Mild Stroke Detection

Si Min, Hongwei, Irfan
DSI-42

Table of Contents

About Stroke & TIA

01

Stroke is a leading cause of death

Every 40s

Someone has a stroke

In the United States

Every 3min 14s

Someone dies of a stroke

In the United States

Source: cdc.gov

Mild Strokes (TIA) foreshadow Full-blown Strokes

~240,000

people in the United States experience a TIA every year

~1 in 5

people who have a suspected TIA **will have a stroke within 90 days**

Symptoms

can mimic other neurological symptoms, so it's best to get a detailed evaluation

TIA – Transient Ischemic Attack

Source: stroke.org

Symptoms don't point to one conclusion

~240,000
people in the United States experience a TIA every year

~1 in 5
people who have a suspected TIA **will have a stroke within 90 days**

Symptoms

- Partial numbness/ paralysis
- Slurred speech
- Blurred vision
- Dizziness or headache

Source: stroke.org

Severity of consequences differ greatly

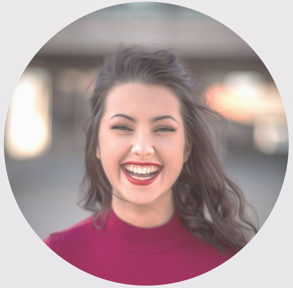
TEMPORARY SYMPTOMS

- Partial numbness/ paralysis
- Slurred speech
- Blurred vision
- Dizziness or headache

WHAT ELSE COULD IT BE?	TREATMENT PROCESS
Migraines	Painkillers
Low Blood Sugar	Dietary adjustments, Urinalysis
High Blood Sugar	Dietary adjustments, Urinalysis
Pinched Nerve	CT scans, neurology tests
Anxiety or Panic attacks	Psychiatric referral

Source: nih.gov, cdc.gov, yalemedicine.org, healthline.com, healthline.com

Profile



Nancy Jones

Age: 30

Occupation: Business Development Manager

A week ago, Nancy **experienced partial numbness**. In a state of panic her family did some brief online research and **suspected that she was experiencing a TIA**. Worried about the life-threatening consequences, they rushed her to the ER.

She was informed that she had to **wait for 4 to 6 hours** before a doctor assesses her situation. While waiting, she continued reading up online on her symptoms and realised that there **could be multiple causes to her symptoms.** If it was a TIA, she would have required more immediate medical intervention.

The long wait **drove her anxiety further** as she impatiently waited in a hospital feeling uncertain of her diagnosis.

Should she leave and visit a GP instead?
Or stay and wait hours to get diagnosed in a hospital?
Would anyone even be able to give her an immediate diagnosis?

She understands that the protocol to immediately visit the ER in such cases is to account for the worst-case scenarios, but she still **wishes that there was a way to tell if it was a TIA** or not, to quell her concerns.

Problem Statement

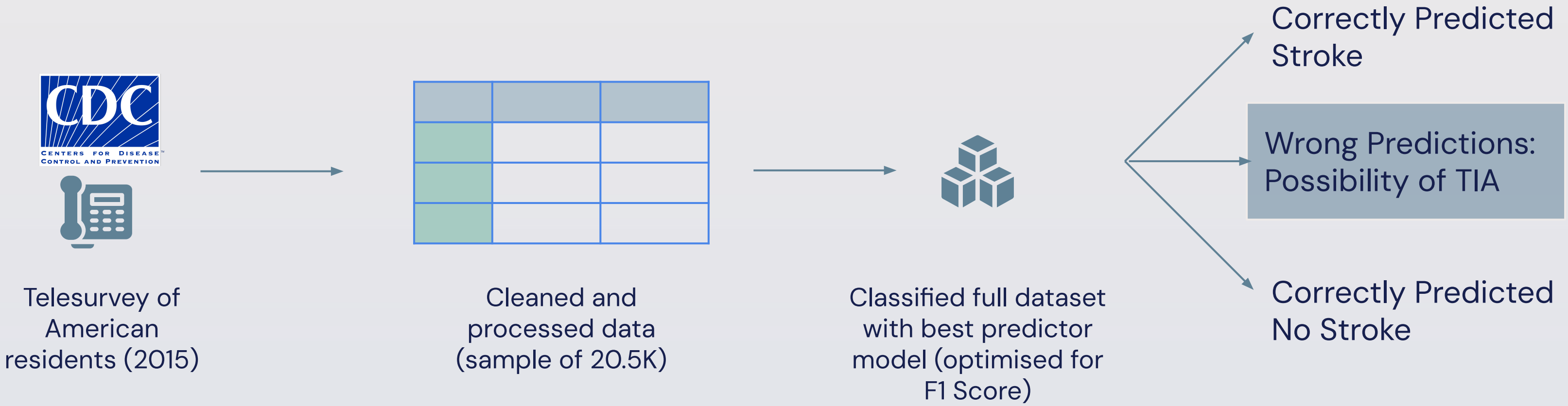
02

How might we help patients confidently assess if they experienced a TIA, using a precise and sensitive classification model?

Our Methodology

03

Step 1:
Supervised Learning – Binary Classification



Data retrieved from: kaggle

Step 2:
Unsupervised Learning: Clustering

Our Hypothesis for wrong classification (in this context):

Model Error	TIA
- Model did not perfectly predict stroke/no stroke due to random error - Actual values are accurate	- Model's inaccuracy due to a third class (TIA) not being accounted for - Actual values are inaccurate

Stroke

No Stroke

Mild Stroke

Step 3:

Supervised Learning: Multi-Classification Model

Train

Test

STROKE

NO STROKE

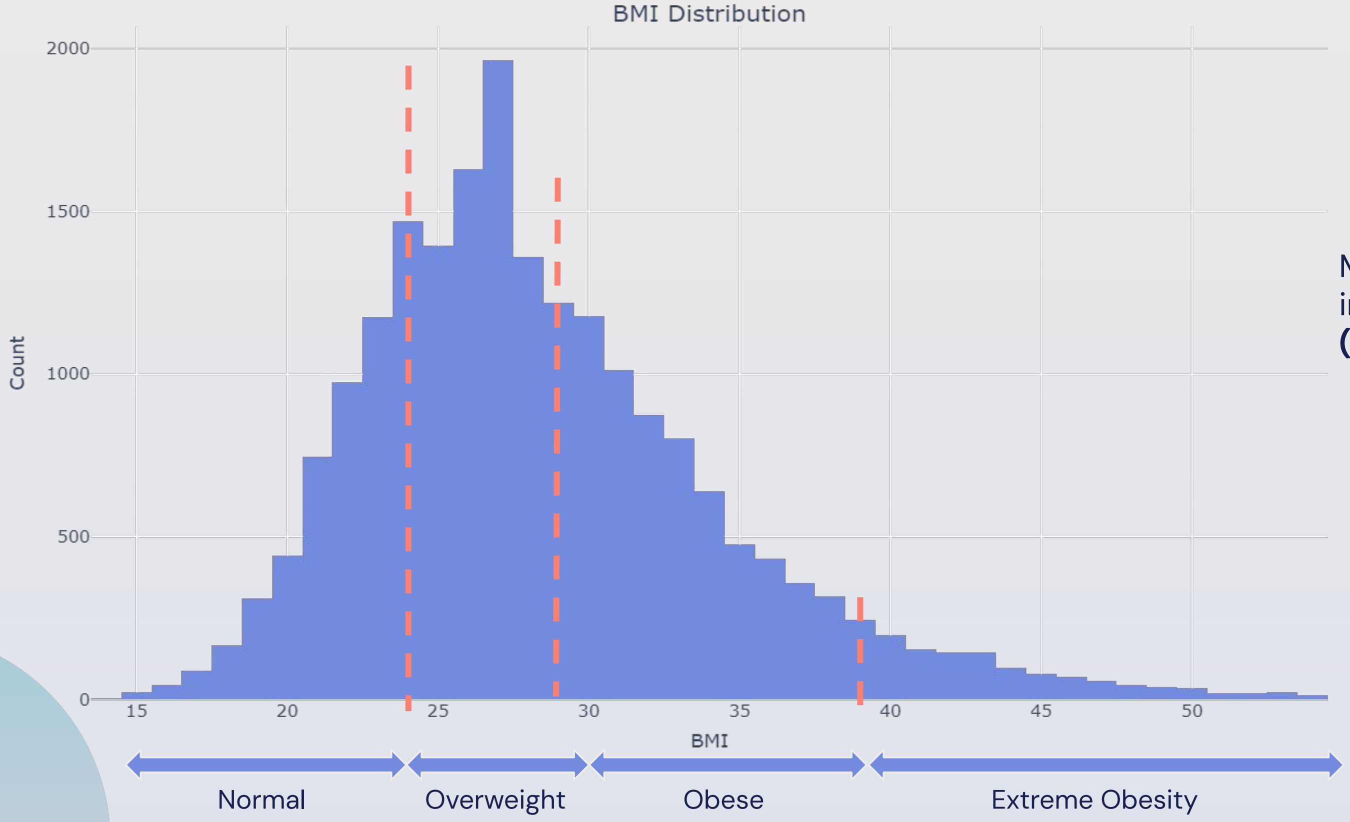
MILD STROKE

Multi-Class Predictor

Initial Findings

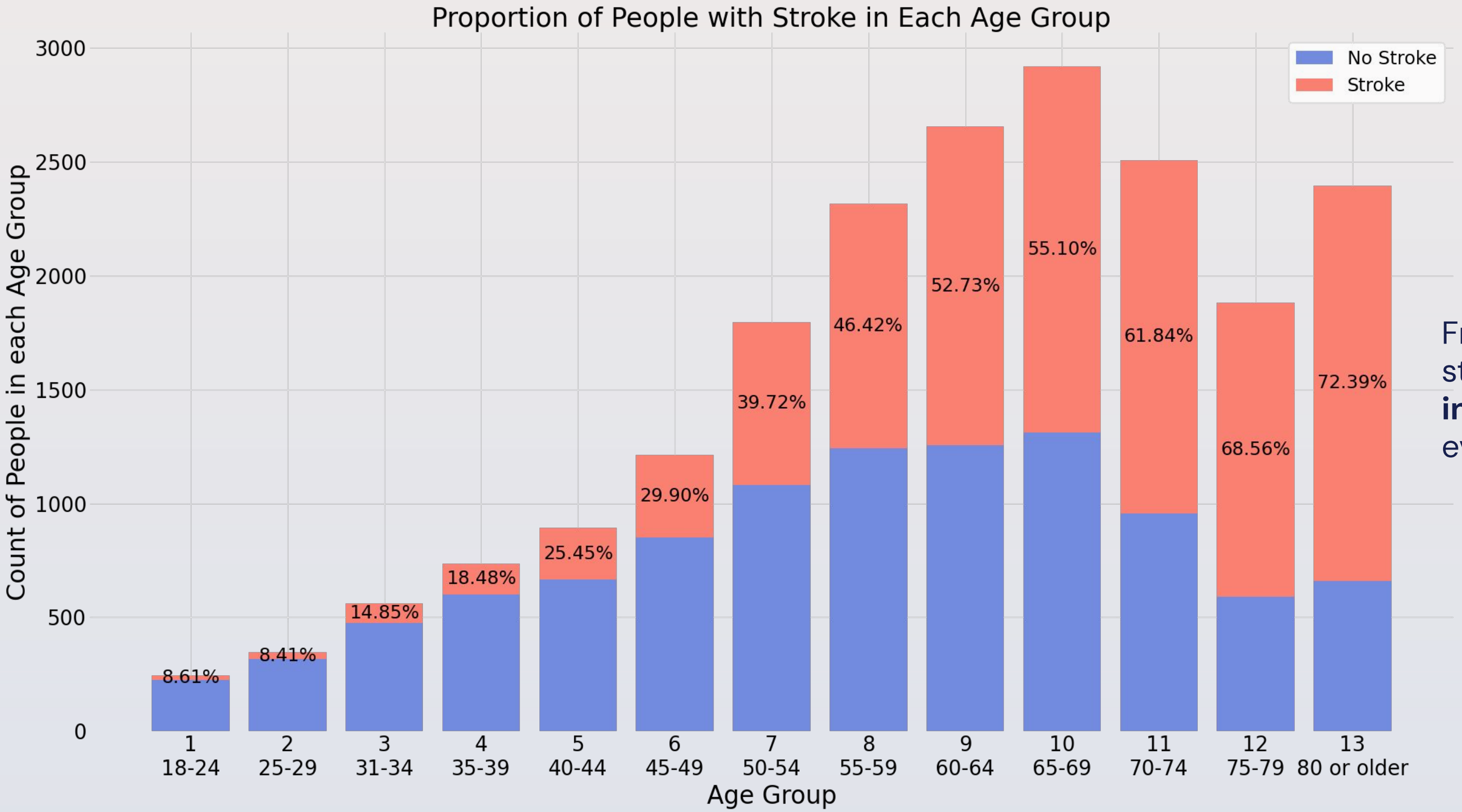
04

BMI of Sample Population is normally distributed



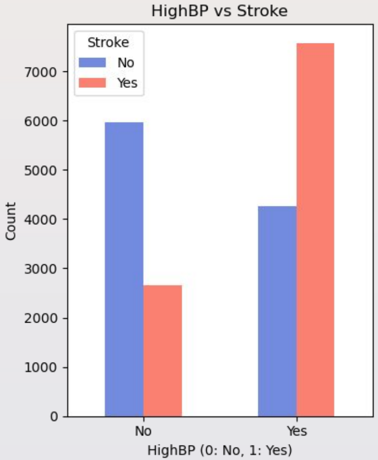
Most of the people are in the **overweight (24-29) category**

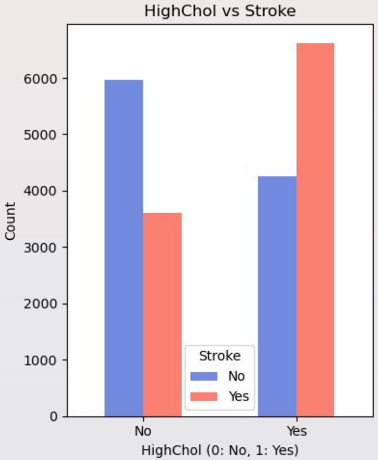
Stroke occurrence increases with Age

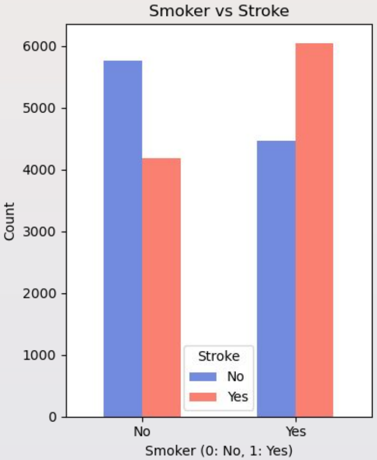


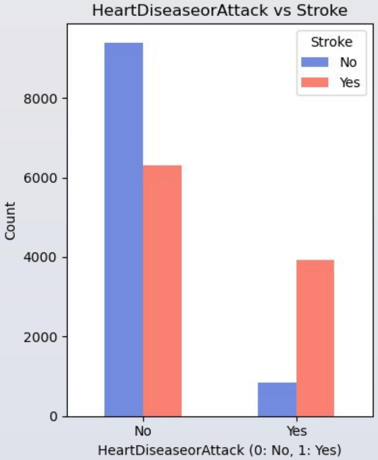
From Age 50, stroke occurrence **increases by 10%** every 10 years

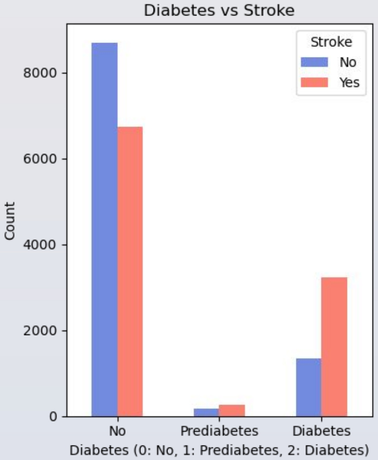
High Risk Factors Relating to Stroke











Presence of these factors increases chance of stroke

- High Blood Pressure
- High Cholesterol
- Smoker
- Heart Disease or Attack
- Diabetes

Modelling & Evaluation

05

RECAP: Our Hypothesis

Using the most optimised binary classification model, we hypothesise that the **cases of TIA** belong to the set of **data that the model falsely predicted**.

Metric Optimisation

Precision

Minimize False Positives:

- Reduce wrong predictions of stroke

Consequences:

- Inefficient use of medical resources and patient's time and money

✖

Sensitivity

Minimize False Negatives:

- Reduce wrong predictions of not having a stroke

Consequences:

- Loss of lives

➡

F1 Score

Aim for a balance between precision and sensitivity.

Possibility of TIA in both False Positive and False Negative predictions.

Disclaimer: This is not the actual formula for F1 Score, relationship for illustration purposes only.

Base model F1 Scores

Classification Models	Train F1-Score	Test F1-Score
Decision Tree	0.656425	0.662507
Random Forest	0.753842	0.760998
Bagging	0.705192	0.703561
Adaboost	0.748190	0.754440
Support Vector	0.755588	0.765905
Gradient Boost	0.761466	0.767612

Criteria for selection:

- Generalisable
  - Train and Test F1 scores need to be balanced
- Predictable
  - F1 scores across both are high

Selected 3 Models to Hypertune

Models	Train F1-Score	Test F1-Score
Decision Tree Classifier	0.656425	0.662507
Random Forest Classifier	0.753842	0.760998
Bagging Classifier	0.705192	0.703561
Adaboost Classifier	0.748190	0.754440
Support Vector Classifier	0.755588	0.765905
Gradient Boost Classifier	0.761466	0.767612

Selection Evaluation

- Generalisable
  - All models have a balanced F1 score across both sets
- Predictable
  - F1 scores across both train and test sets are highest

Hypertuned Results

Models	Train F1-Score	Test F1-Score
Random Forest Classifier	0.753842	0.760998
Random Forest Classifier (Tuned)		
Support Vector Classifier	0.755588	0.765905
Support Vector Classifier (Tuned)		
Gradient Boost Classifier	0.761466	0.767612
Gradient Boost Classifier (Tuned)		

Hypertuned Results

Models	Train F1-Score	Test F1-Score
Random Forest Classifier	0.753842	0.760998
Random Forest Classifier (Tuned)	0.763569	0.769894
Support Vector Classifier	0.755588	0.765905
Support Vector Classifier (Tuned)	0.764362	0.764184
Gradient Boost Classifier	0.761466	0.767612
Gradient Boost Classifier (Tuned)	0.766811	0.768999

Hypertuned Results

Models	Train F1-Score	Test F1-Score
Random Forest Classifier	0.753842	0.760998
Random Forest Classifier (Tuned)	0.763569	0.769894
Support Vector Classifier	0.755588	0.765905
Support Vector Classifier (Tuned)	0.764362	0.764184
Gradient Boost Classifier	0.761466	0.767612
Gradient Boost Classifier (Tuned)	0.766811	0.768999

Selected Model

To verify the predictability across the entire set, we run the best 2 models against the full dataset.
The **model which has the best F1-Score across the entire dataset** will be our selected classifier

Models	Train F1-Score	Test F1-Score	Full Dataset F1-Score
Random Forest Classifier (Tuned)	0.763686	0.770700	0.765165
Gradient Boost Classifier (Tuned)	0.766811	0.768999	**0.767361**

Classification

True Positives (Correctly predicted stroke) **7421**	**False Negatives** (Predicted no stroke, experienced stroke) **2808**
False Positives (Predicted stroke, never experienced stroke) **2113**	**True Negatives** (Correctly predicted no stroke) **8116**

Data for Clustering

True Positives (Correctly predicted stroke) **7421**	**False Negatives** (Predicted no stroke, experienced stroke) **2808**
False Positives (Predicted stroke, never experienced stroke) **2113**	**True Negatives** (Correctly predicted no stroke) **8116**

Unsupervised Learning: Identifying TIA

06

Three Clustering Algorithms

* Two selected

KMeans Clustering

Divides data into k clusters with minimal variance within each

- Effective for spherical clusters
- Sensitive to placement of the mean data point

Hierarchical Clustering

Builds a tree of clusters where each branch represents a cluster

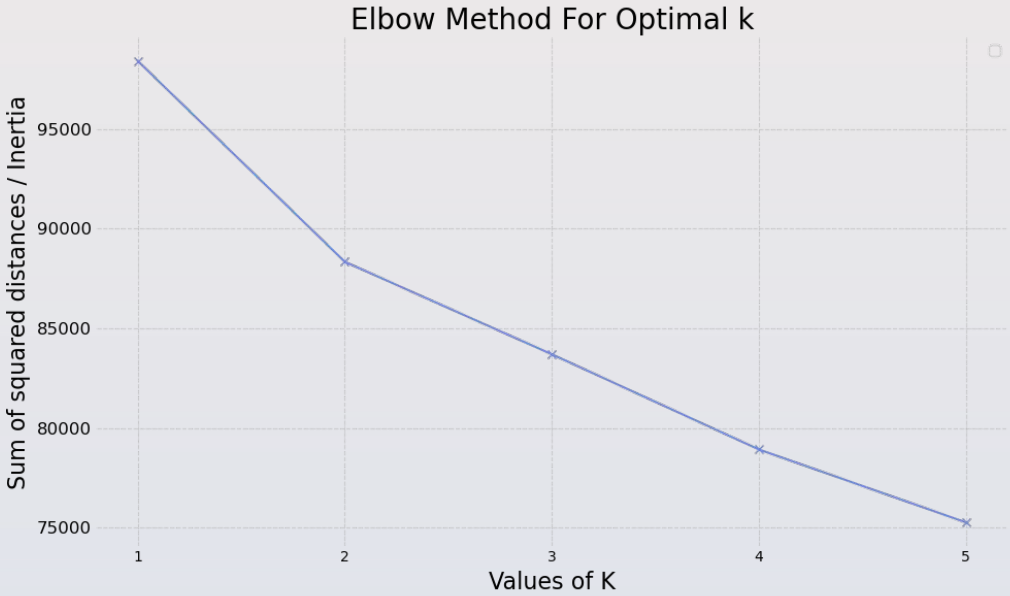
- Offers insights at different levels of granularity
- Computationally expensive

Density-Based Clustering

Identifies clusters based on density

- Capable of detecting outliers
- Struggles with datasets of high dimensionality

K-Means Clustering



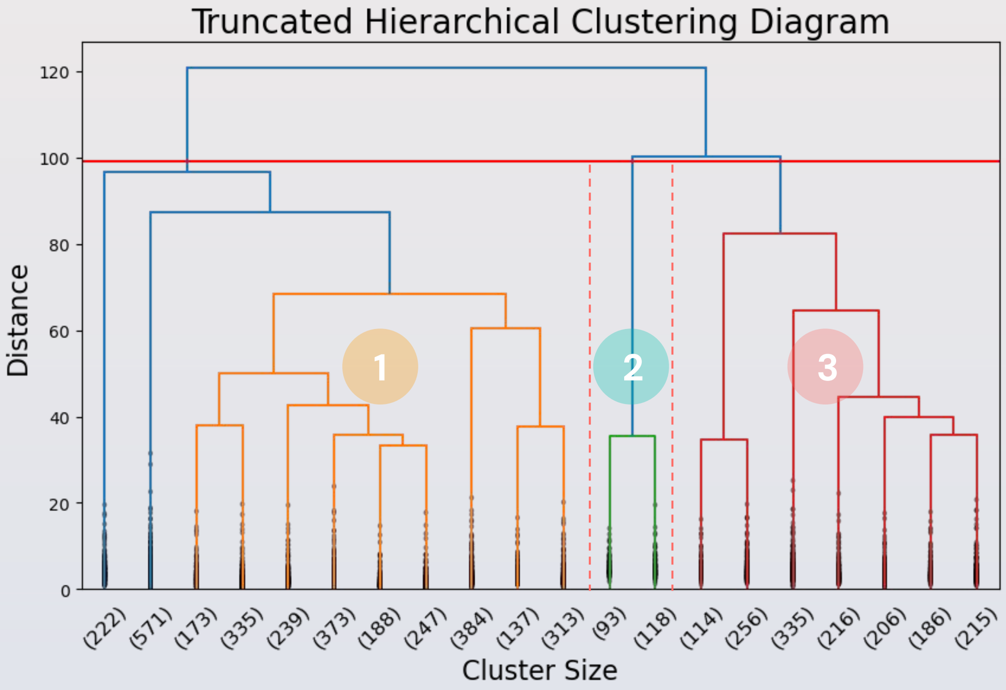
The sum of squared distance shows the variance of each cluster at k clusters

Evaluation:

No clear elbow at k=3

- Variance within each of the 3 clusters is high
- Between 3-4 clusters, the variance reduces as much as between 2-3

Hierarchical Clustering



Each line represents the distance between 2 subclusters forming.

The 3 clusters are separated by dotted lines.

Evaluation:

- Cluster 1: Datapoints are sparse, contains 2 outlier groups
- Cluster 2: Distinct cluster
- Cluster 3: Datapoints are sparse but less so compared to cluster 1

Comparing Silhouette Scores

Silhouette scores range from -1 to 1, with 1 being a perfect clustering.

Clustering Algorithm	Silhouette Score
KMeans Clustering	0.142581
Hierarchical Clustering	0.111030

Choosing Key Features only

- We researched and selected key features that were risk factors to having a stroke
- Created a new dataframe to perform unsupervised clustering

Columns	highbp	highchol	heartdisease /attack	diabetes	BMI
Correlation to stroke	0.33	0.23	0.36	0.23	0.05

KMeans Clustering:
Silhouette score: +0.0006

Hierarchical Clustering:
Silhouette score: No improvement

Reducing to just 5 features
has **no significant impact** on the clustering algorithm

Learnings & Cost-Benefit Analysis

07

Limitations

Binary Data

- Blood Pressure
- Cholesterol
- Heart Disease
- Heavy Alcohol Consumption

Too many binary features **limit performance** of clustering algorithms

Limitations

Binary Data

- Blood Pressure
- Cholesterol
- Heart Disease
- Heavy Alcohol Consumption

Too many binary features **limit performance** of clustering algorithms

Data Collection

In **tele-surveys**, it is **not easy to get continuous variables** for the aforementioned features (*e.g.* asking the surveyee to perform a blood pressure test at home, obtaining their blood cholesterol levels)

Limitations

Binary Data

- Blood Pressure
- Cholesterol
- Heart Disease
- Heavy Alcohol Consumption

Too many binary features **limit performance** of clustering algorithms

Data Collection

In **tele-surveys**, it is **not easy to get continuous variables** for the aforementioned features (*e.g.* asking the surveyee to perform a blood pressure test at home, obtaining their blood cholesterol levels)

Data Formats

Deeper analysis can’t be done on stroke/TIA patients.

With **images** such as **MRI scans** we can conduct deeper analysis

Recommendations

~~Binary~~ Continuous Data

Collect data on patients who have stroke and no stroke, with **continuous variables for key features** such as systolic/diastolic values for blood pressure, and mg/dL values for cholesterol levels.

Recommendations

~~Binary~~ Continuous Data

Collect data on patients who have stroke and no stroke, with **continuous variables for key features** such as systolic/diastolic values for blood pressure, and mg/dL values for cholesterol levels.

Data Collection

These data could be **collected in clinics in presence of medical professionals** and stored when there are stroke or TIA-like symptoms

Recommendations

~~Binary~~ Continuous Data

Collect data on patients who have stroke and no stroke, with **continuous variables for key features** such as systolic/diastolic values for blood pressure, and mg/dL values for cholesterol levels.

Data Collection

These data could be **collected in clinics in presence of medical professionals** and stored when there are stroke or TIA-like symptoms

Data Formats

Process data by **identifying surveyee** to their **MRI scans or carotid ultrasound scans** (if available) and execute **image recognition** prediction models

Cost-Benefit Analysis

While we were not able to create a model that helps to predict the likelihood of TIA, we believe that if we are able to overcome the limitations mentioned earlier, we will be able to create a model that meets our objective.

Most importantly, we need to understand that the cost of creating this model is minimal compared to the millions of lives saved in stroke prevention, and billions of dollars on our healthcare system.

Country Level

Potential Cost-Benefit Analysis **of Stroke Prevention**

Cost: Implementation and maintenance

~$60,000–$150,000

* Est. cost of single feature healthcare app development

Cost: Wrong prediction

~$600,000

* In 2016, a BP predictor app paid the FTC for publicising an inaccurate app

Benefit: Employment productivity

$68.5 billion

*** Indirect cost savings from underemployment and premature death**

Benefit: Medical Resource Allocation

$22.4 billion

*** Direct Cost Savings on Stroke Care**

Individual Level

Potential Cost-Benefit Analysis **of Stroke Prevention**

Cost: Personal Health Data Provision

Collect data for model training

Cost: If stroke occurs

an individual loses

2M brain cells/min

If left unattended/If stroke was not expected

Amount saved with stroke prevention

$3.86B per year from TIA patients

Aggregated cost

Psychological Benefits

- Improved brain health
- Reduces anxiety of diagnosis
- Certainty in treatment

Conclusion

08

Despite the limitations that posed as a barrier to reach our goal, we believe that **there is still value in creating a precise and sensitive multi-class predictor** to identify mild stroke as a means to prevent full stroke.

The **benefits majorly outweighs the cost**, both for our citizens and our country as a whole.

Thanks

Do reach out to us if you have any questions or wish to support our project

Appendix

Reducing Dimensionality for Clustering

1. Transforming data points to principal components

2. Using only key features

Principal Component Analysis

- We transformed our data into principal components to see if most of the variance in our data can be explained within 2-3 principal components
- The aim is to be able to visualise the data and cluster segregation.

Principal Components	1	2	3	4	5	6
Cumulative Explained Variance Ratio	0.15	0.27	0.31	0.37	0.43	0.49

Only 30% of the variance in data can be explained with the **first 3 principal components**

Cost-Benefit Analysis

Country level -
Cost:

1. How much does it cost the government to implement this? (Depends on how much govt want to pay us for the app)
2. How much does it cost to maintain this? (salary of data scientists, salary of software engineer maintaining the app)
3. What happens if it is a wrong prediction?

Benefits:

1. How much does the country save from this? (**https://doi.org/10.1016/j.jns.2019.116643)**
   Based on salary difference, missed workdays, and mortality, indirect cost from under-employment was $38.1 billion and from premature mortality was $30.4 billion.
2. How does this improve our economy?
3.

Cost-Benefit Analysis

Individual level –

Cost:

1. How much does it cost for individuals to use this technology? (We'll be collecting more data from users to improve model accuracy)
2. What happens to the individual if it is a wrong prediction? (If a stroke is untreated for the full 10 hours, the brain ages up to 36 years! With every minute you wait, the brain loses two million brain cells)

Benefits:

1. How much money does an individual save? (USD 23k)
2. How much time does an individual save?
3. What troubles (mental/physical) do we help to keep them from?
   (The rewards of successfully making these changes are great, not only in stroke prevention, but in improving overall brain health, and preventing cognitive decline and allowing patients to remain independent and productive.) link

What happens to the individual if it is a wrong prediction?

Risk of Stroke After TIA

- **20% Chance:** Within 90 days of a TIA, there's a significant risk of a full stroke.
- **Preventive Treatment:** Early intervention for TIA can significantly reduce this risk.

Devastating Impacts of Stroke (costs):

- Speech/language problems
- Vision problems
- Slow, cautious / Quick, inquisitive behavioral style
- Memory loss
- Paralysis on one side of the body
- Death

Treatment for TIA:

- Anti-platelet drugs - make platelets less likely to stick together
- Anticoagulants - lower the risk of blood clots by affecting clotting-system proteins

Other than death, the impacts of stroke bring huge daily inconveniences to the basic lifestyle of a patient. Daily tasks such as getting around, cooking and bathing may be more difficult than before.

Thus, these highlight the benefits of seeking timely preventive treatment for TIA

Source: stroke.org

Calculated by 20% X 800,000

~160,000

Experience a TIA every year (USA)

If these patients continued to get full stroke:

~$23K / patient

Cost of stroke treatment

~$3.68 Billion / year

Total healthcare spendings by TIA patients who experienced full stroke