|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Team\_8 members**   |  |  |  | | --- | --- | --- | | **Name** | **Major** | **student ID number** |  |  |  |  | | --- | --- | --- | | **ANH SUBIN** | **Ai/Software** | **(202334305)** | | **Seo Hyungyo** | **Ai/Software** | **(202334298)** | | **KIM WONJUN** | **Ai/Software** | **(202239866)** | | **CHAE HEEJAE** | **Ai/Software** | **(202135585)** | |

* ***GitHub URL***

<https://github.com/datascienceteam8>

* *dataset URL*

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

* **index explanation ) *Title: Team member introduction, team name, etc.***

***Chapter 1 : Common Project Objective and Motivation***

***Chapter 2 : Individual Contributions and Detailed Tasks***

***Chapter 3 : Sources and References***

***ANH SUBIN (202334305)***

* ***Implemented classification and clustering algorithms (e.g., Decision Tree, K-Means)***
* ***Focused on identifying stroke risk factors and grouping similar individuals based on health attributes***

***SEO HYUNGYO (202334298)***

* ***Designed and documented the data preprocessing pipeline***
* ***Handled missing data, encoding of categorical variables, and feature normalization***

***KIM WONJUN (202239866)***

* ***Wrote the background and outlined the clinical and social significance of the project***
* ***Conducted comparative experiments using multiple classification models (e.g., Logistic Regression, Random Forest, XGBoost)***

***CHAE HEEJAE (202135585)***

* + ***Conducted supplementary research and filled in gaps in insufficient source materials***
  + ***Verified data and supported the validation of key information***
  + ***Contributed to the preparation of the team project report and presentation slides***
  + ***Managed project organization tools such as GitHub for version control and collaboration***

***ANH SUBIN 1)***

***Statistical Description of Dataset***

The dataset comprises **5,110 observations** with **12 features**, categorized as follows:

***1. Numerical Features (7 variables)***

**id**: Unique identifier

**age**: Patient age (years)

**hypertension**: Hypertension status (0: Absent, 1: Present)

**heart\_disease**: Heart disease status (0: Absent, 1: Present)

**avg\_glucose\_level**: Average blood glucose level (mg/dL)

**bmi**: Body Mass Index (obesity metric)

**stroke**: Stroke occurrence (0: No, 1: Yes, target variable)

***2. Categorical Features (5 variables)***

**gender**: Male / Female / Other

**ever\_married**: Marital status (Yes / No)

**work\_type**: Occupation type (Private / Self-employed / Government job / Children / Never worked)

**Residence\_type**: Urban / Rural

**smoking\_status**: formerly smoked / never smoked / smokes / Unknown  
Categorical variables will be used to differentiate lifestyle characteristics during clustering.

***Project Objectives***

This project aims to predict stroke risk and develop personalized prevention strategies using health data analytics.

***1. Early Detection of High-Risk Factors***

Identify key health indicators correlated with stroke (e.g., hypertension, advanced age, elevated blood glucose).

Analyze feature importance using SHAP values and correlation matrices.

***2. Cluster-Wise Feature Analysis***

Group individuals with similar health profiles via clustering algorithms (K-means/DBSCAN).

Compare lifestyle patterns (e.g., smoking rates, BMI distributions) across clusters.

***3. Targeted Prevention Strategy***

Design risk-tiered interventions:

**High-risk cluster**: Mandatory MRI screenings + antihypertensive therapy.

**Moderate-risk cluster**: Biometric monitoring every 3 months.

**Low-risk cluster**: Lifestyle modification programs (diet/exercise plans).

Example: Customized guidelines for the "smoking + high blood sugar + middle-aged males" subgroup.

***Seo Hyungyo 1)***

***For Clinical ,***

***Social Impact***

***Early stroke prevention\_saves lives.***

**Stroke is an unpredictable and life-threatening condition, often causing severe consequences if not promptly addressed. Early detection and prevention are crucial not only for improving individual health outcomes but also for alleviating the strain on healthcare systems and reducing social burdens. By focusing on early intervention, we can mitigate the long-term costs associated with stroke care, improving both personal and public health.**

***Project Objectives***

**Predict Early Stroke Risk and Segment the Population into Actionable Risk Groups**

**The goal is to predict stroke risk early and segment the population into actionable risk groups, enabling more effective and personalized prevention strategies.**

**By employing machine learning algorithms, we can assess each individual's stroke risk and group people with similar characteristics together. This clustering method allows for more precise interventions, such as targeted screenings, personalized counseling, and guidance on lifestyle modifications tailored to each group's specific health needs.**

***Build a Classifier to Identify High-Risk Individuals***

**A predictive classifier will be built to identify individuals at high risk for stroke.**

**Using advanced modeling techniques, such as logistic regression, random forests, and XGBoost, this model will help detect individuals who are most likely to experience a stroke. It will allow healthcare professionals to assess risk quickly and effectively with basic health data, improving early intervention and potentially saving lives by identifying at-risk individuals in a timely manner.**

***Enable Clinicians and Policymakers to Allocate Screening Resources More Efficiently***

**This project will provide healthcare professionals and policymakers with the tools to allocate screening resources more efficiently and strategically.**

**By clearly identifying high-risk individuals, resources can be directed to those who need them most, optimizing the use of limited public healthcare resources. This targeted approach not only improves the efficiency of large-scale health screenings but also contributes to cost savings, better allocation of medical personnel, and improved diagnostic accuracy, leading to more effective stroke prevention and management.**

* ***Checking References2***

***KIM WONJUN 1)***

***backGround***

Stroke is one of the leading causes of death worldwide and a dangerous disease that occurs suddenly, causing severe nerve damage and disability. Since brain cells are difficult to recover once damaged, prompt emergency treatment upon stroke onset is crucial for the patient's survival and recovery. Furthermore, as stroke has a high risk of recurrence, continuous risk factor management and preventive efforts are essential for long-term well-being. Preventive interventions can significantly reduce the risk of a first stroke and minimize the likelihood of recurrence.

***Motivation***

Stroke is characterized by its unpredictability and sudden onset. This makes it difficult for patients to receive timely and appropriate treatment, negatively impacting their prognosis and quality of life. However, by systematically managing well-known risk factors such as high blood pressure, smoking, diabetes, and obesity, the risk of stroke occurrence and recurrence can be effectively reduced.

Recently, vast amounts of medical data, including individuals' health information such as age, hypertension status, and blood sugar levels, have been accumulated. Data analysis and artificial intelligence techniques to analyze this information have been rapidly advancing. This technological progress now allows for more precise, data-driven predictions of stroke risk, enabling personalized prevention and management strategies that are tailored to individual health profiles.

***Objective***

The objective of this project is to analyze individual health data to predict the risk of stroke occurrence and identify groups of individuals with similar health risk characteristics. To achieve this, two main tasks, Classification and Clustering, will be performed.

**Classification**: Develop a model to predict whether an individual is at high risk for stroke.  
**Clustering**: Identify groups with similar health statuses to develop tailored prevention strategies based on specific characteristics.

***ANH SUBIN 2)***

***Algorithms to be Used***

1. **Decision Tree (Classification)**

**Objective**: Predict stroke (0: No stroke, 1: Stroke)

**Learning Method**: Supervised learning

**Key Features**:

Use health information (age, hypertension, heart disease, average glucose level, BMI) to predict stroke occurrence

Split the model based on features in a tree structure

Analyze feature importance to determine which factors strongly influence stroke prediction

1. **K-Means (Clustering)**

**Objective**: Group individuals based on similar health information

**Learning Method**: Unsupervised learning

**Key Features**:

Cluster individuals without using the stroke feature

Optimize clusters by iterative centroid recalculation

Analyze the high-risk clusters based on stroke-related features

***Seo Hyungyo 2)***

**Dataset**

**Stroke Prediction Dataset (Kaggle, by Federico Soriano)**

* Size: 5,110 rows × 12 columns (11 predictors + 1 target)
* Target, stroke (0 = No stroke, 1 = Stroke)
* Licence: CC0 — free for academic & commercial use

**Why this dataset?**

* High clinical & social impact – early stroke prevention saves lives.
* Balanced mix of categorical (6) & numerical (5) variables
* Manageable size for rapid iteration
* Contains real‑world challenges (approx. 4 % positive class, missing BMI, class imbalance).
* CC0 license

**Plan(preprocessing)**

**1. Data Restructuring**

**\* Drop id column(non‑informative)**

**\* Data Value Changes**

**2-1. Impute missing values**

* Address the missing data in the BMI column (about 201 entries are NaN).
* Rather than removing these records, we will fill in missing BMI values with a representative statistic such as the median or mean BMI of the dataset(using pandas fillna).

**2-2. Clean and validate data entries**

* Verify categorical values are consistent (e.g., gender entries are "Male", "Female").
* Confirm data types are correct (e.g., ages are numeric, not strings) and convert types if needed.

**2-3. Handle outliers**

* Identify any extreme values that might skew the analysis.
* For instance, the maximum BMI is 97.6, which is unusually high.
* If a value is an obvious data error or overly influential, remove that record
* Otherwise, if we keep it, we may apply a robust scaling or transformation so that outliers do not affect the model.

**2-4. Normalize numeric features**

* We will apply -score normalization to features like age, avg\_glucose\_level, and BMI using scikit-learn’s StandardScaler.
* To ensure that differences in units or ranges (for example, age in years vs. glucose level) do not bias algorithms.

**2-5. Encode categorical variables**

* We will use one-hot encoding for nominal categories
* Expand work\_type, smoking\_status column into binary indicator columns.
* For binary categories like gender (Male/Female), ever\_married (Yes/No) or Residence\_type (Urban/Rural), map them to 0/1 directly.

**3. Feature Engineering**

* Create age\_group feature by binning the numeric age into categories (e.g. 0–18 = "Child", 19–60 = "Adult", 60+ = "Senior").
* This will help capture age-related risk patterns in a way that’s more interpretable.

**4. Data Reduction(Feature selection)**

* In this dataset, every remaining feature appears relevant (no constant or duplicate columns), so no further feature drop is needed.

**Role**

* Develop pre‑processing pipeline for the Stroke Prediction Dataset.
* Document decision (code comments + report).

***KIM WONJUN 2)***

## **Objective**

The objective of this project is to analyze individual health data to predict the risk of stroke occurrence and identify groups of individuals with similar health risk characteristics. To achieve this, two main tasks, Classification and Clustering, will be performed.

***Classification***

Using individual health data such as age, hypertension status, heart disease status, blood sugar levels, BMI, and smoking status, predict the occurrence of stroke.

Various classification algorithms, including Logistic Regression and Random Forest, will be applied to compare the prediction accuracy.

The main risk factors influencing stroke occurrence will be identified, and through the developed predictive model, high-risk individuals will be classified in advance to enable preventive interventions.

***Clustering***

Utilize personal health data to identify groups with similar health characteristics.

K-means clustering will be used to form the groups and analyze the health characteristic differences between clusters.

This will evaluate whether certain clusters represent high-risk groups for stroke and will help derive cluster-specific health risk profiles.

***Analysis of Clusters and Stroke Risk***

The relationship between the identified clusters and actual stroke occurrence will be carefully analyzed.

Stroke incidence rates will be calculated within each cluster, and the differences in incidence rates between clusters will be compared to assessing the stroke risk level for each cluster.

* ***Checking References1***

***Data***

|  |  |  |
| --- | --- | --- |
| **number** | **Feature** | **describe** |
| 1 | id | A variable that uniquely identifies an individual. |
| 2 | gender | Gender of the individual |
| 3 | age | Age of the individual |
| 4 | hypertension | Whether the individual has hypertension (0 = No, 1 = Yes). |
| 5 | heart\_disease | Whether the individual has heart disease (0 = No, 1 = Yes). |
| 6 | ever\_married | Whether the individual has ever been married (0 = No, 1 = Yes). |
| 7 | work\_type | Indicates a person's work type |
| 8 | residence\_type | Type of residence (Urban, Rural). |
| 9 | avg\_glucose\_level | Average glucose level (mg/dL). |
| 10 | bmi | BMI (kg/m²). |
| 11 | smoking\_status | |  | | --- | |  |  |  | | --- | | smoking status (formerly smoked, never smoked). | |
| 12 | stroke | Whether the individual has had a stroke (0 = No, 1 = Yes). |
|  |  |  |

Among these data, 5 features (age, avg\_glucose\_level, bmi, hypertension, and heart\_disease) were numerical data, while 7 features (gender, ever\_married, work\_type, residence\_type, smoking\_status, id, and stroke) were categorical data.

The data types were all integers except for avg\_glucose\_level and bmi, which were floats. There were null values in the data. Below are the descriptive statistics of the dataset.

***CHAE HEEJAE 1,2)***

***1. Project Purpose***

**This project aims to analyze individual health data to support early prediction and prevention of stroke.**

***2. Background of Stroke***

**Stroke is one of the leading causes of death worldwide.**

**It occurs suddenly and can cause severe neurological damage or permanent disability.**

**Damaged brain cells are difficult to recover, so early emergency treatment is critical.**

**Prevention is more effective than treatment after onset.**

***3. Motivation***

**Stroke’s unpredictability makes timely treatment difficult, worsening outcomes and quality of life.**

**Although risk factors like hypertension, smoking, and obesity are known, their interactions are complex.**

**A data-driven method is needed for precise risk assessment.**

***4. Technological Opportunities***

**The growth of medical data and digitized health records allows for advanced analysis.**

**Machine learning and AI can help identify high-risk individuals.**

**These tools enable personalized prevention strategies.**

**5. Project Objective**

**Build predictive models to estimate stroke risk.**

**Use classification to identify high-risk individuals.**

**Use clustering to group people with similar health patterns.**

**Develop group-specific prevention strategies.**

***6. Clinical & Public Health Impact***

**Enables early detection and personalized intervention.**

**Example: A cluster of middle-aged male smokers with high glucose can receive tailored guidance.**

**Helps allocate healthcare resources efficiently (e.g., targeted screening).**

**Reduces unnecessary costs and increases diagnostic efficiency.**

***7. Conclusion***

**The project offers a practical and scalable model for data-driven stroke prevention.**

**The approach can be adapted to other diseases requiring early detection and targeted care.**

***References1***

**1)** [Korea Disease Control and Prevention Agency (KDCA)](https://health.kdca.go.kr/healthinfo/biz/health/gnrlzHealthInfo/gnrlzHealthInfo/gnrlzHealthInfoView.do?cntnts_sn=5495) / **2)** [Asan Medical Center](https://www.amc.seoul.kr/asan/healthinfo/disease/diseaseDetail.do?contentId=30518) / **3)** [ScienceOn Report](https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=TRKO202300010450)

***References2***

* Stroke Prediction Dataset (Kaggle, by Federico Soriano)   
  <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
* Lecture Slides (by Kim & Loh) , Pre‑processing