**APPLICATION OF SCAN STATISTICS AS AN INFERENTIAL AND MULTIVARIATE TOOL TO ANALYSE CARDIOVASCULAR CASES IN INDIA: A DATA SCIENCE APPROACH**

Saumya Sinha

*Department of Mathematics*

*M.Sc. Data Science*

*Vellore Institute of Technology, Vellore, India*

[Saumya.sinha2023@vitsudent.ac.in](mailto:Saumya.sinha2023@vitsudent.ac.in)

Sriya S Pillai

*Department of Mathematics*

*M.Sc. Data Science*

*Vellore Institute of Technology, Vellore, India*

[sriya.spillai2023@vitstudent.ac.in](mailto:sriya.spillai2023@vitstudent.ac.in)

Dr. Jitendra Kumar

*Associate Professor*

*Department of Mathematics*

*School of Advanced Sciences*

*Vellore Institute of Technology, Vellore, India*

[jitendra.kumar@vit.ac.in](mailto:jitendra.kumar@vit.ac.in)

***Abstract*: Cardiovascular Diseases continue to be the leading causes of death across all regions of India, including less affluent states and rural areas. Medical professionals and experts note that heart-related issues among Indians have nearly doubled over the past decade, affecting even younger individuals. This increase is attributed to several factors, including irregular working hours causing stress, as well as the growing prevalence of smoking and alcohol consumption. Such lifestyles contribute to conditions like high blood pressure, obesity, and diabetes, ultimately leading to heart problems. Early identification of signs of cardiovascular diseases and consistent medical treatment can reduce both mortality rates and the number of individuals affected. In the realm of spatial epidemiology, employing Scan Statistics as an inferential and multivariate analytical tool assists in determining which clusters of alarms warrant further investigation and which are likely random occurrences. Calculations have been performed with SaTScan Version 10.1.2. The research presented in this paper incorporates data from various states and union territories of India from the years 2017-2021.**

***Keywords*: *Cardiovascular Diseases, Spatial epidemiology, Scan Statistics.***

I. INTRODUCTION

In India, Cardiovascular Diseases are emerging as a major public health concern. They are classified as non-communicable diseases and are responsible for a significant number of deaths in India. The age-standardized death rate for CVD in India (282 deaths/100,000 (264–293)) is higher compared with global levels (233 deaths per 100,000 (229–236)). Urban areas have a higher prevalence of CVD compared to rural areas.

Heart attacks occur when there is a blockage in the blood vessels that disrupts blood flow to the brain or heart. The accumulation of fatty deposits on the inner walls of arteries supplying nutrients to the brain or intestines is recognized as a primary cause of CVD. Strokes can also be caused by blood clots or bleeding in the brain's blood vessels. Many researches have shown that the major contributing factors to strokes and heart attacks typically include alcohol and tobacco consumption, an unhealthy diet, lack of physical activity, excessive alcohol use, high blood pressure, and diabetes.

The aim of this research paper is to detect cardiovascular disease hotspots across India. The problem of CVD hotspot detection is modelled using the Spatial Scan Statistics for finding every state/UT where there is higher density of cardiovascular cases than normal at a particular period of time.

For calculations, SaTScan, a software that analyses spatial, temporal and space-time data has been used. The software is used to perform geographical surveillance of disease, to detect space-time disease clusters, test whether a disease is randomly distributed over space and time, and to evaluate the statistical significance of disease cluster alarms.

Further, to analyse relation between different factors affecting CVD, the data has been modelled using linear regression.

The project work is relevant to all stake holders ranging from common people to policy makers. In India, there is an approximate ratio of 17 nursing and midwifery professionals per 10,000 residents, while the doctor-to-patient ratio stands at 9 per 10,000 inhabitants. Hence, it is crucial to find out the feasible solution and exploring the hotspots. It enables in precise decision making and efficient resource allocation.

II. LITERATURE SURVEY

This section discusses some of the related works conducted in the field of cardiovascular disease analysis.

This research [1] works to find out the best prediction system for heart disease detection among the seven different supervised machine learning techniques, namely KNN, Support Vector Machine, Naïve Bayes, Random Forest, AdaBoost, ANN and Logistic Regression. The proposed work found Logistic Regression to be the promising predictive model with classification accuracy 85.3 %, F1- Score 80%, Precision 82.6%, and Recall 85.3%.

The research of authors [Fiaidhi, J.](javascript:__doLinkPostBack('','ss~~AR%20%22Fiaidhi%2C%20J.%22%7C%7Csl~~rl','');) and [Mohammed S.](javascript:__doLinkPostBack('','ss~~AR%20%22Mohammed%2C%20S.%22%7C%7Csl~~rl','');) [2] show that the strength of qualitative analytics lies in data thickness, involving patient characteristics such as socioeconomic status, family background, working conditions, social support, lifestyle risk factors, age group and gender and their weights in a particular clinical practice. A Fuzzy C-Means algorithm is presented as technique for prognostic predictions to identify risk groups associated with CVD conditions, compared to Machine Learning Algorithms: Logistic Regression and Decision Tree.

In this research a stacking ensemble method is used [4] to produce an optimal predictive model, by combining several single classifiers, for the prediction of CVDs. The exploratory data analysis indicates that CVD was more common in males and diabetic individuals. Furthermore, people above the age of 65 were more susceptible to CVDs.

III. METHODOLOGY

DATASET:

For scan statistics, the geographical data, the estimated population and the no. of cardiovascular cases in each state and union territory of India has been collected to locate the hotspots.

For modelling, data for the age group 45 and above has been collected from Longitudinal Ageing

Study in India (LASI) Wave 1 for every state. The LASI Wave 1 is a nationally representative survey of 73,396 adults of age 45 and above across all states and union territories of India. The data for wave 1 was collected from the year 2017-2019, but the data of Sikkim was collected from the year 2020-2021.The data is then gathered and converted into excel form for further processing. The dataset contains cases of different factors relating to cardiovascular diseases such as, blood pressure, diabetes, alcohol and tobacco consumption, along with other necessary details.

For comprehensive study, it is recommended to use the data from each district of the country.

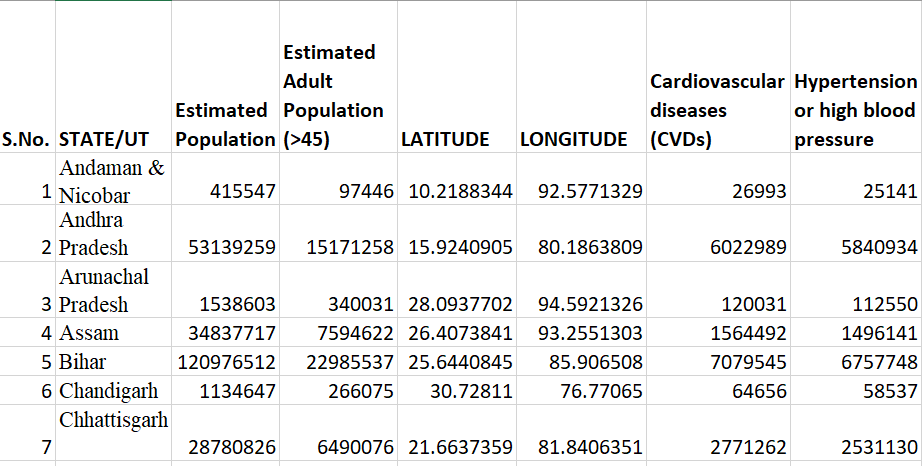


Table1: glimpse of data (37 rows X 16 cols)

After collecting and processing the data, Scan Statistics, Descriptive analysis and Regression analysis is done, using SatScan and data analysis tool in excel.

MULTIPLE LINEAR REGRESSION MODEL:

For a given regressed variable, multiple regression involves two or more regressor variables. A linear regression model is created by calculating the regression coefficients, as indicated in eq1. The inference of the regressor variable on the regressed variable is suggested by these regression coefficients. Afterwards, prediction of regressed variable can be performed for a specific collection of regressor variables.

(1)

where,

y = dependent/regressed/output/response/predict variable (no. of cardiovascular diseases)

= y-intercept

= regression coefficients

= independent/ regressor variables

Model Assumptions:  
1. There is no Autocorrelation. Random error terms or disturbances are iid distributed.

2.

3. There’s assumption of Homoskedasticity. Variance of each disturbance term conditional on the chosen value of independent variable is equal i.e.

Descriptive analysis of various factors i.e. Blood pressure, Diabetes, Alcohol consumption, Smoking, Obesity, High cholesterol and Anaemia, is performed.

The mean and standard deviation is utilised to calculate

Coefficient of Variance of each independent variable. Ranking is done on the basis of CV.

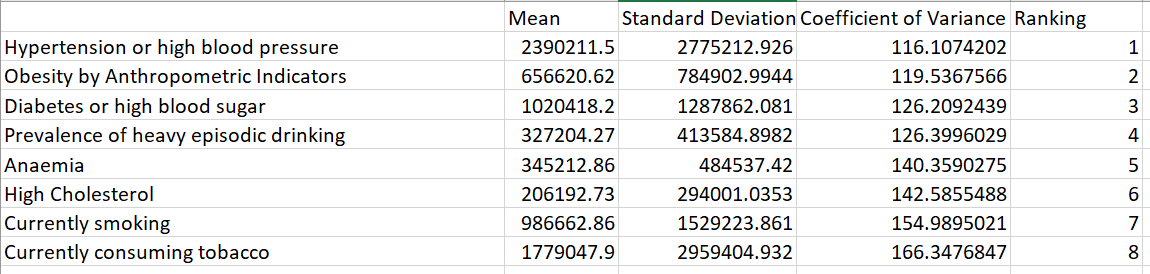


Table2: Ranking variables on the basis of CV

The top 5 variables are considered for regression analysis, namely, number of cases of High BP, Obesity, Diabetes, Alcohol Consumption and existence of Anaemia. The summary output of regression analysis includes regression statistics, ANOVA, residual output and probability output.

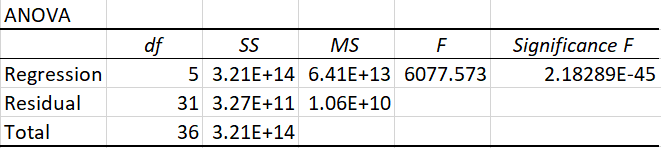


Table3: ANOVA Table-1

Similarly, Proportional Ranking has been calculated for each independent variable.



Table4: Proportional Ranking

The top 5 variables are considered for regression analysis, namely, number of cases of High cholesterol, Alcohol Consumption, existence of Anaemia, Obesity and Smoking.

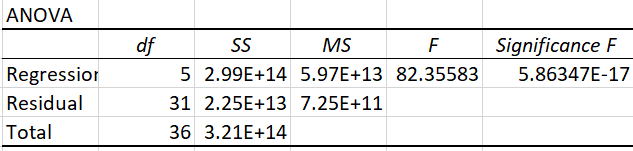
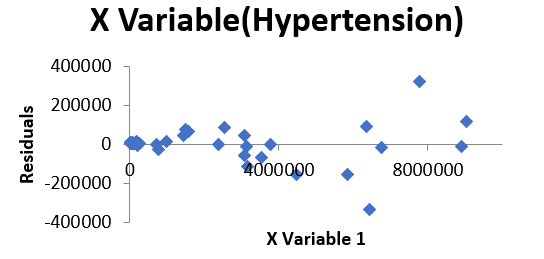
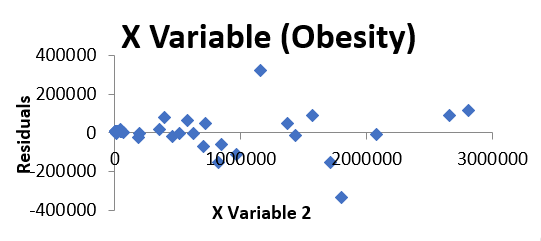
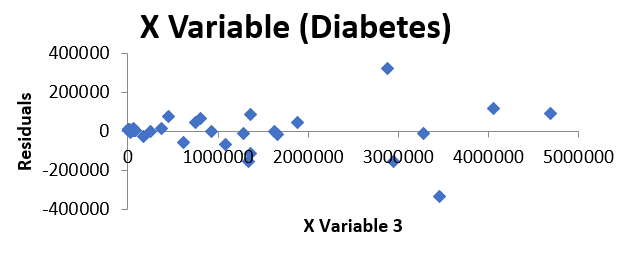
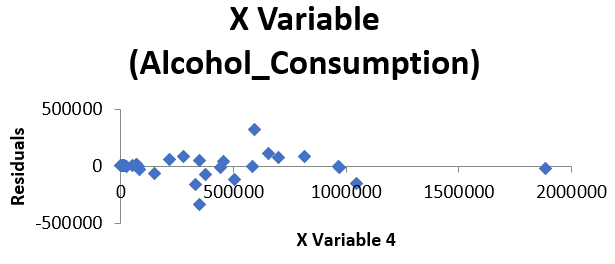
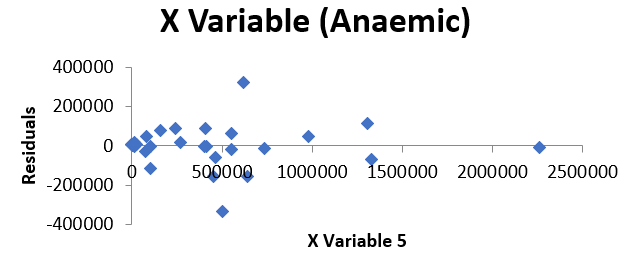
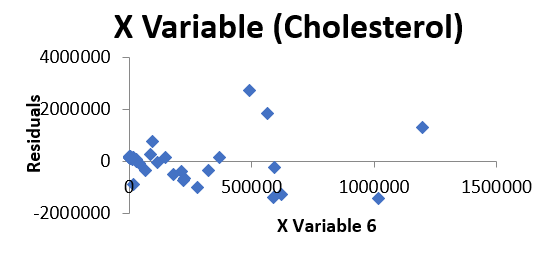


Table5: ANOVA Table-2

Regression Plots of important variables are as follows:

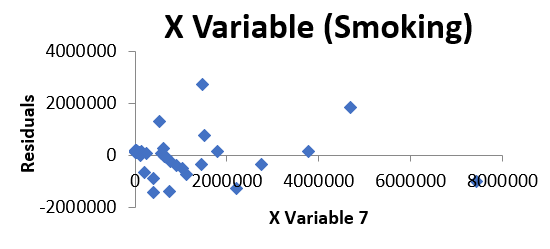
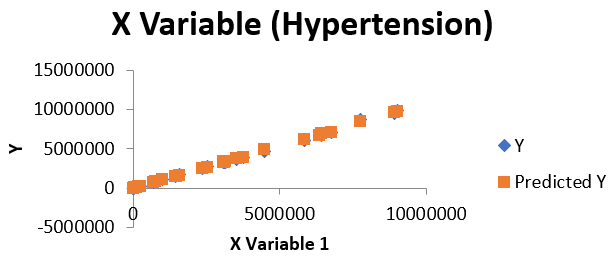
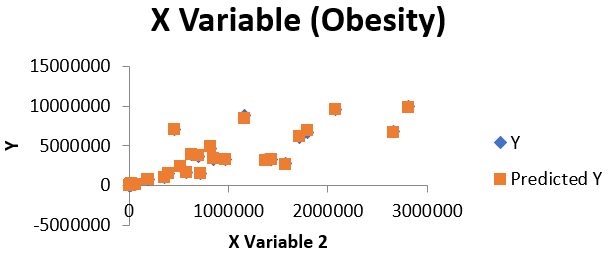
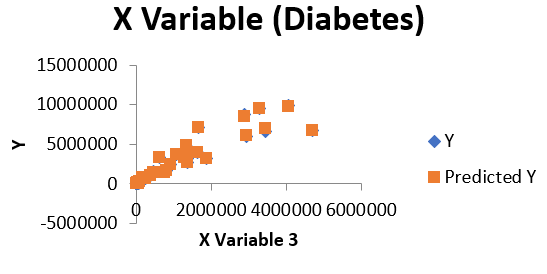
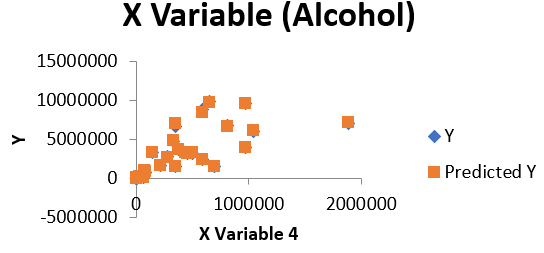
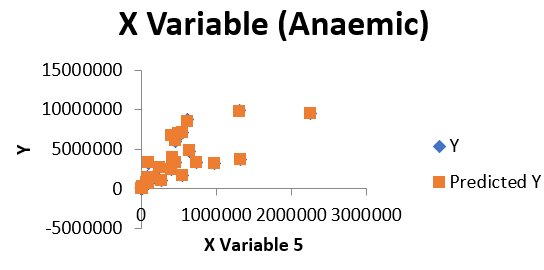
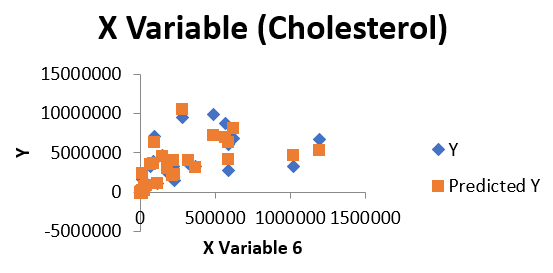


Fig1: Residual Plots

There is no pattern followed in the above error plots, which implies that 3rd assumption is valid.

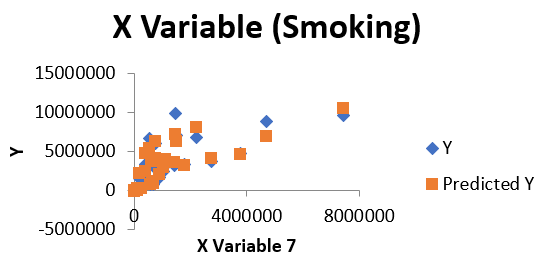


Fig2: Line Fit Plots

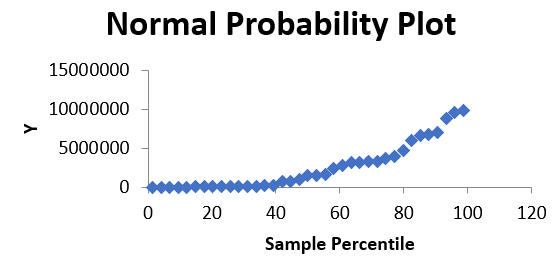


Fig3: Normal Probability plot

Following table highlights the Comparison between models based on CV and Proportional ranking

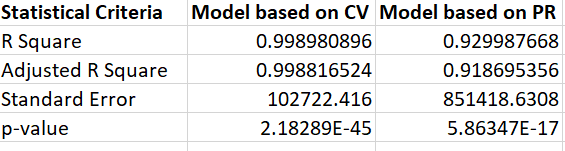


Table6: Comparative Study of models based on Consistency ranking and that based on Proportional ranking.

SCAN STATISTICS:

The excel data has been used as input into SatScan application. Case file includes the number of cases of independent variables, taken separately, coordinate file contains the latitude and longitude, and lastly, population file is also imported which contains the number of cardiovascular cases of every state and union territory. Discrete Poisson based model is used to identify the hotspots.

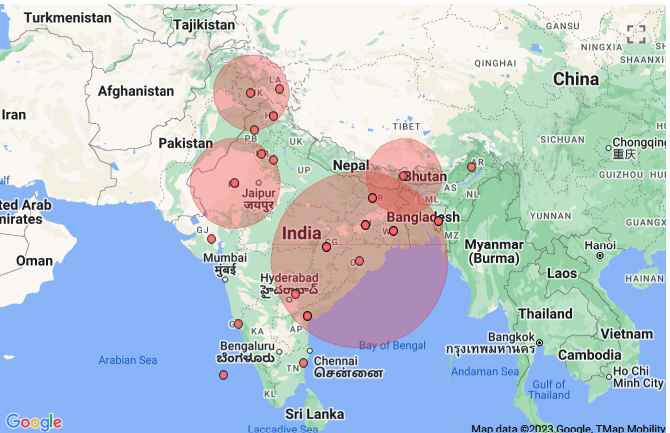


Fig4: Clusters on the basis on Cardiovascular cases

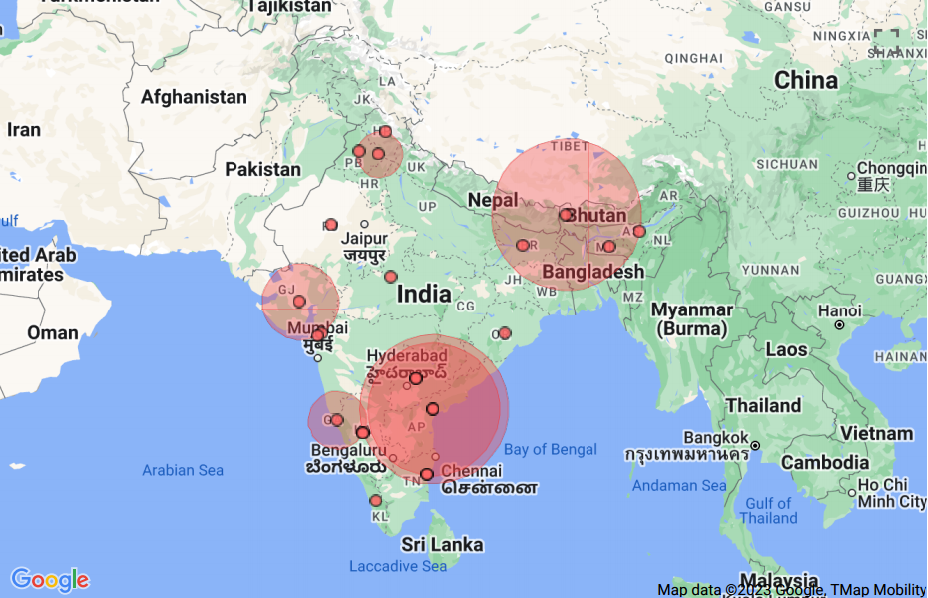


Fig5: SatScan image considering Hypertension as a factor.

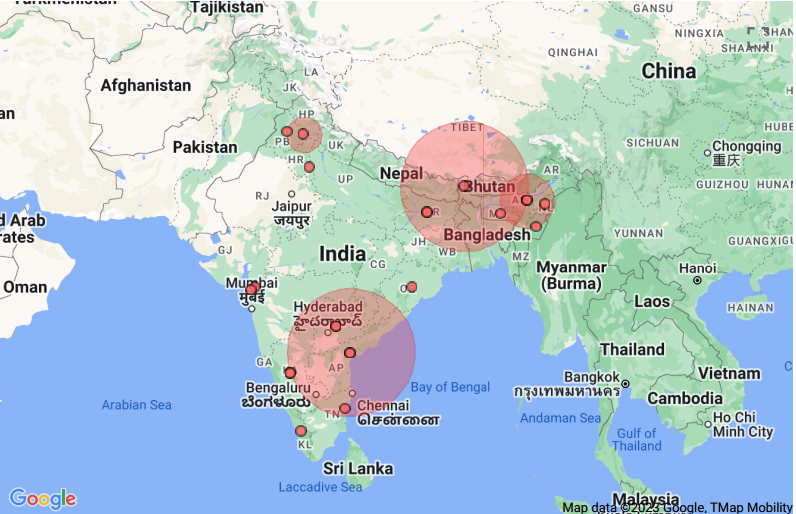


Fig6: SatScan image considering Consumption of Alcohol as a factor

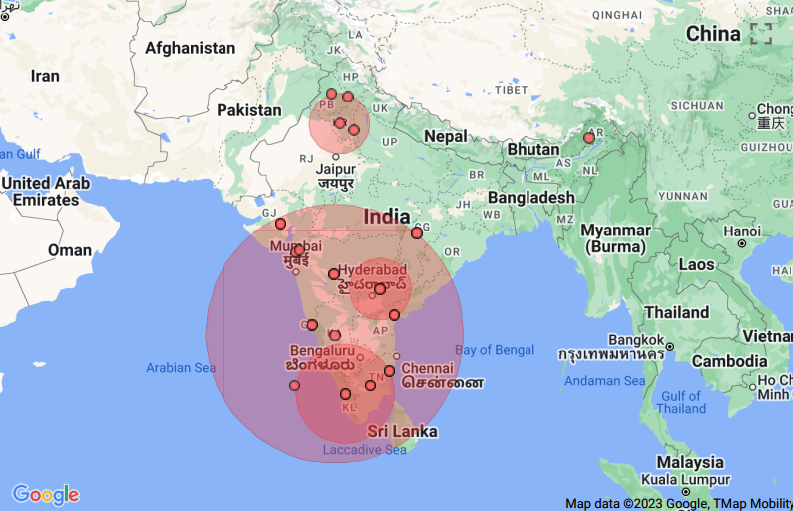


Fig7: SatScan image considering Obesity as a factor

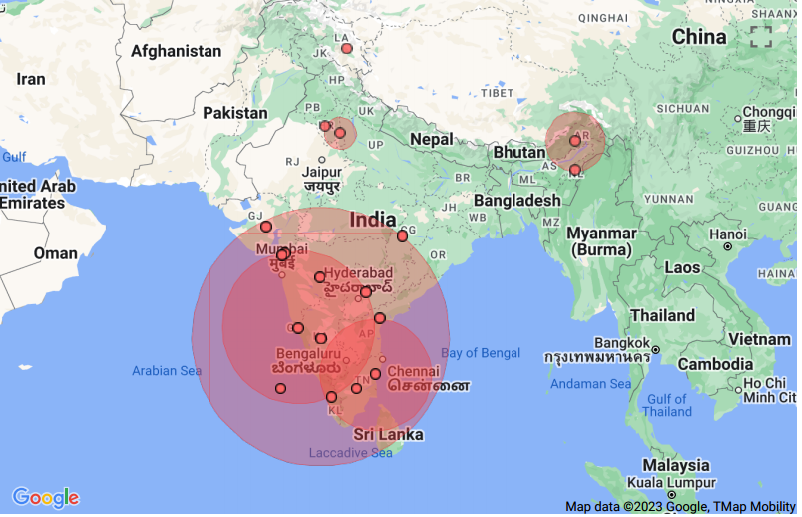


Fig8: SatScan image considering Diabetes as a factor



Fig9: SatScan image considering Anaemia as a factor

IV. RESULT AND DISCUSSION

The main objective of this research is to evaluate the statistical significance of disease cluster alarms and find the hotspots. Here Discrete Poisson based model is used to identify the hotspots.

After the completion of the calculations, a standard text-based results file is automatically generated, which contains information about the clusters detected, summary information about the data, computing time and the analysis parameters chosen. Prominent clusters on the basis of cardiovascular cases are as follows: Primary-Gujarat, Odisha, Chhattisgarh, Jharkhand, West Bengal, Bihar, Telangana, Andhra Pradesh, Tripura. Secondary- Jammu & Kashmir, Ladakh, Himachal Pradesh, Punjab. On the basis of Hypertension- Andhra Pradesh, Telangana, Puducherry, Karnataka. On the basis of Alcohol Consumption- Sikkim, Bihar, Meghalaya, Assam. On the basis of Obesity- Karnataka, Goa, Kerala, Maharashtra, Tamil Nadu, Andhra Pradesh, Lakshadweep, Telangana, Puducherry, Daman & Diu, Dadra Nagar Haveli, Gujarat, Chhattisgarh. On the basis of Diabetes- Karnataka, Goa, Kerala, Maharashtra, Tamil Nadu, Andhra Pradesh, Lakshadweep, Telangana, Puducherry, Daman & Diu, Dadra & Nagar Haveli, Gujarat, Chhattisgarh. On the basis of Anaemia- Delhi, Haryana, Chandigarh, Uttarakhand, Punjab, Himachal Pradesh, Rajasthan, Uttar Pradesh, Madhya Pradesh, Jammu &Kashmir.

The multiple linear regression equation considering top 5 variables based on CV:  
y = -6990.78+ 1.130775+( -0.02882)+( -0.037)+ (-0.23315)+ (-0.0423)(2)

The multiple linear regression equation considering top 5 variables based on PR:  
y = -155527+ 1.438701+ 2.478916+(-0.05643)+ 1.444067+ 0.676285 (3)

From the normal probability plot (Fig3) we can see that it has a sharp change in the direction of the trend in an upward direction from the mid. This indicates that the underlying distribution is positively skewed.

MODEL VALIDATION:

Hypothesis testing

1.

2.Atleast one variable is significant

Let

Test statistic:

Suppose we apply p-value criteria of the models. In both the cases p-value is less than (Table5,6). Hence, we reject

Similarly, suppose we apply F-statistic criteria.

for both the models respectively (Table5,6). Since , we reject our null hypothesis .

This implies that both our models are valid.

Comparing the 2 models (Table5), model based on coefficient of variance gives better results as R square value i.e. 0.998 is higher than that of the counterpart. Even the standard error is lesser in comparison.

V. CONCLUSION

This project involves examining cardiovascular data from 37 states and union territories. The dataset's descriptive analysis reveals a predominantly positive skew. Regression analysis indicates a strong fit with 99% accuracy when considering major factors such as Hypertension, Diabetes, Obesity and Anaemia. It has yielded a multiple linear regression equation (eq2). Hotspots, identified through SatScan, majorly include regions like Gujarat, Andhra Pradesh, Telangana, Bihar, Karnataka, Maharashtra, Kerala, Chhattisgarh, Punjab and Tamil Nadu. These analyses can be used to design a model for proper resource allocation in these critical regions and prevent increase in cardiovascular diseases.

VI. REFERENCES

[1] [Kumar, Vinod](javascript:__doLinkPostBack('','ss~~AR%20%22Kumar%2C%20Vinod%22%7C%7Csl~~rl','');), [Lalotra, Gotam Singh](javascript:__doLinkPostBack('','ss~~AR%20%22Lalotra%2C%20Gotam%20Singh%22%7C%7Csl~~rl','');). “*Predictive Model Based on Supervised Machine Learning for Heart Disease Diagnosis”.* 2021 IEEE International Conference on. :1-6 Dec, 2021

[2] [Fiaidhi, J.](javascript:__doLinkPostBack('','ss~~AR%20%22Fiaidhi%2C%20J.%22%7C%7Csl~~rl','');), [Mohammed S.](javascript:__doLinkPostBack('','ss~~AR%20%22Mohammed%2C%20S.%22%7C%7Csl~~rl','');) *“Prognosis analysis of thick data: Clustering heart diseases risk groups case study”.* Department of Computer Science, Lakehead University, Ontario, Canada: June 2021

[3] [Lohaj, Oliver](javascript:__doLinkPostBack('','ss~~AR%20%22Lohaj%2C%20Oliver%22%7C%7Csl~~rl','');), [Pella, Zuzana](javascript:__doLinkPostBack('','ss~~AR%20%22Pella%2C%20Zuzana%22%7C%7Csl~~rl','');), [Paralic, Jan](javascript:__doLinkPostBack('','ss~~AR%20%22Paralic%2C%20Jan%22%7C%7Csl~~rl','');). *“Data analytics methods for analyzing the impact of factors on early detection of cardiovascular risk”*. 2022 IEEE 20th Jubilee World Symposium (SAMI): Mar, 2022

[4] [Rustamov, Zahiriddin](javascript:__doLinkPostBack('','ss~~AR%20%22Rustamov%2C%20Zahiriddin%22%7C%7Csl~~rl','');), [Rustamov, Jaloliddin](javascript:__doLinkPostBack('','ss~~AR%20%22Rustamov%2C%20Jaloliddin%22%7C%7Csl~~rl','');), [Sultana, Most Sarmin](javascript:__doLinkPostBack('','ss~~AR%20%22Sultana%2C%20Most%20Sarmin%22%7C%7Csl~~rl','');), [Ywei, Jeanne](javascript:__doLinkPostBack('','ss~~AR%20%22Ywei%2C%20Jeanne%22%7C%7Csl~~rl','');), [Balakrishnan, Vimala](javascript:__doLinkPostBack('','ss~~AR%20%22Balakrishnan%2C%20Vimala%22%7C%7Csl~~rl','');), [Zaki, Nazar](javascript:__doLinkPostBack('','ss~~AR%20%22Zaki%2C%20Nazar%22%7C%7Csl~~rl','');). *“Cardiovascular Disease Prediction using Ensemble Learning Techniques: A Stacking Approach”*. 2023 19th IEEE International Colloquiu m(CSPA) : Mar, 2023