-fit distribution were in close agreement with observed data.

Rainfall is a significant climate parameter that affects cropping patterns, soil erosion, and sedimentation. The nature and status of agriculture in an area particularly depend on the total rainfall, its duration, and distribution. The rainfall distribution varies greatly over time and space. Inadequate rainfall also causes serious problems. Agriculture is the backbone of our economy, and millions of people in the farming sector depend on rainfall. Reliable rainfall forecasting can help farmers and policymakers to take important decisions. Long-range forecasts are used for agriculture, transport, and water management. Analysis of rainfall data focuses on its distribution pattern and the development of a probability model for rainfall yield like rainfall of the state Haryana.

In conclusion, polynomial regression and quantile regression are both extensions of linear regression that address different aspects of the relationship between variables. Polynomial regression allows for non-linear relationships, while quantile regression focuses on estimating quantiles of the response variable but During classification task (i) on the basis of independent features class labels or discrete target variables can be predicted (ii) Decision boundaries can also be found to separate the classes in target variable.

In classification technique, the basic problems to decide whether the data is into binary discrete labels or to multiple discrete labels.

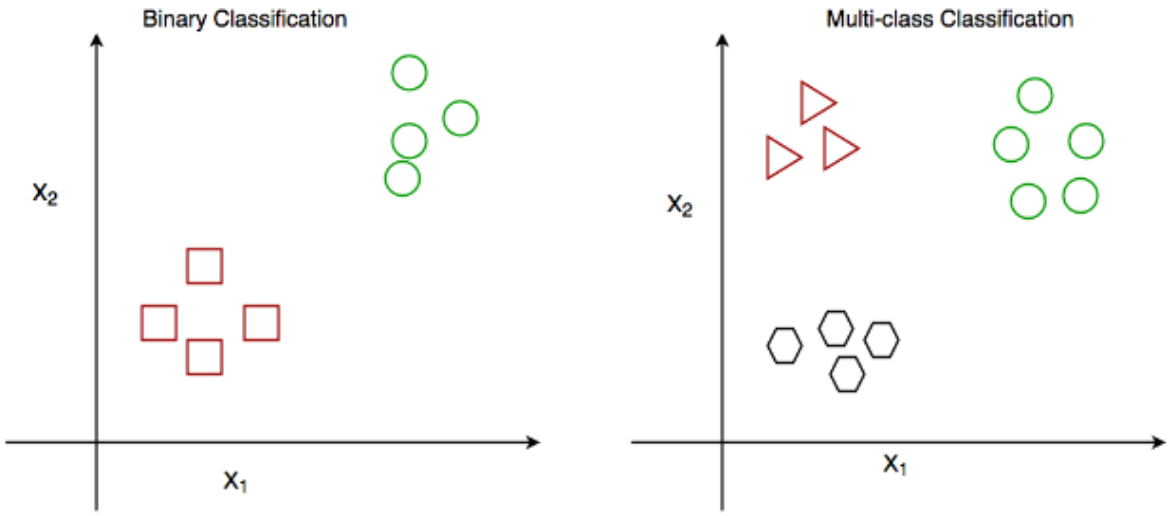


Figure 4.1 Binary Classification and Multi Classification

There are many techniques like bagging and boosting to improve the basic classification techniques. The basic classification techniques/ algorithms are Random forest classifier, K-nearest neighbours, Support vector machine and decision tress. The metrics for evaluation are Recall, Precision and F1-score to check the performance/ accuracy of the algorithms.

4.2 Preliminaries for the Development of Model

4.2.1 List of symbols

1. $z(F)$: Quantile distribution function
2. F : Annual non-exceedance probability
3. Z : Random variable

4. $M_{p,r,s}$: Probability weighted moments
5. $\beta_r (r = 0,1,2)$: special probability-weighted moments
6. E[.] : Expected value
7. $\lambda_r, r = 1,2,3,4$: First four population L-moments
8. $l_r, r = 1,2,3,4$: First four sample L-moments
9. t: Sample L-coefficient of variation
10. t_3: Sample L-coefficient of skewness
11. t_4 : Sample L-coefficient of kurtosis
12. t^R, t_3^R, t_4^R: Rregional average L-moments ratios
13. D_i : Discordancy value of ith station
14. u_i : Vector consist of first three sample L-moment ratios of the ith station
15. $\bar{u}$: Vector containg average of u_i
16. Z^{DIST}: Goodness-of-fit value of selected distribution
17. n_i: Record length at ith station
18. $Q_i(F)$: Quantile function at given probability F
19. μ_i: index-flood (or scaling factor) for station i
20. $q(F)$: regional quantile
21. θ_k^R: k-th regional parameter

4.2.2 Materials and Methods

In this study, daily rainfall data (1970–2017) of 29 rain gauge stations in Haryana were obtained from the National Data Center, India Meteorological Department (IMD) Pune. The annual total rainfall series were constructed for each rain gauge station for the analysis. The geographical locations of rain gauge stations considered for this study in map of Haryana are shown below.

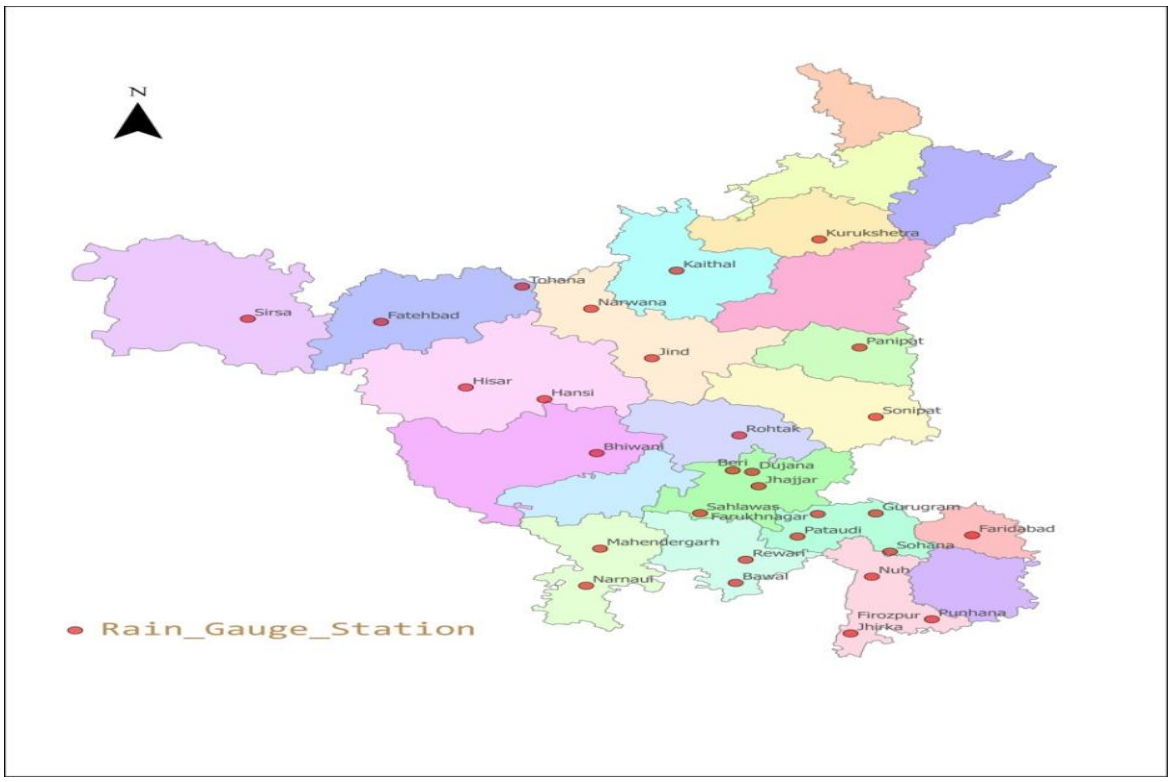


Figure 4.2 Map of Haryana

4.3 Machine learning Regression technique for forecasting: The process to find the continuous values rather than division in classes. It uses historical data to identify the movement of distribution and in these existing techniques bagging and boosting techniques can be employed to get the best results for regression. Regression techniques of forecasting are based on statistical models that use historical data to predict future trends. It is a type of quantitative forecasting method that attempts to identify and quantify the relationships between different variables, such as sales and customer demand, and then uses those relationships to make predictions about future values. It is a powerful forecasting tool that can be used to provide insight into the future.

Regression forecasting uses data from the past to generate predictions about the future. It can be used to predict sales, customer demand, and other future trends and there are many regression techniques like Ridge, Lasso regression and XGBoost, LGBM Regressor. In this the MSE,R_2 score and MAPE can be used to check the accuracy of model.

4.3.1 Expectation Maximization method: Since there is wealth of health data which demands the accurate prediction so there are few methods which works for the latent variables. The exactness of prediction depends on various factors and still there lying various factors in those factors so these methods are quite apt for these types of data where unobservable variables influence the observable data.

4.3.2 Theoretical L-Moments

The theoretical L-moments for a real-valued random variable Z with a Quantile distribution function (QDF) $z(F)$ are defined from the expectations of order statistics. For the random variables $Z_1, Z_2, \dots Z_n$ of sample size n let $Z_{1:n} \leq Z_{2:n} \dots \dots \leq Z_{n:n}$ be the order statistics of the sample and the theoretical L-moments for $r \geq 1$ are defined by

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(Z_{r-k:r}), \quad r = 1,2, \dots, \tag{4.1}$$

where r is the r^{th} L-moment, and $E(Z_{r-k:r})$ is the expectation value of the $r-k$ order statistic of a sample of size r, and this equation is commonly expressed in terms of the QDF.

The L-moments can be defined as:

$$\lambda_1 = E(Z_{1:1})$$

$$\lambda_2 = \frac{1}{2} E(Z_{2:1} - Z_{1:2})$$

$$\lambda_3 = \tfrac{1}{3} E \begin{pmatrix} Z_{3:3} \\ -2Z_{2:3} + Z_{1:3} \end{pmatrix} \tag{4.2}$$

$$\lambda_4 = \tfrac{1}{4} E(Z_{4:4} - 3Z_{3:4} + 3Z_{2:4} - Z_{1:4})$$

The expected value of $(Z_{r:n})$ can be find as:

$$E(Z_{r:n}) = \frac{n!}{(r-1)(n-r)} \int_0^1 z(F)\, F^{r-1} (1-F)^{n-r}\, dF \tag{4.3}$$

L-moments, also calculated from r^{th}-shifted Legendre polynomials $P_r^*(F)$ defined as:

$$P_r^*(F) = \sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}\binom{r+k}{k}F^k \quad (4.4)$$

Thus for random variable Z having quantile function $z(F)$, L-moments to be the quantities

$$\lambda_r = \int_0^1 z(F)\, P_{r-1}^*(F)dF \quad (4.5)$$

The probability-weighted moments are given by Greenwood *et al.* [10] as:

$$E(Z_{r:n}) = \frac{n!}{(r-1)(n-r)}\int_0^1 z(F)\, F^{r-1}(1-F)^{n-r}\, d \quad (4.6)$$

Where p, r, and s are real numbers. One very useful relation of probability-weighted moments is $\beta_r = M_{1,r,0}$. Therefore

$$\beta_r = \int_0^1 z(F)\, F^r\, dF \quad (4.7)$$

Where $z(F)$ is a quantile function and r is a non-negative integer value. Using probability-weighted moments, L-moments are calculated as :

$$\lambda_{r+1} = \sum_{k=0}^{r} P_{r,k}^*\, \beta_k$$

where $P_{r,k}^*$ is a set of orthogonal polynomials described as:

$$P_{r,k}^* = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!} \quad (4.8)$$

where $P_{r,k}^*$ is the r^{th} shifted Legendre polynomials (Hosking [2]).

The first four L-moments, from the equation (8) are given by:

$$\begin{aligned} \lambda_1 &= \beta_0 \\ \lambda_2 &= 2\beta_1 - \beta_0 \\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= 20\,\beta_3 - 30\,\beta_2 + 12\,\beta_1 - \beta \end{aligned} \quad (4.9)$$

Where λ_1 is simply L-location (mean) and λ_2 is the L-scale of the distribution respectively. The L-moments ratios (Hosking) are identified by the L-coefficient of variation (τ), L-coefficient of skewness (τ_3) and L-coefficient of kurtosis (τ_4), are computed as:

$$\tau = \frac{\lambda_2}{\lambda_1}, \tau_3 = \frac{\lambda_3}{\lambda_2} \; and \; \tau_4 = \frac{\lambda_4}{\lambda_2}.$$

4.3.3 Sample L-Moments

In practice, theoretical L-moments defined above need to be estimated from a finite sample. Let $z_{1:n} \leq z_{2:n} \leq \ldots \leq z_{n:n}$ be the ordered sample of the n observations. An unbiased estimator of β_r is given by (Landwehr *et al.* [11]) as:

$$b_r = \frac{1}{n}\binom{n-1}{r}^{-1} \sum_{i=r+1}^{n} \binom{i-1}{r} z_{i:n} \tag{4.10}$$

This may alternatively be written as:

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} z_{i:n}$$

$$b_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{i-1}{n-1} z_{i:n}$$

$$b_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(i-2)}{(n-1)(n-2)} z_{i:n}$$

and generally, we can write

$$b_n = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(i-2)\ldots(i-r)}{(n-1)(n-2)\ldots(n-r)} z_{i:n} \tag{4.11}$$

Hence, the four sample L-moments are defined by

$$\begin{aligned} l_1 &= b_0 \\ l_2 &= 2b_1 - b_0 \\ l_3 &= 6b_2 - 6b_1 + b_0 \\ l_4 &= 20\, b_3 - 30\, b_2 + 12\, b_1 - b_0 \end{aligned} \tag{4.12}$$

The sample L-moments ratios given as:

$$t_2 = \frac{l_2}{l_1}, t_3 = \frac{l_3}{l_2} \; and \; t_4 = \frac{l_4}{l_2} \tag{4.13}$$

Wheret_2, t_3, t_4 are the sample L-coefficient of variation, L-coefficient of skewness, and L-coefficient of kurtosis respectively.

4.4 Regional Frequency Analysis Procedure

4.4.1 Description of Dataset

In this study, daily rainfall data (1970–2017) of 29 rain gauge stations in Haryana were obtained from the National Data Center, India Meteorological Department (IMD) Pune. These rain gauge stations include Sirsa, Fathebad, Hisar, Jhajjar, Jind Sonipat, Rohtak, Panipat, Gurgaon, Faridabad, Kurukshetra, Narnaul, Farukhnagar, Pataudi, Hansi, Punahana, Beri, Salhawas, Bhiwani, Tohana, Sohana, Bawal, Dujana, Narwana, FirozpurJhirka, Nuh, Mahendragarh, Kaithal, and Rewari. The annual total rainfall series were constructed for each rain gauge station for the analysis.

4.4.2 Formation of the Homogeneous Region/Groups Using Cluster Analysis

There are various grouping methods to construct the homogeneous region in a regional frequency analysis procedure. The homogeneous region can be decided based on geographical convenience, clustering techniques, and Subjective partitioning. Cluster analysis is very popular in regional frequency analysis for identifying homogenous groups. Many different methods can be used to carry out a cluster analysis but in this Ward′s method has been used for clustering in this study.

4.4.3 Heterogeneity test

A heterogeneity test was used to test the homogeneity of formed groups. In this test 500 equivalent groups are generated by Monte Carlo simulations by fitting Kappa distribution to the regional average L-moments ratios $(1, t^R, t_3^R, t_4^R)$. This test compares the variability of L-statistics of the real group to the simulated group. The H test defined as:

$$\mathrm{H} = \frac{V - \mu_{\mathrm{v}}}{\sigma_{\mathrm{v}}} \qquad (4.14)$$

where μ_{v} and σ_{v} represent the mean and standard deviation of simulated V_j values, respectively, V-statistic as follows:

$$V = \left\{ \frac{\sum_{i=1}^{N} n_i \left(t^i - t^R \right)}{\sum_{i=1}^{N} n_i} \right\}^{\frac{1}{2}}$$

If the value of H is less than 1 i.e. <1; the group can be considered homogeneous, if $1 \leq H < 2$; the group can be considered heterogeneous and if $H > 2$; the group is definetly heterogeneous.

4.4.4 Discordancy measure

A discordancy measure (D_i) given has been used in this study to identify discordant sites/stations within a group. The discordancy measure statistic D_i for the i[th] site/station is calculatedas $D_i = \frac{1}{3} N (u_i - \bar{u})^T (\sum_{i=1}^{N} (u_i - \bar{u}) (u_i - \bar{u})^T)^{-1} (u_i - \bar{u})$ (4.15)

where u_i is a vector consist of first three sample L-moment ratios of the i[th] site, N is the total sites in the region, and $\bar{u}$ is unweighted regional average of L-moments ratio i.e.

$u_i = (t^i \ \ t_3^i \ \ t_4^i)^T$ and $\bar{u} = \frac{1}{N} \sum_{i=1}^{N} u_i$.

Thus i[th] site is discordant if its D_i > critical value (as shown in Table 4.1)

Table 4.1: Critical values of D_i with Number of Sites/Stations (N) in a group

N	5	6	7	8	9	10	11	12	13	14	≥ 15
Critical value	1. 33	1. 65	1. 92	2. 14	2. 33	2. 49	2. 63	2. 76	2. 87	2. 97	3. 00

4.5 Choice of Regional Frequency Distribution

4.5.1 L-moment Ratio Diagram

The average sample L-moment ratios (t_3^R, t_3^R) are plotted as a scatter plot with the theoretical L-moment ratios (τ_3, τ_4) of candidate distributions in an L-moment ratio diagram. Chose the distribution that gives the closest approximation to the point (t_3^R, t_3^R). The construction of the

L-moments ratio diagram can be achieved by following the equation given as

$$\tau_4 = \sum_{k=0}^{8} \Upsilon_k \, \tau_3^k \qquad (4.16)$$

where Υ_k are the coefficients of polynomial approximations for different distributions as shown in Table 4.2.

Table 4.2: Coefficients of Polynomial approximations of τ_4 as a function of τ_3 (dash indicates the absence of coefficients)

	Υ_0	Υ_1	Υ_2	Υ_3	Υ_4	Υ_5	Υ_6	Υ_7	Υ_8
GEV	0.1070	0.1109	0.8484	-0.0667	0.0057	-0.0421	0.0376	-	-
GLO	0.1667	-	0.8333	-	-	-	-	-	-
GNO	0.1228	-	0.7752	-	0.1228	-	-0.1364	-	0.1137
PE3	0.1224	-	0.3012	-	0.9581	-	-0.5749	-	0.1938

4.5.2 Goodness-of-fit test (Z^{DIST})

After visual inspection by L-moment ratio diagram, additionally a quantitative measure of goodness is calculated to confirm whether the candidate distribution is the best fit for the observed data. This test determines how closely the average regional L-kurtosis fits the theoretical L-kurtosis of the candidate distribution. The value of the Z^{DIST} is calculated as:

$$Z^{DIST} = \frac{\tau_4^{DIST} - t_4^R}{\sigma_4} \qquad (4.17)$$

Where, τ_4^{DIST}- theoretical L-kurtosis of the candidate distribution, t_4^R – group average L-

kurtosis, σ_4 -standard deviation of t_4^R. Calculate Z^{DIST} for all candidate distributions and the distribution is selected if its value of $|Z^{DIST}| < 1.64$ at 90% confidence interval. If more distributions qualify the criteria of Z^{DIST} statistic, then select the distribution which has minimum value of $|Z^{DIST}|$.

There are many probability distributions to model extreme events in hydrological studies such as rainfall, floods, etc. However, there is no particular model that may be considered superior for all practical applications. After reviewing the literature, we selected the following three parameters probability distribution:

- GEV −Generalized Extreme Value,
- GLO −Generalized Logistic,
- GNO −Generalized Normal and
- PE-3−Pearson type-3

4.6 Rainfall Quantile Estimation

Consider that there are N rain gauge stations in a homogeneous region and station i having record length n_i.Then $Q_{i,j}, i = 1,2,\ldots..N$ and $j = 1,2,\ldots..n_i$ denote the observed data and $Q_i(F), 0 < F < 1,$ be the quantile function of frequency distribution at the i^{th} rain gauge station. Then index-flood method assumes that frequency distributions of N sites are identical except for a scaling factor or index flood. Formally we can write

$$Q_i(F) = \mu_i q(F), \quad i = 1,2,\ldots N \tag{4.18}$$

where μ_i is the index-flood (or scaling factor) for station i, usually estimated by sites' mean and $q(F)$ is the dimensionless quantile function of non-exceedance probability F. The sample mean at i^{th} site is estimated by $\hat{\mu} = \sum_{i=1}^{N} Q_i/n$ and the dimensionless rescaled data are computed by $q_{i,j} = Q_{i,j}/\hat{\mu}_i$ which are the basis for estimating $q(F)$.

As per Hosking & Wallis regional parameters can be estimated as:

$$\hat{\theta}_k^R = \frac{\sum_{i=1}^{N} n_i \hat{\theta}_k^i}{\sum_{i=1}^{N} n_i} \tag{4.19}$$

where N is the number of rain gauge stations,$\hat{\theta}_k^i$ is the i[th] L-moment to be estimated, and n_i is the data length at station i. Substitution of these estimates into $q(F)$ gives the estimated regional quantile.

$\hat{q}(F) = \hat{q}(F; \theta_1^R, \theta_2^R, \ldots\ldots, \theta_p^R)$. Thus quantiles at site i with non-exceedance probability F or different return can be obtained combining the estimates of μ_i and $q(F)$:

$$\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F).$$

Software used: R Studio version 1.2.1335 has been used for all statistical analysis and graphical displays. The R package lmom RFA has been used for L-moment analysis.

4.7 Results and Discussion:

4.7.1 Cluster Analysis

Before cluster analysis, all stations were considered as a single large homogeneous group, and the heterogeneity measures H was calculated for this.

The H value was found 1.86 for this, which means 29 stations together are considered as possibly heterogeneous since 1<H<2. Thus, there was a need for further subgrouping of stations. The cluster analysis technique using Ward's method was used. The resulting dendrogram with three groups/clusters is shown in Fig.4.4. Group G1 contains 10 stations, group G2 contains 8 stations and group G3 contains 11 stations. In further text, we have used only these notions (G1, G2, and G3) to denote a group or cluster.

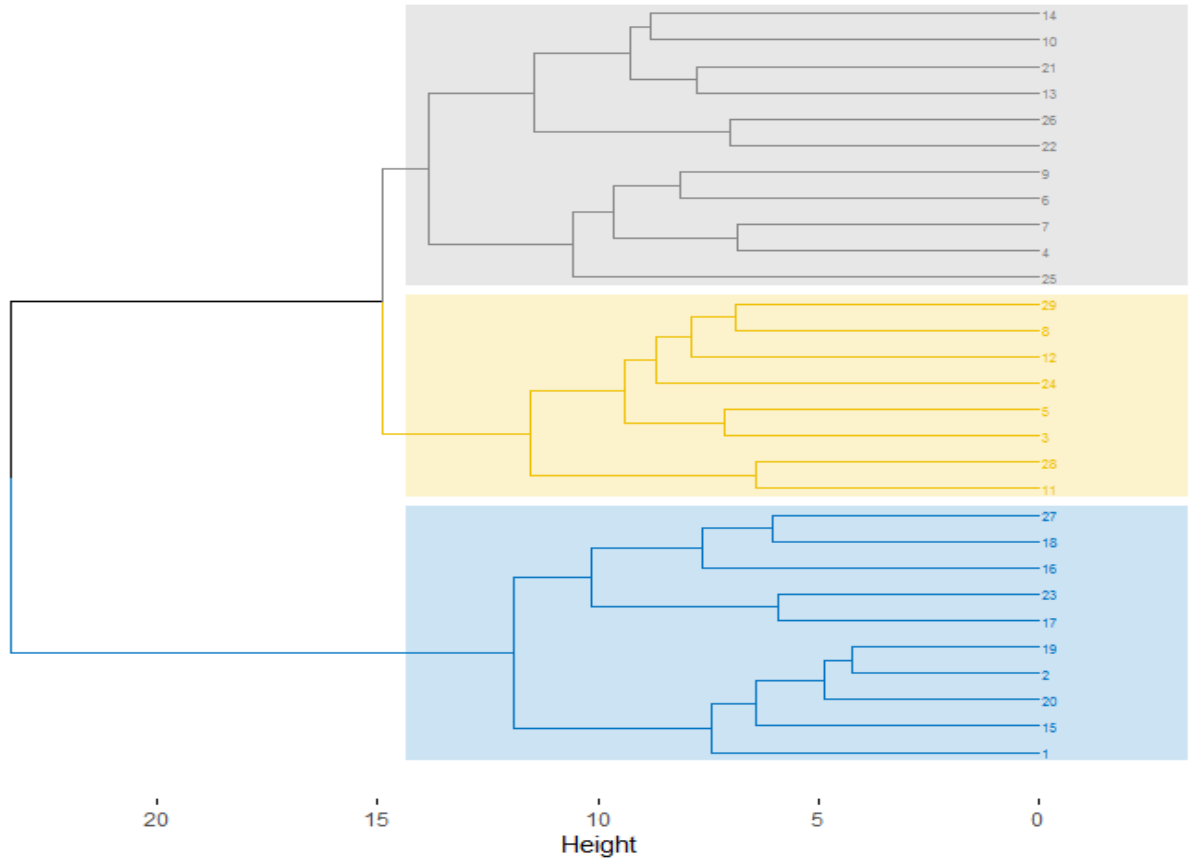


Figure 4.3 Dendrogram by Ward's method

L-moments ratios and discordancy measure values are computed and summarized in Table 4.3. Discordancy measure (D_i) was computed for each station in all groups. The critical values of D_i have been presented in table 1. D_i values range from 0.51 to 2.21 for all stations in group G1, which are less than the critical value of 2.49 for this region. In group G2 and G3 also no discordant station was found. The regional average L-moment ratios were computed using the L-moment ratios in Table 3, and these regional average L-moment ratios were then used in the kappa distribution for simulation in homogeneity test (H) for each homogenous group, i.e. G1, G2, and G3. The results of heterogeneity measures (H) indicated that all three groups can be considered homogeneous because none of these regions have H >1 (Table4.5).

Table 4.3 Rain gauge Stations and their L-Moments Statistics of Annual Total Rainfall

Groups	Mean	l-coefficient of variation(L-	L-coefficient of	L-coefficient	D_i valu

G1	(l_1)	Cv)	Skewness(L-Cs)	of Kurtosis (LCk)	e
Sirsa	344.921	0.252	0.094	0.138	0.51
Fathebad	314.728	0.246	0.189	0.108	0.68
Hansi	258.11	0.27	0.149	0.034	0.99
Punahana	405.931	0.291	0.100	0.174	1.04
Salhawas	359.259	0.309	0.207	0.175	2.21
Bhiwani	360.655	0.232	0.187	0.09	1.04
Tohana	367.033	0.257	0.149	0.137	0.12
Mahendergarh	387.082	0.221	0.098	0.165	1.51
Dujana	471.97	0.301	0.108	0.083	0.71
Beri	424.981	0.289	0.088	0.048	1.19
G2					
Kurukshetra	543.858	0.289	0.255	0.172	1.68
Kaithal	528.510	0.256	0.192	0.165	0.15

Hisar	444.492	0.210	0.033	0.081	1.23
Jind	463.214	0.248	0.110	0.121	0.82
Panipat	519.669	0.206	0.119	0.107	1.15
Narwana	492.087	0.256	0.207	0.208	1.13
Rewari	526.546	0.220	0.131	0.087	0.94
Narnaul	526.056	0.256	0.154	0.190	0.91
Firozpur jhirka	610.875	0.259	0.215	0.227	1.30
G3					
Jhajjar	573.078	0.232	0.068	0.026	1.71
Rohtak	545.302	0.233	0.005	0.106	0.79
Sonipat	642.489	0.225	0.092	0.087	0.37
Gurgaon	625.962	0.232	0.207	0.147	1.20
Faridabad	524.72	0.226	0.002	0.164	1.24
Farukhnagar	393.03	0.299	0.138	0.184	1.99

	8				
Pataudi	471.052	0.244	0.111	0.103	0.30
Sohana	496.998	0.202	0.082	0.155	0.62
Nuh	608.158	0.24	0.145	0.173	0.22
Bawal	583.160	0.183	0.024	0.134	1.26

The Z^{DIST} statistic values for the candidate three-parameter distributions GLO, PE3, GEV, and GNO were computed (Table 4.5). The best fit distribution for each group according to $|Z^{DIST}| \leq 1.64$ is presented in Table 4.8. For group G1; GNO, GEV, and PE3 are good fit distributions and the most suitable is PE3 because it has the lowest $|Z^{DIST}|$ value. For group G2; GNO, GEV, and PE3 are good fit distribution but the best fit is GEV (smallest $|Z^{DIST}|$). In case of group G3 among three fitted distributions (GEV, GNO and PE3) GNO is the best fit distribution due to the lowest $|Z^{DIST}|$ value. The estimated parameters of all fitted distributions have been presented in Table 4.9

Table 4.4 Heterogeneity measure for three groups

Groups	G1	G2	G3
H-value	0.78	0.6	0.84

Table 4.5 Z^{DIST} Value for Candidate Distributions

Groups	Z^{GLO}	Z^{GNO}	Z^{GEV}	Z^{PE3}	Best fit
G1	4.44	1.48*	1.54*	0.90*	PE3
G2	2.35	-0.12*	0.02*	-0.69*	GEV
G3	2.42	-0.47*	-0.72*	-0.78*	GNO

*indicate level of significance which is 10%

For group G1 (Figure 2(a)), the regional average of L-skewness and L-kurtosis lie closet to the PE3 distribution curve, so PE3 is considered a suitable candidate regional distribution for this region. In case of group G2 (Figure 2 (b)) and in the case of group G3 (Figure 2 (c)), the regional average L-skewness and L-kurtosis lie closet to the GEV and GNO distribution curves, so, for group G2 and G3; GEV and PE3 can be considered suitable candidate regional distributions.

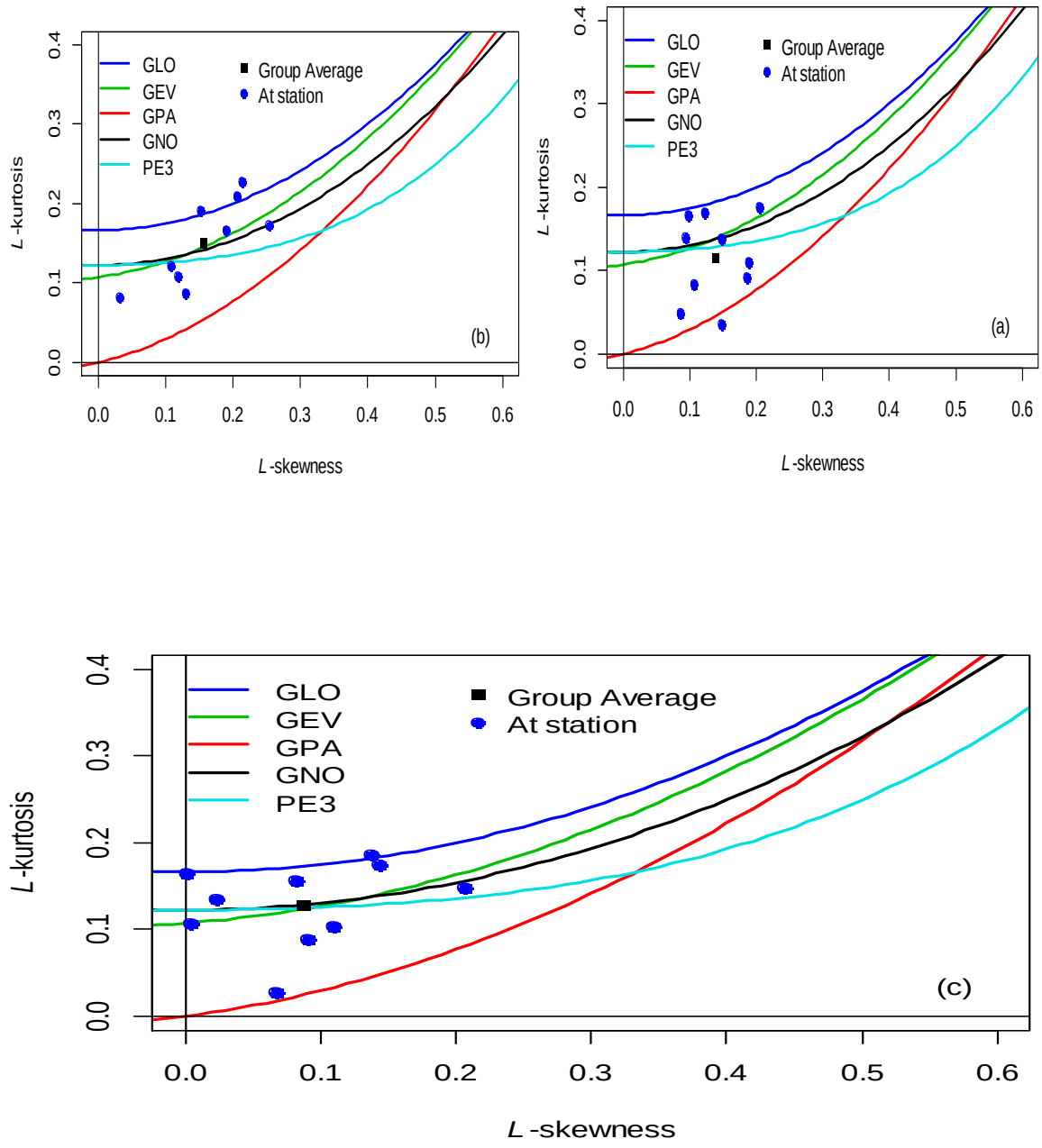


Figure 4.4 L moment ratio diagrams for three groups; (a) for group G1, (b) for group G2, and (c) for group G3

Table 4.6 Estimated parameters (at 0.90 Level)

Groups	Distributions	parameters		
		α	ξ	k
G1	GEV	0.788	0.404	0.053
	GNO	0.935	0.457	-0.280
	PE3	1.000	0.483	-0.280
G2	GEV	0.804	0.360	-0.306
	GNO	0.935	0.360	0.033
	PE3	1.000	0.441	0.906
G3	GEV	0.824	0.371	-0.103
	GNO	0.958	0.408	-0.203
	PE3	1.000	0.420	0.605

Table 4.7 Estimated regional quantiles using best fit distributions with accuracy measure.

Groups	**Dist.**	**F**	**Quantile estimates**	**RMSE**	**$LEB_{(0.05)}$**	**$UEB_{(0.95)}$**
G1	PE3	0.5	0.934	0.012	0.913	0.952
		0.8	1.375	0.011	1.356	1.393
		0.9	1.646	0.024	1.607	1.689
		0.95	1.891	0.040	1.830	1.964
		0.98	2.191	0.064	2.096	2.309
		0.99	2.406	0.082	2.282	2.558

		0.995	2.613	0.101	2.460	2.802
		0.5	0.935	0.011	0.915	0.954
G2	GEV	0.8	1.330	0.012	1.309	1.350
		0.9	1.584	0.022	1.547	1.620
		0.95	1.822	0.041	1.754	1.893
		0.98	2.122	0.075	1.996	2.253
		0.99	2.340	0.107	2.164	2.529
		0.995	2.553	0.145	2.317	2.809
		0.5	0.958	0.009	0.943	0.973
G3	GNO	0.8	1.333	0.010	1.316	1.349
		0.9	1.555	0.019	1.525	1.587
		0.95	1.755	0.032	1.706	1.809
		0.98	1.998	0.051	1.918	2.084
		0.99	2.171	0.068	2.067	2.282
		0.995	2.339	0.084	2.205	2.482

LEB= Lower Error Bound, UEB= Upper Error Bound, Dist= Distribution

The estimated regional quantiles with RMSE and 90% error bounds for each homogeneous region have been presented in Table 7. For group G1, the RMSE value varies from 0.012 to 0.101 when the return period is 200 years. For group G2, the RMSE value ranges from 0.011 to 0.145, and in group G3 it ranges from 0.009 to 0.085. Group G1 and G3 have relatively lower RMSE values as compared to group G2. It indicates that regional quantiles produced by GNO in group G3 are more accurate than quantiles in group G1 and G2 due to low RMSE and narrower error bounds. Also, the PE3 in group G1 produces less RMSE as compared to GEV in group G2. Thus quantile estimates by GNO in group G3 are more reliable than others. These estimated rainfall quantiles can be further extended based on requirements.

4.7.2 Regional Growth Curve

The frequency distribution at all stations in a homogeneous region is summarised by the growth curve.The regional growth curve for each group is developed using best-fit distribution have been illustrated in Figure 3. Groups G1 and G2 growth curves are looking similar to the return period of 100 years but provide high quantiles as compares to group G3.

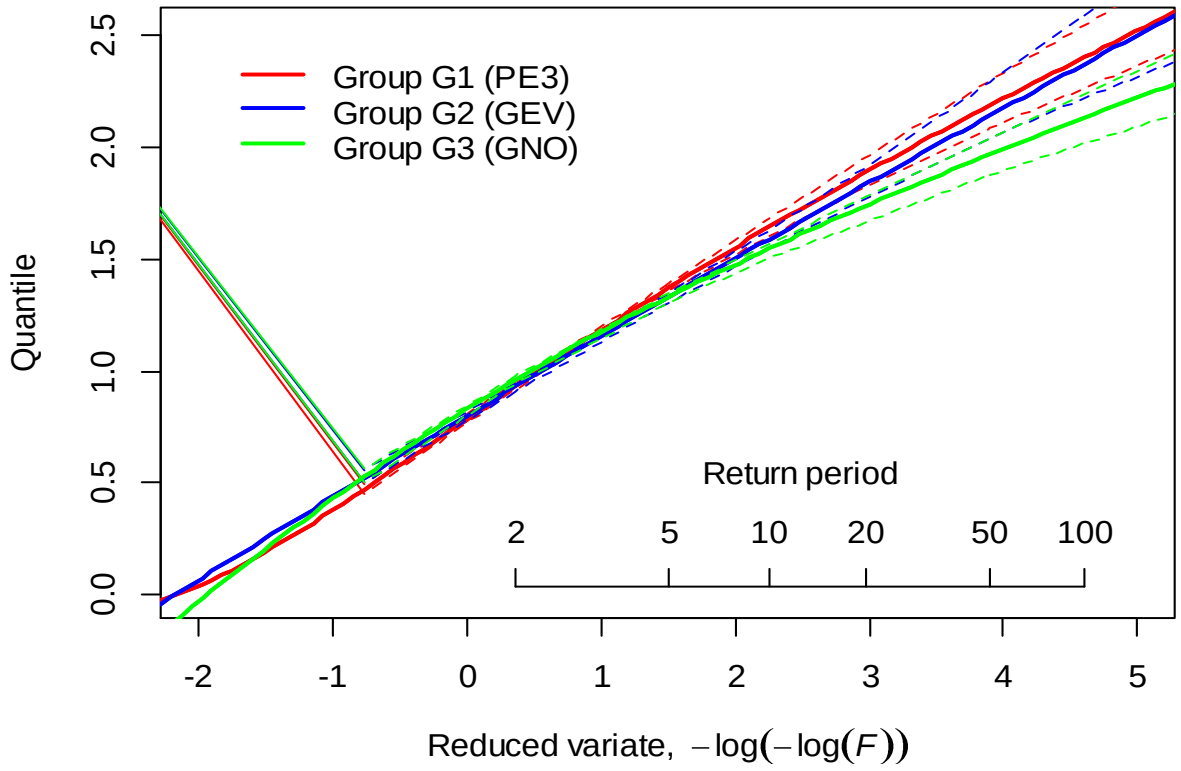


Figure 4.5 Regional growth curves (with 90% error bounds) for the three groups

Quantiles are calculated at each station using best-fitted distributions and their extreme value plots are also drawn. For the sake of brevity, The extreme value plots for only six stations have been presented in Fig. 4(a-f) i.e. two stations from each group, respectively. The rainfall quantiles estimated using the best-fitted distributions are found to be close agreement to the observed annual total rainfall series as shown in these plots.

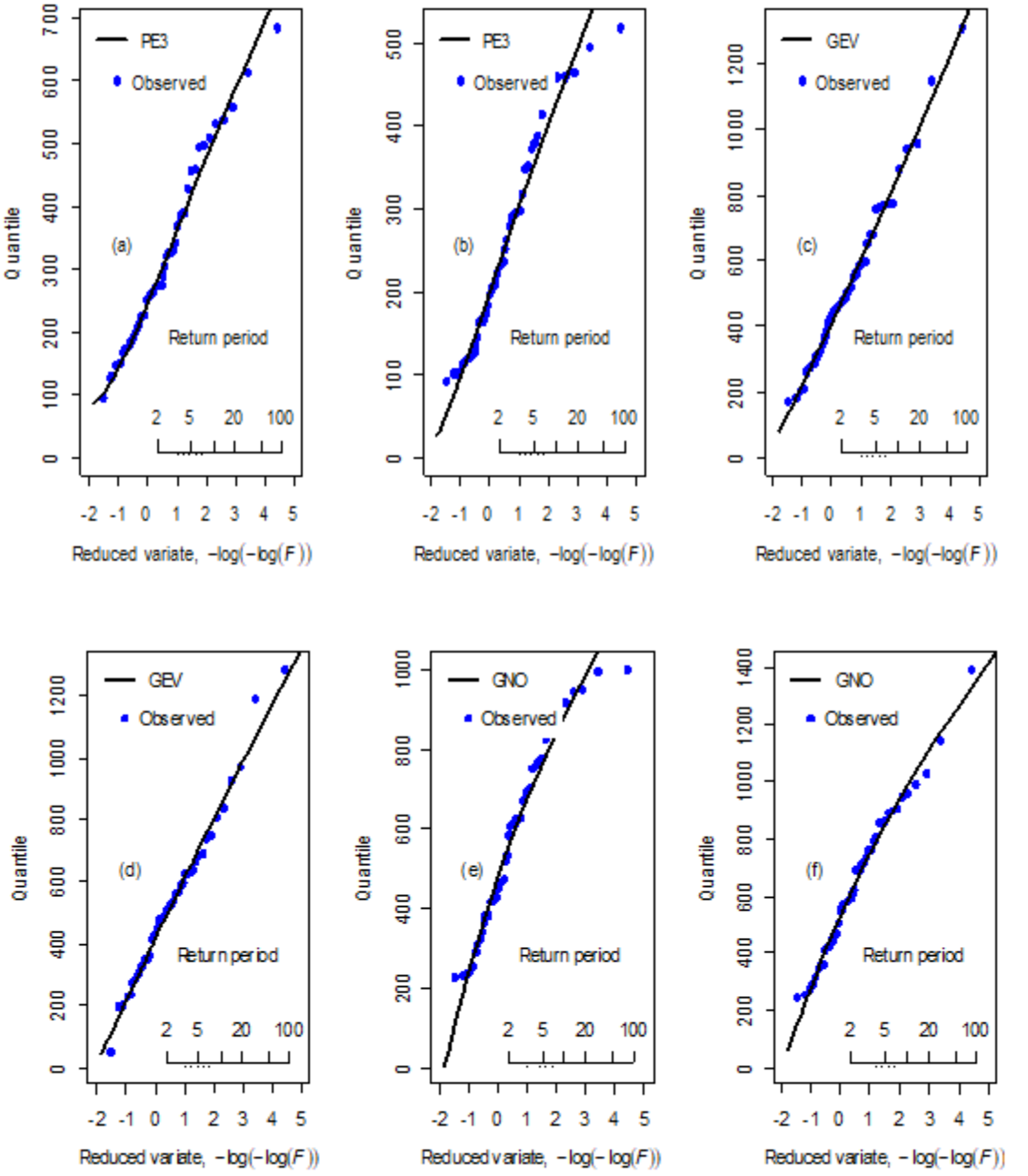


Figure 4: Extreme value plot for a) Fathebad station, b) Hansi station, c) Kaithal station, d) Naural station, e) Jhajjar station, f) Sonipat station

Figure 4.6 Extreme value plot of six stations

In Figure 4.5, 2 to 200-year rainfall return levels of all stations are illustrated. The vertical axis represents the annual total rainfall. These return levels are obtained using best-fit distributions. As an example, for the station Dujana, which is best-fitted by the PE3 distribution, the rainfall amounts of 2, 5, 10, 20, 50, 100, and 200-year return periods are

440.4 mm, 647.7 mm, 775 mm, 891 mm, 1032 mm, 1134, and 1232 mm respectively. These estimated rainfall quantiles are very useful and maybe a rough guideline for policymakers and structural engineers for planning and designing hydraulic structures.

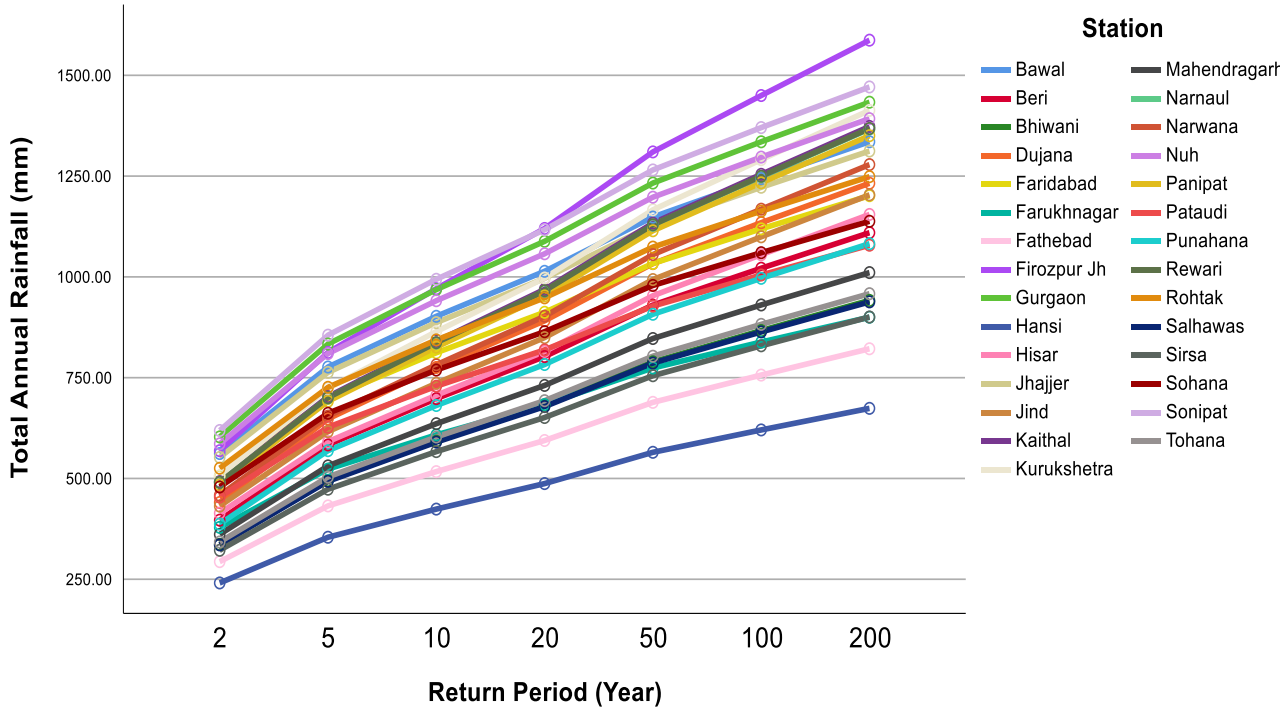


Figure 4.7 Estimated annual total rainfall quantiles at each station using best fit distribution

4.8 Conclusion

In this study, a regional frequency analysis of the annual total rainfall of 29 rain gauge stations was carried out. Using ward's method 29 stations were divided into three homogeneous groups using cluster analysis. It was observed that PE3 distribution was best fit for group G1, GEV distribution was best fit for group G2 while for group G2; GNO was found best distribution. The regional quantiles with their accuracy measures showed that GNO in group G3 has a relatively low RMSE value and 90% error bounds. Furthermore, for higher return periods, rainfall quantiles showed uncertainty and should be treated with warnings. Using the index-flood method rainfall quantiles are obtained for each station for 200-year return periods. Identifying the best-fit distribution amount of annual total rainfall data could have a wide range of applications in agriculture like crop planning and water-related projects in the state. It

is recommended that at least these three distributions (PE3, GEV, and GNO) be considered for the final selection of distribution for annual total rainfall in Haryana.

Results of H test showed that all three groups were homogeneous and indicated that there was no discordant station in each group. Best-fit distribution was chosen using a goodness-of-fit measure (Z^{DIST}) and L-moment ratios diagram. Results of these tests showed that Pearson type-3 (PE-3) was the best-fitted distribution for group G1 and Generalised Extreme Value (GEV) and Generalised Normal (GNO) were the best for group G2 while for group G3; GNO was the best-fitted distribution. Using the L-moments method, the parameters of the fitted distributions were estimated, and regional growth curves were generated for each homogenous group. Regional growth curves provided regional quantiles at various return periods for each group.

Chapter V

REGRESSION MODELS IN THE FIELD OF AGRICULTURE

5.1 Introduction

Forecasting with regression models can be a powerful tool for predicting future trends. However, the accuracy of the forecasts depends on the quality of the data and the accuracy of the model. If the data is unreliable or the model is poorly constructed, the forecasts will not be accurate. Regression models use linear equations to predict a response variable given a set of predictor variables. Linear regression is a statistical technique used to predict numerical values. It is one of the oldest and most widely used predictive techniques in both academic and business settings. The agriculture is the backbone of our country. This is the field on which our economy lies and none of the country can think their development without agriculture but agriculture depends on many climatic conditions for the crop yield and production so this study is for the rainfall analysis of Haryana state which contributes 15% of agricultural production in India

The basic mathematics behind linear regression is that a straight line can be used to describe the relationship between two variables $\gamma = \beta_0 + \chi$ where β_0 is the intercept and χ is the slope.

The formula for linear regression is a generalization of the two-variable linear equation. It is written as $\gamma = \beta_0 + \beta\ \chi_0\ + \beta_2\ \chi_2 + \ldots\ldots + \beta_n\ \chi_n$ Where β_0 is the intercept, slopes and the independent variables are β_i, χ_i respectively for all i = 1,2, ……n.

Multiple linear regression models can be used to make forecasts based on past data, to identify relationships between independent and dependent variables, and to test hypotheses about the relationships between the variables. It can also be used to identify non-linear relationships between variables.

5.1.1 VARIABLES: DEPENDENT AND INDEPENDENT

Regression models are a type of statistical model used to assess the relationship between one or more independent (predictors) and one or more dependent variables (outcomes). The independent variables are the variables that are used to predict the dependent variable, and the dependent variables are that being predicted.

Independent variables are the predictors or inputs of the regression model. They are the causes that are used to explain the variation in the dependent variable. Independent variables can be continuous, discrete, or categorical. Dependent variables are the outcomes or responses of the regression model. They are the variables that are being predicted by the independent variables. Dependent variables are usually continuous but can also be discrete or categorical.

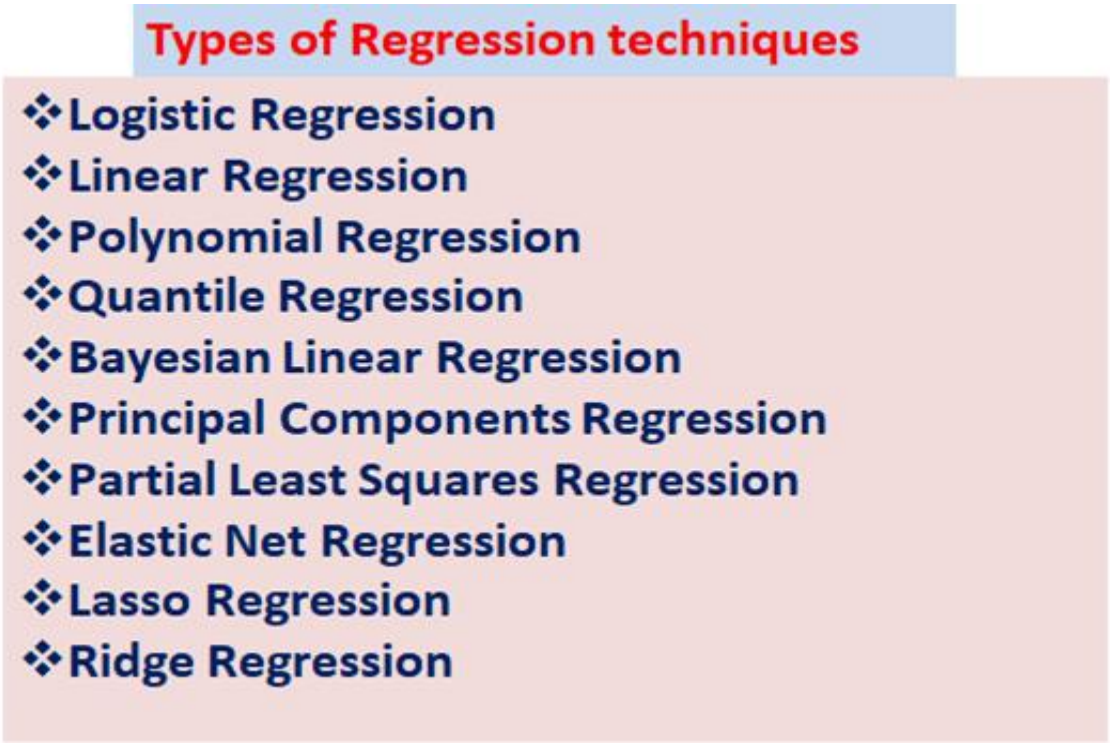


Figure 5.1: Regression models and their types

Logistic regression and linear regression are both machine learning algorithms used for predictive modeling. Linear regression is a supervised learning algorithm while Logistic regression is also a supervised learning algorithm but is used for predicting binary categorical outcomes. It models the relationship between the dependent variable and independent

variables using a logistic function, which calculates the probability of the dependent variable belonging to a particular category. Logistic regression uses maximum likelihood estimation to find the coefficients that maximize the probability of the observed data.

Polynomial regression: It is a type of linear regression that allows for non-linear relationships between the dependent and independent variables. It involves fitting of a polynomial function for a specified degree to the data points, thereby capturing more complex patterns and curvature.

Quantile regression: It is a regression technique that focuses on estimating the conditional quantiles of the dependent variable instead of the mean. It models the relationship between the predictors and various quantiles of the response variable, allowing for a more flexible analysis of the data. This technique is particularly useful when studying the distributional properties of the data or when the response variable is not normally distributed.

Elastic Net Regression, Lasso Regression, and Ridge Regression are all regularization techniques used in linear regression to combat over fitting and improve the model's predictive performance.

Ridge Regression: It is also known as L2 Regularization. Ridge Regression adds a penalty term to the linear regression objective function which is the sum of the squares of the coefficients multiplied by a regularization parameter (alpha). This penalty term helps to shrink the coefficients towards zero, reducing their impact on the model and preventing overfitting. Ridge Regression is particularly effective when dealing with multi collinearity, as it forces the model to select a subset of correlated variables rather than relying on all of them equally.

Lasso Regression: It is known as L1 Regularization Similar to Ridge Regression, Lasso Regression also adds a penalty term to the objective function. However, Lasso Regression uses the sum of the absolute values of coefficients multiplied by a regularization parameter (alpha) as the penalty term. This penalty term encourages sparsity in the model, meaning it tends to set some coefficients to zero, effectively selecting only the most relevant features. Lasso Regression is particularly useful in feature selection, as it can automatically perform variable selection by eliminating irrelevant or redundant predictors.

Elastic Net Regression: It is a combination of Ridge Regression and Lasso regression. So it

is a hybrid method combination of both techniques to improve the accuracy by reducing error. Linear regression is used for continuous numeric predictions, while logistic regression is used for binary classification problems.

5.2 Forecasting model for crop yield:

Classical methods for time series forecasting are a set of well-established techniques used to make forecasts about future values based on past data. These methods include ARIMA models, exponential smoothing, and Holt-Winters seasonal decomposition. ARIMA models are used to identify and quantify the relationships between past values and future values, while exponential smoothing and Holt-Winters decomposition are used to identify and quantify seasonal patterns. These methods are widely used by organizations and companies to gain insights into future trends and anticipate changes in demand, inventory, and prices of data and in this chapter the focus is on the factor which affects the crop.

The first step is to know about the data and the Fig 5.2 shows the time plot series of data.

```
import matplotlib.pyplot as plt
%matplotlib inline
```

```
df.columns=["Year","Yield_1"]
df.head()
df.describe()
df.set_index('Year',inplace=True)

from pylab import rcParams
rcParams['figure.figsize'] = 15, 7
df.plot()
```

Figure 5.2 Data import and analysis of crop

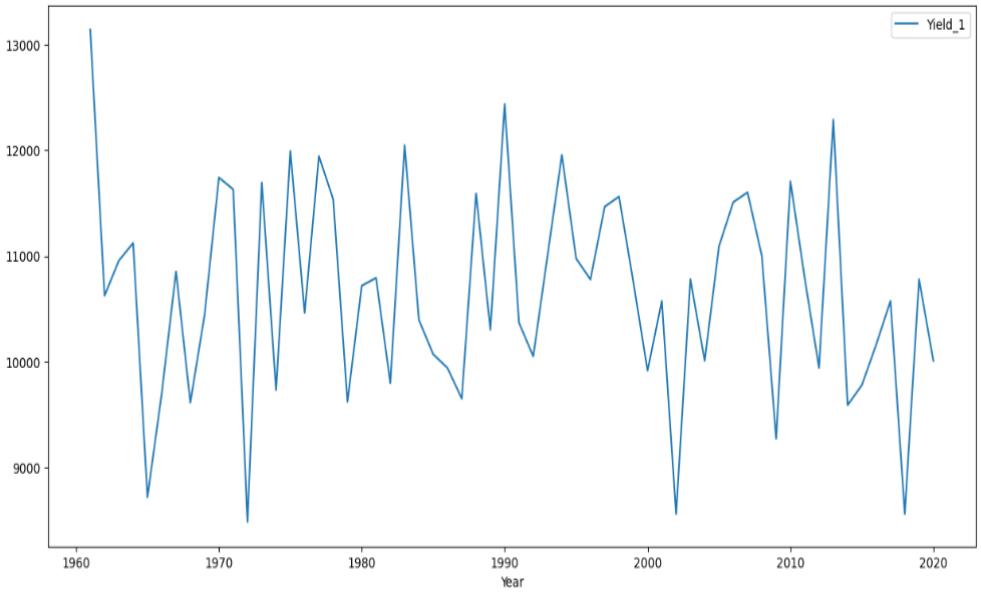


Figure 5.3 Time Plot series of the crop yield data

Time plot is also to check that whether the data is of stationary type or not and then by using various coefficients, PACF and ACF identify the values of p, d and q then on the basis of this Table1 shows various values on different (p,d,q) which is suitable for different years on the basis of various independent factors so based AIC (Akaike information criteria), Adj R^2 and Mape values different Arima models choosen for the crop yield in respective years which is shown in Table5.1 by following the various steps as shown in Fig 5.2

```
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf
import statsmodels.api as sm
fig = plt.figure(figsize=(12,8))
ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(df['Seasonal First Difference'].dropna(),lags=40,ax=ax1)
ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(df['Seasonal First Difference'].dropna(),lags=40,ax=ax2)
```

Algorithm

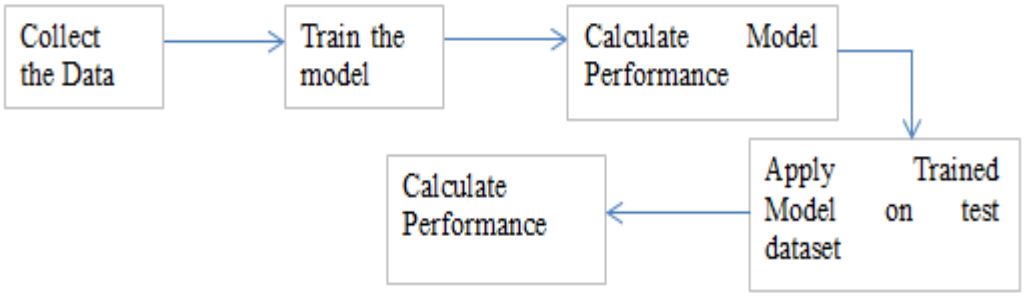


Figure 5.4 Multiple Regression methodology for crop yield prediction

Table5.1 shows the various orders of partial autocorrelation coefficients and autocorrelation coefficients for crop yield production in Jammu and Kashmir. The 95% confidence interval fall outside and ACF's are different from zero. It indicates crop yield production is far from stationary. For data n = 60 implying variance is 1/60 or standard error (S.E) of ($\sqrt{60}^{-1}$) = 0.064 then at 95% confidence interval ρ_k will be $\pm(1.96)\times(0.064)$ = 0.12544. It can be in both side of the zero i.e., the confidence will be (-0.12544, 0.12544). The Fig 5.1 plots the data for crop yield and the time series plot also shows that this time series data is not stationary and PACF's are statistically significant after Lag 2. In this to make data stationary, the differencing is required.

Table 5.1 Forecasted Crop yield and evaluation by the ARIMA models.

ARIMA (p,d,q) Model	Adj R^2	MAPE	AIC values	Year for forecasting	Forecasted value
(0,1,2)	0.56	14.57	13.12	2021	39586
(1,1,2)	0.40	11.52	14.31	2022	41045
(1,1,2)	0.33	9.68	16.85	2023	42064.8
(2,1,1)	0.14	10.45	11.22	2024	42118
(0,1,2)	0.42	9.54	21.43	2025	44689.08

5.3 Diagnostic checking:

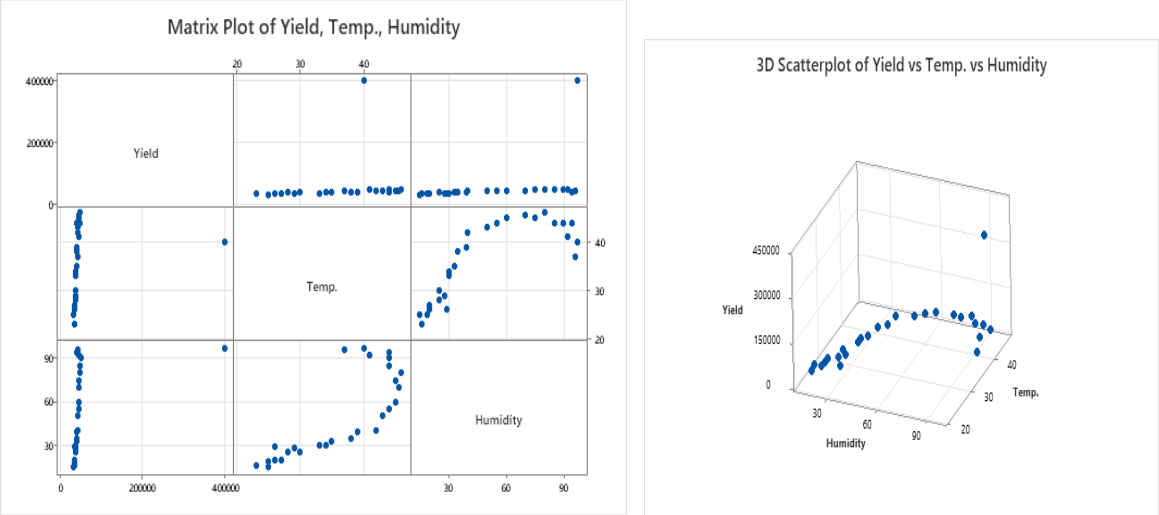


Figure 5.5 Matrix Plot and 3D scatterplot of yield with Temperature and humidity

Firstly to ensure that is it applicable for multiple linear regression model, Plot multiple scatter diagram which is known as the matrix plot in Minitab software and the above Fig 5.3 shows Matrix plot and 3D scatterplot of yield with other variables. Hence, From the above Fig 5.3 it can be accurately said that multi linear regression model can be applied on this data. Moreover, ANOVA (Analysis of Variance) is a statistical method used to compare the means of multiple groups or treatments. It determines whether there are any statistically significant differences between the means of the groups being compared. ANOVA calculates the variability between the groups and within the groups and compares them to determine if the differences observed are likely due to chance or if they are statistically significant. The F-test is commonly used in ANOVA to assess the significance of these differences. ANOVA can be used for both balanced and unbalanced designs, and there are different types of ANOVA tests such as one-way ANOVA, factorial ANOVA, and repeated measures ANOVA.

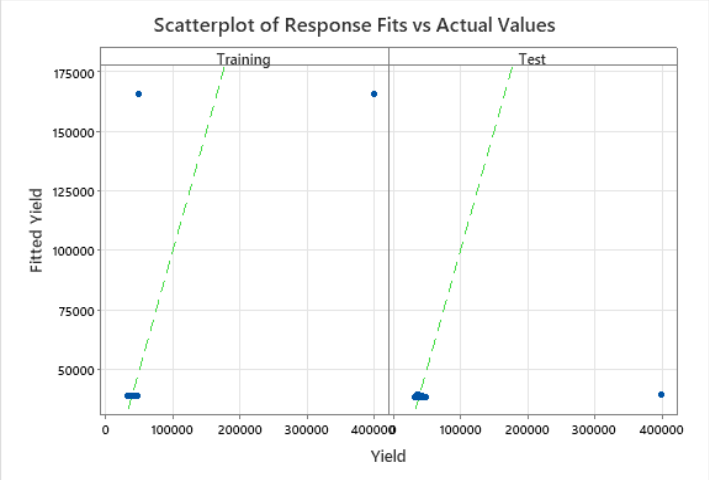


Figure 5.6 Scatterplot of responses of residuals with fitted and actual values on training and testing data.

In Fig 5.4 the response variable has been taken as Yield and in this the confidence interval that is alpha has been chosen as 95% and it is shown that the residuals are properly scattered and follows the same pattern for testing also.

In this $Z_i = \gamma_0 + \gamma_i x_{i1} + \gamma_2 x_{i2} + \epsilon_i$, I = 1, 2, …,40

n=40, K=3

In general it is expressed as $Z = X\gamma + \in$ where X is the matrix of order $n \times k$

γ is the K ×1 matrix of regression coefficients and ∈ is the n×1 matrix of random variable.

Z is n ×1 matrix and in this n = 40 and K = 3 including intercept term

$\gamma = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$ here γ_0 is intercept term and other two are slope parameters. First of all, only two factors chosen for crop yield prediction thus, the Regression equation with respect to 2 independent variables is

Crop Yield = 21361 +412.3 Temperature + 77.5 (humidity)

Table 5.2 Analysis of Variance and evaluation of model with 2 independent factors

Coefficients term	Coef	SE coef	95% cr	T-value	P-value	VIF
Constant	21361	1206	(18871,23851)	17.71	0.000	
Temperature	412.3	46.6	(316.2, 208.5)	8.85	0.000	3.41
Humidity (%)	77.5	12.6	(52.6, 103.5)	6.18	0.000	3.41

Table 5.3 Model summary with 2 studied factors

S	R-sq	R-sq (adj)	Press	R-sq (Pred)
0.58444	99.93%	99.92%	7712.53	99.92%

For Hypothesis

$H_0: \gamma_1 = \gamma_2 = 0$

In this the $\alpha = 0.05$ and P value is 0.00 which is < α

Hence H_0 Rejected.

Error 0.07% Total contribution 100%. In this four independent variables has taken for the study and thus the regression equation is based on 4 independent variables.

Regression equation with respect to 4 variables

Crop Yield=14.9 + 1.911γ_1 + 299.358γ_2 -7.314γ_3 -3.442γ_4

Table 5.4 Analysis of Variance and evaluation of model with 4 independent factors

Coefficients term	Coeff.	SE Coeff.	95% cr	T-value	P-value	VIF
Constant	14.9	13.8	(12.5,42.3)	1.08	0.283	
X_1	1.911	0.899	(316.2, 208.5)	8.85	0.000	1.01
X_2	299.358	0.844	(52.6, 103.5)	6.18	0.000	1.03
X_3	-7.314	0.905	(-9.11, -5.517)	2.13	0.000	1.03
X_4	-3.442	0.857	(-5.14, -1.740)	04.01	0.036	1.01

Table 5.5 Model summary with 4 studied factors

S	R-sq	R-sq (adj)	Press	R-sq (Pred)
994.608	96.73%	96.46%	30407053	95.81%

For Hypothesis

$H_0: \gamma_1 = \gamma_2 = 0$

In this the $\alpha = 0.04$ and P value is 0.00 which is < α Hence H_0 Rejected.

With Error 3.27% and Total contribution 100%

From the above it is concluded that number of independent variables and accuracy follows inverse Proportion. It shows the constant coefficient, standard error coefficients and number of variables which is taken place for multiple linear regression model Table 2 and Table 4 along with the model summary in Table 3 and Table 5. Table 2 shows that with two independent variables the accuracy of 99.52% can be achieved by this model which is shown

by predicted R square value in Table 3 but it is inferred that as the number of independent variable increases the accuracy will decrease and thus the predicted R square become 95.81% in Table 5.5.

5.4 Conclusion

Crop yield prediction with different variables has been taken with two different regression models i.e., Arima and multiple linear regression model. Both the models shown indispensable results, the Arima model given the forecasted values and the multiple regression methods shown the accuracy of 99.52% and can be applied upto the number of terms. ARIMA model has been applied in this paper to forecast crop yield along with effect of various factors affecting it. The data has been taken and all the necessary steps of ARIMA followed systematically for the forecasting of 5 years from 2021 onwards. Moreover, the crop yield depends on various factors and while considering ceteris paribus assumptions the forecasting is accurate.

Chapter VI

STOCHASTIC MATHEMATICAL MODELS FOR FORECASTING INDIA'S ELECTRICITY GENERATION

6.1 Introduction: The world's aggregate energy consumption is increasing in lockstep with the economy's rapid development like wealth, health etc. it plays the major role for the development of a country. Nowadays, due to many factors like high temperature, dry airs the consumption of electricity is growing exponentially. Furthermore as India is having large population so it is quite challenging to meet the demand of whole India. Thus, it is mandatory to forecast the electricity demand of country and to pre-plan the system for electricity production as According to population estimates, India may become the most populous country by leaving behind China and it consists of almost 1.304 billion people. Furthermore it is said that 300 million Indians were without power when there is the least demand of power or it is of less use and it is also concluded that electricity demand may get doubled by 2035 so the country must devote itself to the exploitation of power resources to meet the sufficient demand of people for the power. Our country India heavily dependent on coal as the primary fuel which resulted in low energy efficiency, not only caused a shortage of energy but the environmental pollution also caused by these factors. All the problems discussed here are for India's long-term development.

The world's 18% population is in India but the world's energy reserves which are only 6% indicates that the counry was suffering from severe energy poverty. According to the indian government order to adapt to population growth and economic development order to make balance between power consumption and power demand. To address these energy challenges, India's current development prioritized renewable energy generation, environmental improvement, and energy poverty alleviation. According to the Indian government, 16% contribution will be meet by wind energy for India's electricity demand in 2030.

6.1.1 Research Implications: On the electricity generation data of India which is taken from IAE and the data is in GWH. This study applied five modified models of grey model for the prediction of power generation from 2023 to 2026.Moreover this study also gives the general directions to follow the trend of the India's electricity demand.

(1)Methods can vary according to the data and here regression and classification models are used Then all the forecasted results are related with each other .To make research more systematic and influential every factor is analyzed in detail.

(2)MSE, MAPE and MSPE are parameters to check the accuracy of forecasted data i.e., for the validity of model.

6.1.2 Mathematical description of basic grey method: The concept of grey system is totally based on multidisciplinary theory and in this not all the information is known but in this some is known and partial may be unknown. The grey forecasting model is a mathematical technique that uses a set of past data points to predict the future trend of a given process. It is based on the assumption that the future trends of a process are related to its past trends and that the future can be predicted by extrapolating from past trends. The grey forecasting model is used in fields such as economics, engineering, and meteorology. It is particularly useful for predicting long-term trends that are difficult to predict using traditional forecasting models. The grey forecasting model is a simplified forecasting model that relies on historical data and trends to make predictions. To better understand grey model, the explanation of few notations are: **Table 6.1** List of symbols

Symbol	explanation	Symbol	explanation
$\zeta^{(0)}$	Original sequence	$\zeta^{(1)}$	One accumulation sequence
$\widehat{\zeta}^{(0)}$	Forecasting of $\zeta^{(0)}$	$\hat{\zeta}^{(1)}$	Forecasting of $\zeta^{(1)}$
$\mathcal{T}$	Time sequence	γ	Data matrix

a, b	constant		

6.2 GM (1,1): In this notation (1,1) means the first order and first variable. In this GM denotes grey model. It is used only for positive sequences only.

The basic model GM (1, 1) has time varying coefficients.

Let $\boldsymbol{\zeta}^{(0)} = (\zeta^{(0)}(1), \zeta^{(0)}(2), \ldots., \zeta^{(0)}(n)$ (6.1)

be a sequence then the accumulation sequences can be written as

$\boldsymbol{\zeta}^{(1)} = (\zeta^{(1)}(1), \zeta^{(1)}(2), \ldots. \zeta^{(1)}(n)$ Then

$$\zeta^{(0)}(\varkappa) + a\zeta^{(1)}(\varkappa) = b \quad (6.2)$$

The equation (2) is the original form of GM (1, 1) model and it is known as grey model of 1st order and 1st variable

If $T^{(1)} = (t^{(1)}(2), t^{(1)}(3), \ldots. t^{(1)}(n)$ and

$t^{(1)}(\varkappa) = (\frac{1}{2}\zeta^{(1)}(\varkappa) + \zeta^{(1)}(\varkappa - 1)$ with k = 2,3,. . . .n

Then $\zeta^{(0)}(\varkappa) + at^{(1)}(\varkappa) = b$ (6.3)

and it can be Considered as the basic form of this model.

6.3 Prerequisites results to develop the model

Theorem 1: Let $\zeta^{(0)}$, $\zeta^{(1)}$ and $T^{(1)}$ be the same as above defined with one condition that $\zeta^{(0)}$ is non negative and if $\hat{c} = (c, d)^{\mathcal{T}}$ is a sequence of parameters and

$$\gamma = \begin{bmatrix} \zeta^{(0)}(2) \\ \zeta^{(0)}(3) \\ \cdot \\ \cdot \\ \cdot \\ \zeta^{(0)}(n) \end{bmatrix}, \qquad D = \begin{bmatrix} -T^{(1)}(2)1 \\ -T^{(1)}(3)1 \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ -T^{(1)}(2)1 \end{bmatrix} \quad (6.4)$$

Then equation (3) satisfies $\hat{c} = (D^{\mathcal{T}}D)^{-1} D^{\mathcal{T}} \gamma$

so from notations of theorem 1, If $(c,d)^{\mathcal{T}} = (D^{\mathcal{T}}D)^{-1}\,D^{\mathcal{T}}\,\gamma$

Then $\frac{d\zeta}{dt} + c\zeta^{(1)} = d,$ which became as the whitenization equation of (6.3).

Theorem 2. Let C, γ, $\hat{c}$ be the same as in theorem 1. If $\hat{c}$ = [c, d] $^{\mathcal{T}}$=$(D^{\mathcal{T}}D)^{-1}\,D^{\mathcal{T}}\,\gamma$, then

I. The solution of $\frac{d\zeta}{dt} + c\zeta^{(1)} = d,$ is given by

$$\widehat{\boldsymbol{\zeta}}^{(1)}(t) = \left(\boldsymbol{\zeta}^{(1)}(1) - \frac{d}{a}\right)e^{-ct} + \frac{d}{a}, \tag{6.5}$$

II. Then the time response sequence of $\frac{d\zeta}{dt} + c\zeta^{(1)} = d,$ is given as follows

$$\widehat{\boldsymbol{\zeta}}^{(1)}(\varkappa+1) = \left(\boldsymbol{\varkappa}^{(0)}(1) - \frac{d}{a}\right)e^{-ck} + \frac{d}{a}, \qquad \boldsymbol{\varkappa} = 1, 2, \ldots, n \tag{6.6}$$

The marked * equations are the equations of restored values of $x^{(0)}(\varkappa)$, which is as

$$\widehat{\boldsymbol{\zeta}}^{(0)}(\varkappa+1) = \alpha^{(1)}\,\widehat{\boldsymbol{\zeta}}^{(1)}(\varkappa+1) = \alpha^{(1)}\,\widehat{\boldsymbol{\zeta}}^{(1)}(\varkappa+1) \tag{6.7}$$

6.3.1 Verhulst Model: Pierre franois Verhulst was the first German biologist who introduced the model whose aim to limit the whole system and also helps to describe some increasing processes like S- curve which has a saturation region then by its name this model is named as Verhulst model. The Verhulst model is an equation used to model growth over time. It was developed by Pierre Verhulst in the 19th century and is a special case of the logistic equation, which is a nonlinear differential equation. The equation is used to predict the future growth for whole population given a certain set of parameters such as initial population, growth rate, and carrying capacity. These parameters are all used to determine how the basic need for whole population will grow over time and the future demand can be predicted with this.

Therefore, $\zeta^{(0)}(\varkappa) + at^{(1)}(\varkappa) = b\left(\zeta^{(1)}(\varkappa)\right)^{\alpha}$ (6.8)

the equation (6.8) is established as Power model equation for GM(1,1) and

$$\frac{d\zeta}{dt} + c\zeta^{(1)} = c\left(\zeta^{(1)}\right)^{\alpha} \tag{6.9}$$

The equation (6.9) is known as the whitenization equation of equation (8) and in this $\zeta^{(0)}$ is considered as sequence of raw data and $\zeta^{(1)}$ is the accumulation generation of $\zeta^{(0)}$ and $T^{(1)}$ is adjacent neighbor mean of $\zeta^{(1)}$.

Theorem 3. $\zeta^{(1)}(s) = \left\{e^{-(1-c)cs}[(1-c)\int de^{-(1-c)cs}ds + a]\right\}^{\frac{1}{(1-a)}}$ (6.10)

is the solution of equation (9)

Theorem 4. with $\zeta^{(0)}$, $T^{(1)}$ and $\zeta^{(1)}$ as defined above , Let

$$\mathrm{D} = \begin{bmatrix} -t^{(1)}(2) & \left(t^{(1)}(2)\right)^{\alpha} \\ -t^{(1)}(3) & \left(t^{(1)}(3)\right)^{\alpha} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ -t^{(1)}(n) & \left(t^{(1)}(n)\right)^{\alpha} \end{bmatrix}, \qquad \gamma = \begin{bmatrix} \zeta^{(0)}(2) \\ \zeta^{(0)}(3) \\ \cdot \\ \cdot \\ \cdot \\ \zeta^{(0)}(n) \end{bmatrix} \tag{6.11}$$

Then the least square estimate of the parametric sequence $\hat{c} = [\mathrm{c}, \mathrm{d}]^{T}$ of equation (6.8) is $\hat{c} = (D^{T}D)^{-1}D^{T}\gamma$

When in equation (6.8), the power of $\alpha = 2$, the resultant model will become

$$\zeta^{(0)}(\varkappa) + at^{(1)}(k) = b\left(\zeta^{(1)}(\varkappa)\right)^{2} \tag{6.12}$$

finally the equation (12) is the required grey Verhulst model and

$$\frac{d\zeta}{dt} + c\zeta^{(1)} = b\left(\varkappa^{(1)}\right)^{2} \tag{6.13}$$

Equation (13) is the whitenization equation of grey Verhulst method.

Theorem 5. The solution of equation (6.12) is

$$\zeta^{(1)}(s) = \frac{1}{e^{cs}\left[\frac{1}{\zeta^{(1)}(0)} - \left(\frac{d}{c}\right)(1-e^{-cs})\right]}$$

$$= \frac{c\zeta^{(1)}(0)}{e^{cs}[c - d\zeta^{(1)}(0)(1 - e^{-cs})]}$$

$$= \frac{c\zeta^{(1)}(0)}{d\zeta^{(1)}(0) + (c - d\zeta^{(1)}(0))e^{cs}} \tag{6.14}$$

The time response sequence of the grey verhulst model is

$$\hat{\zeta}(s+1) = \frac{c\zeta^{(1)}(0)}{d\zeta^{(1)}(0) + (c - d\zeta^{(1)}(0))e^{ck}} \tag{6.15}$$

Theorem 6. Let $\boldsymbol{\zeta}^{(o)} = \zeta^{(0)}(1), \zeta^{(0)}(2), \ldots, \zeta^{(0)}(n)$ be a non negative sequence and its accumulation generation $\boldsymbol{\zeta}^{(1)} = \{\zeta^{(1)}(1), \zeta^{(1)}(2), \ldots, \zeta^{(1)}(n)\}$. If $\hat{\beta} = (\beta_1, \beta_2)^{\mathcal{T}}$ is the parametric sequence and

$$\mathcal{Y} = \begin{bmatrix} \zeta^{(0)}(1) \\ \zeta^{(0)}(2) \\ \cdot \\ \cdot \\ \cdot \\ \zeta^{(0)}(n) \end{bmatrix}, \quad D = \begin{bmatrix} \zeta^{(1)}(1) & 1 \\ \zeta^{(1)}(2) & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \zeta^{(1)}(n-1) & 1 \end{bmatrix} \tag{6.16}$$

find the least square estimates of the parameters of the discrete model

$\zeta^{(1)}(\varkappa + 1) = \beta_1 \zeta^{(1)}(\varkappa) + \beta_2$ satisfy $\hat{\beta} = (D^{\mathcal{T}} D)^{-1} D^{\mathcal{T}} \gamma$.

Proof: In grey differential equation

$$\zeta^{(1)}(\varkappa + 1) = \beta_1 \zeta^{(1)}(\varkappa) + \beta_2$$

substitute the data sequence , which produces

$$\begin{aligned} \zeta^{(1)}(2) &= \beta_1 \zeta^{(1)}(1) + \beta_2, \\ \zeta^{(1)}(3) &= \beta_1 \zeta^{(1)}(2) + \beta_2, \\ &\cdot \\ &\cdot \\ &\cdot \\ \zeta^{(1)}(n) &= \beta_1 \zeta^{(1)}(n-1) + \beta_2. \end{aligned} \tag{6.17}$$

Here is the matrix $D\hat{\beta} = \gamma$.

For the pair of estimated values , β_1, β_2 using $\beta_1 \zeta^{(1)}(s) + \beta_2$ to substitute

$\zeta^{(1)}\,(\varkappa+1)$, $k = 1,2,\dots,n-1$, then there will be a error sequence as $\mathcal{E} = \gamma - D\hat{\beta}$.

Assume that

$$\begin{aligned} \mathrm{U} &= \mathcal{E}^{\mathcal{T}}\mathcal{E} = (\gamma - D\hat{\beta})^{\mathcal{T}}(\gamma - D\hat{\beta}) \\ &= \sum_{k=1}^{p-1}\left(\zeta^{(1)}\,(\varkappa+1) - \beta_1\zeta^{(1)}\,(\varkappa) - \beta_2\right)^2 \end{aligned} \tag{6.18}$$

So, let us find the values of $\beta_1\ and\ \beta_2$ to make U smallest as much as possible and satisfies

$$\begin{aligned} \frac{\mathfrak{d}U}{\mathfrak{d}\beta_1} &= -2\left(\left(\zeta^{(1)}\,(\varkappa+1) - \beta_1\zeta^{(1)}\,(\varkappa) - \beta_2\right).\zeta^{(1)}\,(\varkappa) = 0\ and \right. \\ \frac{\mathfrak{d}U}{\mathfrak{d}\beta_1} &= -2\left(\left(\zeta^{(1)}\,(\varkappa+1) - \beta_1\zeta^{(1)}\,(\varkappa) - \beta_2\right) = 0\right. \end{aligned} \tag{6.19}$$

By solving system of equations it produces,

$$\begin{aligned} \beta_1 &= \frac{\zeta^{(1)}\,(\varkappa+1)(\varkappa) - \left(\frac{1}{p-1}\right)\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa+1))\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa))}{\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa))^2 - \left(\frac{1}{p-1}\right)\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa))^2} \\ \beta_2 &= \frac{1}{n-1}\left[\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa+1) - \beta_1\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa))\right] \end{aligned} \tag{6.20}$$

From $$D^{\mathcal{T}}\,\gamma = D^{\mathcal{T}}D\,\hat{\beta}\ ,\ \hat{\beta} = (D^{\mathcal{T}}D\,)^{-1}D^{\mathcal{T}}\gamma \tag{6.21}$$

Also from $$D^{\mathcal{T}}D = \begin{bmatrix} \zeta^{(1)}(1) & 1 \\ \zeta^{(1)}(2) & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \zeta^{(1)}(n-1) & 1 \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} \zeta^{(1)}(1) & 1 \\ \zeta^{(1)}(2) & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \zeta^{(1)}(n-1) & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^{p-1}(\zeta^{(1)}\,(\varkappa))^2 & \sum_{k=1}^{p-1}(\zeta^{(1)}\,(\varkappa)) \\ \sum_{k=1}^{p-1}(\zeta^{(1)}\,(\varkappa)) & p-1 \end{bmatrix},$$

$$(D^{\mathcal{T}}D)^{-1} = \frac{1}{(p-1)\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa))^2 - [\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\,(\varkappa))]^2} \begin{bmatrix} p-1 & -\sum_{k=1}^{p-1}\zeta^{(1)}\,(\varkappa) \\ -\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}\ (\varkappa) & \sum_{k=1}^{p-1}(\zeta^{(1)}\ (\varkappa))^2 \end{bmatrix},$$

$$D^{\mathcal{T}}\gamma = \begin{bmatrix} \zeta^{(1)}(1) & 1 \\ \zeta^{(1)}(2) & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \zeta^{(1)}(n-1) & 1 \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} \zeta^{(1)}(2) \\ \zeta^{(1)}(3) \\ \cdot \\ \cdot \\ \cdot \\ \zeta^{(1)}(n) \end{bmatrix}$$

$$\begin{bmatrix} \sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa).\zeta^{(1)}(\varkappa+1) \\ \sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa+1) \end{bmatrix} \tag{6.22}$$

It follows that

$$\hat{\beta} = (D^{\mathcal{T}}D)^{-1}D^{\mathcal{T}}\gamma = \frac{1}{(p-1)\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa))^2 - [\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa))]^2}$$

$$\times$$

$$\begin{bmatrix} (p-1)\sum_{k=1}^{p-1}\zeta^{(1)}(\varkappa)\zeta^{(1)}(\varkappa+1) - \sum_{k=1}^{p-1}\zeta^{(1)}(\varkappa)\sum_{k=1}^{p-1}(\zeta^{(1)}(\varkappa+1) \\ -\sum_{k=1}^{p-1}(\zeta^{(1)}(\varkappa))\sum_{k=1}^{p-1}(\zeta^{(1)}(\varkappa)(\zeta^{(1)}(\varkappa+1) + \sum_{k=1}^{p-1}(\zeta^{(1)}(\varkappa+1)\sum_{k=1}^{p-1}(\zeta^{(1)}(\varkappa))^2 \end{bmatrix}$$

$$\begin{bmatrix} \frac{\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa)(\zeta^{(1)}(\varkappa+1) - \frac{1}{(p-1)}\sum_{\varkappa=1}^{p-1}\zeta^{(1)}(\varkappa+1)\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa)}{\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa))^2 - \frac{1}{(p-1)}\sum_{k=1}^{p-1}(\zeta^{(1)}(\varkappa))^2} \\ \frac{1}{(p-1)}\sum_{\varkappa=1}^{p-1}\zeta^{(1)}(\varkappa+1) - \beta_1\sum_{\varkappa=1}^{p-1}(\zeta^{(1)}(\varkappa) \end{bmatrix}$$

$$= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{6.23}$$

Theorem 7. Let D, γ, $\hat{\beta}$ be the variables same as defined in theorem 1 and

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = (D^{\mathcal{T}}D)^{-1}D^{\mathcal{T}}\gamma$$

Then the following result holds:

If $\zeta^{(1)}(1) = \zeta^{(0)}(1)$, the recurrence relation is

$$\hat{\zeta}(\varkappa+1) = \beta_1^{\varkappa}\zeta(1) + \frac{1-\beta_1^{\varkappa}}{1-\beta_1} * \beta_2; \varkappa = 1,2,\dots,n-1; \tag{6.24}$$

or

$$\hat{\zeta}^{(1)}(\varkappa+1) = \beta_1^{\varkappa}\left(\zeta^{(0)}(1) - \frac{\beta_1}{1-\beta_1}\right)\zeta^{(0)}(1) + \frac{1-\beta_1^{\varkappa}}{1-\beta_1} * \beta_2; \varkappa = 1,2,\ldots,n-1; \qquad (6.25)$$

6.3.2 Proposed Multifarious GM (1 ,1) Modified Forecasting Models:

This model is the modification of basic grey model and it removes the long term forecasting effect in this 1 term is forecasted and then again it will be used as input for the forecasting of next term .It uses the iteration process and it will be iterated over time to forecast the desired number of terms. In this it can predicts with partial information also .During the modeling of MGM(1,1) Firstly accumulate the original data and then by use first order differential equation to achieve short term prediction. The cyclic model for MGM (1,1) forecasting model is

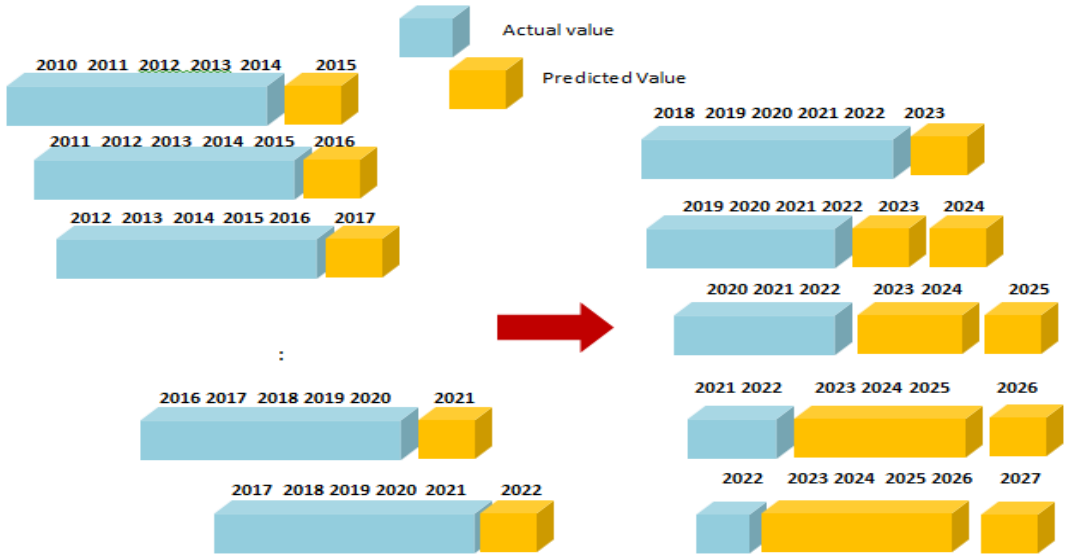


Figure 6.1 Cyclic model of MGM(1,1)

Cyclic model of MGM(1,1) in Figure 1 explains that by taking actual five values, one term can be predicetd and then in next step it excludes the actual value and use predicted value as input to forecast the electric demand for next consecutive year. The iteration will not stops until the forecasted results will not get.From the above Fig. it is seen that the forecasted value of 2023 is used for the forecasting of electricity demand for 2024 so on and so forth to 2026. Further here, the Figure 1 shows that in the same way it can be forecasted for the next years also.

The values of parameter c and d is as shown in the figure 2(a) which makes the basic grey model different from other to improve the forecasting and reduce the error.

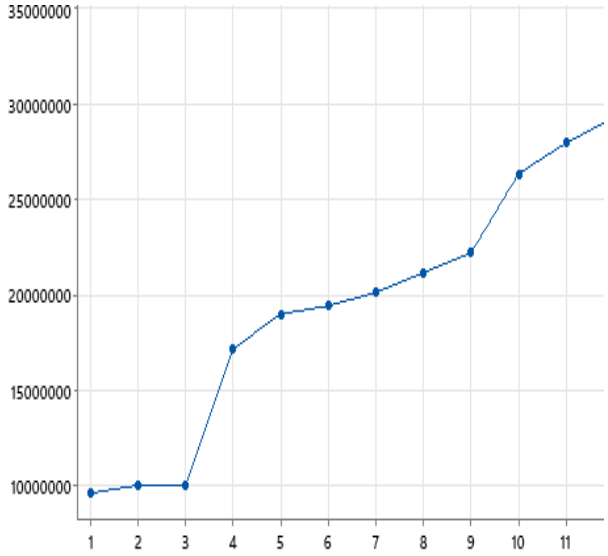


Figure 6.2(a) The value of parameter 'c'

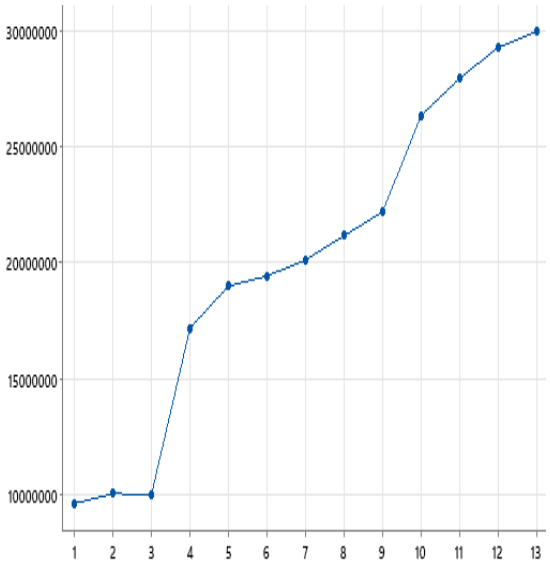


Figure 6.2(b) The value of parameter ' d'

6.3.3 Algorithm for Multifarious grey model: The instructions to run the model is explained in Figure 3 which is known as the algorithm and with the help of Figure 3 this model can run easily by any one and can be applied in any other field also.

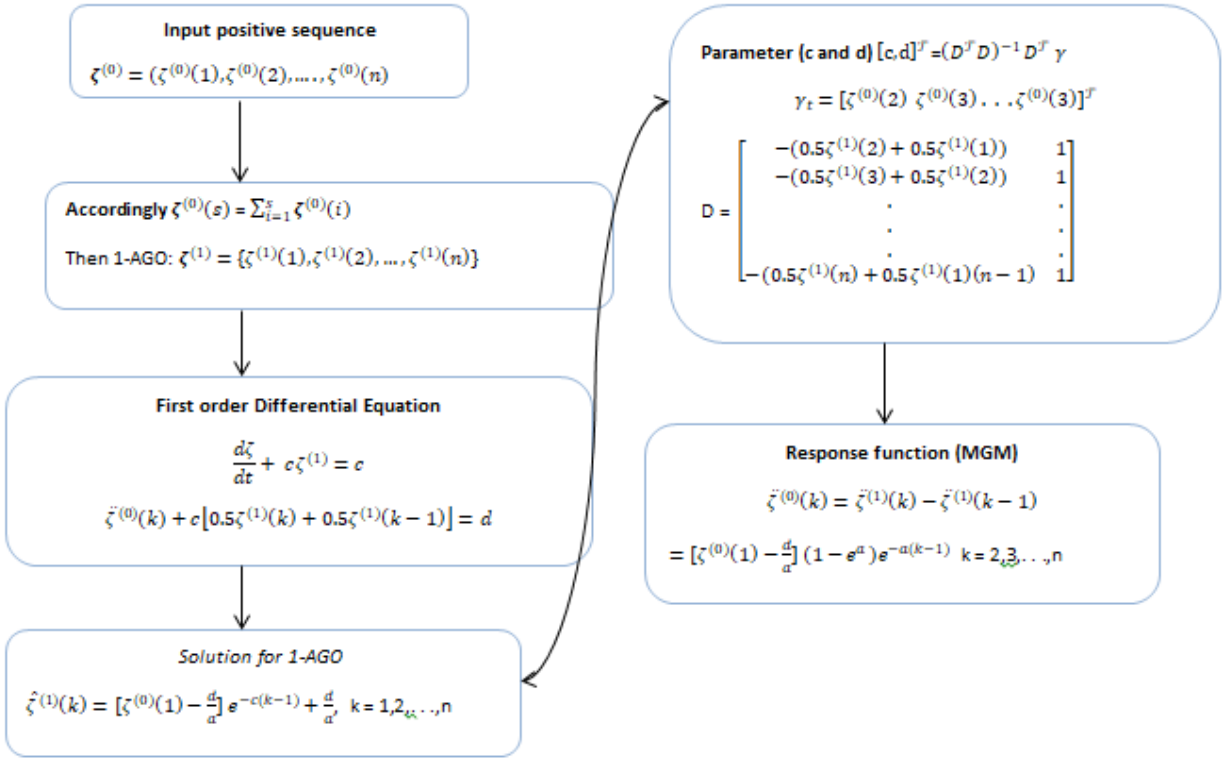


Figure 6.3 Algorithm for multifarious grey model

6.3.4 Arima:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_1 \quad (6.26)$$

$$Y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_q \varepsilon_{t-q} \quad (6.27)$$

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \varepsilon_t \quad (6.28)$$

$$Y_{(t-1)} = \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + .. + \beta_0 Y_0 + \varepsilon_{t-1} \quad (6.29)$$

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + .. + \phi_q \varepsilon_{t-q} \quad (6.30)$$

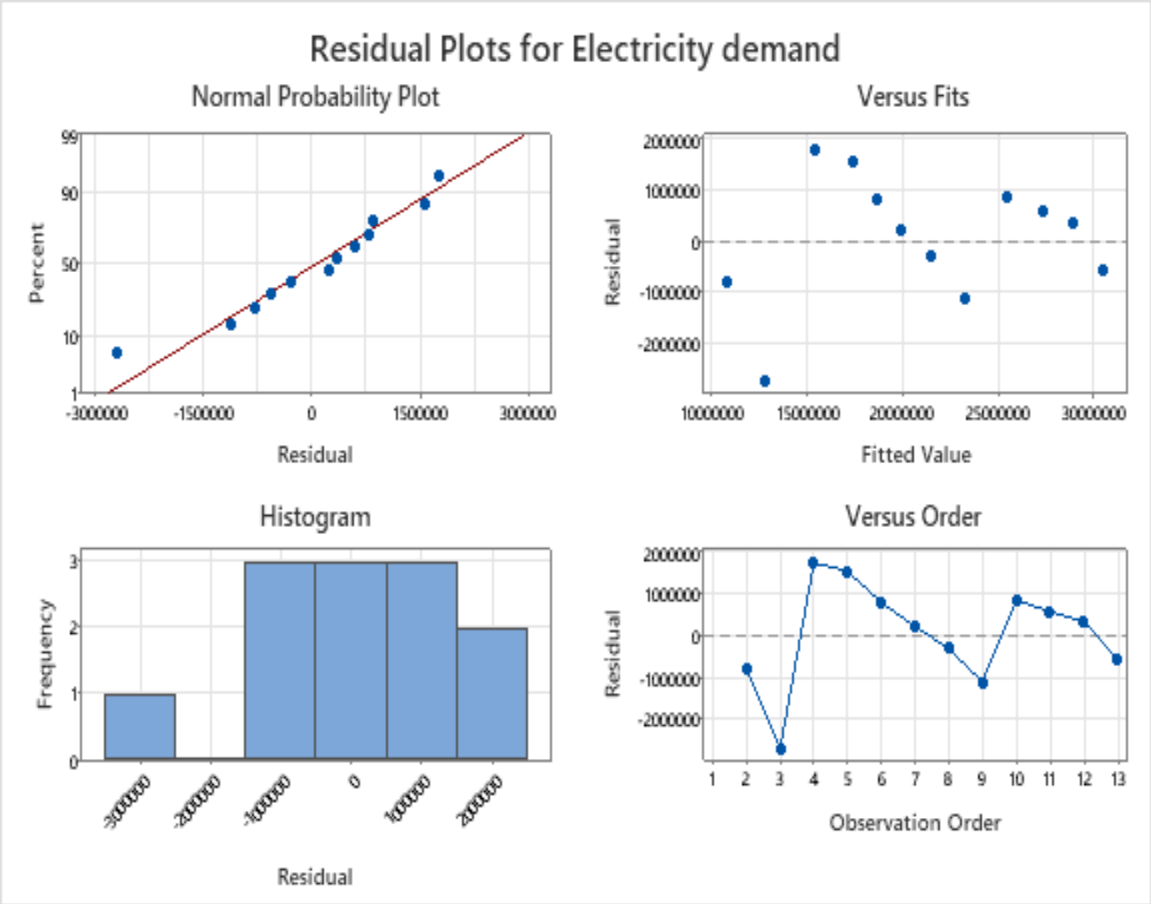


Figure 6.4 Residual Plots for electricity demand in the form of Normal Probability Plot, versus fits in the form of histogram and fitted values of ARIMA model.

Table 6.2 The model statistics for ARIMA.

Model	R^2 Stationary	R^2	RMSE	MAPE
Electricity generation	0.720	0.928	15,583	1.573

To check the stationarity of data the differencing is required and thus the stationary result is shown in Figure 6.5 and Table 1with ACF ,PACF and coefficient map of Arima model. Figure 6.6 shows the fitting of the ARIMA model which is inferred while running the model.

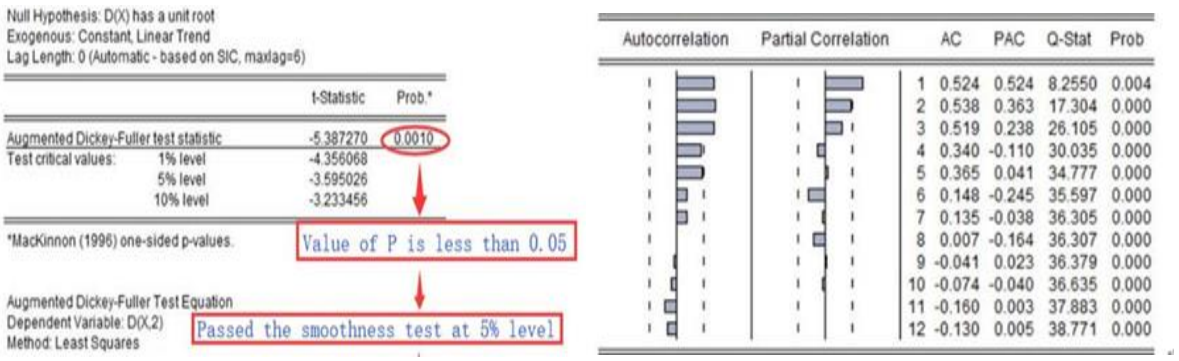


Figure 6.5 Results of Q-stat, Autocorrelation (AC), Partial Autocorrelation PAC, Stationary test and augumented Dickey Fuller statistic

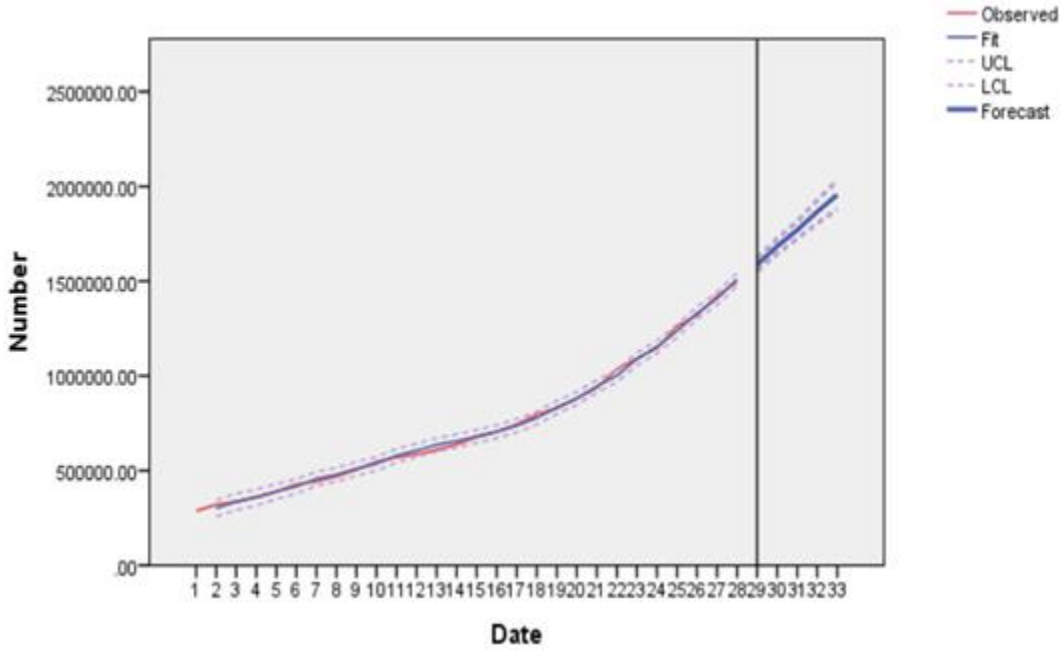


Figure6.6 Fitting of the ARIMA model.

6.3.5 MGM-ARIMA Model: The residual plots which are shown in Figure 6.4 can be removed or reduced by using MGM model so the next model which may outperformed than this will be MGM-ARIMA model. The forecasted result of MGM-ARIMA model is shown in Figure 6.7

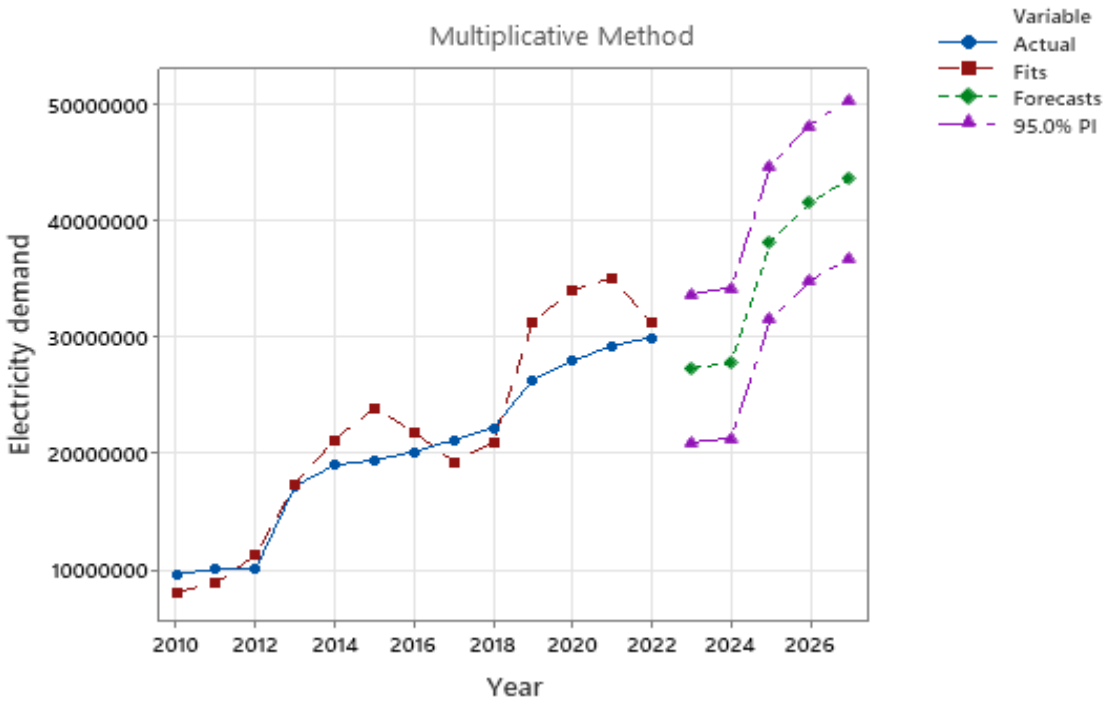


Figure 6.7 Forecasted electricity generation in india and the actual electricity generation by MGM-ARIMA model and degree of coincidence in between.

6.3.6. Nonlinear Multifarious Grey Model: This model works like the same process as the above multifarious grey model. In this the only difference is of the constant parameter 'c' and 'd'. In this model by setting the constant parameters c and d the errors can be reduced which can be found in basic non linear grey model and hence the prediction accuracy can be improved.

$$\zeta^{(0)} = (\zeta^{(0)}(1), \zeta^{(0)}(2), \ldots, \zeta^{(0)}(n)$$

considered as the original sequence which is usually consists of raw data so the sequence must be represented as 1-AGO

i.e., $\zeta^{(1)} = \{\zeta^{(1)}(1), \zeta^{(1)}(2), \ldots, \zeta^{(1)}(n)\}$ for the prediction.

In second step,To establish first order differential equation Deform

$$\hat{\zeta}^{(0)}(k) + c\left[0.5\zeta^{(1)}(k) + 0.5\zeta^{(1)}(k-1)\right] = d \qquad (6.31)$$

and get the exponential growth form of first oder differential equation i.e.,

$\frac{d\zeta}{dt} + c\zeta^{(1)} = d$ then from the above equation $\zeta^{(1)}$ satisfied the linear regression formula

In third step on solving first order differential equation

$$\hat{\zeta}^{(1)}(k) = [\zeta^{(0)}(1) - \frac{d}{a}]\, e^{-c(k-1)} + \frac{d}{a},\ k = 1,2,\ldots,n. \tag{6.32}$$

It is seen that the constant coefficient parameters c and d can be achieved which plays main role to solve the formula.The constant parameters c and d satisfies

$[c, d]^{\mathcal{T}}=(D^{\mathcal{T}} D)^{-1}\, D^{\mathcal{T}}\, \gamma$

$$\gamma_t = [\zeta^{(0)}(2)\ \ \zeta^{(0)}(3) \ldots \zeta^{(0)}(n)]^{\mathcal{T}} \tag{6.33}$$

$$D = \begin{bmatrix} -(0.5\zeta^{(1)}(2) + 0.5\zeta^{(1)}(1)) & 1 \\ -(0.5\zeta^{(1)}(3) + 0.5\zeta^{(1)}(2)) & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ -(0.5\zeta^{(1)}(n) + 0.5\zeta^{(1)}(1)(n-1) & 1 \end{bmatrix} \tag{6.34}$$

Replace c and d by bringing γ and D in .then in the last step, the prediction formula for grey model is which is known as response function of MGM (1,1)

$$\widehat{\zeta}^{(0)}(k) = \hat{\zeta}^{(1)}(k) - \hat{\zeta}^{(1)}(k-1)$$

$$= [\zeta^{(0)}(1) - \frac{d}{a}]\, (1 - e^{c}) e^{-c(k-1)} \quad k = 2,3,\ldots,n \tag{6.35}$$

By using Runge-Kutta method of fourth – order with the help of MATLAB software and the whole process of this model is well explained in Figure 8.

Thus, the Fourth order Runge –Kutta is:

$$\frac{d\zeta}{dt} = F(t, \zeta)$$

$$k_1 = F(t_n, \zeta_n)$$

$$k_2 = F\left(t_n + \frac{h}{2}, \zeta_n + \frac{h}{2} K_1\right)$$

$$k_3 = F\left(t_n + \frac{h}{2}, \zeta_n + \frac{h}{2}K_2\right)$$

$$k_4 = F\left(t_n + \frac{h}{3}, \zeta_n + \frac{h}{3}K_3\right)$$

$$\zeta_{n+1} = \zeta_n + \frac{h}{6}[k_1 + 2k_2 + 2k_3 + k_4] \qquad (6.36)$$

Nonlinear Multifarious grey model (NMGM) Algorithm

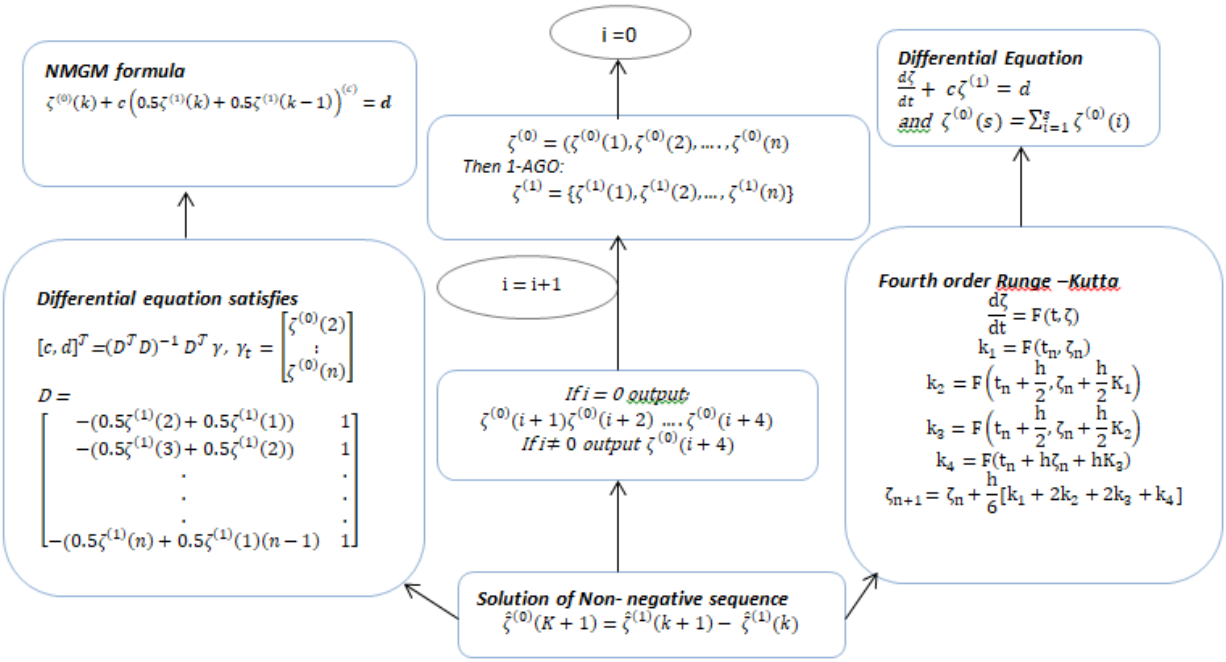


Figure 6.8 Algorithm for non-linear multifarious grey model

6.3.7 NMGM-ARIMA: These all models are developed from the basic grey model and it works on the same principle as the multifarious grey model.This model is he combination of both linear as well as non-linear models and hece try to dilute the errors formed by both the methods. In this the residuals formed for NMGM method will b parsed by ARIMA model and thus in conclusion the whole process of NMGM- ARIMA model can be explained as follows in three steps:

I. NMGM model can be used to obtain the forecasted value and to find the residuals.
II. ARIMA model can reduce or correct the residual value and it can help to achieve the desired residual value.

The following Figure 6.9 and Table 6.2 shows the values for the power coefficients obtained from the constants “c” “d” along with average relative error and model statistics fitting of the model respectively. Figure 10 shows the forecasted values obtained by using non-linear multifarious grey model along with comarison and contrast between actual and forecasted value.

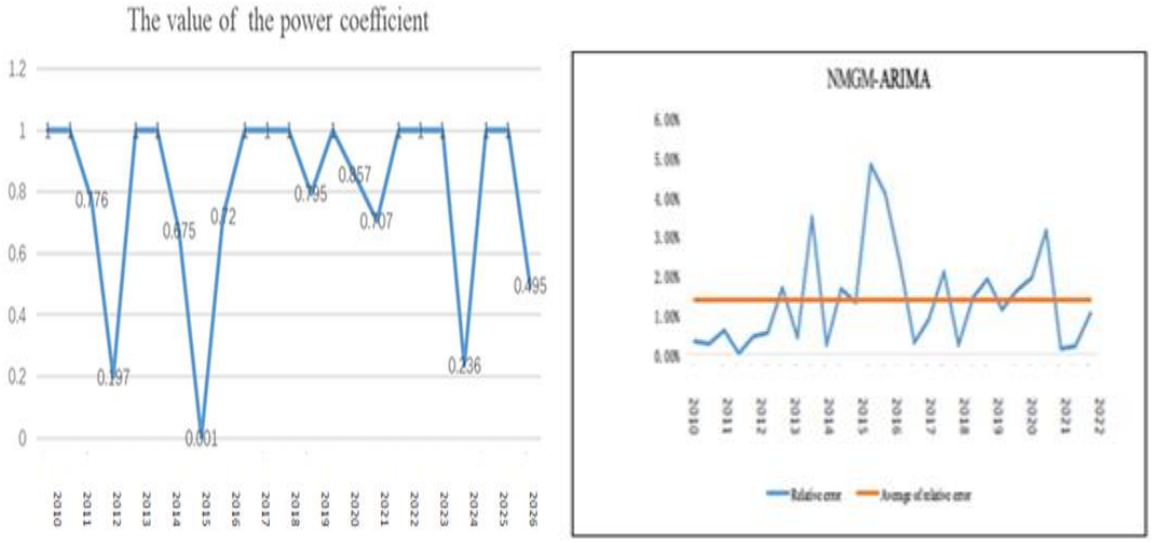


Figure 6.9 (a)The Power coefficients of NMGM-ARIMA model on the basis of constants “c” and “d” 6.9(b) Relative error of NMGM-ARIMA model.

Table 6.3 Statistics of Model Fitting for NMGM –ARIMA model

Model	R^2 Stationary	R^2	RMSE	MAPE
Electricity generation	0.621	0.621	17,863	2.658

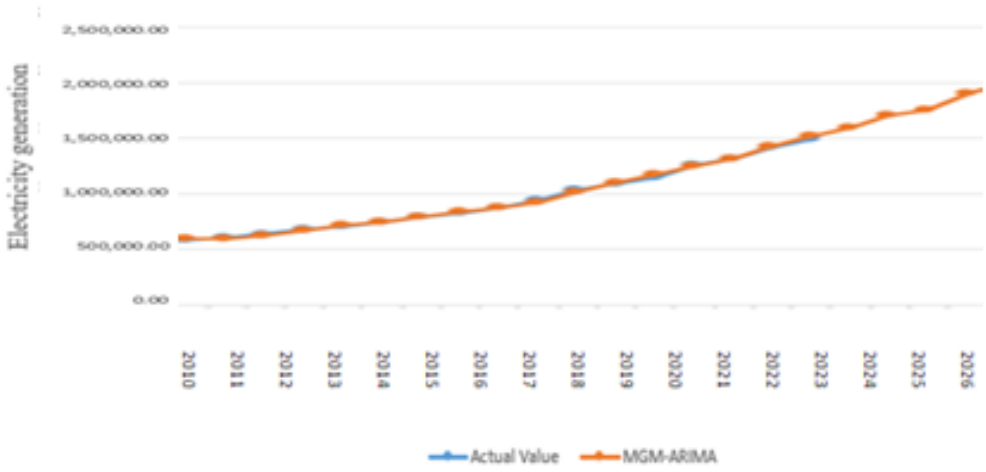


Figure 6.10 Forecasting using MGM-ARIMA model in contrast to actual electricity Production

6.4 Result and Discussion: This study is based on five deterministic methods for forecasting india's electricity demand.In this study the whole system is based on two parts, The original part and the forecasted part by the system. The data taken from internataional energy agency (Data and statistics) which is in Gwh has been used to find out the changes in electricity generation or electricity production and population growth shown in Figure6.11 . In this both sides are used and the power generation and growth rate shown in left and right side respectively.

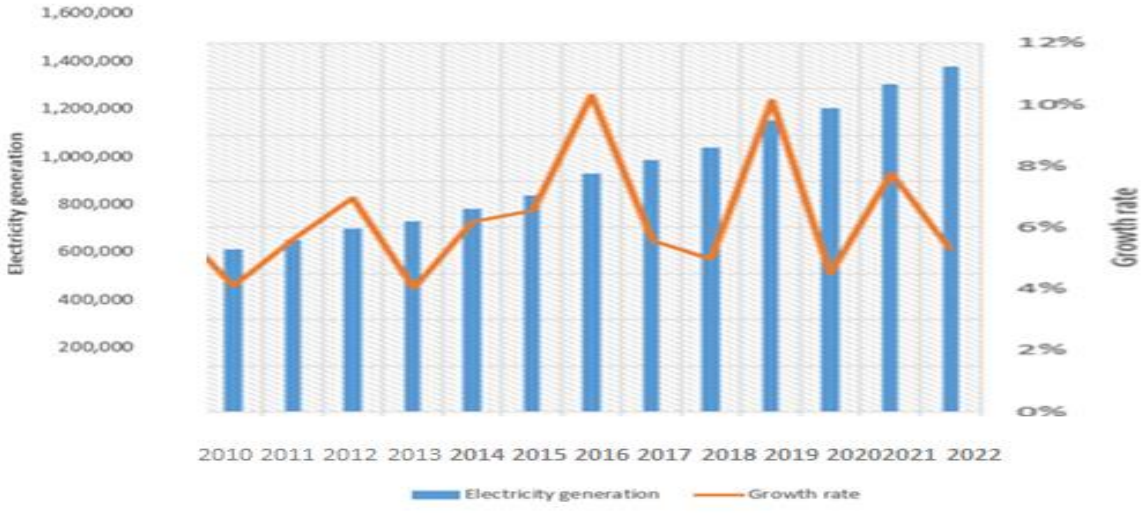


Figure 6.11 Comparison and contrast between growth rate and the power generation from 2010-2022

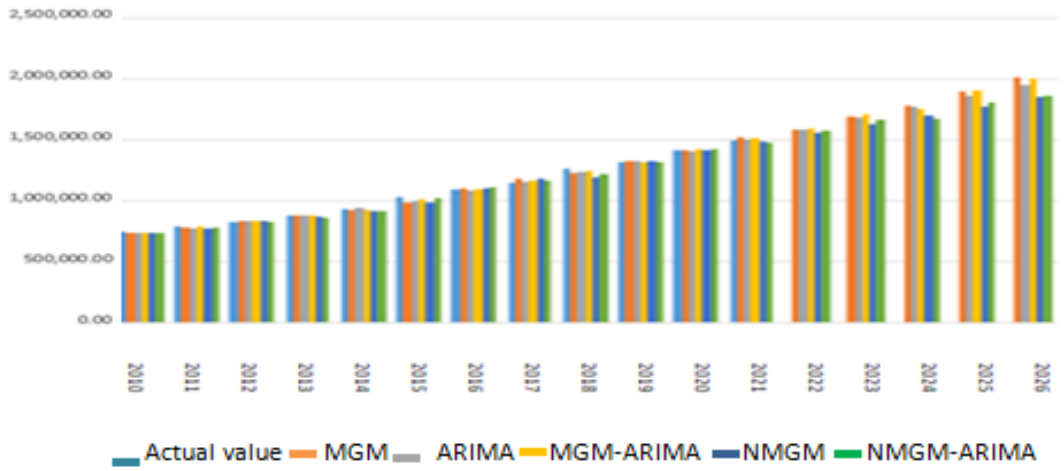


Figure 6.12 Actual electricity productions in comparison with the proposed deterministic models.

Table 6.4 Statistics of Model Fitting for deterministic models

Parameters	MGM	NMGM	ARIMA	MGM-ARIMA	NMGM-ARIMA
MSE	224,078,886.8	41,642,5674.6	188,854,755.6	91,855,859.86	211,602,535.2
MAPE	1.6586%	1.7353%	1.5906%	0.7346%	1.4237%
MSPE	0.4364%	0.4271%	0.4342%	0.2019%	0.2641%

6.5 Conclusion: This study is the combination of linear as well as nonlinear regression methods. Among these five methods only one is nonlinear and in this one is the combination of linear and non-linear model to enhance the accuracy so that the forecasting may approach to the exactness. During the forecasting, The relative errors for five models are shown in Table 6.3 which shows that the errors are very small and hence the selected models are quite appropriate for this study. In this all the models shows that the power generation trend is upgrading then also even among the five models NMGM-ARIMA model outperformed than all. Thus, this study deals with forecasting India's power. Based on time series analysis, the

forecasting is done by using Data mining. In this, the basic grey model has been used to develop MGM, NMGM, ARIMA, MGM-ARIMA, NMGM-ARIMA models for the forecasting of India's power generation and the MSE values for India's electricity demand is 224,078,886.8, 41,642,5674.6, 188,854,755.6, 91,855,859.86, 211,602,535.2 respectively. The statistical models used for relative errors are MAPE and MSPE. In this study, the Python 3 and Matlab software has been used for forecasting. India's electricity generation. It is forecasted that it will continue to grow with an average annual growth rate of 8.17% in the next four years (2023– 2026).

Chapter VII

SUMMARY AND CONCLUSION

The three main research questions addresses in this thesis are:

1. How to know the best fit forecasting technical or non- technical model based on the type of data or available data.
2. How to forecast or classification or analysis can be done in various fields like pandemic spread, rainfall analysis or in agricultural field and many more by using these improvised models?
3. How to do extension from individual model to mixed model for accurate forecasting?

To answer these above questions, the Present thesis reports a study on forecasting models based on classification, clustering, association, long term memory, short term memory, machine learning and many more. The detail of all the technical models and non-technical models are well explained. The models based on machine learning and artificial intelligence much supports to the data mining which is discussed in detail in the thesis which consists of seven chapters.

Chapter I: In the first chapter, introduction about forecasting and its models, Along with their application and uses in various fields. It explains, work flow for forecasting with dealing science which is known as data science and the types of data on which it works like big data . There are many confluxes of data mining and various multidisciplinary fields where it can be benefited as well as the limitations in various fields. While, dealing with forecasting using data mining there are many outliers to deal with along with their different approaches by using various tools like software which can be either paid or free software can be used through interfaces.

Chapter II: To fulfil the objectives and to reach the problem, the existing literature review is discussed in chapter two which examines its existing applications in various fields like healthcare, agriculture with various methods along with their practical implementation by using various machine learning algorithms.

Chapter III: The role of data mining in healthcare for forecasting is discussed as forecasting proves like boon during pandemic. It prepares organisers to deal with flaws and benefits and incase of flaws to make strategies to reduce the losses of lives. In this chapter, the two case studies has been taken (I) is of the malarial disease which was also like pandemic and another is of corona pandemic. For the forecasting of malarial disease various models like feed forward neural network, LSTM model are discussed and it also shows that climate also affects the transmission of disesea while the study for prediction of COVID-19 shows that the propsed model has highest accuracy comparatively to others and in this it can be well said that hybrid model is far better than individual model which was the combination of random forest and decision tree method and shows comparably more accurate results.

Chapter IV deals with "Regional frequency analysis approach for rainfall analysis" and in this the theoretical L- moments, regional frequency analysis approach has been used which is the most emphasis on mathematical models for forecasting. In this chapter it was observed that PE3 distribution was best fit for group G1, GEV distribution was best fit for group G2 while for group G2; GNO was found best distribution. The regional quantiles with their accuracy measures showed that GNO in group G3 has a relatively low RMSE value and 90% error bounds. Furthermore, for higher return periods, rainfall quantiles showed uncertainty and should be treated with warnings.

Chapter V regression models in agriculture are discussed. The regression models can be of linear, non- linear type or can be of multi linear regression type. In this the time series forecasting also considered for the price of crop forecasting and for the crop yield forecasting and various regression techniques has been applied. In this, it is concluded that number of independent variables and accuracy follows inverse Proportion. It shows the constant coefficient, standard error coefficients and number of variables which is taken place for multiple linear regression but it shows that with two independent variables the accuracy is 99.52% and it is inferred that as the number of independent variable increases the accuracy will decrease and then the predicted R square become 95.81%.

Chapter VI The combination of mathematical models, linear as well as non-linear grey models discussed in this chapter. In this the combination of linear and non-linear methods with grey models gives absolute results which motivates other researchers also to implement

in various fields for forecasting. In this among the entire non-linear multifarious grey model with combination of ARIMA shows marvellous result and which shows that on an average it will grow annually 8.17%.

Scope for Future Work: Overall, the future of forecasting using data mining holds great potential for advancements that can revolutionize decision-making processes across industries and domains. In this the parameters taken from various fields are of numerical type only so in this sometimes dataset can be of various formats like text, images, speech etc. can be included. In this the other machine learning algorithms can be used for the analysis. It would be also interesting to create an objective metric for the particular problem along with to create a particular model for the particular field by using various mathematical equations on the basis of defined parameters.

REFERENCES:

[1] Lafferty, K. D. (2017). Marine infectious disease ecology. Annual Review of Ecology, Evolution, and Systematics, 48, 473-496.

[2] Tai, K. Y., & Dhaliwal, J. (2022). The machine learning model for malaria risk prediction based on mutation location of large-scale genetic variation data. Journal of Big Data, 9(1), 1-22.

[3] Garrido-Cardenas, J. A., Cebrián-Carmona, J., González-Cerón, L., Manzano-Agugliaro, F., & Mesa-Valle, C. (2019). Analysis of global research on malaria and Plasmodium vivax. International journal of environmental research and public health, 16(11), 1928

[4] Kumar, V., Mangal, A., Panesar, S., Yadav, G., Talwar, R., Raut, D., & Singh, S. (2014). Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis. Malaria research and treatment, 2014.

[5] Gawande, R., & Patil, D. (2022). Prediction of Frequent Hospitalisation of Diabetic Patients. Prediction of Frequent Hospitalisation of Diabetic Patients (February 25, 2022).

[6] Talapko, J., Škrlec, I., Alebić, T., Jukić, M., & Včev, A. (2019). Malaria: the past and the present. Microorganisms, 7(6), 179.

[7] Eswari, T., Sampath, P., & Lavanya, S. J. P. C. S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208.

[8] Butwall, M., & Kumar, S. (2015). A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. International Journal of Computer Applications, 120(8).

[9] Faruque, F M., Asaduzzaman, I., Sarker, H., "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2021.

[10] Woldemichael F.G. and Menaria S., "Prediction of Diabetes Using Data Mining Techniques," *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 414-418, doi: 10.1109/ICOEI.2018.8553959.

[11] Hina, S., Shaikh, A., & Sattar, S. A. (2017). Analyzing diabetes datasets using data mining. Journal of Basic and Applied Sciences, 13, 466-471.

[12] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning logarithms. Procedia Computer Science, 165, 292-299.

[13] Tisdell, C.C. Technology, Alternate solution to generalized Bernoulli equations via an integrating factor: An exact differential equation approach. Int. J. Math. Educ. Sci. Technol. 2017, 48, 913–918. Energies 2019, 12, 2574 22 of 24

[14] Ning, X.; Dang, Y.; Gong, Y.J.E. Novel grey prediction model with nonlinear optimized time response method for forecasting of electricity consumption in China. Energy 2017, 118, 473–480

[15] Markovics, D., & Mayer, M. J. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renewable and Sustainable Energy Reviews*, *161*, 112364.

[16] Saravanan, S.; Kannan, S.; Thangaraj, C. Forecasting India's electricity demand using Artificial Neural Network. In Proceedings of the IEEE International Conference on Advances in Engineering, Nagapattinam, Tamil Nadu, India, 30–31 March 2012.

[17] Kumar, R.; Jilte, R.; Nikam, K.C.; Ahmadi, M.H. Innovation, Status of carbon capture and storage in India's coal fired power plants: A critical review. Environ. Tech. Innov. 2018, 13, 94–103.

[18] Bhowte, Y.W. Forecasting the load of demand and supply of Electricity in India. 2016 International Conference on Computation of Power. In Proceedings of the IEEE Energy Information and Commuincation (ICCPEIC), Chennai, India, 20–21 April 2016.

[19] Saravanan, S.; Kannan, S.; Thangaraj, C. Forecasting India's electricity demand using Artificial Neural Network. In Proceedings of the IEEE International Conference on Advances in Engineering, Nagapattinam, Tamil Nadu, India, 30–31 March 2012.

[20] Singh, D.P.; Gadakh, P.J.; Dhanrao, P.M.; Mohanty, S.; Swain, D.; Swain, D. An Application of NGBM for Forecasting Indian Electricity Power Generation. Adv. Int. Syst. Comput. 2017, 556, 203–214.

[21] Bedi, J.; Toshniwal, D.J.I.A. Empirical Mode Decomposition Based Deep Learning for Electricity Demand Forecasting. IEEE Access 2018, 6, 49144–49156.

[22] Inglesi, R.J.A.E. Aggregate electricity demand in South Africa: Conditional forecasts to 2030. Appl. Energy 2010, 87, 197–204.

[23] Jiang, P.; Zhou, Q.; Jiang, H.; Dong, Y. Applied Analysis, An Optimized Forecasting Approach Based on Grey Theory and Cuckoo Search Algorithm: A Case Study for Electricity Consumption in New South Wales. Abstr. Appl. Anal. 2014, 9, 1–13.

[24] Wang, Q.; Song, X. Forecasting China's oil consumption: A comparison of novel nonlinear-dynamic grey model (GM), linear GM, nonlinear GM and metabolism GM. Energy 2019, 183, 160–171. [CrossRef]

[25] Xie, N., & Wang, R. (2017). A historic Review of Grey Forecasting Models. *Journal of grey System*, *29*(4).

[26] Emodi, N.V.; Emodi, C.C.; Murthy, G.P.; Emodi, A.S.A.J.R.; Reviews, S.E. Energy policy for low carbon development in Nigeria: A LEAP model application. Renew. Sustain. Energy Rev. 2017, 68, 247–261.

[27] Ning, X.; Dang, Y.; Gong, Y.J.E. Novel grey prediction model with nonlinear optimized time response method for forecasting of electricity consumption in China. Energy 2017, 118, 473–480

[28] Liu, S.; Tao, L.; Xie, N.; Yang, Y. On the new model system and framework of grey system theory. In Proceedings of the IEEE International Conference on Grey Systems & Intelligent Services, Leicester, UK, 18–20 August 2015.

[29] Dharmaraja, S., Jain, V., Anjoy, P., & Chandra, H. (2020). Empirical analysis for crop yield forecasting in india. Agricultural Research, 9, 132–138.

[30] Salpekar, H. (2019). Design and implementation of mobile application for crop yield prediction using machine learning. 2019 Global Conference for Advancement in Technology (GCAT), 1–6.

[31] Bang, S., Bishnoi, R., Chauhan, A. S., Dixit, A. K., & Chawla, I. (2019). Fuzzy Logic based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA, and ARMAX models. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1–6

[32] Hu, Y., Li, J., Man, Y., & Ren, J. (2022). The dynamic hydrogen production yield forecasting model based on the improved discrete grey method. International Journal of Hydrogen Energy, 47(42), 18251–18260.

[33] Choudhury, A., & Jones, J. (2014). Crop yield prediction using time series models. Journal of Economics and Economic Education Research, 15(3), 53–67.

[34] Li, A., Liang, S., Wang, A., & Qin, J. (2007). Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. Photogrammetric Engineering & Remote Sensing, 73(10), 1149–1157.

[35] Chen C, Mcnairn H. A neural network integrated approach for rice crop monitoring. International Journal of Remote Sensing. 2006; 27(7):1367–93.

[36] Hamjah, M. A. (2014). Rice production forecasting in Bangladesh: An application of Box-Jenkins ARIMA model. Mathematical Theory and Modeling, 4(4), 1–11

[37] Conradt, T. (2022). Choosing multiple linear regressions for weather-based crop yield prediction with ABSOLUT v1. 2 applied to the districts of Germany. International Journal of Biometeorology, 66(11), 2287–2300.

[38] Liu, Y., Heuvelink, G. B., Bai, Z., He, P., Xu, X., Ding, W., & Huang, S. (2021). Analysis of spatio-temporal variation of crop yield in China using stepwise multiple linear regression. Field Crops Research, 264, 108098.

[39] Spiertz, J. H. J., & Ewert, F. (2009). Crop production and resource use to meet the growing demand for food, feed and fuel: opportunities and constraints. NJAS: Wageningen Journal of Life Sciences, 56(4), 281-300.

[40] Balakrishna, K., Mohammed, F., Ullas, C. R., Hema, C. M., & Sonakshi, S. K. (2021). Application of IOT and machine learning in crop protection against animal intrusion. Global Transitions Proceedings, 2(2), 169-174.

[41] Sadenova, M. A., Beisekenov, N. A., Rakhymberdina, M. Y., Varbanov, P. S., & Klemeš, J. J. (2021). Mathematical Modelling in Crop Production to Predict Crop Yields. Chemical Engineering Transactions, 88, 1225-1230.

[42] Mishra, S., Mishra, D., & Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper. Indian J. Sci. Technol, 9(38), 1-14.

[43] Chen C, Mcnairn H. A neural network integrated approach for rice crop monitoring. International Journal of Remote Sensing. 2006; 27(7):1367–93.

[44] Co HC, Boosarawongse R. Forecasting Thailand's Rice Export: Statistical Techniques vs. Artificial Neural Networks, Computers and Industrial Engineering. 2007; 53(4):610–27.

[45] Garov, A. K., & Awasthi, A. K. A computational mathematical model for forecasting of Indian crop.

[46] Nischitha, K., Dhanush Vishwakarma, M. N., & Ashwini, M. M. (2020). Crop prediction using machine learning approaches. International Journal of Engineering Research & Technology (IJERT), 9(08).

[47] Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.

[48] Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019, November). Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE. Gabriele S., Arnell N, "A hierarchical approach to regional flood frequency analysis," Water Resources Research, 27 no. 6, pp. 1281-1289, 1991.

[49] Gabriele S., Arnell N, "A hierarchical approach to regional flood frequency analysis," Water Resources Research, 27 no. 6, pp. 1281-1289, 1991.

[50] Hosking J. R. M., " L-moments: Analysis and estimation of distributions using linear combinations of order statistics." Journal Royal Statistical Society, Series B, no. 52, pp. 105-124, 1990.

[51] Hosking J. R. M., Wallis J. R., "Some statistics useful in regional frequency analysis" Water Resources Research**,** 29 no.2, pp. 271-281,1993.

[52] Stedinger J. R., Vogel R. M., Georgiou E.F., "Frequency analysis of extreme events," Maidment, D.J.(Ed.), Handbook of Hydrology, McGraw Hill,1993 pp 18.1 -18.66.

[53] Hooda B. K., " Probability Analysis of Monthly Rainfall for Agricultural Planning at Hisar" *Indian Journal of Soil Conservation,* vol**.** 34**,** no.1, pp. 12-14, 2006.

[54] Kumar M., Shekhar C., Manocha V., " Maximum rainfall probability distributions pattern in Haryana–A case study"Journal of Applied and Natural Science, vol. 8, no. 4, pp. 2029-2036, 2016.

[55] Babu V.B., Hooda B.K., "Fuzzy Majority Approach for Modeling Spatial and Temporal Distributions of Daily Rainfall in Western Zone of Haryana," International Journal of Agricultural and Statistical Sciences, vol.14, no.1, pp. 57-67, 2018.

[56] Nain M., and Hooda B.K., "A Probability and Trend Analysis of Monthly Rainfall in Haryana" International Journal of Agricultural and Statistical Sciences, vol. 15 no.1, pp. 221-229, 2019.

[57] Nain M., Hooda B. K., "Regional Frequency Analysis of Maximum Monthly Rainfall in Haryana State of India Using L-Moments," Journal of Reliability and Statistical Studies, pp. 33-56, 2021.

[58] Greenwood J., Landwehr J., Matalas C., Wallis J., " Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form," Water Resources Research, vol. 15, no. 5, pp. 1049-1054, 2013.

[59] Landwehr, J. M., Matalas, N. C, & Wallis, J. R., "Probability-weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles," Water Resources Research, vol. 15, no.5, pp. 1055-1064, 1979.

[60] Ward J.H., "Hierarchical grouping to optimize an objective function," Journal of American Statistical Association, Vol. 58, pp. 236-244, 1963.

[61] Yashwant S., Sananse S. L., "Comparisons of different methods of cluster analysis with application to rainfall data," International Journal of Innovative Research in Science, Engineering and Technology, vol 4 no. 11, pp. 10861-10872, 2015.

[62] Nain M., Hooda B. K., "Identification of homogeneous rainfall stations in Haryana," *JEBAS*, vol.7, no.5, pp. 452-462, 2019.

[63] Dalrymple T., "Flood frequency analysis. Manual of hydrology: Part 3. Flood flow techniques," United States Geological Survey, v.80, no. 1543. https://doi.org/10.3133/wsp1543A.

[64] Hosking J. R. M., Wallis J. R., "Regional frequency analysis: An approach based on L-Moments," Journal of the American Statistical Association, vol. 93, pp.1233, 1997.

[65] Hosking J. R. M., "Regional frequency analysis using L-moments," lmomRFA R package, version 2.2. , 2009. http://CRAN.R-project.org/package=lmomRFA.

[66] A.K. Awasthi and M. Sharma, Comparative analysis of forecasting models in healthcare (COVID19),International Journal of Health Sciences, 6(S3), pp. 8649–8661, (2022).

[67] Sharma, M. (2022, May). Data Mining Prediction Techniques in Health Care Sector. In Journal of Physics: Conference Series (Vol. 2267, No. 1, p. 012157). IOP Publishing.

[68] A. K. Garov, and A.K., Awasthi, "Case Study-Based Big Data and IoT in Healthcare," in Machine Learning, Deep Learning, Big Data, and Internet of Things for Healthcare, Chapman and Hall/CRC, pp. 13-35 (2022)

[69] Si, X., Bambrick, H., Zhang, Y., Cheng, J., McClymont, H., Bonsall, M. B., ... & Nevels, M. (2021). Weather variability and transmissibility of COVID-19: a time series analysis based on effective reproductive number. Experimental Results, 2, e15.

[70] Ahmad, A., Garhwal, S., Ray, S. K., Kumar, G., Malebary, S. J., & Barukab, O. M. (2021). The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. Archives of Computational Methods in Engineering, 28, 2645-2653.

[71] Siddiqui, M. K., Menendez, R. M., Gupta, P. K., Hussain, F., Khatoon, K., & Ahmad, S. (2020). Correlation between temperature and COVID-19 (suspected, confirmed and death) cases based on machine learning analysis.

[72] Atef, O. M., Nassir, Q., Nassif, A. B., & AbuTalib, M. (2020). Death/Recovery Prediction for Covid-19 Patients using Machine Learning. Recovery Prediction for Covid-19 Patients using Machine Learning,[online] Available: https://www. who. int/emergencies/diseases/novel-coronavirus-2019.

[73] Namasudra, S., Dhamodharavadhani, S., & Rathipriya, R. (2021). Nonlinear neural network based forecasting model for predicting COVID-19 cases. Neural processing letters, 1-21.

[74] Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, solitons & fractals, 135, 109864.

[75] Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals, 134, 109761.

[76] Maleki, M., Mahmoudi, M. R., Heydari, M. H., & Pho, K. H. (2020). Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. Chaos, Solitons & Fractals, 140, 110151.

[77] A. K. Awasthi, M. Sharma, A. K. Garo and P. Chaudhary, "Presentation of futuristic Malarial Disease through a Hybrid Model of A.I. and Big data," 2023 International

Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 1044-1050, doi: 10.1109/AISC56616.2023.10085407.

[78] Rudra Kumar, M., Pathak, R., & Gunjan, V. K. (2022). Diagnosis and Medicine Prediction for COVID-19 Using Machine Learning Approach. In Computational Intelligence in Machine Learning: Select Proceedings of ICCIML 2021 (pp. 123-133). Singapore: Springer Nature Singapore.

[79] Rahimi, I., Chen, F., & Gandomi, A. H. (2021). A review on COVID-19 forecasting models. Neural Computing and Applications, 1-11.

[80] Vijiya K., Kumar, Lavanya, B., Nirmala, I., Sofia, S, C., "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

List of Publications

- Minakshi sharma, Rashi Jarial, Mohit Nain ,A.K. Awasthi , Owais Aalam ,Bhanu Pratap Environment and Ecology Research Vol. 11(4). **(Scopus)**
- Awasthi, A. K., Garov, A. K., Sharma, M. and Sinha, M. (2023). GNN Model Based On Node Classification Forecasting in Social Network, 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, pp. 1039-1043, doi:10.1109/AISC56616.2023.10085118. **(Scopus)**
- Awasthi, A. K., Sharma, M., Garov, A. K., & Chaudhary, P. (2023).Presentation of futuristic Malarial Disease through a Hybrid Model of A.I. and Big data, 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, pp. 1044-1050, doi:10.1109/AISC56616.2023.10085407. **(Scopus)**
- Awasthi, A. K., & Sharma, M. (2022). Comparative analysis of forecasting models in healthcare (COVID-19). *International Journal of Health Sciences*, 8649-8661.
- Sharma, M. (2022, May). Data Mining Prediction Techniques in Health Care Sector. In *Journal of Physics: Conference Series* (Vol. 2267, No. 1, p. 012157). IOP Publishing. **(Scopus)**
- A.K. Awasthi, Minakshi Sharma and Arun Kumar Garov, Forecasting Analysis of COVID-19 Patient Recovery Using RF–DT Model, AIP Conference Proceedings 23 June 2023; 2768 (1): 020009. **(Scopus)**

Book Chapter Published:

Arun Kumar Garov, A.K. Awasthi, Minakshi Sharma.Artificial Intelligence and Data Science in Various Domains in Advancement of Data Processing methods for Artificial and computing Intelligence.

Doi: https://www.taylorfrancis.com/chapters/edit/10.1201/9781032630212-10/artificial-intelligence-data-science-various-domains-arun-kumar-garov-awasthi-minakshi-sharma?context=ubx

List of conferences/workshops attended

- Presented a paper title “Presentation of Futuristic Malarial Disease through a Hybrid Model of A.I. and Big Data” in the AICTE-sponsored International Conference on Computational Methods in Science & Technology, 19th -20th January 2023 at Chandigarh Engineering College, CGC Landran.
- Presented a paper title “GNN Model Based on Node Classification Forecasting in Social Network” in the AICTE-Sponsored International Conference on Computational Methods in Science & Technology, 19th -20th January 2023 at Chandigarh Engineering College, CGC Landran.
- Presented a paper title Data mining techniques for prediction of COVID 19 cases in the 4th International E-conference on incipient research in information technology commerce, management and linguistics 2021 on 17th and 18th September 2021.
- Presented a paper title “Comparative analysis of forecasting models in Healthcare (COVID-19) in Two days international conference on recent Trends and Technologies in soft computing (ICRTSC 22) organised by the department of Computer Science and applications, St. Peter’s Institute of Higher Education and Research, Avadi, Chennai 600054 on 21st and 22nd April 2022.
- Presented a paper on ‘Data mining prediction techniques in Healthcare sector” in the international conference on recent advances in fundamental and Applied Sciences (RAFAS 2021) held on June 25-26, 2021 organised by School of chemical engineering and Physical Sciences, lovely Faculty of Technology and Sciences Lovely Professional University, Punjab.
- Attended short time course on “Real time big data injection and analytics with Apache Kaka Hadoop and Spark” organised by Ravi Lovely Professional University w.e.f October 25,2021 to November 02,2021.
- Presented a paper titled “Development of Regression models for Crop Yield Forecasting of Jammu Kashmir, India: A case study” in the sixth international conference on advances in agriculture technology and allied Sciences (ICAATAS 2023) on June 19 to 21, 23 held at Loyola Academy Secunderabad, Telangana 500010 India.
- Presented a paper titled “Weather forecasting using data mining” in three days international conference on new frontiers in engineering and Social Sciences INFES 2022 organized by Eudoxia research University.

Published Matter

Horizon Research Publishing Corporation

Home Browse Journals Resources Online Submission Books About Us Contact Us

Environment and Ecology Research

Environment and Ecology Research Vol. 11(5), pp. 779 - 791
DOI: 10.13189/eer.2023.110508
Reprint (PDF) (1012Kb)

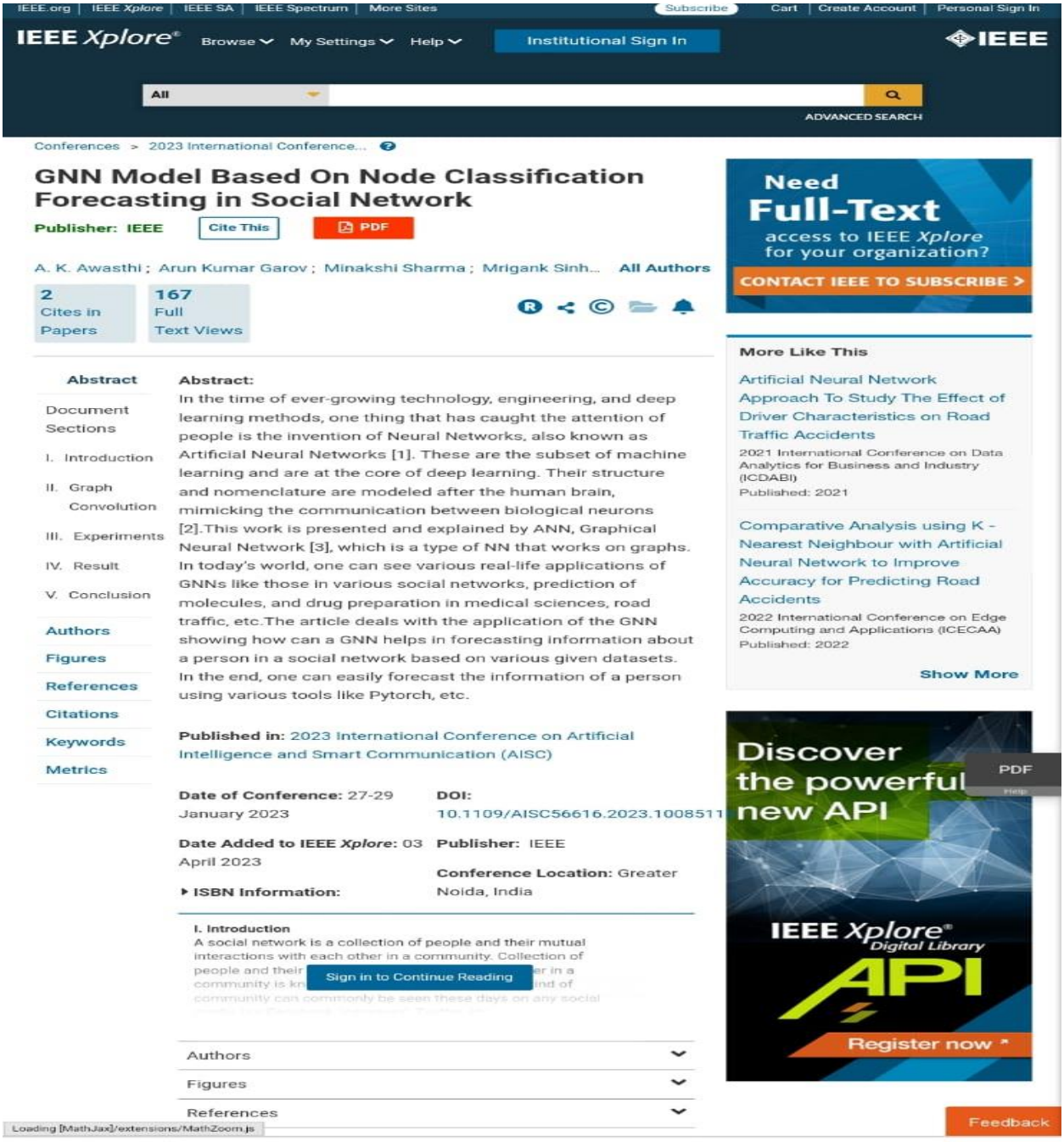
Rainfall Analysis of Haryana State, India: A Regional Frequency Analysis Approach

Minakshi Sharma [1], Rashi Jarial [1], Mohit Nain [1], A. K. Awasthi [1,*], Owais Aalam [1], Bhanu Pratap [2]
[1] School of Chemical Engineering and Physical Sciences, Lovely Professional University, India
[2] Haryana Space Applications Centre (HARSAC), CCS HAU, Hisar, India

ABSTRACT

The present study deals with the regional frequency analysis approach to analyzing rainfall data of Haryana state. From all over Haryana, 29 rain gauge stations were selected based on data availability. These stations were clustered into three groups namely, G1, G2, and G3 with the help of Ward's method. L-moments-based Heterogeneity test (H) was used to test the region's homogeneity and for each station, a discordancy measure (D_i) was calculated. Results of H test showed that all three groups were homogeneous and indicated that there was no discordant station in each group. Best-fit distribution was chosen using a goodness-of-fit measure (Z^{DIST}) and L-moment ratios diagram. Results of these tests showed that Pearson type-3 (PE-3) was the best-fitted distribution for group G1 and Generalized Extreme Value (GEV) and Generalized Normal (GNO) were the best for group G2, while for group G3, GNO was the best-fitted distribution. Using the L-moments method, the parameters of the fitted distributions were estimated, and regional growth curves were generated for each homogenous group. Regional growth curves provided regional quantiles at various return periods for each group. Further, quantiles were estimated at each station using regional quantiles by taking the station's mean as a scaling factor. The results showed that estimated quantiles using the best-fit distribution were in close agreement with observed data.

KEYWORDS
Annual Total Rainfall, Haryana, L-Moments, Quantiles, Growth Curve


IEEE.org | IEEE Xplore | IEEE SA | IEEE Spectrum | More Sites
Subscribe
Cart | Create Account | Personal Sign In
IEEE Xplore
Browse
My Settings
Help
Institutional Sign In
IEEE
All
ADVANCED SEARCH
Conferences > 2023 International Conference...
GNN Model Based On Node Classification Forecasting in Social Network
Publisher: IEEE
Cite This
PDF
A. K. Awasthi ; Arun Kumar Garov ; Minakshi Sharma ; Mrigank Sinh... All Authors
2 Cites in Papers
167 Full Text Views
Abstract
Document Sections
I. Introduction
II. Graph Convolution
III. Experiments
IV. Result
V. Conclusion
Authors
Figures
References
Citations
Keywords
Metrics
Abstract:
In the time of ever-growing technology, engineering, and deep learning methods, one thing that has caught the attention of people is the invention of Neural Networks, also known as Artificial Neural Networks [1]. These are the subset of machine learning and are at the core of deep learning. Their structure and nomenclature are modeled after the human brain, mimicking the communication between biological neurons [2].This work is presented and explained by ANN, Graphical Neural Network [3], which is a type of NN that works on graphs. In today's world, one can see various real-life applications of GNNs like those in various social networks, prediction of molecules, and drug preparation in medical sciences, road traffic, etc.The article deals with the application of the GNN showing how can a GNN helps in forecasting information about a person in a social network based on various given datasets. In the end, one can easily forecast the information of a person using various tools like Pytorch, etc.
Published in: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC)
Date of Conference: 27-29 January 2023
DOI: 10.1109/AISC56616.2023.1008511
Date Added to IEEE Xplore: 03 April 2023
Publisher: IEEE
ISBN Information:
Conference Location: Greater Noida, India
I. Introduction
A social network is a collection of people and their mutual interactions with each other in a community. Collection of people and their
Sign in to Continue Reading
community is kn
community can commonly be seen these days on any social
Authors
Figures
References
Loading [MathJax]/extensions/MathZoom.js
Need Full-Text access to IEEE Xplore for your organization?
CONTACT IEEE TO SUBSCRIBE >
More Like This
Artificial Neural Network Approach To Study The Effect of Driver Characteristics on Road Traffic Accidents
2021 International Conference on Data Analytics for Business and Industry (ICDABI)
Published: 2021
Comparative Analysis using K - Nearest Neighbour with Artificial Neural Network to Improve Accuracy for Predicting Road Accidents
2022 International Conference on Edge Computing and Applications (ICECAA)
Published: 2022
Show More
Discover the powerful new API
IEEE Xplore Digital Library API
Register now
PDF
Feedback

Presentation of futuristic Malarial Disease through a Hybrid Model of A.I. and Big data

Publisher: IEEE

A.K. Awasthi ; Minakshi Sharma ; Arun Kumar Garo ; Prayanshu Ch... All Authors

Abstract:

Transmission of parasites of female Anopheles mosquito is a major cause for the fatal disease like Malaria and most common symptoms are high fever, headache, abdominal pain, muscle pain, vomiting, diarrhea, Anemia, etc.Healthcare is a field where Forecasting is most beneficial and helpful for the cure of diseases and it can be possible only by using models of Neural Networks, Regression, and LSTM for predicting that how many confirmed cases will occur, how many people will get recovered and among them how many get dead cause of any disease based on past and present data for forecasting the future trend of these cases. Even though forecasting can be done by traditional methods but traditional methods are time consuming. The machine learning techniques are fast and accurate then also for adding efficiency and accuracy to these methods and techniques, Artificial Intelligence takes place in the field of Data Sciences. Artificial Intelligence came into the picture 100 years ago and shows remarkable growth in every field in the past few years. There are a lot of models proposed for different types of diseases in the field of healthcare. The visible patterns of symptoms and cases related to the disease can help in the medical field to prevent and cure it. This work is going to deal with the applications of Artificial Intelligence in Data Sciences which is to make an early reliable prediction of Malarial disease using Neural Networks, Regression, and LSTM model. It Compares the trends of different models and comparative results will demonstrate that the machine learning techniques practiced to forecast the malarial disease along with visible patterns gives information related to the patient's data.

Published in: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC)

Data Mining Prediction Techniques in Health Care Sector

Minakshi Sharma

Department of Mathematics, Lovely Professional University, Jalandhar, Punjab, 144411 India
Email: minakshisoharcya@gmail.com

Abstract. Knowledge discovery in databases (KDD) is another name of Data mining. It is an interdisciplinary area which focuses on extraction of useful knowledge from data in every sector

Volume 2768, Issue 1
23 June 2023

RESEARCH ARTICLE | JUNE 23 2023

Forecasting analysis of COVID-19 patient recovery using RF-DT model

A. K. Awasthi ✉; Minakshi Sharma; Arun Kumar Garov

— Author & Article Information

a) Corresponding author: dramitawasthi@gmail.com
b) minakshisohareya@gmail.com
c) arunkumar170@yahoo.com

AIP Conf. Proc. 2768, 020009 (2023)
https://doi.org/10.1063/5.0148356

This paper is to analyse and to foretell the effect of COVID-19. This virus constantly changes due to mutation and hence there are many variants like Delta, Omicron, Alpha, Beta, and Gamma. In this paper, we are using the RF-DT model by machine learning technique for the forecasting analysis of COVID-19 patient recovery data. The proposed result of forecasting analysis of COVID-19 patient recovery data compares with other data mining techniques like Naïve Bayes, linear regression, SVM. The proposed RF-DT model shows 99.5% accuracy. Moreover, a python programming language is used for the analysis and graphical presentation along with statistical summarization.

Topics
Data mining, Machine learning, Programming languages, Coronaviruses, Regression analysis

How to Cite:
Awasthi, A. K., & Sharma, M. (2022). Comparative analysis of forecasting models in healthcare (COVID-19). *International Journal of Health Sciences*, *6*(S3), 8649–8661. https://doi.org/10.53730/ijhs.v6nS3.8055

Comparative analysis of forecasting models in healthcare (COVID-19)

A. K. Awasthi
Department of Mathematics, School of Chemical Engineering & Physical Sciences, Lovely Professional University, Phagwara, Punjab, 144411, India
Email: dramitawasthi@gmail.com, amit.25155@lpu.co.in

Minakshi Sharma
Department of Mathematics, School of Chemical Engineering & Physical Sciences, Lovely Professional University, Phagwara, Punjab, 144411, India
*Corresponding author email: minakshisohareya@gmail.com

Abstract---Knowledge discovery in databases (KDD) is another name of Data mining. It is an interdisciplinary area which focuses on extraction of useful knowledge from data in every sector like health, education, business etc. There are many fields to explore like business, health care, e-commerce etc but nowadays, as covid pandemic is affecting everyone and due to surge in coronavirus cases causing shortage of hospital beds, oxygen supplies, vaccine and turning away patients from hospitals, put creaky health infrastructure in spotlight. The plenty of data is available in the medical field of these conditions. To analyse the problems, there are many data mining approaches which can be used to extract useful patterns from these types of data to follow the upcoming trends. This study is to compare the various models like KNN, improved RF model and multilayer perceptron by using SPSS and python software. The data of COVID-19 has been taken from Kaggle's website which is based on the symptoms and the forecasted results has been shown. In results and conclusion, the performance of every model has shown along with this, it also shows the models and mathematical algorithms in various fields of healthcare accordingly which can be used and benefitted in medical industries.

Keywords---data mining, knowledge discovery in databases, healthcare analysis, spss, python.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 27 March 2022, Manuscript revised: 9 April 2022, Accepted for publication: 18 May 2022

Conferences and workshops

Certificate No. 225495

LOVELY PROFESSIONAL UNIVERSITY
Transforming Education Transforming India

RAFAS

Certificate of Participation

This is to certify that Prof./Dr./Mr./Ms. **Ms. Minakshi Sharma**

of **Lovely Professional University**

has given poster presentation on **Data Mining Prediction Techniques in Health Care Sector**

in the International Conference on "Recent Advances in Fundamental and Applied Sciences" (RAFAS 2021) held on June 25-26, 2021, organized by School of Chemical Engineering and Physical Sciences, Lovely Faculty of Technology and Sciences, Lovely Professional University, Punjab.

Date of Issue : 15-07-2021
Place of Issue: Phagwara (India)

Prepared by
(Administrative Officer-Records)

Organizing Secretary
(RAFAS 2021)

Convener
(RAFAS 2021


CHANDIGARH
GROUP OF COLLEGES
Certificate
This is to certify that Prof./ Dr./ Mr./ Ms./ Minakshi Sharma
Research Scholar/ Professional from LPU, Punjab has participated/ presented/
attended/ co-authored a paper titled "GNN Model based on Node classification Fore-
-casting in Social Network"
in the AICTE-sponsored International Conference on Computational Methods in Science &
Technology held on 19th-20th January 2023 at Chandigarh Engineering College, CGC Landran.
Dr. Sushil Kumar
General Co-Chair
Dr. Sukhpreet Kaur
General Co-Chair
Dr. Rajdeep Singh
Executive Chair

HUMAN RESOURCE DEVELOPMENT CENTER

[Under the Aegis of Lovely Professional University, Jalandhar-Delhi G.T Road, Phagwara (Punjab)]

Certificate No. 236075

Certificate of Participation

This is to certify that **Ms. Minakshi Sharma** D/o Sh. Uttam Raj, participated in **Online Short Term Course on Real Time Big Data Ingestion and Analytics with Apache Kafka, Hadoop and Spark** organized by Lovely Professional University w.e.f. **October 25, 2021** to **November 02, 2021** (08 Days) and obtained "**B**" Grade.

Date of Issue : 02-11-2021
Place : Phagwara (Punjab), India

Prepared by	Checked By	Head
(Administrative Officer-Records)	Program Coordinator	Human Resource Development Center


EUDOXIA
USA
Certificate
of
Participation
MRD - 2022
This certificate is presented to
Minakshi Sharma
Lovely Professional University, India
for successful participation in two days
International Conference on Advanced Research,
Tools Techniques for Multidisciplinary Research Design
organised by
Eudoxia Research University - USA
and Eudoxia Research Centre India
In collaboration with
Department of Periodontology, Rama Dental College
Hospital and Research Centre, Rama University, UP, India
and Franklin International Business School, Nigeria
Date of Conference
9th and 10th December 2022
Prof. Catalino N. Mendoza
PHILIPPINES
Patron
Prof. Dr. Rajesh G Konnur
INDIA
Convener
Dr. Caleb Moyo
QATAR
Organizing Secretary
MRD - 2022
EUDOXIA
Dr. Pratisha Kumari
INDIA
Organizing Secretary
Marie Therese A. Villa
PHILIPPINES
Coordinator
Md. Adu Sayed Haque
BANGLADESH
Coordinator

No:ATAL/2020/1608056074

ALL INDIA COUNCIL FOR TECHNICAL EDUCATION

Nelson Mandela Marg, Vasant Kunj, New Delhi – 110 070

AICTE Training And Learning (ATAL) Academy

Certificate

This is certified that MINAKSHI SHARMA, Research scholar of Lovely professional University participated & completed successfully AICTE Training And Learning (ATAL) Academy Online FDP on "Mathematics for Machine learning" from 2021-1-18 to 2021-1-22 at Sri Jayachamarajendra College of Engineering, JSS Science and Technology university.

Director ATAL Academies

Coordinator


RAMAS
2021
EUDOXIA
CERTIFICATE
OF PRESENTATION
THIS CERTIFICATE IS PROUDLY PRESENTED TO
Minakshi Sharma
Research Scholar, LPU
for participating and presenting the paper titled
Data Mining Techniques for Weather Prediction
in two days
"International Conference on Recent Advances in Social Sciences, Humanities, Management and Scientific Research"
(RAMAS-2021)
Organized by:
Eudoxia Research Centre, India
ISO 9001:2015 Certified, IAF Accredited and Reg. under MCA Govt. of India
Dr. Jo-Ann Rollo
USA
PATRON
Dr. Worakamol Wisetsri
Thailand
CONVENER
Prof. Dr. Ramesh Ch. Rath
India
ORGANIZING SECRETARY
Vusi Maseko
South Africa
ORGANIZING SECRETARY
Chinnah Promise Chiweke
Nigeria
COORDINATOR
Keke Collins Obinna
Nigeria
COORDINATOR
Date: 5th and 6th June, 2021


St.PETER'S INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Deemed to be University U/S 3 of the UGC Act 1956)
NAAC Accredited, AICTE Approved & ISO 9001:2015 Certified
Avadi, Chennai – 600 054, Tamil Nadu, India
ISBN - 978-81-956262-9-8
CERTIFICATE ID : ICRTSC014
CERTIFICATE
OF APPRECIATION
This is to certify that
Dr.Minakshi Sharma
has participated and presented a research paper entitled "Comparative Analysis of Forecasting Models in Healthcare (Covid-19)" in "Two Days International e-Conference on Recent Trends and Technologies in Soft Computing (ICRTSC 22)" organized by the Department of Computer Science and Applications, St.Peter's Institute of Higher Education and Research, Avadi, Chennai – 600 054, on 21st & 22nd April 2022.
ICRTSC 2022
Dr.R.Latha
Organizing Chair
Maj.Dr.M.Venkatramanan
Dean (AS&MS)
Dr.L.Mahesh Kumar
Registrar

International Conference on
"Computational Applied Sciences and its Applications" (ICCASA-2022)
Department of Applied Sciences
University of Engineering and Management, Jaipur

AIP Conference Proceedings
VIRTUAL MODE

ICCASA 2022

CERTIFICATE
OF APPRECIATION

This is to certify that Ms./ Mr./ Dr./ Prof. MINAKSHI SHARMA has participated and presented paper entitled **"Forecasting Analysis of COVID-19 Patient Recovery Using RF-DT Model"** in **International Conference on "Computational Applied Sciences and its Applications" (ICCASA- 2022)** from April 28-29, 2022 conducted by University of Engineering and Management, Jaipur, Rajasthan (India).

Dr. Praphull Chabbra
Convener

Ms. Pallavi Malik
Convener

Prof. (Dr.) Biswajoy Chatterjee
Vice Chancellor, UEM, Jaipur

CHANDIGARH GROUP OF COLLEGES
Building Careers. Transforming Lives.

Certificate

This is to certify that Prof./ Dr./ Mr./ Ms./ Minakshi Sharma
Research Scholar/ Professional from LPU, Punjab has participated/ presented/ attended/ co-authored a paper titled Presentation of Futuristic Malarial Disease through a Hybrid Model of A.I. and Big data
in the AICTE-sponsored International Conference on Computational Methods in Science & Technology held on 19th-20th January 2023 at Chandigarh Engineering College, CGC Landran.

Dr. Sushil Kumar
General Co-Chair

Dr. Sukhpreet Kaur
General Co-Chair

Dr. Rajdeep Singh
Executive Chair