**Topic : “Machine‑Learning Analysis of Stroke Awareness and Risk Patterns in Community Survey Data”**

**Description :** In this mini‑project we propose to conduct a data‑driven analysis of the stroke‑awareness survey data. We will begin by cleaning and integrating the raw questionnaire responses to build a structured dataset that includes demographic variables, medical histories, lifestyle factors and respondents’ knowledge of stroke symptoms and risk factors. Using this dataset, we will compute awareness scores and perform exploratory statistics and visualizations to highlight knowledge gaps across different demographic groups. To deepen our technical contribution, we plan to apply unsupervised clustering (e.g., k‑means or k‑prototype) to identify clusters of participants with similar risk‑factor profiles and awareness levels, and then train supervised machine‑learning models such as logistic regression, random forest or gradient boosting to predict high versus low awareness from demographic and lifestyle features. We will evaluate model accuracy, interpret feature importance using techniques like SHAP, and present the findings in a written report and interactive dashboard. This project will allow us to apply data‑science and machine‑learning skills to a real public‑health problem while producing insights that could inform targeted stroke‑education campaigns.

**Workflow :**

1. Data Cleaning & Integration

Develop a reproducible pipeline to clean and merge survey responses from multiple sources into a unified analysis-ready dataset.

2. Awareness Scoring & Visualization

Create an algorithm to compute awareness scores and generate interactive visualizations to present patterns and gaps.

3. Machine-Learning Classification

Train and evaluate at least three classification models (e.g., logistic regression, random forest, gradient boosting) to predict awareness levels.

4. Population Clustering

Apply k-means or k-prototype clustering to identify subgroups with similar risk-factor profiles and awareness levels.

5. Feature-Importance Analysis

Use model interpretability tools to determine which variables most influence awareness prediction.

6. Interactive Dashboard

Build a web-based dashboard to explore survey insights, model predictions, and cluster profiles for academic and public-health audiences.

**Objectives :**

**1. Assess Community Awareness of Stroke**

* Quantitatively measure the level of knowledge regarding stroke symptoms, risk factors, and emergency responses among various groups within the surveyed community.

**2. Identify and Analyze Risk Patterns**

* Analyze demographic, lifestyle, and medical factors in the community to determine which groups have higher risk profiles for stroke.

**3. Segment Population Using Machine Learning**

* Apply clustering techniques (e.g., k-means, k-prototypes) to segment the community based on combined risk and awareness profiles, revealing hidden subgroups.

**4. Predict Awareness Levels from Risk and Demographic Factors**

* Build classification models to predict an individual's stroke awareness level using survey inputs, identifying which features are most influential.

**5. Interpret and Explain Machine Learning Findings**

* Use explainable AI techniques (like SHAP) to clarify which factors contribute most to low or high awareness, supporting actionable insights.

**6. Guide Targeted Public Health Interventions**

* Provide evidence-based recommendations for health authorities to design focused education and outreach programs for the most vulnerable and least aware groups.

**7. Develop an Interactive Dashboard (Optional)**

* Create a user-friendly interface for exploring analysis results, allowing stakeholders to visualize awareness gaps and risk group distributions easily.

**Literature Survey :**

**Paper One : "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm"**

**What the Paper Presents**

* Proposes using the random forest (RF) machine learning algorithm to predict stroke incidence based on demographic and behavioral data.
* Compares RF with logistic regression (LR) and decision tree (DT), finding RF achieves the highest accuracy (~94%) and best balance between precision and recall.
* Highlights age and body mass index (BMI) as the two most significant predictors for stroke incidence.
* Uses a dataset of 5110 patients from Bangladesh, addressing class imbalance with SMOTE oversampling.
* Emphasizes the importance of model interpretability and robustness in medical applications, preferring RF due to its higher transparency than deep learning "black box" models.
* Suggests the model can help clinicians predict stroke risk early, supporting preventative care.
* Provides a detailed methodology for data preprocessing, model training, evaluation, and interpretation.
* Concludes that RF is a state-of-the-art interpretable model suitable for stroke prediction in healthcare settings.

**Gaps :**

* The paper focuses on predicting stroke disease occurrence using demographic and behavioral data with a random forest model.
* It achieves high accuracy (~94%) in identifying individuals at risk of stroke.
* Key predictors identified include age and body mass index (BMI).
* The study uses clinical and demographic data but does not measure stroke awareness or knowledge.
* It does not apply clustering methods to identify subgroups based on risk or awareness.
* The study is limited to a dataset from Bangladesh, lacking local relevance to New Bombay.
* There is no focus on emergency response knowledge or lifestyle awareness in the model.
* No interactive dashboards or tools are developed for public health education.
* Your project differs by focusing on measuring stroke awareness and knowledge in the community.

**Paper 2 :** **Community Awareness regarding Stroke: A Survey-based Study Involving Family of Patients Presenting to an Indian Emergency Medicine Department**

What the Paper Presents

* Conducts a prospective survey of family members of patients at an Indian emergency department to assess stroke awareness.
* Uses an 18-question survey covering demographic details, stroke knowledge, risk factors, symptoms, and treatment awareness.
* Out of 1,118 respondents, 80.7% had heard of stroke, but only 0.8% were completely aware of stroke symptoms, risk factors, and treatment.
* Majority gained stroke knowledge from family (27%), education (26%), and healthcare providers (15%), with media awareness lower.
* Stroke awareness was poor irrespective of education level, age, or gender.
* Highlights that early recognition and timely treatment of stroke are critical but awareness remains deficient.
* Suggests public education campaigns via healthcare providers and media for mass dissemination of stroke awareness.
* Emphasizes the need for targeted community stroke education, especially for elderly and less educated groups.
* Uses cluster analysis to categorize respondents by awareness level and presents detailed demographic and awareness statistics.

Limitations / Gaps

* Study conducted in a single urban tertiary care center; may not represent rural or wider Indian populations.
* Convenience sampling could introduce selection bias.
* Survey responses might be influenced by the stress of relatives in an emergency department setting.
* Limited by self-reported knowledge, which might not perfectly reflect true awareness.
* No machine learning or advanced data modeling deployed to predict awareness based on demographics.
* The influence of media was reported to be low, suggesting gaps in public health communications.
* Does not provide an interactive tool or dashboard for visualization and application of insights.
* Focuses primarily on awareness but does not explore risk factors or behavior modification related to stroke.

**Paper 3 :** **Stroke awareness in the general population: knowledge of stroke risk factors and warning signs in older adults**

**Summary**

* The study surveyed community-dwelling older adults (65+) in Ireland to assess knowledge of stroke warning signs and risk factors.
* Conducted as part of a large population survey with 2,033 participants.
* Less than half of respondents could recognize established stroke warning signs (e.g., weakness, slurred speech) and risk factors (e.g., smoking, hypertension).
* Awareness was poorest among those with lower education and in Northern Ireland compared to the Republic of Ireland.
* Having personal risk factors like hypertension or prior stroke did not correlate with better knowledge.
* Mass media campaigns and public education have improved knowledge in younger groups, but awareness remains low in older adults who are at higher risk.
* Study highlights the importance of targeted education for older adults to reduce delays in seeking treatment and improve outcomes.

**Gaps Identified**

* Knowledge of stroke warning signs and risk factors remains insufficient, especially in older populations at greatest risk.
* Awareness does not necessarily translate into timely medical care or behavior changes.
* Education campaigns show limited long-term effectiveness, particularly for older adults.
* Differences between geographic regions need further exploration.
* Study uses list recognition format, which might overestimate knowledge compared to open-ended questions.
* No focus on awareness measurement tools or detailed subgroup clustering.
* Absence of interactive or advanced data-driven tools for public health interventions.

**Relevance to our Project**

This paper underscores the challenge of improving stroke awareness in vulnerable populations, supporting our project's focus on measuring and enhancing community knowledge using data-driven methods and demographic profiling. The gaps point to opportunities for our project to develop more detailed awareness indices, clustering of at-risk groups, and targeted education supported by interactive visualizations.

**Paper 4 :** **Knowledge structure and global trends of machine learning in stroke over the past decade: A scientometric analysis**

* Analyzed global trends of machine learning (ML) research in stroke over the past decade.
* Found rapid growth in ML applications for stroke diagnosis, prediction, treatment, and prognosis.
* Key ML methods include Random Forest, SVM, and Deep Neural Networks.
* Most research originates from developed countries; less focus on developing nations.
* Challenges remain in clinical adoption due to interpretability, data quality, and ethics.
* Limited integration of lifestyle and awareness data in ML stroke research.
* Your project complements these gaps by focusing on stroke awareness, community knowledge, and building interactive tools for public health education.

**Paper 5 :** **Assessing Public Awareness of Stroke: Knowledge of Warning Signs, Risk Factors, and Treatment Responses**

Link : <https://healthcare-bulletin.co.uk/article/assessing-public-awareness-of-stroke-knowledge-of-warning-signs-risk-factors-and-treatment-responses-2523/>

**What is There in the Paper**

* Conducted a 2-month cross-sectional survey with 500 participants aged 18-80 in Ahmedabad, India.
* Assessed public awareness of stroke warning signs, risk factors, and appropriate treatment responses.
* 95% of participants knew the term “stroke,” but only 25% could correctly identify warning signs.
* Females showed significantly higher awareness of warning signs and treatment protocols compared to males.
* Moderate awareness (~54%) of risk factors like hypertension, smoking, stress, and heart disease.
* 41% knew the correct treatment guidelines for stroke.
* Found strong associations between awareness of risk factors, warning signs, and treatment response.
* Emphasizes importance of timely recognition and emergency care for reducing stroke-related disability and mortality.
* Highlights need for targeted public education and better dissemination of information through media and healthcare providers.

**Gaps in the Paper**

* Low overall recognition of stroke warning signs and urgent treatment steps.
* Gender disparities in awareness remain unexplained or unaddressed.
* Limited understanding of how to translate awareness into immediate emergency action.
* Influence of cultural beliefs on stroke knowledge and behavior was not studied.
* No advanced data analysis, clustering, or predictive modeling used to understand awareness patterns.
* Geographic limitation to Ahmedabad; results may not generalize broadly.
* Lack of interactive or data-driven public health education tools based on findings.

**Paper 6 :** **Study on the public awareness of stroke risk factors, symptoms, treatment and rehabilitation: A community based cross sectional survey from a tertiary referral centre in South India**

**What is There in the Paper**

* Community-based cross-sectional online survey conducted over 3 months in Kerala, India.
* Surveyed 277 participants aged 18+ on stroke risk factors, symptoms, treatment, and rehabilitation awareness.
* Majority (85-90%) recognized common risk factors like hypertension, dyslipidemia, smoking, and diabetes.
* Common stroke symptoms such as facial weakness (94%) and speech disturbance (87%) were well known.
* 95% correctly identified thrombolysis treatment time window (4.5 hours) and hospital selection.
* Rehabilitation awareness was moderate (~60%) with knowledge of speech therapy and physiotherapy benefits.
* Good knowledge about home care for stroke survivors like feeding posture, prevention of bed sores, catheter care, and complications such as aspiration pneumonia and depression.
* Highlights Kerala’s higher literacy and community education efforts compared to other regions.
* Uses simple, direct questions with real-time data collection using KoBo Toolbox and analyzed with SPSS.

**Gaps in the Paper**

* Sample size is limited and may not represent broader Indian population.
* Online survey may exclude less literate or tech-unfamiliar participants.
* Some respondents chose "don't know" possibly without proper consideration.
* Leading questions may inflate correct answer rates.
* Few studies compare awareness about rehabilitation and home care.
* Needs more emphasis on uncommon stroke symptoms and transient ischemic attacks (TIA).
* Calls for greater education on immediate transportation to equipped centers and discouraging ill-advised first aid.
* Does not use clustering or machine learning to segment awareness patterns.
* Geographic scope limited to Kerala.

**Paper 7 :** **Awareness of stroke, its signs, and risk factors: A cross‐sectional population‐based survey in Ghana**

**What is There in the Paper**

* Cross-sectional survey of 1000 participants in Kumasi, Ghana, assessing stroke awareness.
* Evaluated knowledge of stroke signs, symptoms, risk factors, and associated perceptions.
* Participants showed limited overall knowledge: awareness of stroke signs and symptoms ranged from 25.9% to 47.2%; awareness of causes and risk factors ranged from 24% to 39%.
* Most recognized symptom was paralysis (70.8%), and diet (59.9%) was the most recognized risk factor.
* Sociodemographic factors (age, religion, education, employment, marital status) were significantly associated with stroke knowledge.
* High fatality rates from stroke in Ghana highlight the public health relevance.
* Findings support urgent need for intensified public health education and resource allocation in stroke prevention.

**Gaps in the Paper**

* Survey limited to a single urban market area, limiting generalizability.
* Lack of detailed clustering or machine learning analysis to profile awareness subgroups.
* Cultural beliefs and myths (e.g., attributing stroke to demons) influence misperceptions but are not deeply explored.
* Doesn't assess specific awareness of emergency response timing or treatment options.
* No interactive or data-driven tools proposed for knowledge dissemination or follow-up.

**Paper 8 :** **Public Awareness of Stroke Recognition, Risk Factors, and Access to Appropriate Treatment (Southern India, 2022)**

* **Nature**: Hospital-based cross-sectional survey (700 participants, Kerala)
* **Findings**:
  + 43.3% could not identify brain as organ affected in stroke; 32% thought it was the heart.
  + Only 9% identified all six stroke warning symptoms. Most recognized symptom: difficulty speaking (59.4%).
  + Hypertension was the best-known risk factor (77.7%), while awareness of diabetes and smoking was low.
  + 82.4% knew immediate treatment was required, but only 15.9% knew about the stroke helpline, and 55.9% were aware of ambulance services.
  + Higher awareness was associated with age >45, higher income, higher education, and urban residence.
* **Gaps**:
  + Limited understanding of diabetes and smoking as stroke risk factors.
  + Poor awareness of helpline, ambulance services, and need for rapid hospital response.
  + Most preferred visiting local doctors rather than emergency departments/hospitals.
  + Awareness was higher than some other Indian regions, possibly due to higher literacy, but still much lower than needed.
  + Limitations: hospital setting may overestimate awareness; multiple choice format may encourage guessing.

**Paper 9 :Non-cognitive Factors Influencing Early Stroke Symptom Recognition Among Korean Working-class Males with Hypertension and Diabetes (2025)**

* **Nature**: National health survey data, machine learning and regression analysis (4125 working-class males with diabetes)
* **Findings**:
  + 72% of participants could recognize early stroke symptoms.
  + Key predictors of poor recognition: younger age, high BMI, hypertension, elevated cholesterol/triglycerides, depression, stress, smoking, alcohol use.
  + Positive predictors: regular physical activity, participation in diabetes education.
  + Blue-collar workers less likely to recognize symptoms than managers/professionals.
  + Gradient boosting model achieved high accuracy (~85%), feature importance: age, depression/stress, BMI, hypertension, lifestyle factors.
* **Gaps**:
  + Limited to male workers, restricting generalizability.
  + Cross-sectional, cannot prove causality.
  + Reliance on self-reported knowledge, possible bias.
  + No interaction analysis for combined effects of factors.
  + Lack of longitudinal data and external validation.