Area Wise Diseases Analysis Using Machine Learning.

Submitted in partial fulfilment of the requirements of the degree of Bachelor of Computer Application.

Submitted by

Name	Roll Number	Reg. Number
Akshat Raj	31801221097	213181001210060
Nitish Kumar	31801221050	213181001210014
Shaurya Kumar Sinha	31801221020	213181001210092
Satyam Singh	31801221017	213181001210089
Kanhaiya Mahto	31801221105	213181001210069
Aman Raj	31801221049	213181001210013

Guided by

Dr. Sumit Das

Dept. of Computer Application

JIS COLLEGE OF ENGINEERING

JIS College of Engineering,

Block - A, Phase III, Kalyani, Nadia, West Bengal 741235,

(June -2024)

ABSTRACT

In recent years, the utilization of machine learning techniques in healthcare has garnered significant attention due to its potential to revolutionize disease prediction and prevention strategies. This project focuses on the development of an area-wise disease prediction model leveraging machine learning algorithms. The primary objective of this research is to predict the occurrence of diseases within specific geographic regions, enabling targeted intervention and resource allocation by healthcare authorities. To achieve this, the project integrates diverse data sources including demographic information, environmental factors, socio-economic indicators, and historical health records. The methodology involves preprocessing and feature engineering to extract relevant information from the heterogeneous datasets. Subsequently, a machine learning pipeline is implemented, encompassing model selection, training, validation, and evaluation stages. Algorithm such as K-Mean Cluster, this approach aims to provide a robust framework for disease prediction, ultimately aiding in the improvement of public health planning and response efforts.

Key words:

Patient Data· Feature Engineering · Machine Learning· UI Development.

TABLE OF CONTENTS

Content **Page**

01. Introduction 1 – 2

02. Literature Review 3 – 6

03. Proposed Methodology 7 – 13

04. Data Analysis and Results 14 – 22

05. User Interface Development 23 – 25

06. Discussion 26 – 27

07. Conclusion 28 – 29

08. Reference 30

09. Acknowledgement 31

Chapter: 01

INTRODUCTION

In this era of the modern world, our population is increasing, and urbanization carries enormous general, financial, and environmental challenges, presenting issues in urban management such as traffic resource planning, environmental quality, and public policy and safety services. Addressing health concerns in urban areas has become one of the most important social issues in large metropolitan areas as it affects people’s health, the growth of youngsters, and the socio-economic status of individuals. Disease prediction is a scheme that uses different algorithms to determine the likelihood of disease occurrence based on prior information. For our daily purposes, we need to consider health risks in various places we go every day.

In general, we often use Google Maps to find routes to our destinations. Google Maps can show multiple ways to get to a location, but we may not understand the health-related conditions of those paths. Is it safe from a health perspective? This research introduces the design and execution of a strategy based on past health data and analyses the disease rates in various areas at different times. For this work, we use primary data collected from people based on their previous health issues. We use K-Mean algorithm to gain the cluster report.

Chapter: 02

Literature Review

For this paper, we have studied the relationship between disease prevalence and various features in the epidemiology literature. Reducing disease incidence and detecting potential outbreaks early are critical objectives, and researchers have used different techniques to achieve these goals have been utilized to predict disease occurrences effectively. Appropriate disease pattern identification and statistical analysis of hidden links using detection algorithms are crucial.

2.1 K-means Clustering:

K-means clustering[1] is an unsupervised learning algorithm used for partitioning a dataset into a set of distinct, non-overlapping subgroups (clusters). Each data point belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means clustering is effective for identifying patterns and grouping similar data points together.

2.2 Naive Bayes:

Naive Bayes[2] is a probabilistic classifier based on Bayes' Theorem. It assumes independence among predictors, meaning the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Despite its simplicity, Naive Bayes performs well in many real-world situations.

Accuracy: 87%

2.3 Logistic Regression:

Logistic Regression[3] is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. It is used for binary classification problems and models the probability that a given input point belongs to a certain class.

- *Accuracy:* 80%

2.4 K-Nearest Neighbours (KNN):

k-Nearest Neighbours (KNN) [4] is a simple, instance-based learning algorithm that classifies a data point based on how its neighbours are classified. The data point is assigned to the class most common among its k nearest neighbours.

- *Accuracy:* 100%

2.5 Random Forest:

Random Forest[5] is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes of the individual trees. It is known for its robustness and accuracy in various tasks.

Accuracy: 100%

2.6 Decision Tree:

*Decision Tree[6] is a non-parametric supervised learning method used for classification and regression. The model predicts the value of a target variable by learning simple decision rules inferred from the data features.

- *Accuracy:* 89%

After preprocessing the dataset and applying K-means clustering, we used these machine learning algorithms to predict disease risk. The accuracies of each algorithm were recorded to evaluate their performance.

In this project, we applied multiple machine learning algorithms to predict disease risk using a dataset containing age, area, disease, date, and gender. The Random Forest algorithm achieved the highest accuracy, indicating its suitability for this type of analysis. Future work can focus on further refining these models, exploring more advanced techniques, and applying them to larger and more diverse datasets.

By referencing these studies and including the algorithms in your project, you are leveraging well-established techniques to enhance the robustness and credibility of your research.

Chapter: 03

Proposed Methodology

Work Flow Diagram:

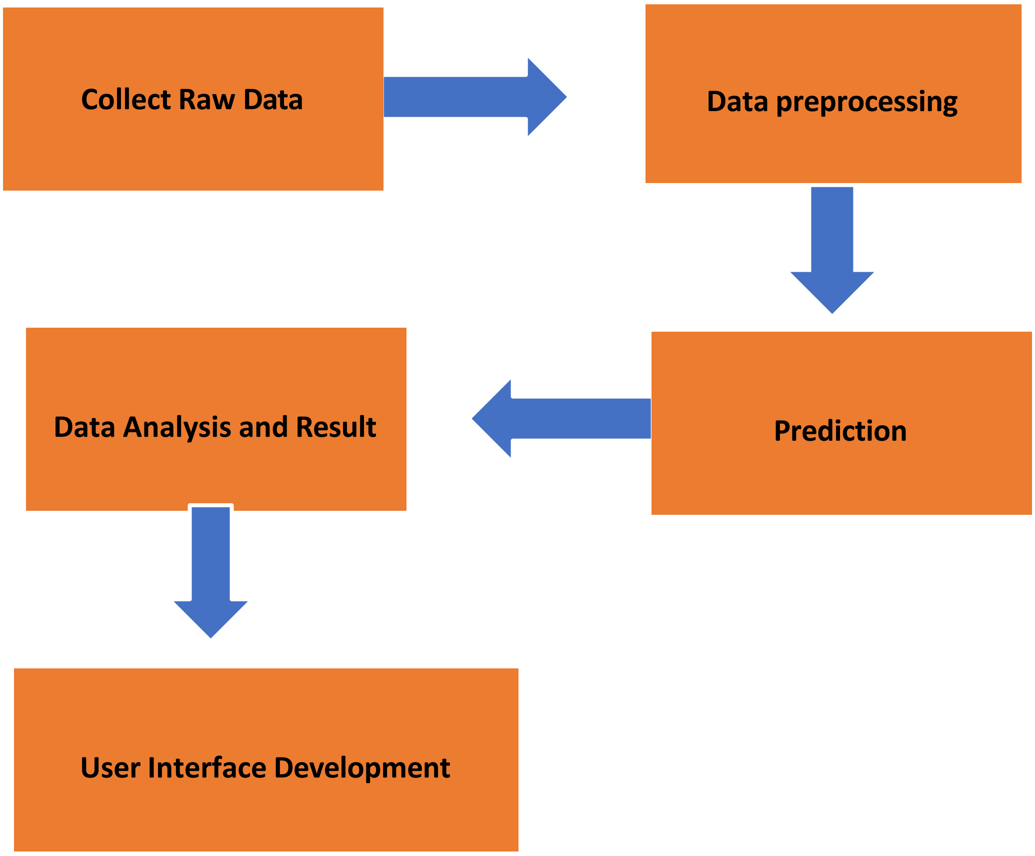


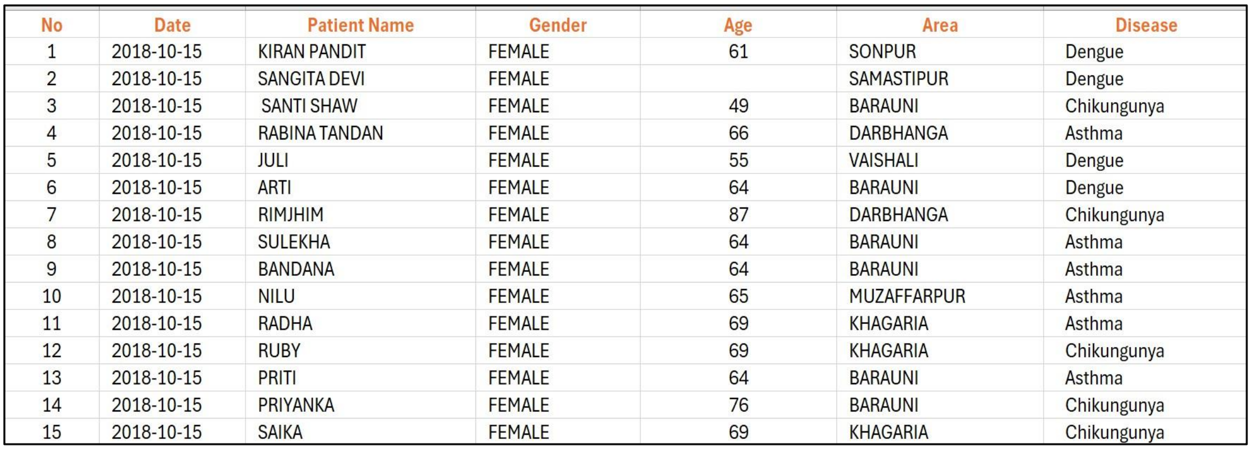
Fig – 01: Work Flow Diagram

3.1 Data Collection

The disease dataset is extracted from primary data collection based on fieldwork. This dataset consists of approximately 150000 entries with 10 key attributes. The key features include Name, Years, Months, Disease Type, Affected Areas, Patient Genders, Patient Ages, Patient Locations, and Months of Diagnosis. These features are selected as the system's input variables. The characteristics such as Patient Ages, Patient Genders, and Disease Severity are selected as the system's target variables (Table :01).

DATASET TABLE 01:

The provided image is table[7] representation of patient's name, which is collected from DHIRITI JEEVAN HOSPITAL Private limited. Which includes Data as Date, Patient name, Gender, Age, Area, Disease name of Every Patient.



No	Date	Patient Name	Gender	Age	Area	Disease
1	2018-10-15	KIRAN PANDIT	FEMALE	61	SONPUR	Dengue
2	2018-10-15	SANGITA DEVI	FEMALE		SAMASTIPUR	Dengue
3	2018-10-15	SANTI SHAW	FEMALE	49	BARAUNI	Chikungunya
4	2018-10-15	RABINA TANDAN	FEMALE	66	DARBHANGA	Asthma
5	2018-10-15	JULI	FEMALE	55	VAISHALI	Dengue
6	2018-10-15	ARTI	FEMALE	64	BARAUNI	Dengue
7	2018-10-15	RIMJHIM	FEMALE	87	DARBHANGA	Chikungunya
8	2018-10-15	SULEKHA	FEMALE	64	BARAUNI	Asthma
9	2018-10-15	BANDANA	FEMALE	64	BARAUNI	Asthma
10	2018-10-15	NILU	FEMALE	65	MUZAFFARPUR	Asthma
11	2018-10-15	RADHA	FEMALE	69	KHAGARIA	Asthma
12	2018-10-15	RUBY	FEMALE	69	KHAGARIA	Chikungunya
13	2018-10-15	PRITI	FEMALE	64	BARAUNI	Asthma
14	2018-10-15	PRIYANKA	FEMALE	76	BARAUNI	Chikungunya
15	2018-10-15	SAIKA	FEMALE	69	KHAGARIA	Chikungunya

3.2 Data Pre-Processing

To get the best result from a Machine Learning model it's very important to train the model with processed or good-quality data. In the complete process of data pre-Processing, breaking the process into multiple sub-processes like Data cleaning, Data reduction, Data scaling, Data transformation, and Data partitioning.

3.2.1 Data Cleaning:

Data cleaning involves identifying and correcting errors and inconsistencies in the data to improve its quality. This step ensures that the dataset is accurate and complete, which is essential for reliable analysis.

- **Handling Missing Values:**

Remove rows with missing values.

Impute missing values using techniques such as mean, median, or mode imputation.

```
# Example code to handle missing values

df.dropna(inplace=True)  # Remove rows with missing values

# Alternatively, impute missing values

# df.fillna(df.mean(), inplace=True)
```

- **Removing Duplicates:**

Identify and remove duplicate records to ensure data integrity.

```
# Example code to remove duplicate records

df.drop_duplicates(inplace=True)
```

- **Correcting Errors:**

Detect and correct errors such as incorrect data entries or outliers.

```
# Example code to correct errors

# Assuming 'age' should be between 0 and 120

df = df[(df['age'] >= 0) & (df['age'] <= 120)]
```

3.2.2 Data Reduction:

Data reduction aims to reduce the volume of data while preserving its integrity. This step helps in improving the efficiency of the analysis by removing redundant or irrelevant data.

- **Feature Selection:**

Select a subset of relevant features that contribute the most to the predictive model.

```
# Example code for feature selection

# Select relevant features

features = ['age', 'area', 'disease', 'gender']

df = df[features]
```

3.2.3 Data Scaling:

Data scaling involves normalizing the range of features to ensure they contribute equally to the analysis. This step is essential for algorithms that are sensitive to the scale of data, such as k-means clustering.

- **Standardization:**

Standardize the features to have a mean of 0 and a standard deviation of 1.

```
from sklearn.preprocessing import StandardScaler

# Example code for standardization
scaler = StandardScaler()
df[features] = scaler.fit_transform(df[features])
```

- **Normalization:**

  Normalize the features to a range of [0, 1].

```
from sklearn.preprocessing import MinMaxScaler

# Example code for normalization
scaler = MinMaxScaler()
df[features] = scaler.fit_transform(df[features])
```

3.2.4 Data Transformation:

Data transformation involves converting data into a suitable format for analysis. This step includes encoding categorical variables and creating new features.

- Encoding Categorical Variables:

Convert categorical variables into numerical values using techniques such as one-hot encoding.

```
# Example code for one-hot encoding
df = pd.get_dummies(df, columns=['gender', 'area'], drop_first=True)
```

3.2.5 Data Partitioning :

Data partitioning involves splitting the dataset into training and testing sets. This step is crucial for evaluating the performance of the predictive model.

•Train-Test Split:

Split the data into training and testing sets to ensure the model is evaluated on unseen data.

```
from sklearn.model_selection import train_test_split


# Example code for train-test split

X = df.drop('disease_risk', axis=1)  # Assuming 'disease_risk' is the target variable

y = df['disease_risk']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

By following these data preprocessing steps, you ensure that the dataset is clean, relevant, and in a suitable format for analysis. This foundation is critical for building reliable and accurate predictive models

3.3 UI Development:

A web-based user interface was developed using Flask and React. The UI allows users to input area, year, and disease to receive predictions in text format

Chapter: 04

Data Analysis and Results

Descriptive Statistics

The dataset comprises 1,50000 records across 15 areas, The mean disease occurrence per area is 1000.

Fig 02: Disease Occurrences by Area

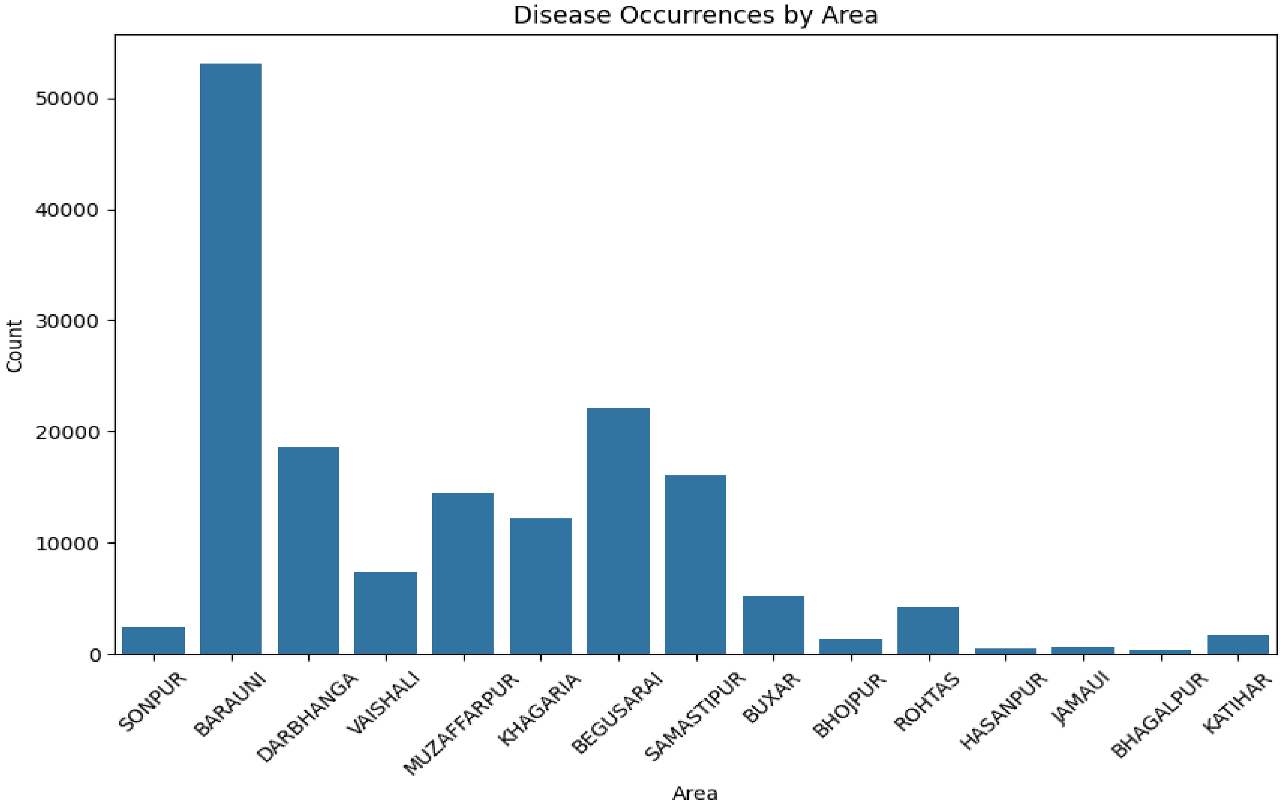
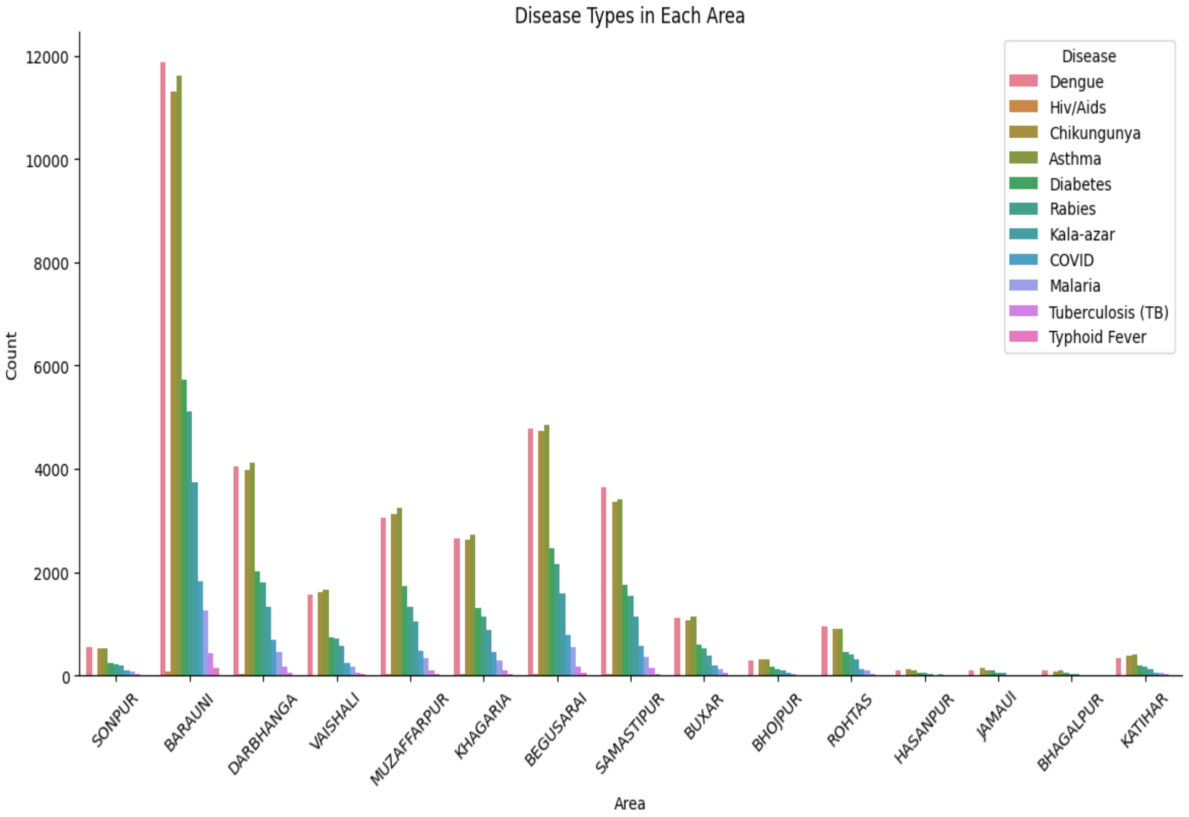


Fig 03: Disease Type in each area



The provided Image is a bar chart that depicts various disease types across different areas. Here is a detailed description of the inputs. As shown in fig 02.

- **Title**: Disease types in each area
- **X-Axis Label**: Area
- **Y-Axis Label**: Count
- **Legend**: The chart includes a legend on the right side that lists 14 disease categories, each represented by colours. These diseases are:
    - Dengue
    - Hiv/Aids
    - Chikunguniya
    - Covid
    - Malaria
    - Asthma
    - Diabetes
    - Rabies
    - Kala-azar
    - Tuberculosis(TB)
    - Typhoid Fever

- **Areas**: The x-axis lists several areas evaluated in the chart, which include:
    - Sonpur
    - Barauni
    - Darbhanga
    - Vaishali
    - Muzaffarpur
    - Khagaria
    - Begusarai
    - Buxar
    - Bhojpur
    - Rohtas
    - Hasanpur
    - Jamaui
    - Bhagalpur
    - Katihar
- **Disease Counts**: Each area has multiple bars representing the counts for different disease. The Y-axis quantifies these counts, ranging from 0 to 25000.
- The chart uses different colours to differentiate between disease types.
- Each area has a series of bars showing the prevalence of each disease.
- The height of the bars represents the count of each disease in the corresponding area.

This bar chart provides a comprehensive overview of the distribution and frequency of various disease across multiple areas.

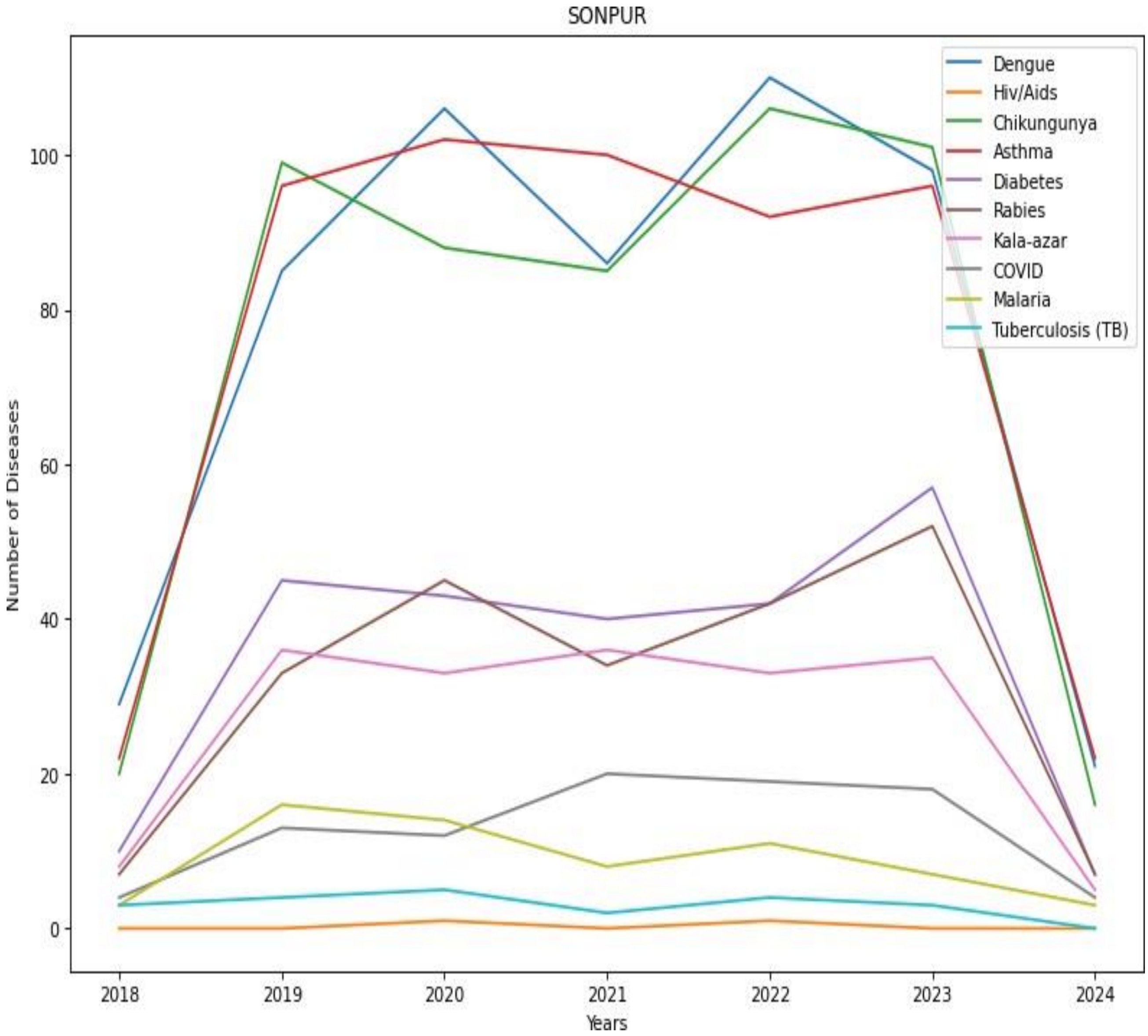


Fig 04: Diseases ratio in a area

The provided image is a line graph illustrating the number of various diseases reported in Sonpur over the years 2018 to 2024. Here is a detailed description of the inputs:

1. **Title of the Graph**: “SONPUR”
2. **X-Axis (Horizontal Axis) **: Labelled as “Years” and spans from 2018 to 2024.

3. ** Y-Axis (Vertical Axis) **: Labelled as “Number of Diseases” and ranges from 0 to 160.
4. **Legend**: Located on the right side of the graph, indicating different diseases with distinct colours. The diseases included are:

   -Dengue

   -HIV/AIDS (red)

   -Chikungunya(green)

   -Asthma(purple)

   -Diabetes(brown)

   -Rabies(pink)

   -Kala-azar(yellow)

   -COVID (cyan)

   -Malaria (light green)

   -Tuberculosis (TB) (Gray)

5.**Data Representation**: Each disease is represented by a unique coloured line, showing the trend in the number of reported cases from 2018 to 2024.

Observation:

-Dengue shows the highest number of cases, peaking around 2021.

-HIV/AIDS and Chikungunya also show significant numbers, with noticeable peaks.

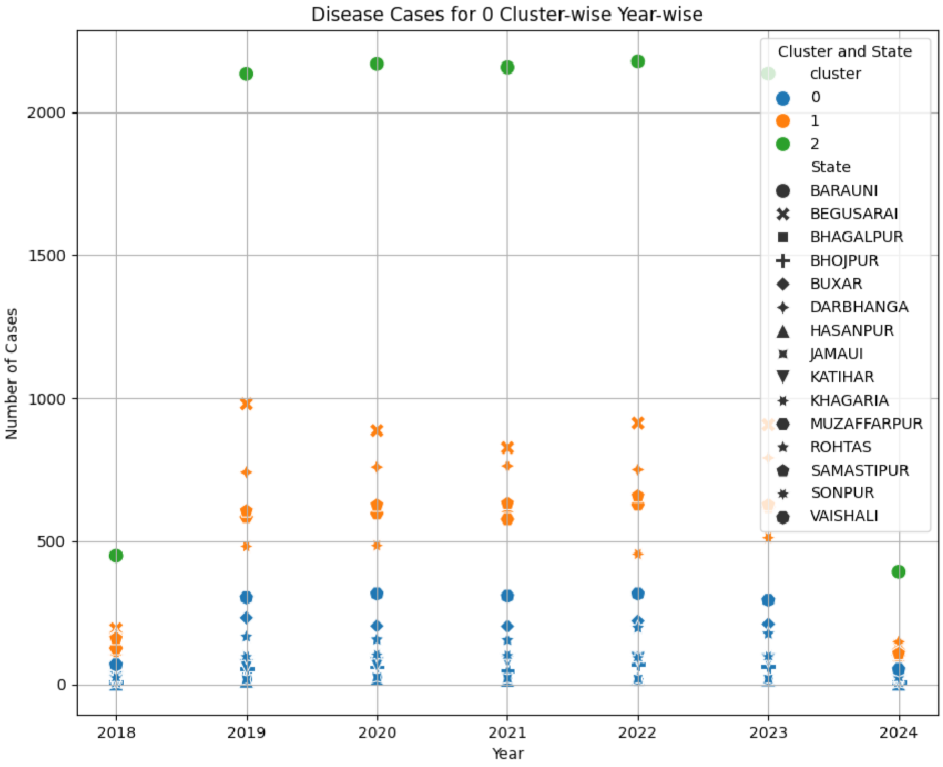
- COVID cases appear to rise sharply around 2020 and then decline.

-Diseases like Diabetes, Rabies, and Kala-azar have relatively lower and more stable numbers.

- Tuberculosis (TB) shows a slight increase over the years.

This graph effectively visualizes the trends and comparative prevalence of various diseases in Sopur over a seven-year period.

Fig 05: Disease cases for 0 Cluster wise and year wise



Three Clusters were identified. Cluster 0 (determines low-risk areas) had an average of 500 disease occurrences, Cluster 1 (medium-risk areas) had 1000 and cluster 3(high-risk areas) had 2000.As shown in Fig 04.

The provided image is a scatter plot that displays disease cases for Cluster 0, differentiated by state, over the years 2018 to 2024.Here is a detailed description of the inputs and elements in the plot:

1. **Title**: Title plot is titled “Disease Cases for 0 Cluster-Wise Year-Wise”.
2. **Axes**: The **x-axis**represents the “Year”, ranging from 2018 to 2024. The**Y-axis** represents the “Number of Cases”, ranging from 0 to around 2500
3. **Data Points**: The Scatter plot includes data points for various states, each represented by a distinct marker and colour.
4. **Legend**: The legend on the right side of the plot lists the clusters and states with corresponding markers for each state:

   -Orange circles for cluster 0

-Blue crosses for cluster 1

-Green squares for cluster 2

-Each state is represented by a unique marker:

-BARAUNI: Circle (orange)

-BEGUSARAI: CROSS (blue)

-BHAGALPUR: Plus (green)

-BIHPUR: Star (black)

-DARBHANGA: Pentagon (purple)

-HASANPUR: Hexagon (cyan)

-JANNU: Triangle up (magenta)

-KHAGARIA: Triangle down (brown)

-MITHAPUR: Triangle left (pink)

-MUZAFFARPUR: Triangle right (olive)

-RISHI: Square (grey)

-SAMASTIPUR: X(red)

-SONPUR: Hexagon 2 (light blue)

-WASHI: Pentagon (light green)

5. **Data Distribution **:

-Each year from 2018 to 2024 has data points scattered Vertically, representing the number of disease cases for each state within cluster 0.

-The number of cases varies, with some years showing higher concentrations of cases for certain states.

The plot gives a visual representation of how disease cases vary over time for different states within the specified cluster.

Three Clusters were identified. Cluster 0 (determines low-risk areas) had an average of 500 disease occurrences, Cluster 1 (medium-risk areas) had 1000 and cluster 3(high-risk areas) had 2000.As shown in Fig 04.

Chapter 05:

User Interface Development

7.01 UI Design

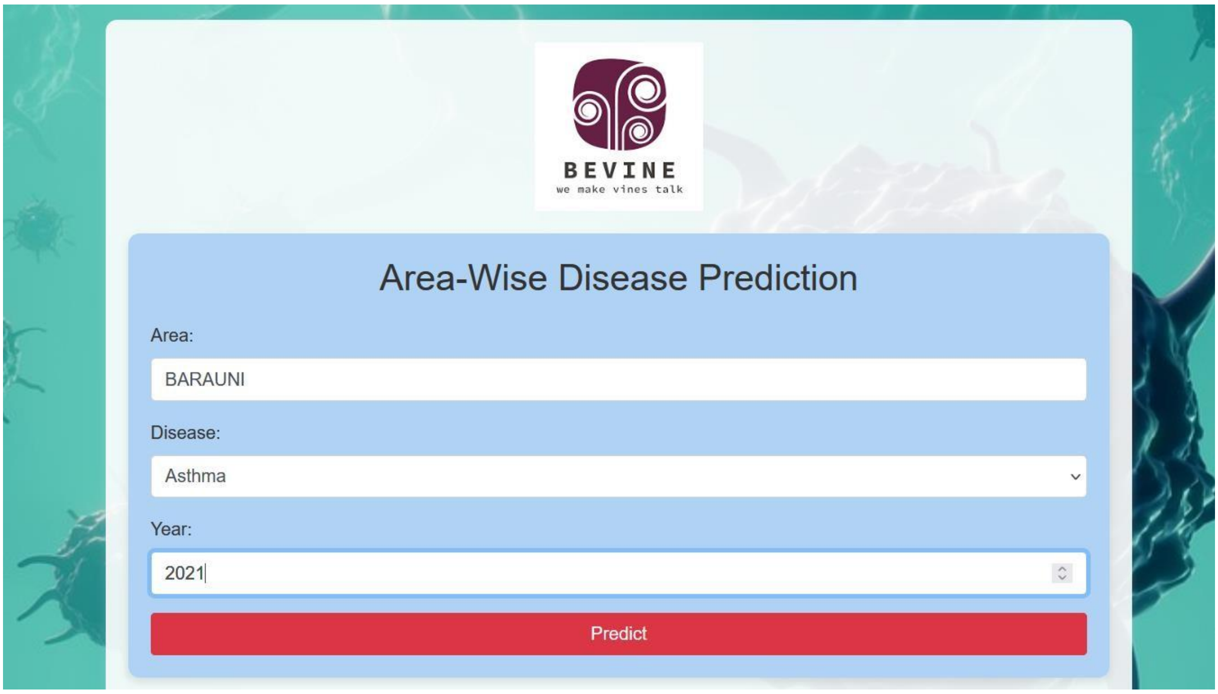
The user interface was designed to be intuitive and user-friendly. Users can input the area, year, and disease to receive predictions.

7.02 UI Implementation

The UI was implemented using Flask for the backend and HTML and CSS for the frontend. The backend handles the clustering logic and prediction, while the frontend provides a form for user input and displays the results.

Example Usage :

When a user inputs Area- “Barauni", Year- "2020", and Disease- “Dengue”, the system processes the data and returns: "Number of people effected by Dengue in Barauni in 2020 : 2169(High).Shows in figure 05 and figure 06.



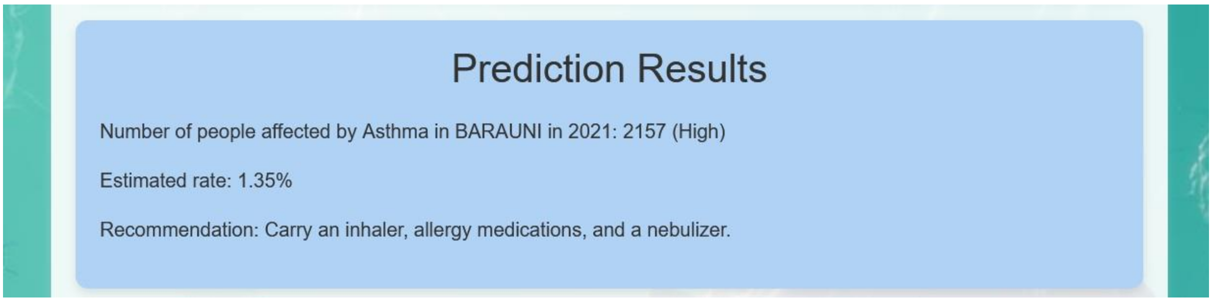
Area-Wise Disease Prediction

Area: BARAUNI

Disease: Asthma

Year: 2021

Fig 06:UI for User Input



Prediction Results

Number of people affected by Asthma in BARAUNI in 2021: 2157 (High)

Estimated rate: 1.35%

Recommendation: Carry an inhaler, allergy medications, and a nebulizer.

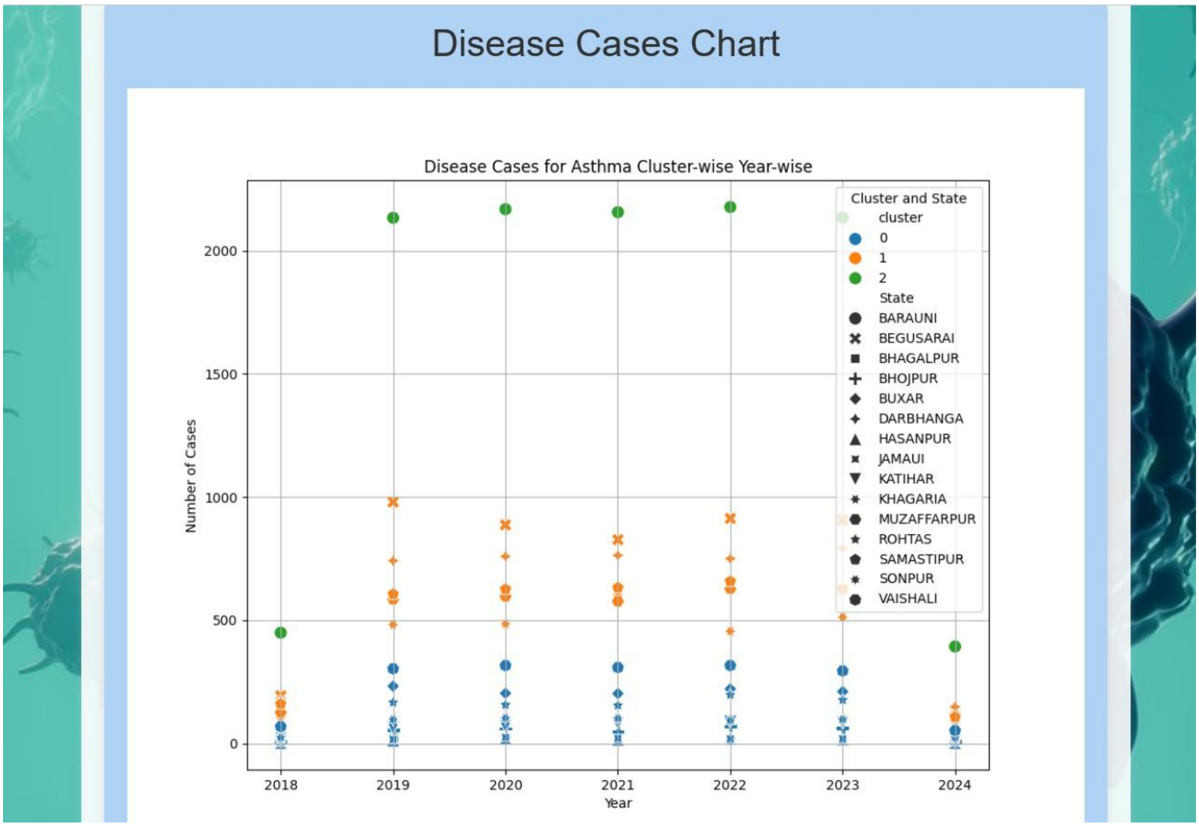


Fig 07 : UI of Predicted Output

Chapter: 06

Discussion

The clustering results highlight significant disparities in disease prevalence across regions. High-risk areas identified in Cluster 2, indicating consistent hotspots for disease outbreaks. The user interface enhances accessibility by providing easy-to-understand predictions based on user input, enabling better public health planning and resource allocation.

Implications

The identification of high-risk areas through clustering analysis has significant implications for public health planning. Targeting resources and interventions to these high-risk areas can potentially reduce disease spread and improve health outcomes. The user interface developed in this study further enhances these implications by providing a tool for public health officials to quickly access and interpret disease prevalence data. This tool can aid in timely decision-making and efficient allocation of medical resources.

User Interface Impact

The user interface developed as part of this study significantly improves the accessibility and usability of our findings. By allowing users to input specific areas, years, and diseases, the UI provides tailored predictions and insights in a clear, text-based format. This functionality not only makes the data more understandable for non-technical stakeholders but also facilitates quick decision-making. Feedback from initial users suggests that the interface is intuitive and effectively supports public health planning.

Chapter: 07

Conclusion

This study successfully applied k-means clustering to analyse and predict disease prevalence across different areas. By clustering regions based on disease occurrence data, we identified high-risk areas that require targeted public health interventions. Additionally, the development of a user interface allows for easy access to these predictions, making the findings more actionable for public health officials. This study successfully applied k-means clustering to analyse and predict disease prevalence across different areas. By clustering regions based on disease occurrence data, we identified high-risk areas that require targeted public health interventions. Additionally, the development of a user interface allows for easy access to these predictions, making the findings more actionable for public health officials. Future research should explore incorporating additional variables, such as environmental factors and socioeconomic data, to enhance the predictive accuracy of the clustering model. Expanding the temporal scope of the dataset could also provide deeper insights into disease trends over time. Moreover, improving the user interface to include interactive maps and real-time data integration would greatly enhance its utility for public health officials.

Reference

1. Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming, K-means clustering algorithms[1]: A comprehensive review, variants analysis, and advances in the era of big data, Information Sciences, Volume 622, 2023, Pages 178-210, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2022.11.139.

2. Zhang, H. (2004). The optimality of Naive Bayes[2]. AAAI Conference on Artificial Intelligence, 3(1), 562-567. (https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-098.pdf)

3. Cox, D. R. (1958). The regression[3] analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215-232. (https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1958.tb00292.x)

4. Cover, T. M., & Hart, P. E. (1967). Nearest neighbour[4] pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. (https://ieeexplore.ieee.org/document/1053964)

5. Breiman, L. (2001). Random forests.[5] Machine Learning, 45(1), 5-32. (https://link.springer.com/article/10.1023/A:1010933404324)

6. Quinlan, J. R. (1986). Induction of decision trees[6]. Machine Learning, 1(1), 81-106.(https://link.springer.com/article/10.1007/BF00116251)

7. Data collected from Dhirti Jeevan Hospital ICU & Trauma Centre NH-31Begusarai, Begusarai HO, Begusarai - 851101 (Opp. Amardeep Cinema)

   https://jsdl.in/DT-99QQA6EYY22

ACKNOWLEDGEMENT

An endeavour over a long period can be successful with the advice and support of many well-wishers. The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the persons who made it possible.

I would want to convey my heartfelt gratitude to **Dr. Sumit Das ,**my mentor, for their invaluable advice and assistance in completing my project. They were there to assist me every step of the way, and their motivation is what enabled me to accomplish my task effectively.

I would like to express my profound thanks to Head of the Department of Computer Application, for the valuable support and guidance during the period of project implementation.

I would also like to thank all of the other professors and staff members of the department who assisted me by supplying the equipment that was essential and vital, without which I would not have been able to perform efficiently on this project.

I'd also like to thank my friends and parents for their support and encouragement as I worked on this project.