# Predictive Modeling and Spatial Hotspot Detection of Cardiovascular Diseases Using Machine Learning and Statistical Methods

*Submitted in partial fulfillment of the requirements for the degree of*

## Master of Science
In
## Data Science

*by*

**Mahajan Hemant Pandharinath**
**24MDT0203**

**Under the guidance of**

**Dr. Jitendra Kumar**
**School of Advanced Sciences**
**VIT - Vellore**


VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

November, 2025

# DECLARATION

I hereby declare that the thesis entitled "**Predictive Modeling and Spatial Hotspot Detection Using Machine Learning and Statistical Methods**" submitted by me, for the award of the degree of *Master of Science in Data Science* to VIT is a record of bonafide work carried out by me under the supervision of **Dr. Jitendra Jumar.**

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

**Place : Vellore**
**Date :** **(Signature of the Candidate)**

**Mahajan Hemant Pandharinath**

# CERTIFICATE

This is to certify that the thesis entitled "**Predictive Modeling and Spatial Hotspot Detection Using Machine Learning and Statistical Methods**" submitted by **Mahajan Hemant Pandharinath (Reg. No.: 24MDT0203)**, **School of Advanced Sciences**, VIT, for the award of the degree of *Master of Science in Data Science,* is a record of bonafide work carried out by him under my supervision during the period, **09.07.2025** to **14.11.2025**, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore

Date :

Signature of the Internal Guide

**Department of Mathematics**
**SAS, VIT- Vellore**

**Signature of the Internal Examiner**

**Dr. KHADARBABU S K**
**Head, Department of Mathematics**
**SAS, VIT-Vellore**

# ACKNOWLEDGEMENT

With immense pleasure and deep sense of gratitude, I wish to express my sincere thanks to my guide **Dr. Jitendra Kumar**, School of Advanced Sciences, VIT, Vellore without his motivation and continuous encouragement, this research would not have been successfully completed.

I am grateful to the Chancellor of VIT, Vellore, Dr. G. Viswanathan, the Vice Presidents and the Vice Chancellor for motivating me to carry out research in the Vellore Institute of Technology, Vellore and also for providing me with infrastructural facilities and many other resources needed for my research.

I express my sincere thanks to Dr. Karthikeyan K, Dean, School of Advanced Sciences, VIT, Vellore, for his kind words of support and encouragement. I like to acknowledge the support rendered by my classmates in several ways throughout my research work.

I wish to thank Dr. KHADARBABU S K**,** Head of the Department of Mathematics, School of Advanced Sciences, VIT, Vellore for his encouragement and support.

I wish to extend my profound sense of gratitude to my parents and friends for all the support they made during my research and also providing me with encouragement whenever required.

**Signature of the Student**

**Mahajan Hemant Pandharinath**

# ABSTRACT

Cardiovascular diseases (CVDs) remain one of the primary causes of mortality in India and it is highly essential to ensure the timely identification of risks and monitor the disease in the area. The current project proposes integrated solution, which involves machine learning-based individual risk prediction and state-scale spatial analysis to gain meaningful information on the trend of CVD in India.To construct predictive models including Logistic Regression and XGBoost, a 70,000 patient record dataset was utilized and then probability calibration was carried out to enhance reliability. The balanced precision, recall, and overall accuracy of the XGBoost model were optimized, and its calibrated model performed the best (with AUC 0.799). To ensure the transparency and interpretability of the results, SHAP explainability methods were used to learn the effect of such features as systolic blood pressure, age, cholesterol, BMI and pulse pressure on risk predictions both at a global and individual level. In conjunction with predictive modelling, the state-wise CVDs count (2018-2022) were evaluated with the help of the statistical measures (mean, standard deviation, CV, IQR), allowing the identification of the stable and volatile states. The SaTScan was used to conduct spatial clustering and important high-risk and low-risk areas were identified in India. CV-based and IQR-based rankings also indicated the most stable and most changeable states, which also allows further epidemiological understanding.The unified findings indicate that machine learning and spatial epidemiology can be mutually enhancing as they assist in the early detection on a personal scale and assist policymakers to track and act on regional disease patterns. The paper concludes with the significance of data-driven solutions to the enhancement of healthcare planning, resource distribution, and cardiovascular disease prevention in India.

# CONTENTS


| | | Page No. |
|---|---|---|
| | **Acknowledgement** | 4 |
| | **Abstract** | 5 |
| | **Table of Contents** | 6 |
| | **List of Figures** | 7 |
| | **List of Tables** | 8 |
| | **Abbreviations** | 9 |
| 1 | **INTRODUCTION** | 10 |
| | 1.1 Objective | 10 |
| | 1.2 Motivation | 10 |
| | 1.3 Background | 11,12 |
| 2 | **PROJECT DESCRIPTION AND GOALS** | 12-14 |
| 3 | **TECHNICAL SPECIFICATION** | 14-15 |
| 4 | **DESIGN APPROACH AND DETAILS (as applicable)** | 15 |
| | 4.1 Materials, Approach And Methods | 15-17 |
| | 4.2 Codes and Standards | 17-20 |
| 5 | **SCHEDULE, TASKS AND MILESTONES** | 21 |
| 6 | **PROJECT OUTPUTS** | 22-41 |
| 7 | **RESULT & DISCUSSION (as applicable)** | 42-46 |
| 8 | **LIMITATIONS** | 47 |
| 9 | **CONCLUSION** | 48-49 |
| 10 | **REFERENCES** | 50-52 |

## List of Figures


| Figure No. | Title | Page No. |
|---|---|---|
| 1. | ROC Curve – Logistic Regression | 22 |
| 2. | ROC Curve – XGBoost | 23 |
| 3. | ROC Curve – Calibrated XGBoost | 24 |
| 4. | Confusion Matrix – Calibrated XGB | 25 |
| 5. | Calibration Curve | 26 |
| 6. | Global SHAP Beeswarm Plot | 27 |
| 7. | Local SHAP Waterfall Chart (Example 1) | 27 |
| 8. | Local SHAP Waterfall Chart (Example 2) | 28 |
| 9. | CVD Clusters — 2018 (2% Risk) | 32 |
| 10 | CVD Clusters — 2018 (5% Risk) | 33 |
| 11 | CVD Clusters — 2019 (2% Risk) | 34 |
| 12 | CVD Clusters — 2019 (5% Risk) | 35 |
| 13 | CVD Clusters — 2020 (2% Risk) | 36 |
| 14 | CVD Clusters — 2020 (5% Risk) | 37 |
| 15 | CVD Clusters — 2021 (2% Risk) | 38 |
| 16 | CVD Clusters — 2021 (5% Risk) | 39 |
| 17 | CVD Clusters — 2022 (2% Risk) | 40 |
| 18 | CVD Clusters — 2022 (5% Risk) | 41 |

# List of tables


| Table No. | Title | Page No. |
|---|---|---|
| 1. | Schedule, Tasks and Milestones | 21 |
| 2. | Accuracy, Precision, Recall, F1-Score for All Models | 25 |
| 3. | Subgroup Performance Analysis | 29 |
| 4. | Top 5 States by CV Ranking | 30 |
| 5 | Bottom 5 States by CV Ranking | 30 |
| 6 | Top 5 States by IQR Ranking | 30 |
| 7 | Bottom 5 States by IQR Ranking | 30 |

# List of Abbreviations

| | |
|---|---|
| **ML** | Machine Learning |
| **CVD** | Cardiovascular Disease |
| **ML** | Machine Learning |
| **ROC** | Receiver Operating Characteristic |
| **AUC** | Area Under Curve |
| **BP** | Blood Pressure |
| **BMI** | Body Mass Index |
| **SHAP** | Shapley Additive Explanations |
| **IQR** | Interquartile Range |
| **TP** | True Postive |
| **TN** | True Negative |
| **FP** | False Positive |
| **FN** | False Negative |

# 1. INTRODUCTION

## 1.1. OBJECTIVE

The main aim of this project is to create a unified system that will be able to facilitate the early identification and monitoring of cardiovascular disease, both done by machine learning and statistical analysis. At the personal level, the goal is to create predictive models with a large-scale cardiovascular dataset of almost 70,000 records, with the most impactful risk factors, and to enhance the consistency of predictions with probability calibration and explainable models like SHAP. In addition to the prediction at the individual level, the project is also aimed at gaining insights into the population-level trends based on five years of state-by-state CVD data in India. The consistency between states is measured by statistical methods like coefficient of variation and inter-quartile range, and the high-risk geographic hotspots are identified with the help of SaTScan. The combination of prediction, interpretation, and spatial insights aims at not only attaining accurate risk classification but also meaningful analytical outputs that could be of value to the decision-making and resource planning of the population in relation to health.

## 1.2. MOTIVATION

Cardiovascular disease has remained among the top causes of mortality in India but a significant number of the cases have yet to be detected until complications set in. This difficulty is even more important in areas where medical accessibility and consciousness is uneven. The rationale of this work is the necessity of methods that extend beyond standard reporting and provide not only early-level risk estimates but also more general epidemiological insights. Machine learning offers the chance to examine large medical datasets and observe subtle risk patterns that might be overlooked with the help of manual evaluation, whereas explainable models can be useful to ensure that predictions become transparent and medically meaningful. Simultaneously, the analysis of state-level CVD trends across the years may show the ways of risks evolution and lead to the recognition of regions that might need special measures. The motivation towards the merging of these two views is the idea that insight that is enhanced by technology when added to the challenges of statistical reasoning can help to result in more informed decisions and eventually assist in contributing to the process of alleviating the burden of cardiovascular disease.

## 1.3. BACKGROUND

One of the most important issues in the 21 st century in terms of public health is cardiovascular diseases (CVDs). It is shown that in world health reports, disorders associated with the heart cause almost a third of all deaths worldwide, and thus they are the main cause of mortality both in developed and developing countries. Although the situation has improved gradually in countries with well-developed healthcare systems through early screening and preventive interventions, numerous low- and middle-income areas are still witnessing an increase in the prevalence of CVD. This has been particularly evident in India where infectious diseases are currently being replaced by lifestyle-related chronic illnesses in the recent decades. Urbanization, sedentary lifestyle, evolving food habits, prolonged lifespan and lack of access to frequent healthcare services have helped to make heart disease to rapidly increase among both rural and urban communities. What is more threatening in the situation is that CVD can proceed unnoticed and is usually detected only when a significant cardiac event takes place, and indicating the necessity of the early detection of the risks.

Historically, clinical scoring schemes like Framingham and WHO charts have been used to assess cardiovascular risks and these were constructed using small groups of the population and mostly make assumptions of the same behavior of risk factors in all groups of demographics. Nonetheless, generalized risk estimates are less dependent on the Indian population, as they are highly diverse in genetic composition, lifestyle exposure, and environmental factors. The development of data collection within the past few years has resulted in the provision of large-scale health data sets that have the ability to capture patient diverse characteristics, such as blood pressure, glucose levels, cholesterol, physical activity and behavioral characteristics, such as smoking and alcohol consumption. With this richness of data, it becomes possible to apply data-driven methods that can learn the non-linear and complex patterns that are not apparent in the statistical modeling approach.

In healthcare, machine learning, specifically, has demonstrated an encouraging future due to its ability to recognize patterns, predict risks, and detect diseases at a very early stage based on the history of the patients. Though positive progress has been made, there are still a number of challenges. A lot of predictive models tend to be black boxes, which give results but do not give reasons as to why a specific prediction should have been made. In a medical context particularly dealing with life-impacting choices, interpretability is as critical as accuracy.

Policymakers and clinicians require not just predictions, but also understanding of the attributes that add the most to the risk, and the difference between the patterns in different groups of patients. Methods like SHAP (Shapley Additive Explanations) allow filling this gap by providing clear-cut justifications to decisions made by a model to guarantee that a data-driven prediction is credible and clinically significant. The next weakness in most studies is the lack of a broader perspective on individual-level prediction beyond relating it to large-scale trends in public health. Cardiovascular disease is not the only medical outcome but an event of the population that is dependent on the geography, social-economic statuses, and access to healthcare in the region.

Knowledge of the disease burden focus may inform resource allocation, screening, and preventive health measures. The current piece of work consolidates two opposing points of view to fill in these gaps. On the personal level, machine learning algorithms are trained on a huge cardiovascular dataset of approximately 70,000 records to anticipate the risk of disease and the parameters of health that can affect it. The reliability of the predictions is improved with the help of probability calibration, and the transparency of model behavior is provided with explainability. On the population level, five-year state-wise CVD data of India (2018-2022) is examined with the help of statistical indicators like coefficient of variation and inter-quartile range to determine the temporal consistency. Spatial clustering is applied with SaTScan to identify geographic hotspots of high risk and give an extra level of understanding compared to the model-based prediction. Through a combination of predictive analytics with spatial and statistical exploration, this research intends to create a more detailed description of cardiovascular disease-not only to detect it among individuals in the early stages but also to plan it effectively at the level of a community health facility.

## 2. PROJECT DESCRIPTION

Cardiovascular disease is considered to be one of the most severe health issues in India and the timely detection can significantly minimize the complexity and avoid damage in the future. This project aims at coming up with an all round analytical model that will look at cardiovascular risk at the individual level and the population level. On the personal level, a big clinical dataset, consisting of nearly 70,000 records of patients, is utilized to create machine learning models that would forecast the existence of cardiovascular disease. The data set comprises key health and lifestyle factors like age, gender, blood pressure, cholesterol, glucose

levels, smoking behavior, alcohol use and physical activity. Once the data is cleaned and its features are engineered, such as the derivation of other indicators, such as BMI and pulse pressure, two predictive models, Logistic Regression and XGBoost models are trained and assessed. The reliability of the predicted risk scores is enhanced with probability calibration, and SHAP explainability is applied to explain which features are used the most by the model. Subgroup analysis is also conducted to evaluate the level of performing the model on the different demographic groups including younger and older individuals and male and female patients.

In order to comprehend how the burden of cardiovascular disease varies throughout India over time, a separate state-wise CVD dataset from 2018 to 2022 is examined in the study's second section. To evaluate consistency and agreement between ranking methods, statistical measures such as coefficient of variation (CV), inter-quartile range (IQR), and Spearman rank correlation are first calculated in Microsoft Excel. The top five most consistent and least consistent states are then determined by importing the processed results into Python. SaTScan software is used for both ranking analysis and spatial cluster detection, which aids in locating areas with noticeably higher disease concentrations at both the 2% and 5% risk. This project offers a more thorough understanding of cardiovascular disease, from identifying individual-level risk to mapping population-level hotspots, by fusing predictive modeling with statistical and spatial insights.

## PROJECT GOALS

i. To carry out the training of different machine learning models and later on evaluate them in terms of efficiency in predicting cardiovascular diseases with the help of a large scale structured dataset.

ii. To guarantee the predicted probabilities are more reliable by means calibration.

iii. To understand and communicate the model behavior with the help of SHAP both at the global and individual levels.

iv. To measure fairness and consistency through the analysis of performance of the subgroups based on age and gender.

v. To study five year state wise CVD data using CV and IQR ranking that were computed in Excel.

vi. To compute Spearman rank correlation for statistical measures to compare the ranking of agreement.

vii. To find high risk spatial clusters and to create their visuals by means of SaTScan.

viii. To combine the predictive and spatial results for the early detection support as well as for the public, health decision planning..

## 3. TECHNICAL SPECIFICATIONS

Jupyter Notebook (Anaconda3 environment) was adopted to execute this project and it was a useful interactive environment to perform data preprocessing, model training, evaluation and visualization. The main code was implemented in Python with the combination of scientific computing and machine learning libraries.

Numpy was applied widely in numerical operations such as derivation of clinical variables like BMI and pulse pressure. Pandas was used in the data processing- loading in the cardiovascular dataset, carrying out preprocessing tasks, creating engineered features, working on abnormal values, and organizing information into structured DataFrames to be used in machine learning.

To create plots, Matplotlib and Seaborn were utilized to draw ROC-AUC curves, calibration curves, and SHAP-based interpretability plots. These visualizations aided the exploratory analysis and interpretation of performance of models. Besides this, confusion matrices and other diagnostic plots were made, in order to assess model behavior.

The scikit-learn library was used to perform machine learning, and it has train-test splitting, scaling with StandardScaler, logistic regression model, evaluation measures (accuracy, precision, recall, F1-score), isotonic probability calibration, and ROC-AUC modules. XGBoost classifier was applied using the xgboost library, which provided a more advanced method of modeling which could be used to elicit non-linear associations in clinical risk factors.

The SHAP (Shapley Additive Explanations) library was used to obtain model interpretability and give a more detailed insight into the contribution of the features at the population level and with a single patient prediction.

To calculate the coefficient of variation (CV), inter-quartile range (IQR), and Spearman rank correlation as part of the statistical aspect of the study, Microsoft Excel was employed to work with a five year state-level CVD dataset (2018-2022). These findings were then imported into Python to rank, filter and visualize consistent and inconsistent states.

SaTScan software was employed in order to determine geographical hotspots of cardiovascular disease throughout India. The scan statistics method used by SaTScan assisted in identifying the important spatial clusters at 2% and 5% levels of risk providing important epidemiological information compared to numerical modeling.

This Python, Excel and SaTScan mix constituted a strong technical base of predictive modeling, statistical analysis and epidemiology of space.

# 4. DESIGN APPROACH AND DETAILS

## 4.1. MATERIALS, APPROACH AND METHODS

This project design is systematic and multi-stage, combining machine learning-based predictive and statistical and spatial analysis to have an overall perspective of the trends in cardiovascular diseases.

- **Preparation and Feature Engineering of Data**

  Pandas was used to load the cardiovascular dataset of around 70,000 patient records to begin with. Primary data cleaning involved the removal of values that were physiologically impossible when it came to height, weight, and blood pressure and transformation of age to a number of years to make it interpretable. Significant derived variables were obtained through NumPy computations- namely Body Mass Index (BMI) and pulse pressure of which the two have high clinical significance. It was also used to create a blood pressure category (bp_cat) variable which is used to categorize patients according to normal, elevated, or high BP.

- **Training and Calibration of a Model.**

The dataset after cleaning was divided into training and testing data, and continuous variables were standardized by StandardScaler to make the models sensitive to scale, including the Logistic Regression. Two predictor models were created:

1. Logistic Regression - can be used as an easy, interpretable baseline.
2. XGBoost Classifier - an indicator of sophisticated, non-linear correlations of clinical variables.

After training, the probability outputs of the XGBoost model were calibrated with isotonic regression with CalibratedClassifierCV of the scikit-learn package. The calibration aims at bringing the predictions of the probabilities closer to the real risk probability, which is critical in the medical context.

Accuracy, precision, recall, F1-score, and AUC-ROC were some of the metrics that were used in model evaluation. Separate ROC curves of Logistic Regression, XGBoost and Calibrated XGBoost were created. Probability predictions were also evaluated by a calibration plot and Brier score evaluation.

- **SHAP Explainability of the Model**

  In order to prevent black-box behavior and maintain transparency the SHAP library was run on the trained XGBoost model. SHAP values generated the specifics on the role of each clinical feature in the prediction outcome. Overall, the most influential features were summarized using global SHAP, whereas local SHAP plots were used to provide the rationale behind the individual patient-level predictions.

- **Subgroup Analysis**

  The subgroup analysis was done on the basis of demographic filters in order to achieve fairness and evaluate the model robustness in various parts of the population. Measures were calculated individually for:

  - individuals aged below 45
  - individuals aged 45 and above
  - male patients
  - female patients

These measures were tabulated allowing to compare the performance of the model in different

groups.

- **Statistics Analysis of CVD State-Wise.**

  Another dataset that included annual cardiovascular cases by state in India (2018-2022) was examined to be able to understand the long-term patterns. The following numbers were calculated using Microsoft Excel:

- Coefficient of Variation (CV)- to reflect the consistency or volatility of five years.
- Inter-Quartile Range (IQR)- to investigate variability in case distributions.
- Spearman Rank Correlation - to estimate conformity between CV-based and IQR-based rankings.

  This processed data was then imported to Python in order to pick the top 5 and bottom 5 states in both ranking systems.

- **SaTScan in the Detection of Spatial Clusters.**

Spatial cluster detection was done using SaTScan in order to supplement the statistical analysis with geographic insight. The applied software used scan statistics to determine the areas of concentration of CVD cases with significantly high percentiles of less than 2 percent and 5 percent of risk. The cluster maps that resulted indicated the regions that were at risk and those that were emerging hotspots which gave a spatial perspective of cardiovascular burden in India.

- **Integrated Analysis**

  The final step synthesized the following:

  - Individual prediction based on machine learning.
  - SHAP-based interpretability
  - Subgroup analysis
  - Statistical five year state assessments.
  - SaTScan spatial cluster analysis.

Such a combined model does not only serve to predict the risk of diseases in a specific person, but also provides a more detailed and significant picture of cardiovascular disease by identifying locations of the geographic area where a specific health intervention is necessary.

### 4.2. CODES AND STANDARDS:

In the present project, machine learning methods, statistical processes and spatial

epidemiology software were utilized to examine patterns of cardiovascular disease (CVD). In order to maintain the methodological consistency, ethical responsibility, and scientific rigor, every step of analysis was based on a clearly defined set of codes, guidelines, and standards that are generally accepted in the field of data science and public-health modelling. The key standards that are followed in this work are listed below:

1. **Data Processing and Ethics.**

   In spite of the fact that the datasets (CVD dataset of 70,000 records and state-wise mortality data) utilized are open and anonymized, the analysis was performed based on several basic ethical principles:

   - There was no effort to distinguish people.
   - Only summarized insight (feature importance, subgroup analysis, cluster positions) was presented.
   - sensitive attributes like gender were converted to fit analytically without changing the semantic meaning.

   This makes it adherent to generic data-protection principles, such as those presented in GDPR and normal epidemiological ethical standards.

2. **Machine Learning Standards**

   Machine learning component of the research is based on generally accepted criteria in scientific modelling:

a. Appropriate Train-Test Split:
   - This was done using a stratified 80-20 train-test split to ensure that both sets had the same disease proportion.
   - Averts information leakage and provides realistic review.

b. Best Practices of Feature Engineering.
   - Calculated BMI, Pulse Pressure and BP Category on the clinical definitions suggested by WHO and AHA (American Heart Association).
   - The outlier removal limits were established around the medically plausible limits.

c. Model Selection Standards
   - Logistic Regression (initial linear model)
   - XGBoost (famous for structured medical dataset)

These options are based on the existing biomedical ML guidelines according to which gradient boosting models are favored on tabular health data.

d. Standards for model Calibration

- Calibration of the probability was done with the help of Isotonic Regression, which is a common method suggested in clinical decision systems to prevent over-confidence.
- On the basis of sklearn guidelines, calibration curves were created.

**3. Standards for Model Evaluation**

Measures of model performance that were suggested in medical predictive modeling were used to evaluate model performance:

- ROC-AUC (It has ability to discriminate between positive and negative classes.)
- Precision, Accuracy, Recall, F1-score were used in evaluation.
- Confusion Matrix was plotted.

SHAP was used for explainability to maintain transparency and interpretability, which is an eminent requirement in healthcare.

**4. Statistical Standards (excel based)**

To analyze CVD at the state level (2018-2022), the general statistical procedures were employed:

a. Coefficient of Variation (CV)

- Ratios of relative state variability.
- It Was used to rank the states based on the consistency of yearly cases of CVD.

**b.** Interquartile Range (IQR)

- It is a non-parametric Measure of variation.
- It helps in determining the states which show high fluctuation for disease cases.

c. Spearman Rank Correlation

- It is a non-parametric correlation measure that was applied to compare ranking based on CV and ranking based on IQR.

These approaches adhere to classical statistical procedures to measure the

variability and the trend in epidemiology.

5. **Standards for detecting clusters using SaTScan**

   The software used in this project is SaTScan, which is widely known for detecting the clusters of disease in spatial and temporal way, and this project adhered to:

   - The Spatial Scan Statistic of Kulldorf.
   - The analysis was conducted at 2 percent and 5 percent risk as recommended in SaTScan documentation.
   - Identification of clusters (2018-2022) by year through Poisson probability model, which is suitable with count-based disease-related data.

   Visualization of all the maps was done according to the instructions suggested by Datawrapper and SaTScan.

   This is so that the spatial analysis is sound, epidemiologically valid and replication ready.

6. **Standards of Reporting and Transparency.**

   To conform to good academic practice:

   - Code, transformations and assumptions are clearly defined.
   - There were no black-box decisions to be made without an explanation, SHAP describes how the model is reasoning.

# 5. SCHEDULE, TASKS AND MILESTONES

*Table 1*

| S.NO | MONTH-WEEK | PLAN |
|---|---|---|
| 1. | JULY- WEEK 1 | Identifiation of research problem. |
| 2. | JULY- WEEK 2, 3 | Done related literature review. |
| 3. | AUGUST- WEEK 1 | Discussion on the aim/targets and objectives decided for reseach. |
| 4. | AUGUST-WEEK 2 | Overview for drafting the abstract. |
| 5. | AUGUST-WEEK 3,4 | Collecting dataset for reaserch. |
| 6. | SEPTEMBER-WEEK 1,2,3 | Adaptation of methodology and exploring SaTScan Softerware. |
| 7. | SEPTEMBER-WEEK 4 | Data processing and basic statistical analysis, understanding the patterns. |
| 8. | OCTOBER- WEEK 1,2 | Build Machine Learning models, did SHAP explainability task and consulted guide. |
| 9. | OCTOBER- WEEK 3 | Cluster Analysis using SaTScan and feedback from guide. |
| 10. | OCTOBER - WEEK 4 | Final documentation and writing report. |
| 11. | NOVEMBER - WEEK 1,2 | Reviwing and refinement of report. |
| 12. | NOVEMBR - WEEK 3,4 | Report submission and final review. |

# 6. PROJECT OUTPUTS

The project has yielded analytical, statistical, machine-learning, and spatial outputs which together provide understanding of patterns of cardiovascular deseases (CVD) at both individual and popuplation level. The outcome of the the project can be divided into two major sections:

1. Outputs from predictive modeling (from CVD dataset with 70,000 records).
2. Outputs of statistical and spatial analysis of state level CVD count data.

Outputs of all the steps are summarized in below sections:

## 6.1 Outputs from predictive modeling section

### 1. Evaluating Models using performance metrics:

Total three models were trained namely-

- Logistic Regression
- XGBoost
- Calibrated XGBoost.

AUC results of these models are:

- Logistic Regression AUC: 0.786

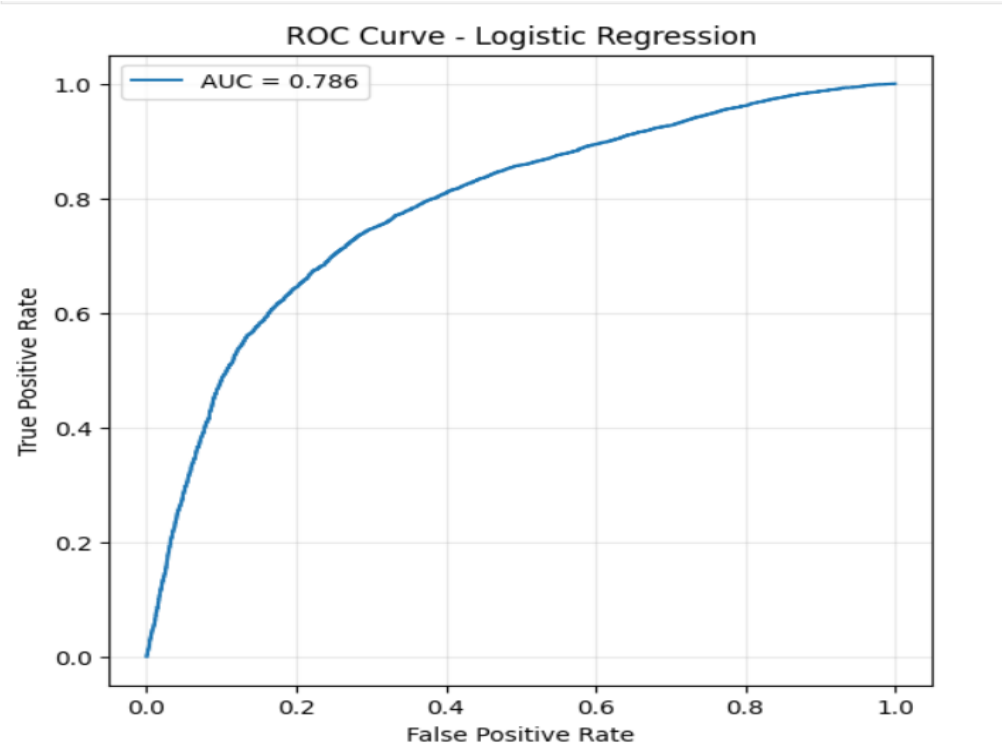


*Figure 1*

Figure 1 shows the ROC curve of Logistic Regression. The curve shows that Logistic Regression provide reasonably good class separation, with AUC value 0.786 indicating the moderate predictive performance.

- XGBoost AUC: 0.799

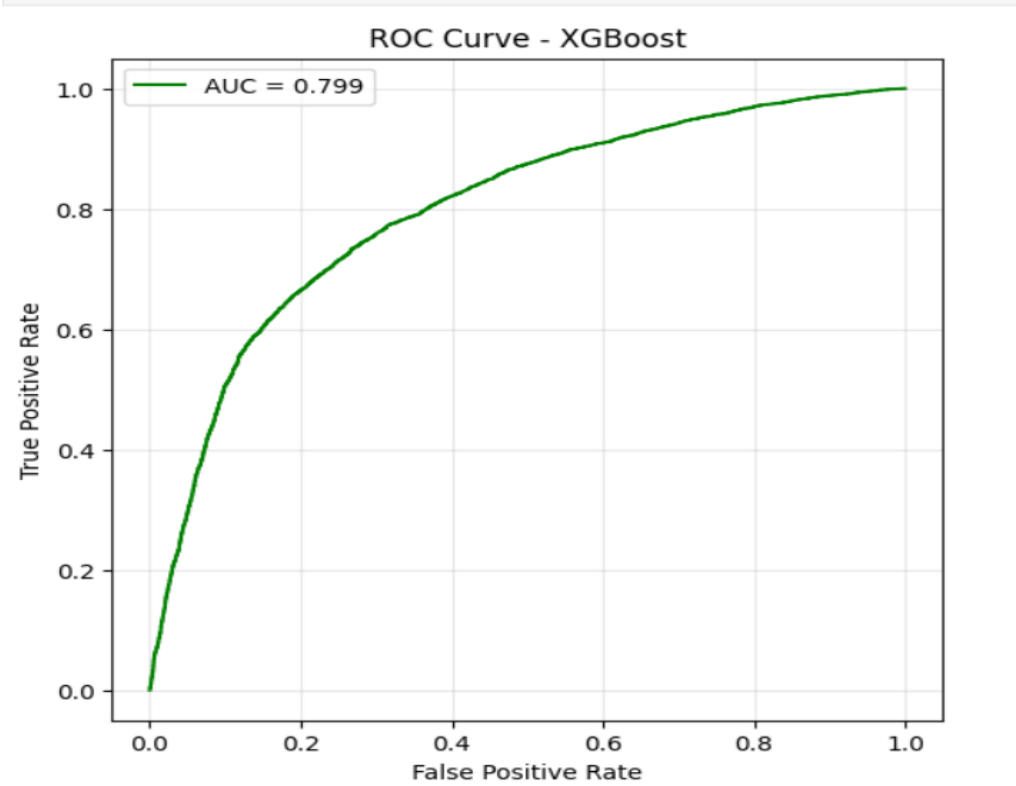


*Figure 2*

Figure 2, which shows ROC curve of XGBoost shows that it performed better than Logistic Regression with an AUC value of 0.799, that means it has stronger discrimination between positive and negative class.

- Calibrated XGBoost: 0.799

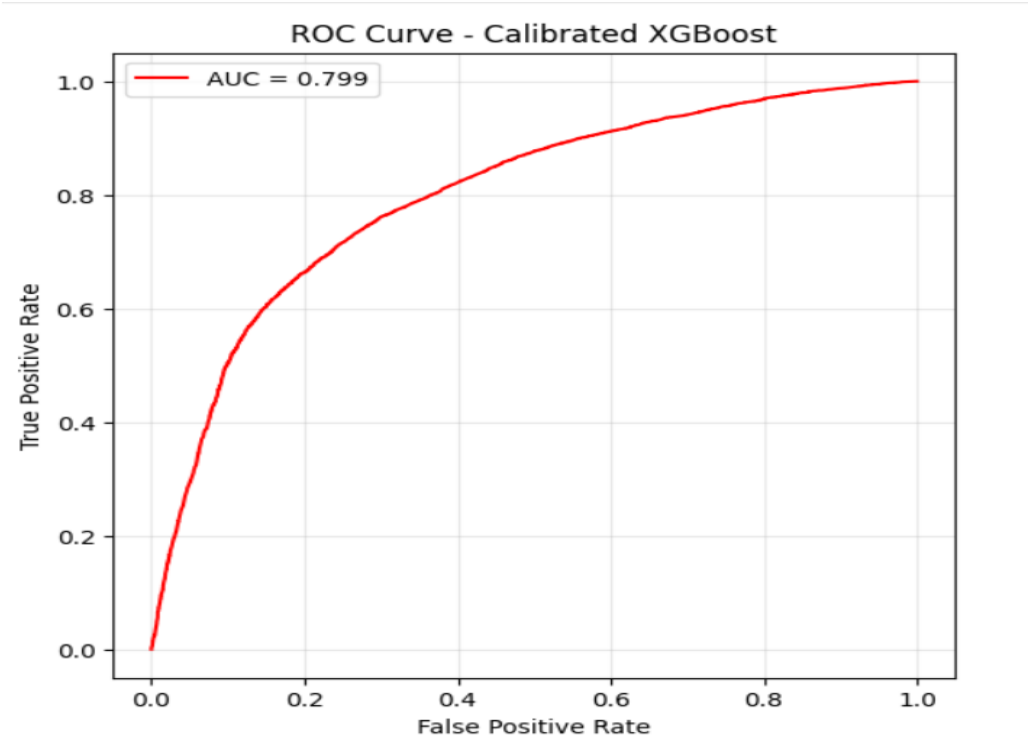


*Figure 3*

Figure 3 shows ROC curve of Calibrated XGBoost. It's value of AUC remains same like that of XGBoost confirming that Calibration improves probability reliability without affecting its overall classification ability.

**2. Performance Metrics used:**

1. **Accuracy:**

   This metric measure the overall number of accurate classess.

2. **Precision:**

   Precision Indicates how many of the cases predicted as 'disease' were actually classified correctly. High precision means fewer false positives.

3. **Recall:**

   It is also known as Sensitivity. Recall shows how model well the model classify the actual number of disease cases. High recall means fewer missed patients.

4. **F1-Score:**

   F1-Score is the harmonic mean of precision and recall. It plays important role when both false positive and false negative classes matter.

5. **AUC-ROC:**

   This metric measures the models capacity to distinguish between positive and

negative classess across different thresholds. Higher AUC means model is more better at separating positive and negative classess

### 3. Accuracy, Precision, Recall and F1-Score for all theree models:

*Table 2*

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 0.725 | 0.748 | 0.670 | 0.707 |
| XGBoost | 0.732 | 0.747 | 0.694 | 0.719 |
| Calibrated XGBoost | 0.733 | 0.754 | 0.685 | 0.718 |

### 4.Confusion matrix:

- True Negatives: 5421
- False Positives: 1522
- False Negatives: 2141
- True Positives: 4660

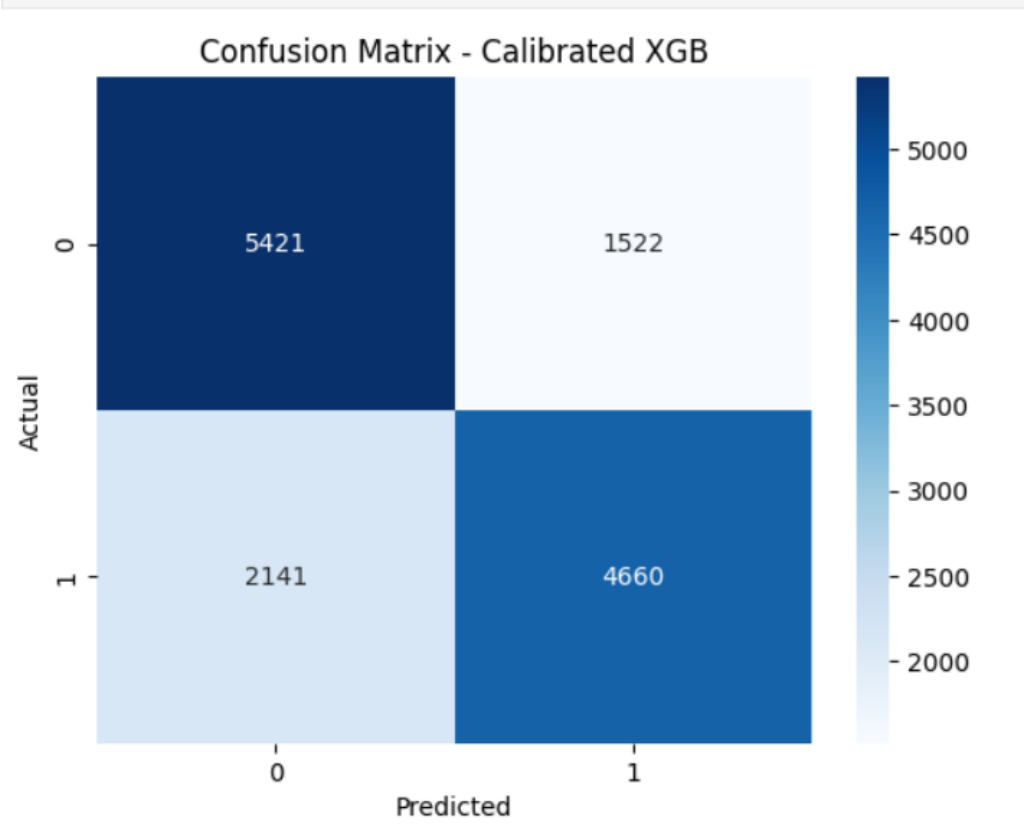


*Figure 4*

**5. Model Calibration Output:**

- Model calibration output is plotted as below
- This plotted curve looks close to the diagonal, which indicates well calibrated risk prediction.
- Also, Brier score obtained here is 0.1815 which means that model's predicted probabilities are fairly accurate and well-calibrated.

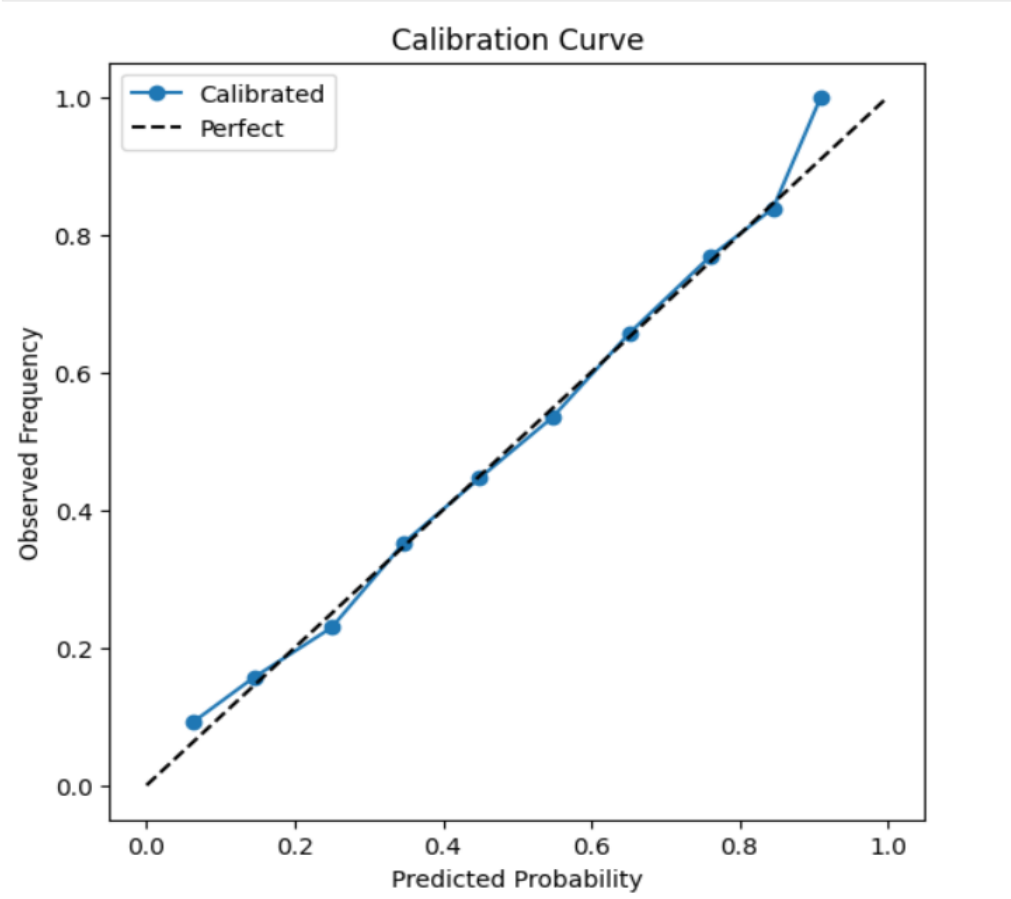


*Figure 5*

**6. SHAP Explainability Outcome:**

a. Global SHAP Beeswarm Plot

It shows overall contribution of features such as:

- ap_hi (systolic BP)
- age
- cholesterol
- pulse pressure
- weight

in CVD prediction.

b. Local SHAP Waterfall Charts:

- This was plotted to show individual patient interpretability

These outputs increase model transparency and help understand why the model makes certain predictions.

**Global SHAP Beeswarm Plot:**

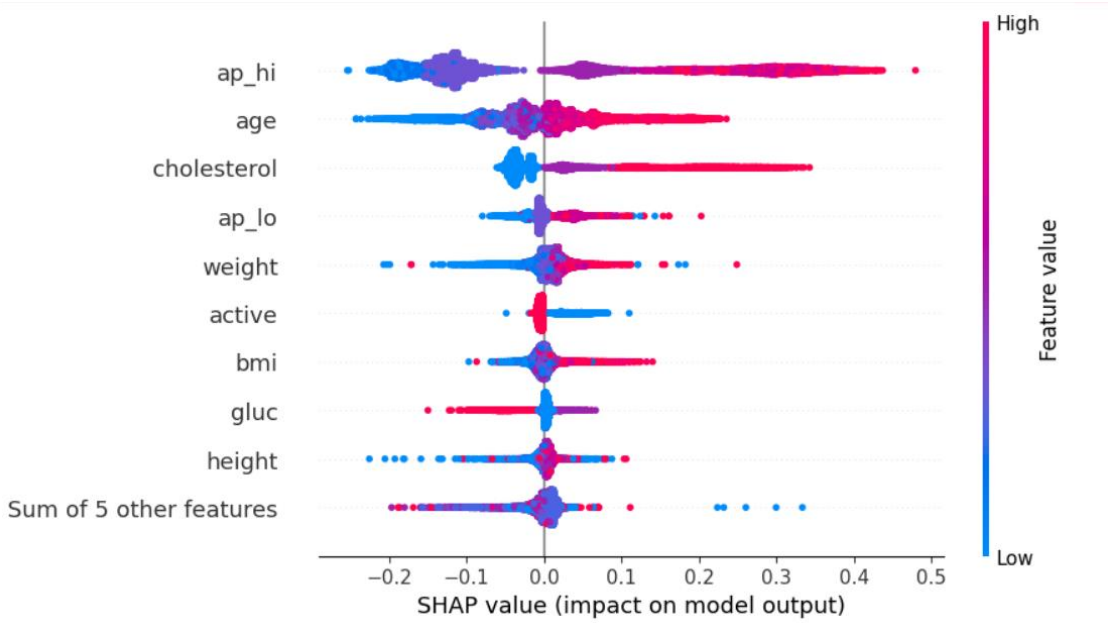


*Figure 6*

**Local SHAP Waterfall Chart (Example 1) :**

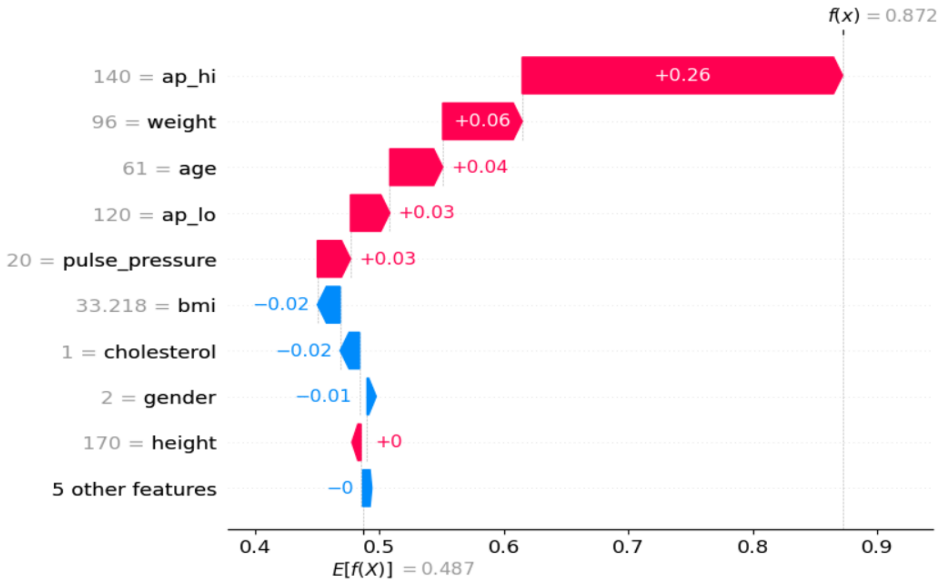


*Figure 7*

**Local SHAP Waterfall Chart (Example 2) :**

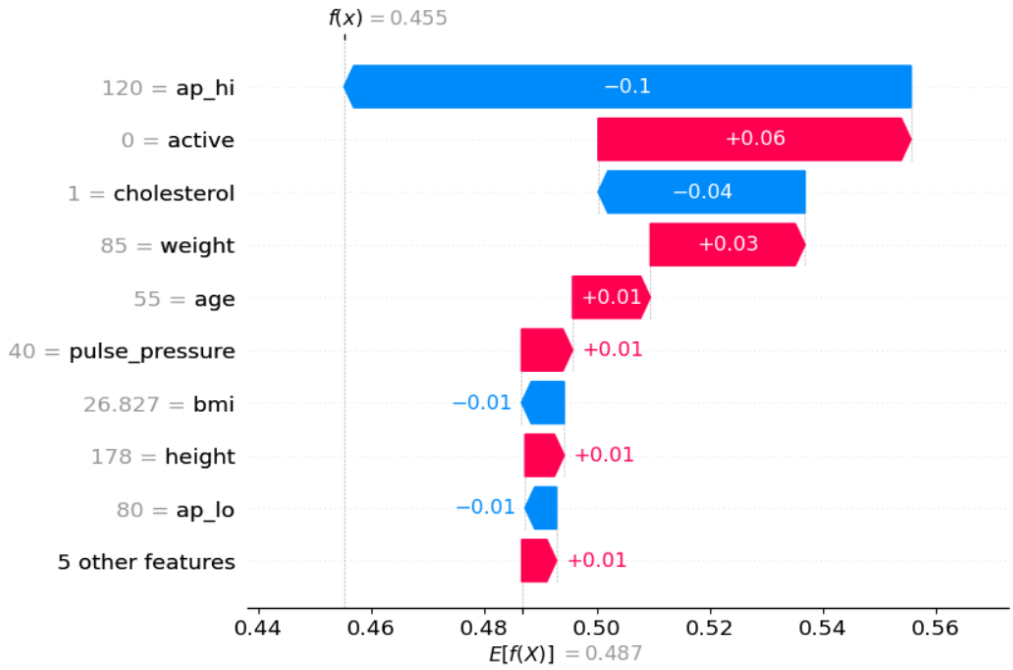


*Figure 8*

## 7. Subgroup Analysis:

A subgroup analysis was done to test if the calibrated XGBoost model's predictive performance kept the same level across various demographic groups. Risk of cardiovascular disease is mostly influenced by age and gender, thus it is very important to check whether the model works equally well in these inherently different groups of people. This action is a measure of the predictive system's fairness, its robustness, and the possibility of bias.

The analysis included the comparison of the two main subgroups:

1. Age-Based Subgroup Analysis
2. Gender-Based Subgroup Analysis

- The subgroup analysis in general supports the conclusion that the calibrated XGBoost model keeps consistent predictive performance for different demographic groups and does not discriminate against any particular patient groups.

**Below is the subgroup performance analysis table:**

*Table 3*

| Group | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Age < 45 | 0.814358 | 0.825013 | 0.666012 | 0.811005 | 0.565000 |
| Age ≥ 45 | 0.782828 | 0.718414 | 0.722273 | 0.749653 | 0.696823 |
| Male | 0.800188 | 0.740230 | 0.728899 | 0.768983 | 0.692786 |
| Female | 0.798154 | 0.729881 | 0.711836 | 0.745572 | 0.681021 |

### 6.2 Outputs of statistical and spatial analysis of state level CVD count data

This part of the project used 5-year state-wise mortality data to analyze spatial patterns and variability.

1. **Computed Statistical Indicators**

   Descriptive statistical indicators were computed for the state, wise CVD data of five years (2018, 2022) to get an idea about the changes in and the regularity of the CVD counts in different states of India. These calculations were initially done in Microsoft Excel and then the results were saved in a workbook, which was later imported into Python for ranking and plotting.

   The following indicators were obtained for each state:

   - **Mean**: Is the total of all CVD counts of the five years divided by the number of years.
   - **Standard Deviation:** Indicates the extent of the variation of yearly values from the mean.
   - **Coefficient of Variation (CV):** A standard measure of variation (SD ÷ Mean). The lower the CV, the more consistent the trend would be and the higher the CV, the trend would be unstable.
   - **Interquartile Range (IQR):** Defines the distance between the first and the third quartile. It is used as a measure of year, to, year changes in a state.
   - **CV Ranking:** Arranged states from the most consistent (lowest CV) to the least consistent (highest CV).

- **IQR Ranking**: In a similar way, ranking depending on IQR values, these pointing to either stable or volatile CVD patterns in the states.

These indicators allowed for:

- Identification of the states where CVD had been stable and predictable and the trends had shown low fluctuation over the years of five.

The top 5 and bottom 5 states according to both CV and IQR ranking are presented in the table below.

1. Top 5 States by CV Ranking:

*Table 4*

| States | cv_rank |
|---|---|
| Andaman and Nicobar Islands | 1 |
| Jammu and Kashmir | 2 |
| Karnataka | 3 |
| Goa | 4 |
| Tamil Nadu | 5 |

2. Bottom 5 states by CV Ranking

*Table 5*

| States | cv_rank |
|---|---|
| Lakshadweep | 35 |
| Telangana | 34 |
| Bihar | 33 |
| Mizoram | 32 |
| Delhi | 31 |

3 Top 5 states by IQR Ranking:

*Table 6*

| States | IQR_rank |
|---|---|
| Andaman and Nicobar Islands | 1.5 |
| Lakshadweep | 1.5 |
| Sikkim | 3 |
| Mizoram | 4 |
| Chandigarh | 5 |

4. Bottom 5 states by IQR Ranking

*Table 7*

| States | IQR_rank |
|---|---|
| West Bengal | 35 |
| Rajasthan | 34 |
| Uttar Pradesh | 33 |
| Madhya Pradesh | 32 |
| Maharashtra | 31 |

**2. Spearman Rank Correlation Output**

Correlation between CV and IQR rankings:

Spearman correlation = -0.18265

Comparison of ranks show that variablility measured by CV and IQR is different for different states as the two ranking systems have a low correlation according to Spearman's correlation coefficient.

**3. Spatial Cluster Analysis (SaTScan)**

With SaTScan, spatial clusters with a high CVD burden were identified for every year (2018, 2022) through:

- Poisson model
- Scanning window sizes of 2% and 5%

In total, 10 cluster maps were created (5 years × 2 risk windows).

The outputs consist of:

- Positions of significant clusters
- Relative risk figures
- The states most affected
- India's geographical map with visual heat, maps

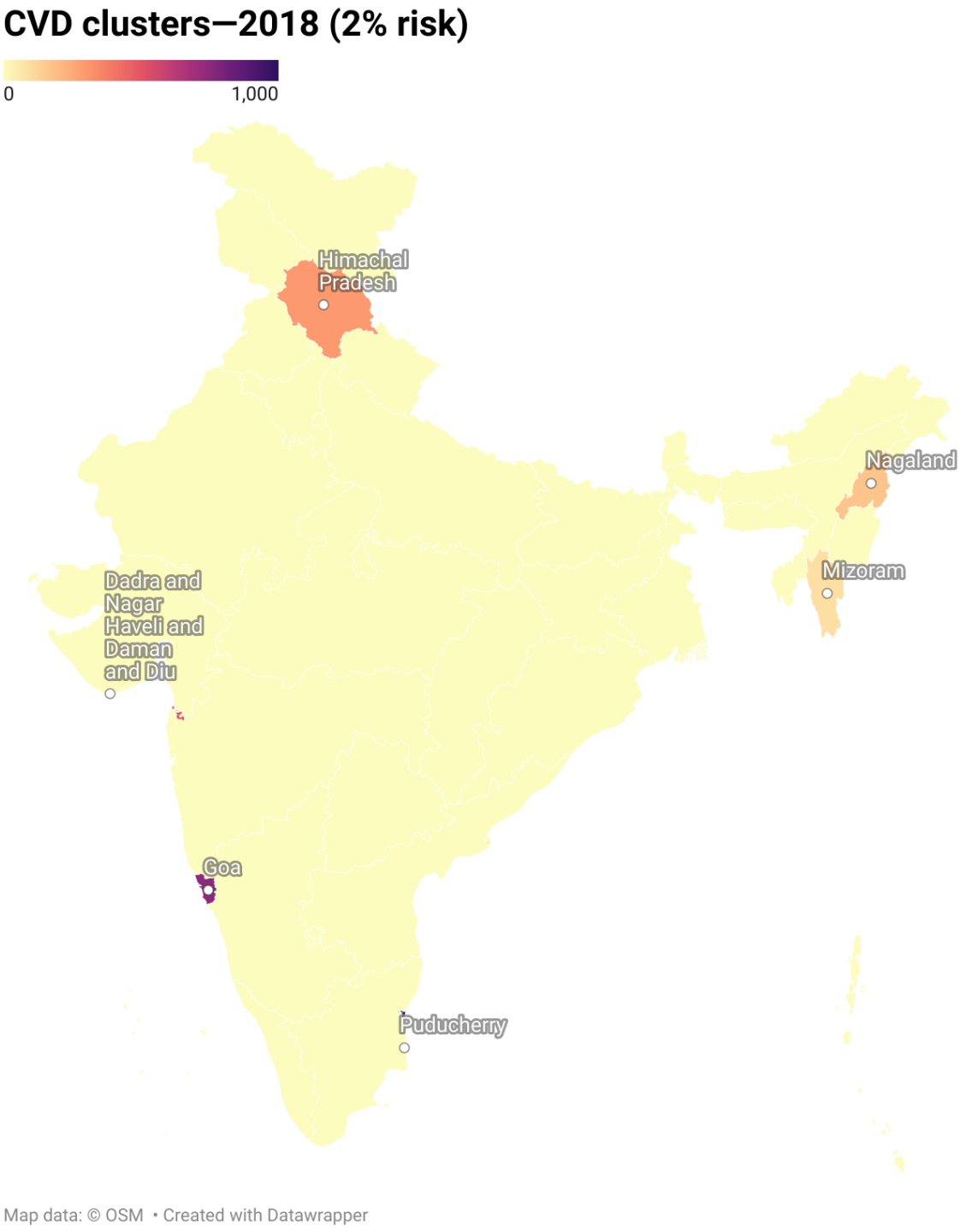
CVD clusters—2018 (2% risk)
0
1,000
Himachal Pradesh
Nagaland
Mizoram
Dadra and Nagar Haveli and Daman and Diu
Goa
Puducherry
Map data: © OSM • Created with Datawrapper

*Figure 9*

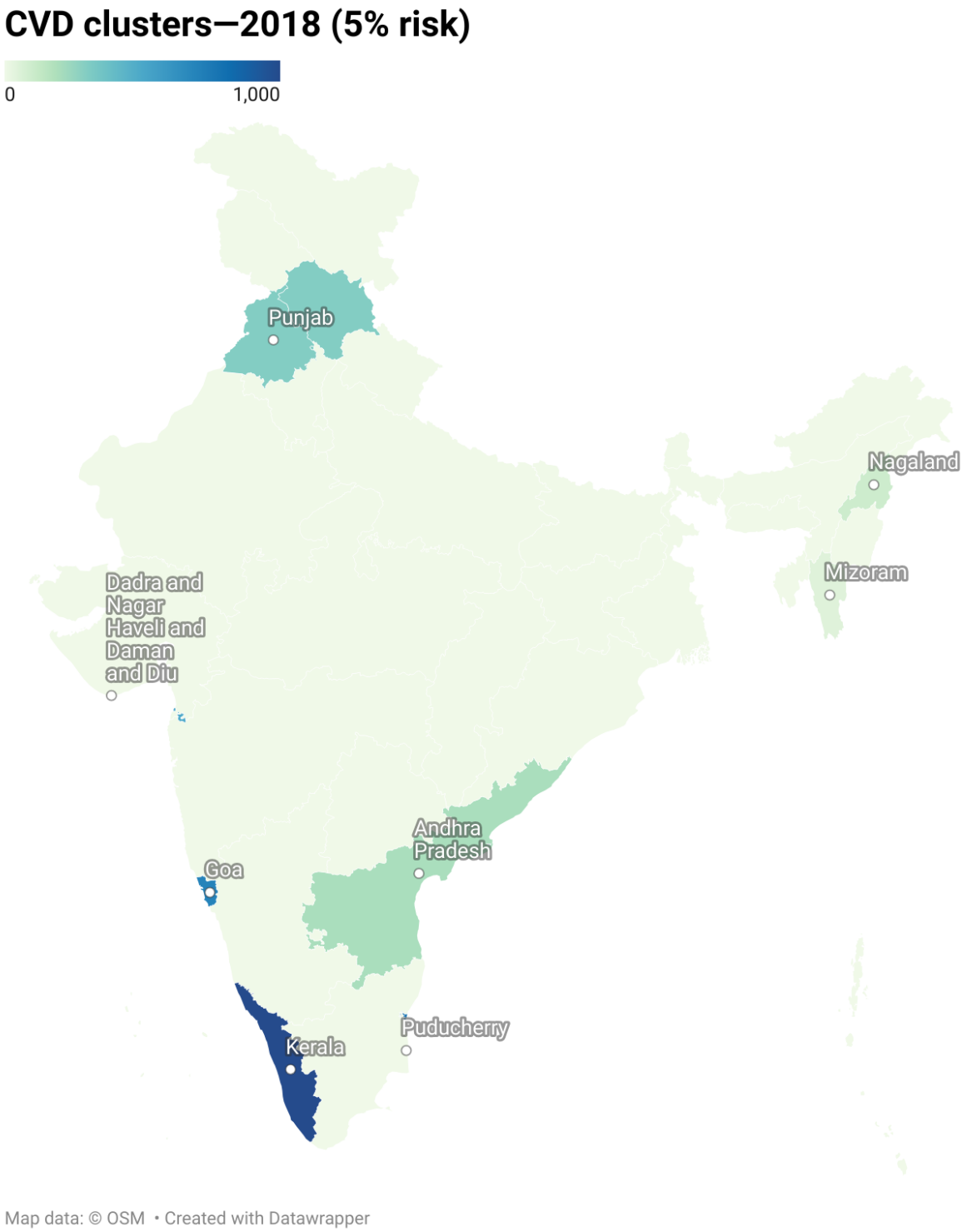
CVD clusters—2018 (5% risk)
0
1,000
Punjab
Nagaland
Mizoram
Dadra and Nagar Haveli and Daman and Diu
Andhra Pradesh
Goa
Puducherry
Kerala
Map data: © OSM • Created with Datawrapper

*Figure 10*

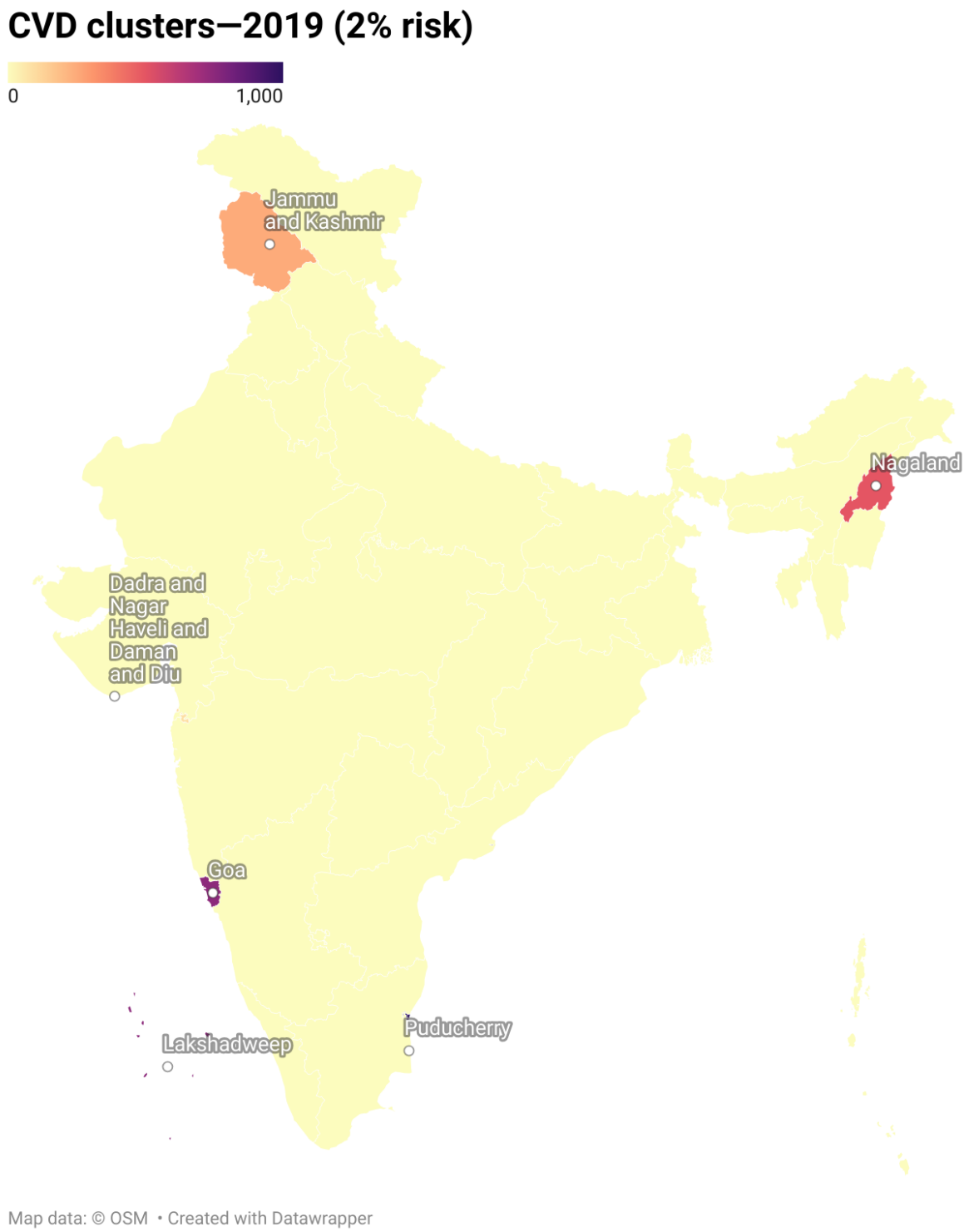
CVD clusters—2019 (2% risk)
0
1,000
Jammu and Kashmir
Nagaland
Dadra and Nagar Haveli and Daman and Diu
Goa
Puducherry
Lakshadweep
Map data: © OSM • Created with Datawrapper

*Figure 11*

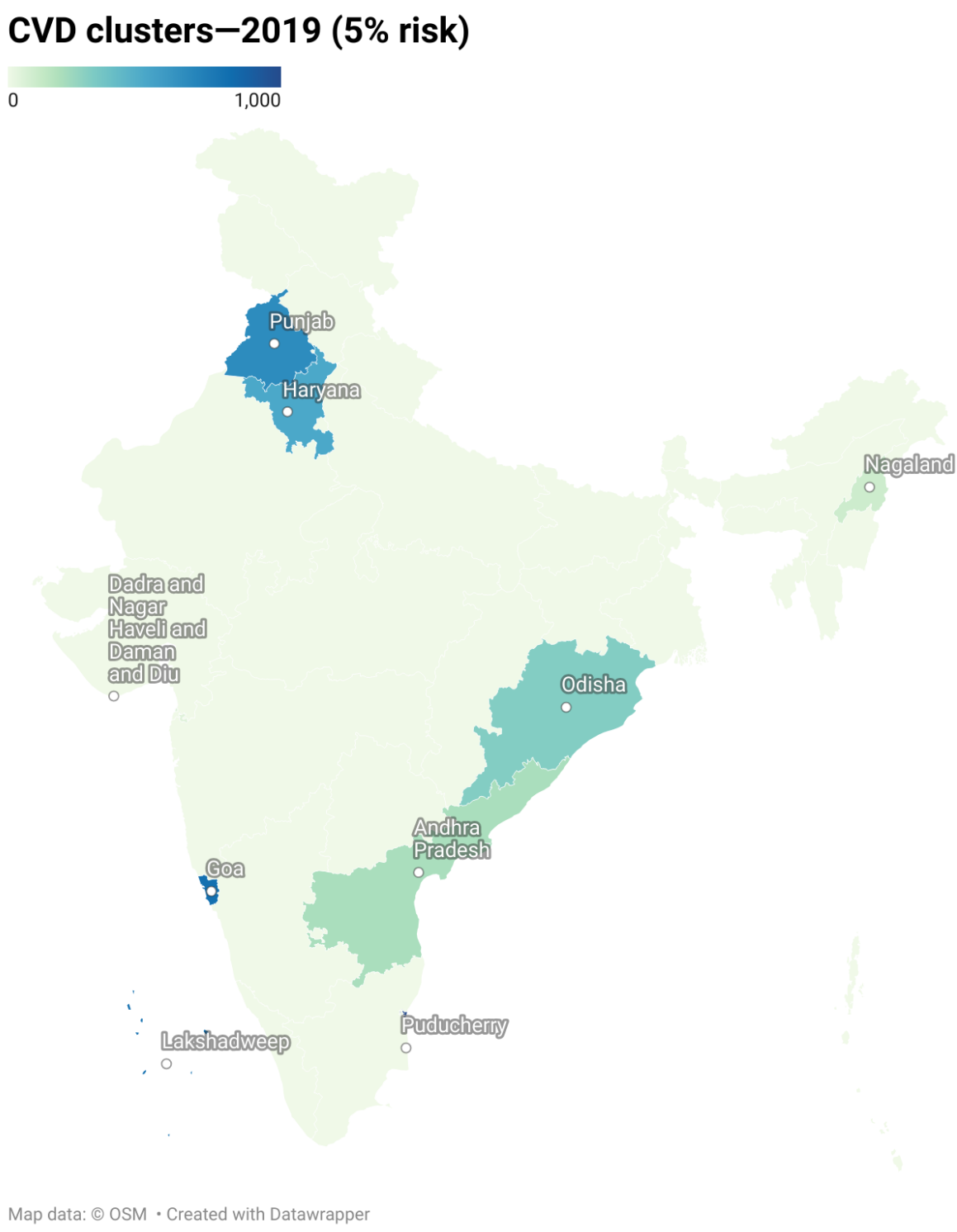
CVD clusters—2019 (5% risk)
0
1,000
Punjab
Haryana
Nagaland
Dadra and Nagar Haveli and Daman and Diu
Odisha
Andhra Pradesh
Goa
Puducherry
Lakshadweep
Map data: © OSM • Created with Datawrapper

*Figure 12*

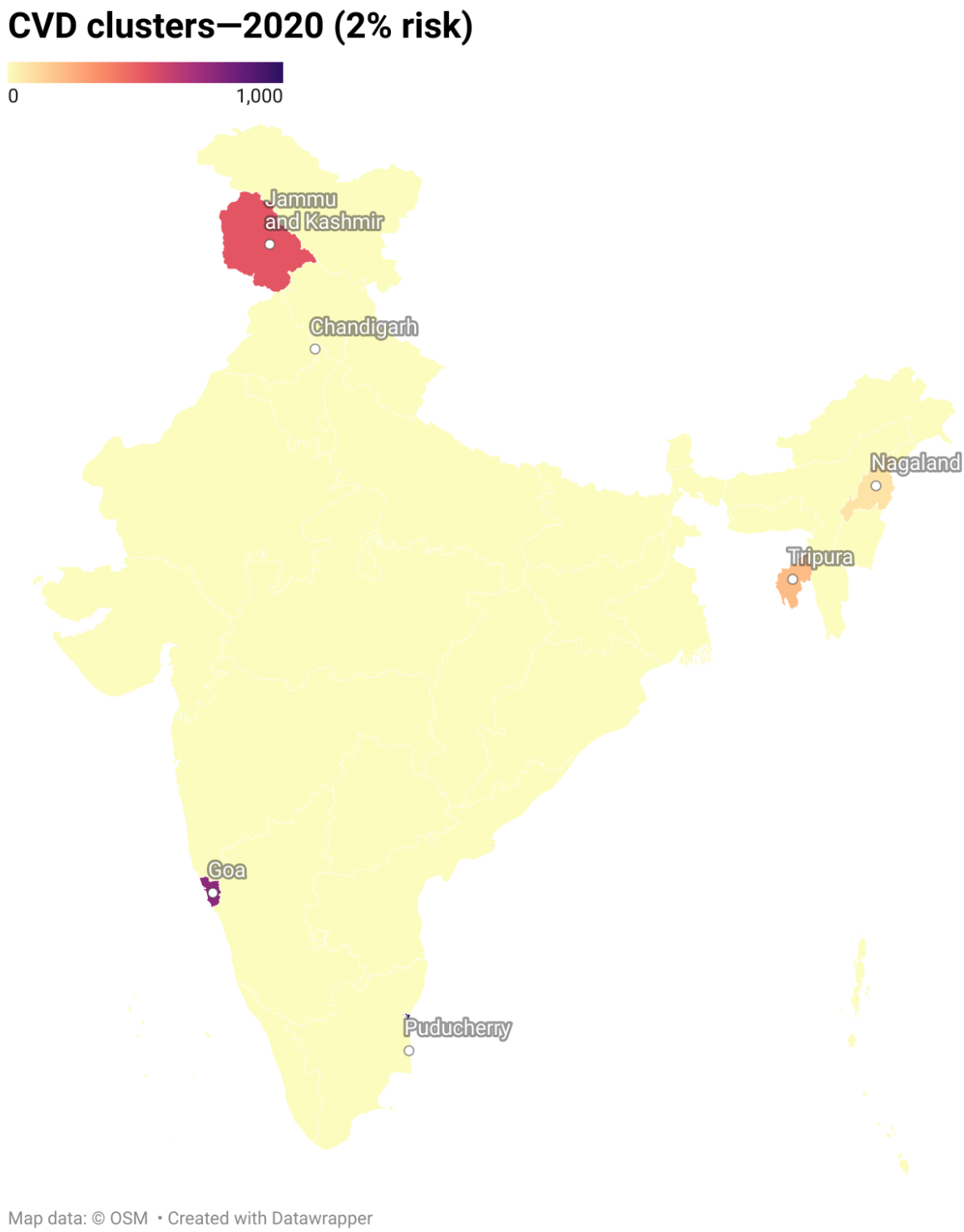
CVD clusters—2020 (2% risk)
0
1,000
Jammu and Kashmir
Chandigarh
Nagaland
Tripura
Goa
Puducherry
Map data: © OSM • Created with Datawrapper

*Figure 13*

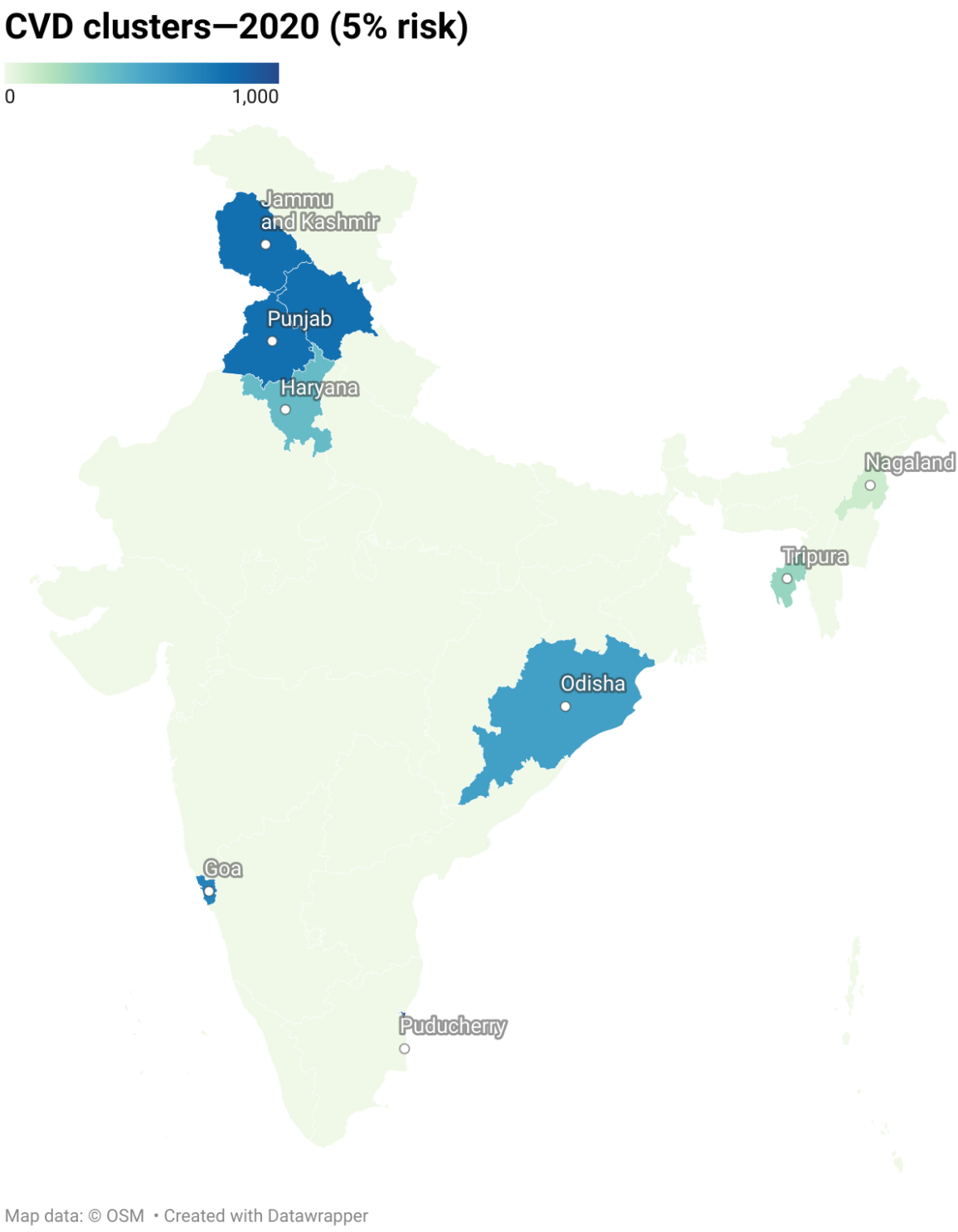
CVD clusters—2020 (5% risk)
0
1,000
Jammu and Kashmir
Punjab
Haryana
Nagaland
Tripura
Odisha
Goa
Puducherry
Map data: © OSM • Created with Datawrapper

*Figure 14*

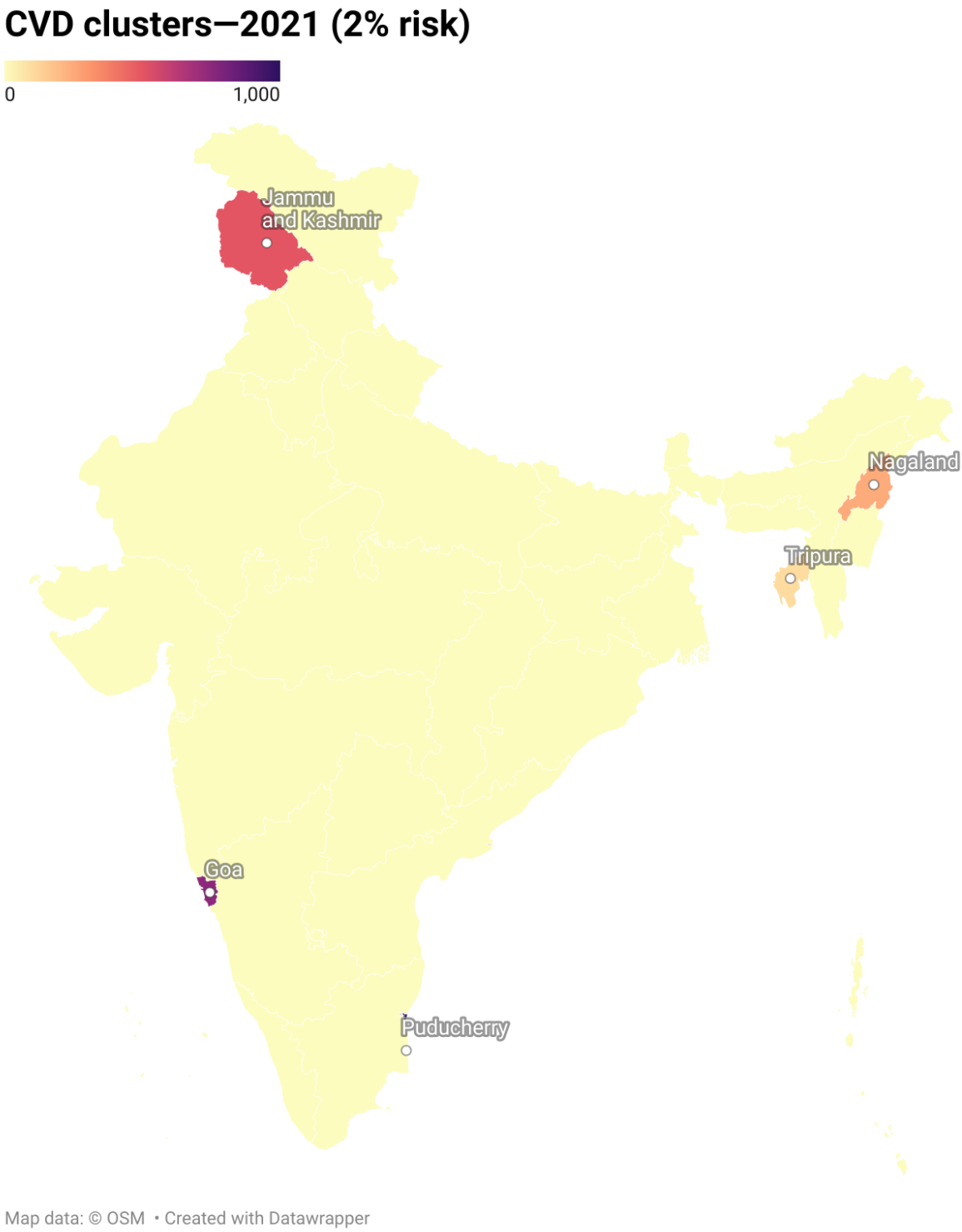
CVD clusters—2021 (2% risk)
0
1,000
Jammu and Kashmir
Nagaland
Tripura
Goa
Puducherry
Map data: © OSM • Created with Datawrapper

*Figure 15*

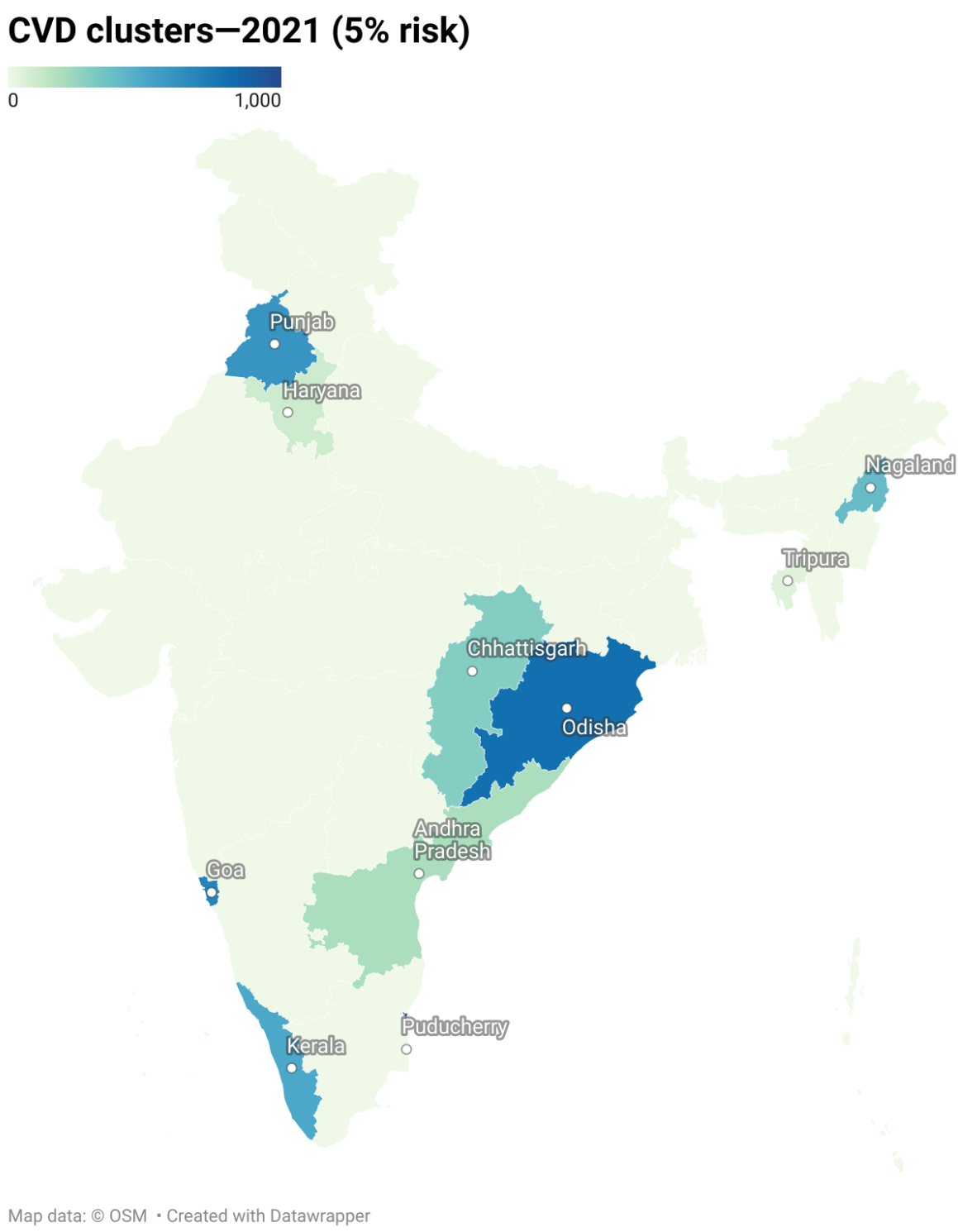
CVD clusters—2021 (5% risk)
0
1,000
Punjab
Haryana
Nagaland
Tripura
Chhattisgarh
Odisha
Andhra Pradesh
Goa
Puducherry
Kerala
Map data: © OSM • Created with Datawrapper

*Figure 16*

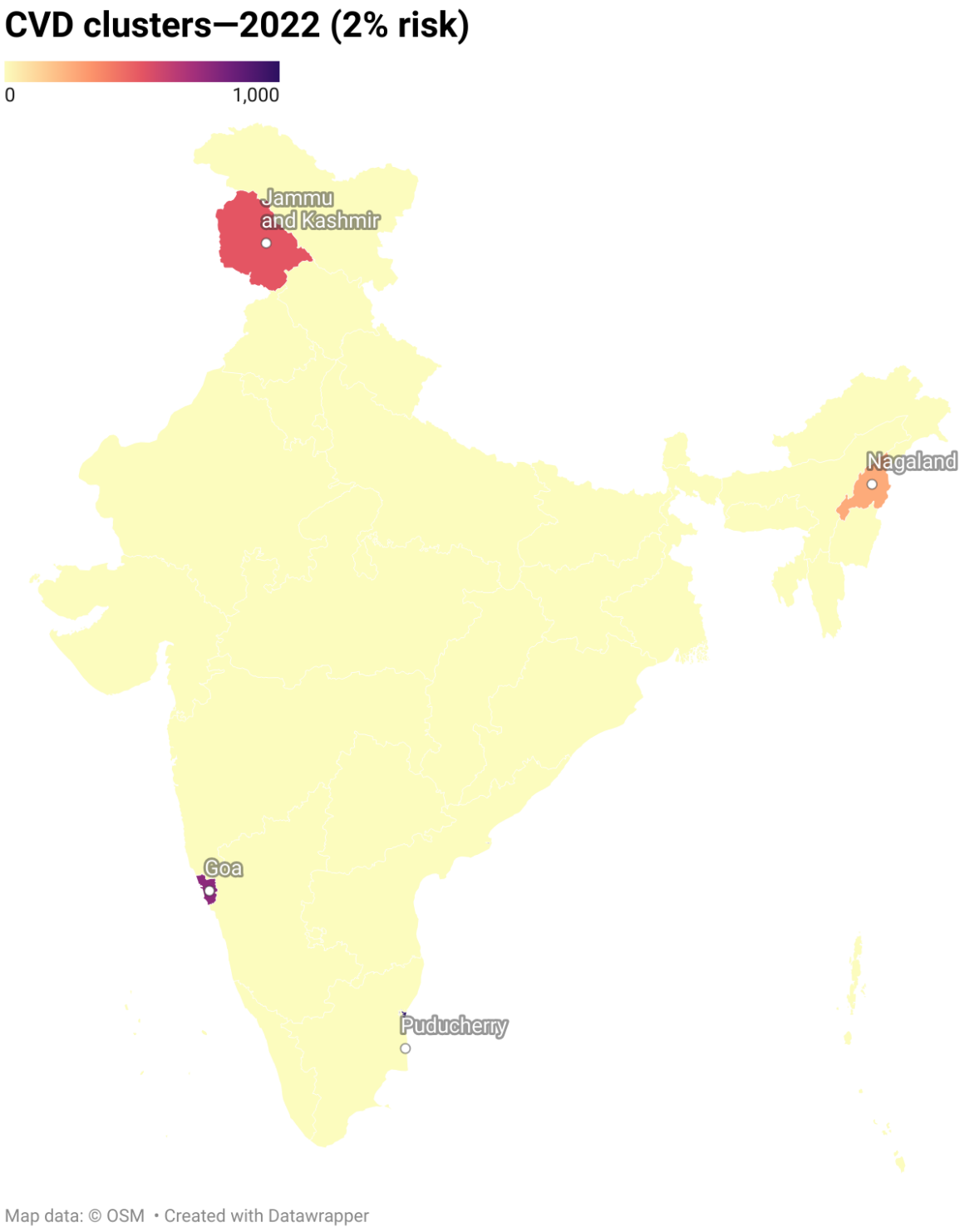
CVD clusters—2022 (2% risk)
0
1,000
Jammu and Kashmir
Nagaland
Goa
Puducherry
Map data: © OSM • Created with Datawrapper

*Figure 17*

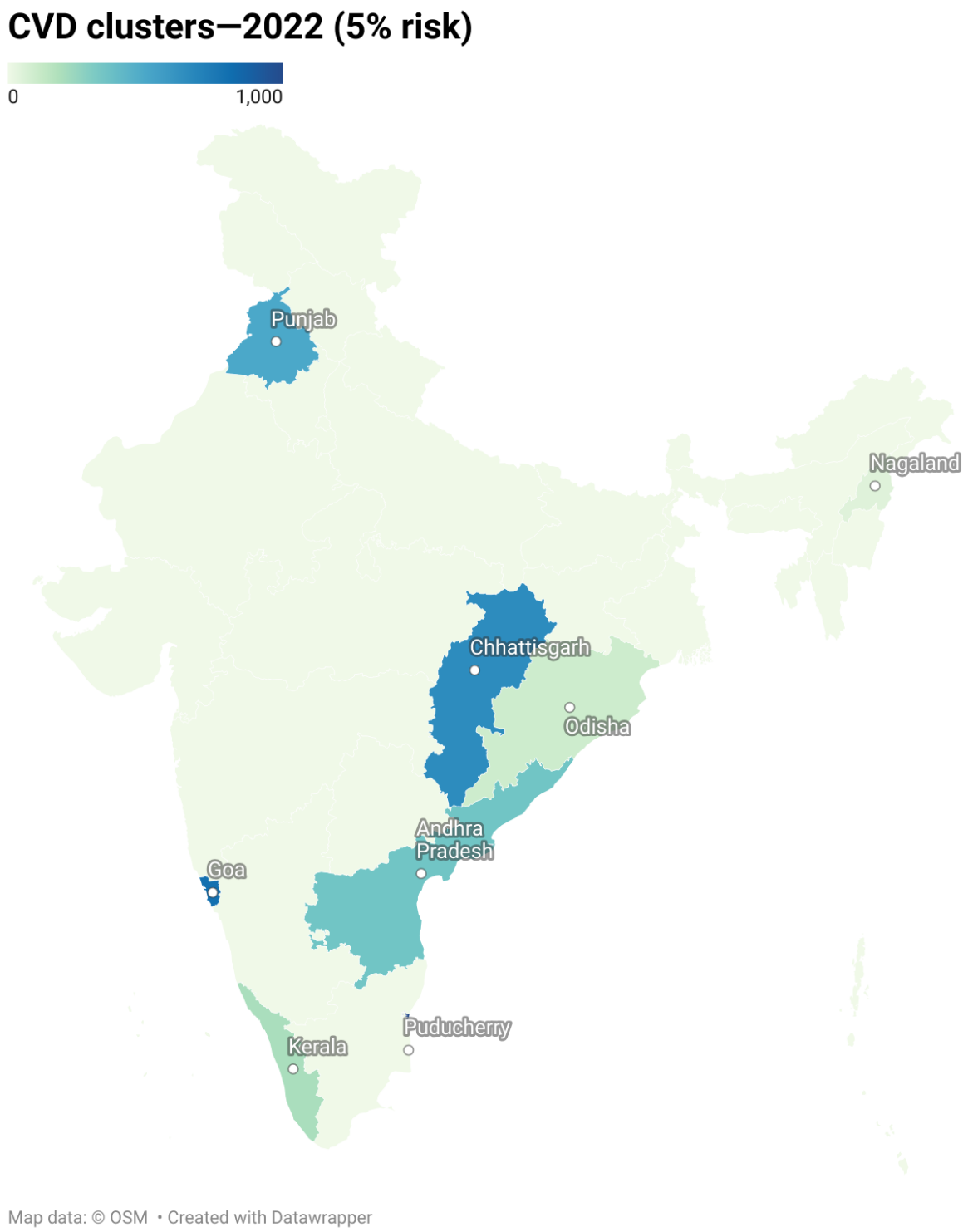
CVD clusters—2022 (5% risk)
0
1,000
Punjab
Nagaland
Chhattisgarh
Odisha
Andhra Pradesh
Goa
Puducherry
Kerala
Map data: © OSM • Created with Datawrapper

*Figure 18*

# 7. RESULTS AND DISCUSSION

In this section, the analytical results of the two significant sections of the research are united. This section is not only intended to replicate the outputs, but to make inferences about them, indicating trends, contrasting approaches, and commenting on their implications to early detection and the planning of public-health.

## 7.1 Predictive Modelling Results Discussion

### 7.1.1 Comparison of performance of the model.

Cleaned and feature-engineered CVD dataset was assessed using three models- Logistic Regression, XGBoost and Calibrated XGBoost to predict cardiovascular disease with an almost 70,000 records of patients. Their ROC curves (Figures 1, 2, 3) show that they are clearly different in the discriminative capability.

- Logistic Regression (AUC = 0.786)
  Averagely competent and establishes a good threshold. Figure 1 demonstrates a moderate level of separation between the classes on its curve.

- XGBoost (AUC = 0.799)
  Performs better than Logistic Regression (Figure 2) and this shows that non-linear interactions between features are better represented with the boosting method like the relationship between blood pressure, cholesterol, and age.

- Calibrated XGBoost (AUC = 0.799)
  Figure 3 demonstrates that AUC does not change following calibration, which confirms the fact that calibration changes probability values and does not affect classification strength.

Altogether, the tendency indicates that XGBoost is better than linear models, and it confirms its applicability to clinical tabular data.

**Interpretation**: (Refer table 2)

- Logistic Regression to XGBoost accuracy is a bit higher.

- Calibrated XGBoost is most precise and therefore it has fewer false positives, which is of essence when screening medical patients.
- Recall (sensitivity) increases 0.670 - 0.694 (XGBoost), which implies that the model will detect additional CVD cases that are true.

The implication of these results is that boosting-based models give the optimal combination of precision and recall and give good probability interpretation following calibration.

### 7.1.2 Understanding Model Errors

Figure 4 (confusion matrix) indicates the following:

- True Negatives: 5,421
- True Positives: 4,660
- False Positives: 1,522
- False Negatives: 2,141

Even the false negatives (missed disease cases) are noteworthy and indicate the necessity of a careful interpretation. In medicine, it is critical to reduce false negatives, and the improvements (rebalancing, threshold tuning) can be made to the model in the future to maximize recall.

### 7.1.3 Calibration Performance

Calibration is essential as the probability results are taken to mean the true risk of the disease. It was demonstrated in the calibration curve (Figure 5 ) that:

- Probabilities that are expected are near the diagonal line,
- Brier Score = 0.1815, which is not large.

This means:

The XGBoost model is not only well-calibrated but also gives sensible percentages on risks, so it can be used in clinical risk estimation tools.

### 7.1.4 Model Explainability (SHAP Analysis)

Healthcare needs explainability and SHAP provides both global and individual understanding.

**SHAP Beeswarm Plot across the world (Figure 6).**

The most significant aspects were:

- ap_hi = systolic blood pressure.
- Age
- Cholesterol levels
- Pulse pressure
- Weight/BMI

(Shown clearly in Figure 6 )

These are known risk factors of CVD that support clinical validity of the model.

**Local SHAP Waterfall Plots (Figure 7 and 8).**

Those are the reasons, patient-by-patient, why a prediction had been high or low:

- There are some characteristics that drive probability in an upward direction (positive contribution).
- Other drive it down (protective factors).

This increases credibility, which enables clinicians to justify predictions, which is absent in conventional ML models.

### 7.1.5 Subgroup Analysis

Possible biases can be determined through the use of subgroup performance (Table 3).

**Age-based**

- The age less than 45 has more AUC (0.814) and accuracy (0.825).
- Age 45 and more group has a slightly lower performance (AUC 0.782), which could be explained by the fact that older adults may be more physiologically variable.

**Gender-based**

- Male AUC: 0.800
- Female AUC: 0.798

Theoretically, the same competence.

Interpretation:

The model obtained after calibration is consistent and does not favor any group of gender or age, and it is not particularly biased. This makes it appropriate to be deployed.

## 7.2 Statistical and Spatial State-level CVD Analysis Discussion.

### 7.2.1 Statistical Measures Coeficient of Variation (CV) and Inter Quartile Range (IQR).

The first step in comprehending differences between the individual states in terms of CVD cases was to rank the states separately using CV and IQR. Tables 4,6 and tables 5,7 reflect the 5 most stable and 5 most fluctuating states respectivly at the bottom of the list according to each statistical indicator. Comparing both rankings discloses the position of a state in terms of variability as well as the magnitude of that variability which is significant because CV and IQR do not always rank states in the same direction. This finding is supported by the weak negative Spearman correlation ($\rho$ =, 0.1826) which shows that both metrics can be used to get a complete picture of the situation.Most consistent (Low CV) states (top 5):

These findings indicate that there are stable annual patterns in some states and big variations in others, possibly because of:

- Mobility and demographic change.
- Under-reporting or over-reporting of some years.
- Alterations in access to healthcare.

### 7.2.2 Spearman Rank Correlation Insight

It is paramount to acknowledge that the Spearman Rank Correlation Insight is one of the insights applicable in the context of the given model as this can help to discover a variety of relationships.

Spearman correlation between the CV and IQR rankings:

$\rho$ = - 0.18265

A negative correlation but weakly indicates:

- The two measures do not always rank the states in a similar manner.
- CV records relative variability, and
- Inter Quartile Range throws light on absolute dispersion.

Therefore, both measures provide a more comprehensive picture of CVD stability among the states.

### 7.2.3 Findings of Spatial Clusters (SaTScan)

Using SaTScan, 10 cluster maps were created (Figures 9-18):

- For 5 years

- At 2 risk levels (2% & 5%)

**Important Conclusions of 2% Risk Maps.**

Across years 2018-2022:

- Jammu & Kashmir is always a high-risk cluster implying that it is always vulnerable.
- Northeastern states (only Nagaland, Tripura, Mizoram) are small and have their spots in between, pointing to the regional sensitivity.
- Goa is mentioned several times, which could have been due to the density of reporting or the demographics.

**Significant Results of 5 percent Risk Maps.**

These broader windows show:

- Even larger clusters in Punjab, Haryana, Odisha, Chhattisgarh, Andhra Pradesh, etc.
- South-western coast like in Karnataka and Goa are seen over several years.

Interpretation:

- Northern clusters (Punjab, Haryana, J&K) conform to already present higher risk factors hypertension, smoking and dietary.
- The problem of healthcare accessibility may be reflected in Eastern clusters (Odisha, Chhattisgarh).
- Small populace results in a large proportional fluctuations that are not steady on small-island territories.

**Overall Spatial Pattern**

Clusters analysis indicates that:

- A high-risk area that is consistently north-India.
- New spots in the northeast.
- The level of risk is relatively low in the coastal southwest states.

This space-time consistency over a few years gives a greater credibility to the hypothesis that geographic risk has close relation with lifestyle, socio-economic status and access to health services.

## 8. LIMITATIONS

Despite the achievement of its main goals in the prediction of models and spatial analysis of cardiovascular disease (CVD), the project had limitations that were realized throughout the process.

Access to and compatibility of external datasets was one of the biggest challenges. Although the primary dataset (70K instances) was large enough, most publicly available datasets of CVDs (including the Cleveland dataset) varied dramatically in the definition of features, scales, and clinical variables. This would not allow valid external validation, and limit the extrapolation of the model to generalize to new data.

The second limitation came up due to the quality and consistency of the self-reported health attributes in the primary dataset. Such characteristics as smoking, alcohol drinking, and physical activity are subjective in nature and can harbor reporting bias. These variations may affect the learning of models and make them less interpretable.

The models of machine learning, particularly, XGBoost and its calibrated version, performed fairly well, yet they were ill-equipped to predict rare or extreme physiological cases, just like models cannot predict extreme events in real-life scenarios. As an example, the found errors in prediction in people whose blood pressure was either abnormally high or low, fell outside the normal ranges of populations. This leads to the larger issue that a model that was trained on previous data might not be able to forecast sudden medical incidents, physiological abnormalities, or the sudden change in disease models.

On the spatial analysis front, such tools as SaTScan demand a lot of data granularity. As state-level data was the only one available regarding the number of Indian CVD (2018-2022), clusters identified represent large-scale regional patterns but not local hotspots. Deeper epidemiological insight would have been made with district level or block level data.

With these restrictions, the project offers a sound basis to predict CVD risks and analyse in large and spatial epidemiological processes in the real world.

## 9. CONCLUSION

The purpose of this project was to combine the predictive modelling and spatial analytics to draw valuable conclusions using cardiovascular disease data on the individual and regional levels. All the key goals were reached by the conclusion of the research.

There were several machine learning models deployed and tested on the predictive analytics side. The best model that was obtained was the calibrated XGBoost model with a balanced score of accuracy, discrimination (with AUC 0.799), and probability calibration. The SHAP model also assisted in revealing the effect of clinical variables (systolic blood pressure, age, cholesterol and BMI) on the predictive CVD risk of both individual and population in general. Such observations render the model not merely true, but also understandable--a crucial feature of health care implementation.

The project also went beyond the traditional modelling to include state level spatial analysis with SaTScan. The clusters of high and low risk CVD were effectively identified in India in 2018-2022. This spatial element was a plus to the study since it gave a regional view of disease burden showing the states with consistently steady CVD trends and the ones with high percentile fluctuations. Also, prioritizing methods such as CV and IQR with subsequent Spearman correlation made it possible to gain a better insight into how the reported cases of CVD in various states were stable over time.

In general, this integrated strategy showed that predictive modelling and spatial epidemiology can be used jointly in supporting the planning of public health. Whereas the risk can be estimated at individual level using models, spatial tools provide the areas that need interventions or policy attention.

The project upholds a valuable conclusion: forecasting in healthcare is strong, yet it can never be perfect. Uncertainties can always come in due to human physiology, the impact of the environment, and the unexpected situations. So, although machine learning is capable of making decisions with high confidence, it is not able to absolutely substitute clinical judgment and actual public health surveillance.

To sum up, the paper managed to demonstrate the ways of combining big health data, machine learning algorithms, and geospatial applications and construct meaningful, interpretable, and

actionable insights to cardiovascular disease risk monitoring and prevention. The learning outcomes of this project did not only enhance technical knowledge but also emphasized on the significance of using data to make decisions in contemporary health care systems.

## 10. REFERENCES


[1] Cao, K., Liu, C., Yang, S., Zhang, Y., Li, L., Jung, H., & Zhang, S. (2025). Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. Scientific Reports, 15(1), 12406. https://doi.org/10.1038/s41598-025-96520-7

[2] Liu, T., Krentz, A., Lu, L., & Curcin, V. (2024). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal - Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080

[3] Hajiarbabi, M. (2024). Heart disease detection using machine learning methods: a comprehensive narrative review. Journal of Medical Artificial Intelligence, 7, 21. https://doi.org/10.21037/jmai-23 152

[4] Narasimhan, G., & Victor, A. (2025). Empirical analysis of predicting heart disease using diverse datasets and classification procedures of machine learning. Ain Shams Engineering Journal, 16(8), 103470. https://doi.org/10.1016/j.asej.2025.103470

[5] S K, H. K., A, P., G, K., T, L., & M, P. K. (2024). Heart Disease Prediction using XGBoost and Random Forest Models. 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 19–23. https://doi.org/10.1109/icmcsi61536.2024.00009

[6] Maach, A., Elalami, J., Elalami, N., & Mazoudi, E. H. E. (2022b). An intelligent Decision support ensemble voting model for coronary artery disease prediction in smart healthcare monitoring environments. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.14906

[7] Wang, Z., Dong, W., & Yang, K. (2022). Spatiotemporal Analysis and Risk Assessment Model Research of Diabetes among People over 45 Years Old in China. International Journal of Environmental Research and Public Health, 19(16), 9861. https://doi.org/10.3390/ijerph19169861

[8] Mena, C., Sepúlveda, C., Fuentes, E., Ormazábal, Y., & Palomo, I. (2018). Spatial analysis for the epidemiological study of cardiovascular diseases: A systematic literature search. Geospatial Health, 13(1), 587. https://doi.org/10.4081/gh.2018.587

[9] Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. TURKISH JOURNAL OF BIOLOGY, 49(5), 600–634. https://doi.org/10.55730/1300-0152.2766

[10] Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering, 2022, 1–9. https://doi.org/10.1155/2022/7351061

[11] Biswas, A., Singh, S. K., & Gupta, J. (2021). Spatial distribution of Cardio-Vascular diseases in India. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-561563/v1

[12] Mena, C., Sepúlveda, C., Fuentes, E., Ormazábal, Y., & Palomo, I. (2018b). Spatial analysis for the epidemiological study of cardiovascular diseases: A systematic literature search. Geospatial Health, 13(1), 587. https://doi.org/10.4081/gh.2018.587

[13] Sukanya, J., Gandhi, D. R., & Palanisamy, V. (2021). Heart Disease Prediction using Cluster Based MapReduce Paradigm. International Journal of Scientific and Research Publications, 11(3), 473–477. https://doi.org/10.29322/ijsrp.11.03.2021.p11167

[14] Miah, J., Ca, D. M., Sayed, M. A., Lipu, E. R., Mahmud, F., & Arafat, S. M. Y. (2023). Improving cardiovascular disease prediction through comparative analysis of machine learning models: a case study on myocardial infarction. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.00517

[15] Chowdhury, E. (2025). Risk Prediction of Cardiovascular Disease for Diabetic Patients with Machine Learning and Deep Learning Techniques. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2511.04971

[16] Assegie, T. A., Sushma, S. J., & Mamanazarovna, S. S. (2023). Explainable Heart Disease Diagnosis with Supervised Learning Methods. ADCAIJ ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL, 12, e31228. https://doi.org/10.14201/adcaij.31228

[17] Wan, S., Wan, F., & Dai, X. (2025). Machine learning approaches for cardiovascular disease prediction: A review. *Archives of Cardiovascular Diseases*, *118*(10), 554–562. https://doi.org/10.1016/j.acvd.2025.04.055