King Saud
University

College of Computer and Information Science
Software Engineering Department

Machine Learning Project

STROKE
ANALYSIS AND PREDICTION

Group 4 | Section 56751
Date of Submission: 5 December 2023

Supervised by: L. Hailah AlMazrua

SWE485 Selected Topics in Software Engineering - 1st Semester 1445

Outline:

-1Problem Definition.

-2Data Set Overview.

-3Machine Learning Techniques:

a- Supervised learning.

b- Unsupervised learning.

-4Results and Findings Discussion.

Problem Definition

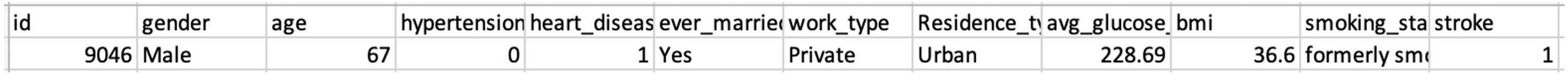
Problem Definition:

In this project, we develop and apply ML algorithms to create predictive models that identify individuals at risk of stroke, improve diagnostic accuracy. So we can positively impact stroke patients' lives by enabling earlier intervention and better healthcare decisions based on data-driven insights.

Data Set Overview

Data Set Overview:

The " ***Stroke Prediction Dataset*** "is a valuable resource available on Kaggle, a popular platform for data science and machine learning enthusiasts. This dataset is utilized to forecast the likelihood of a patient experiencing a stroke by considering input parameters such as gender, age, various diseases, and smoking status .

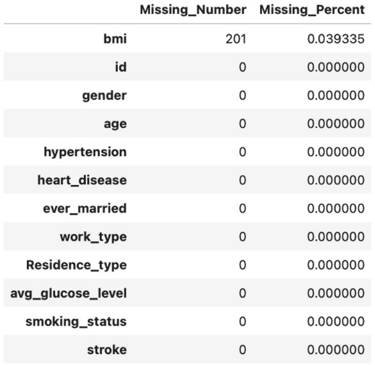


id	gender	age	hypertension	heart_diseas	ever_marrie	work_type	Residence_t	avg_glucose_	bmi	smoking_sta	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly sm	1

Data Set Overview:

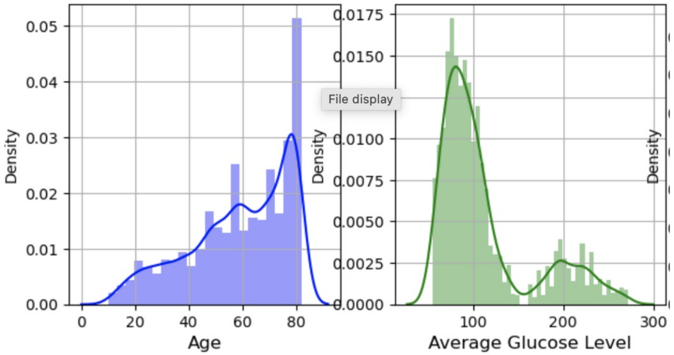
We encounter these problems in the data set :

Missed cells

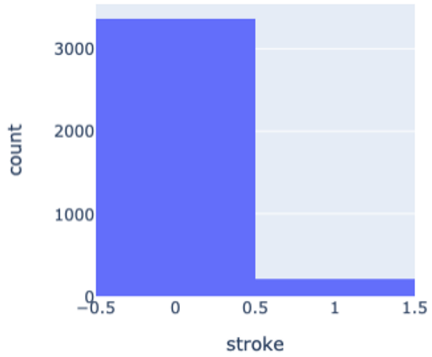


	Missing_Number	Missing_Percent
bmi	201	0.039335
id	0	0.000000
gender	0	0.000000
age	0	0.000000
hypertension	0	0.000000
heart_disease	0	0.000000
ever_married	0	0.000000
work_type	0	0.000000
Residence_type	0	0.000000
avg_glucose_level	0	0.000000
smoking_status	0	0.000000
stroke	0	0.000000

Skewed



Imbalance



Machine Learning Techniques

-Supervised learning

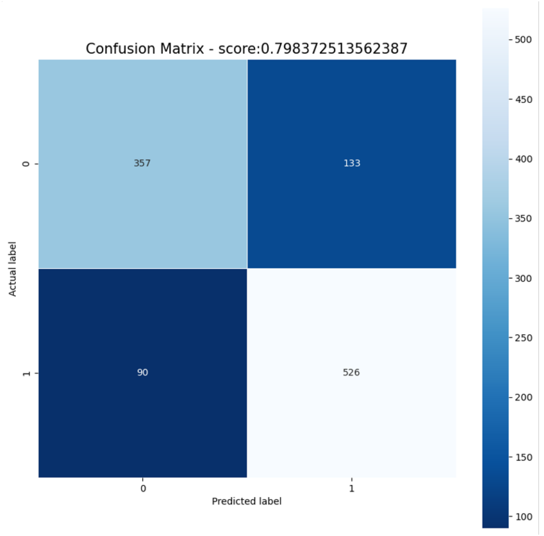
SWE485 Selected Topics in Software Engineering - 1st Semester 1445

-1Logistic Regression:

A logistic function is employed to model the relationship between the input features and the likelihood of the target variable belonging to one of the two categories. The logistic function transforms input values into a range between 0 and 1, which can be interpreted as the probability of the target variable being associated with one of the two specified categories.

Training score: 0.795

Test score0.802 :

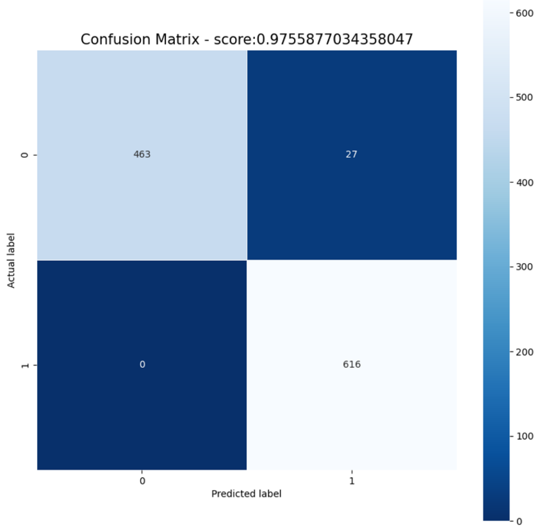


-2Decision Tree:

The decision tree algorithm constructs this tree-like model by iteratively partitioning the dataset into smaller subsets, guided by the values of the input features. At each node of the tree, a decision is made based on the specific feature's value, resulting in the data being divided into two or more subsets according to that feature. This process continues until a predefined stopping condition is met, such as reaching a maximum tree depth or having a minimum number of data points in a leaf node.

Training score: 1.000

Test score0.976 :

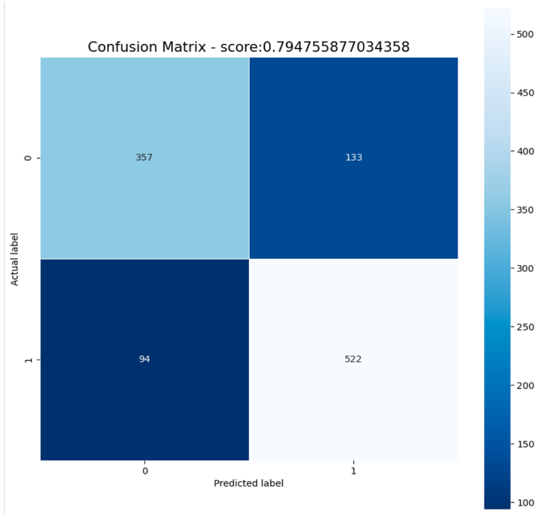

Confusion Matrix - score:0.9755877034358047

-3Support Vector Classifier:

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It is particularly effective for binary classification problems, where the goal is to separate data points into two classes. SVMs can also be extended for multi-class classification and regression tasks.

Training score: 0.796

Test score0.802 :



Model Comparison:

	Model	Accuracy
1	Logistic Regression	0.801989
2	Decision Tree	0.975588
3	Support Vector	0.801989

Machine Learning Techniques

-Unsupervised learning

SWE485 Selected Topics in Software Engineering - 1st Semester 1445

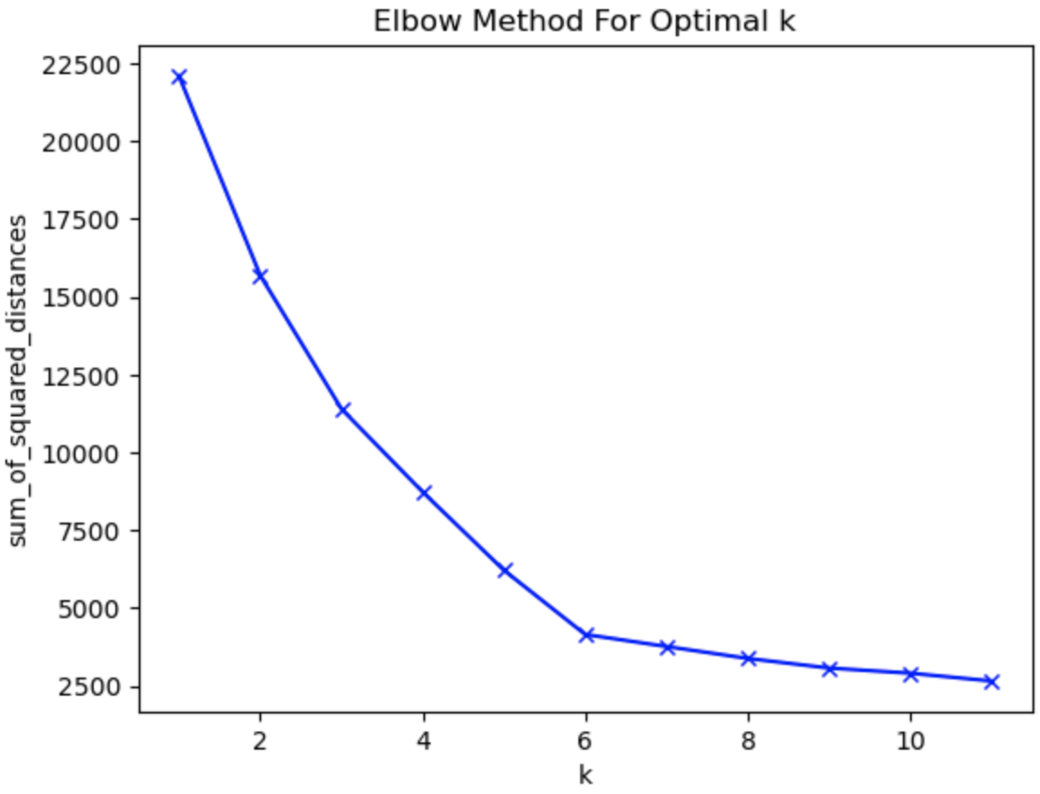
K-means:

K-Means is a clustering algorithm that groups similar data points together based on their features. It works by iteratively assigning data points to the cluster with the nearest centroid (mean of the points in the cluster) and updating the centroids. The algorithm continues this process until convergence. Users specify the desired number of clusters (K). K-Means is efficient for large datasets but assumes clusters of similar sizes and shapes. It's sensitive to initial centroid placement but remains widely used for various clustering applications.

First we have to drop the class attribute which is the
“Stroke”

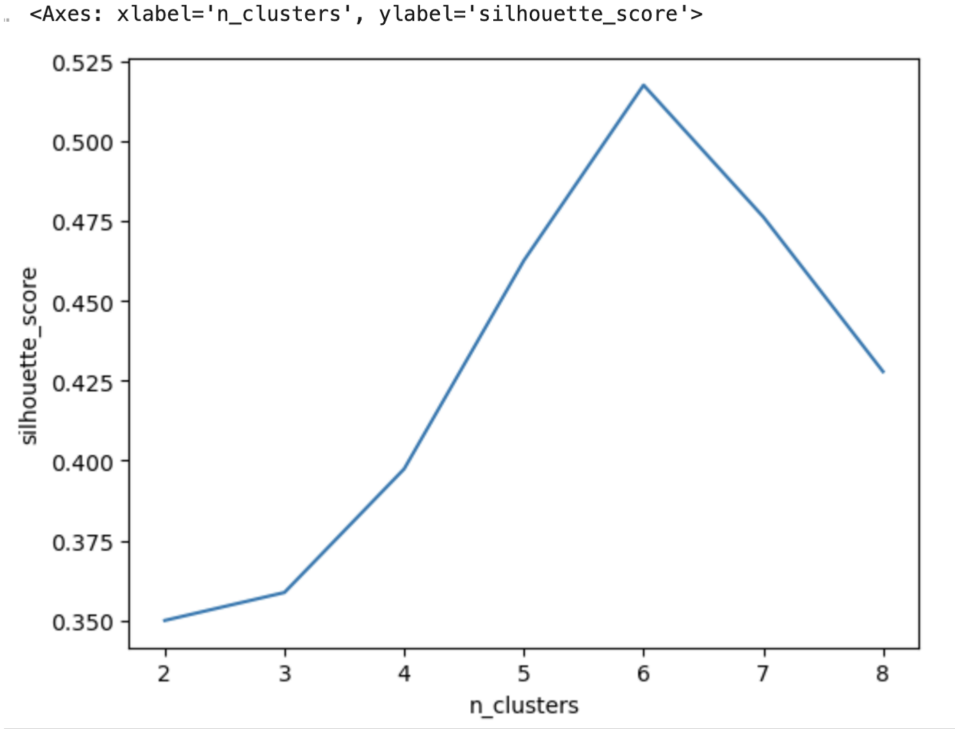
K-means Evaluation:

Sum of squared distances



K-means Evaluation:

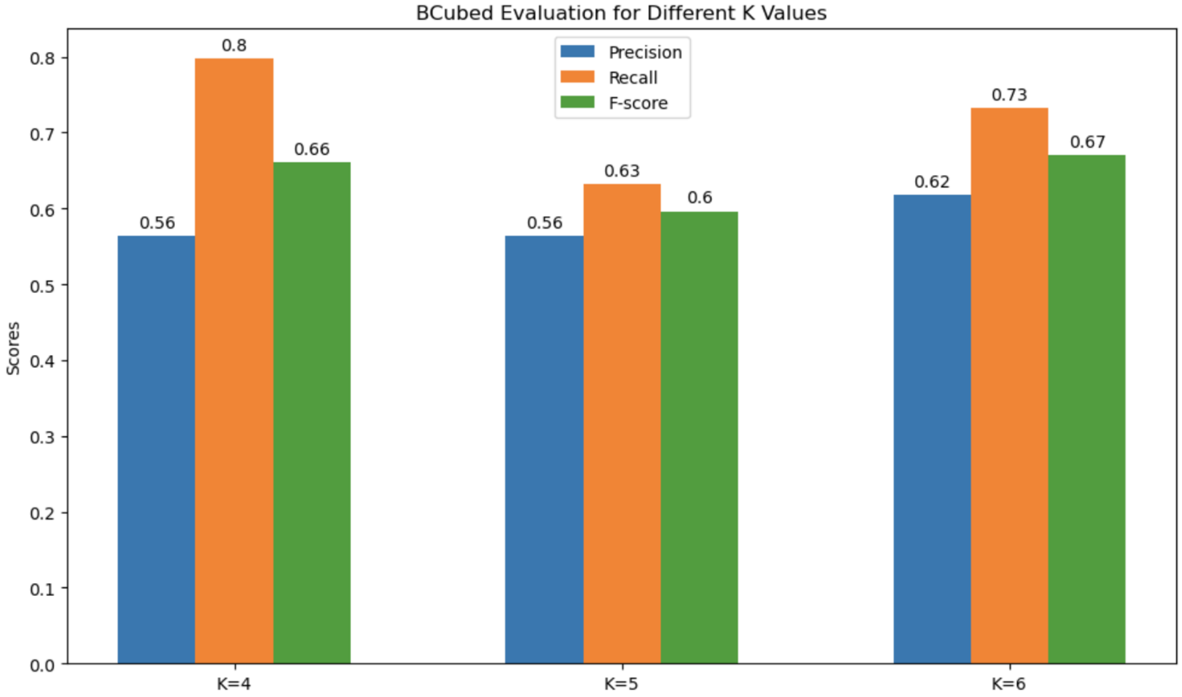
Silhouette coefficient



```
<Axes: xlabel='n_clusters', ylabel='silhouette_score'>
```

K-means Evaluation:

BCubed



Results and Findings Discussion..

Comparison:

Guidance:

- Supervised Learning: Guided by labeled examples to predict stroke occurrence.
- Unsupervised Learning: Self-guided clustering based on inherent patterns in the feature space.

Data Requirement:

- Supervised Learning: Requires labeled data indicating stroke occurrence.
- Unsupervised Learning: Utilizes unlabeled data for clustering.

Applications:

- Supervised Learning: Predicting stroke occurrence for individual patients.
- Unsupervised Learning: Exploring data structure, identifying groups of similar individuals.

Output:

- Supervised Learning: Predicted labels (stroke or no stroke) for individual instances.
- Unsupervised Learning: Cluster assignments, indicating groups of similar individuals.
- investigation or feature engineering.

Comparison:

- Supervised Learning: Best suited for predictive modeling and making individualized predictions, such as identifying individuals at risk of stroke.
- Unsupervised Learning: Useful for exploratory analysis, discovering patterns, and subgroup identification. While not directly involved in prediction, it provides insights that can inform further analysis.

In our stroke prediction project, both aspects are valuable. The supervised learning part is crucial for making predictions, while the unsupervised learning (k-means clustering) contributes to exploratory analysis and potential subgroup identification. Consider the specific objectives of your project and how each approach aligns with those goals. Often, a combination of both techniques can provide a more holistic view of the data.

Future Work:

Data Augmentation:

- Future research could involve data augmentation techniques to address the small sample size limitation. Incorporating additional datasets or synthetic data may enhance the robustness of predictive models.

Longitudinal Studies:

- Conducting longitudinal studies could provide insights into the dynamic nature of stroke risk factors over time. This would allow for a more comprehensive understanding of the temporal evolution of risk.

Explainability and Trust:

- Developing more interpretable models or model-agnostic techniques could enhance trust among healthcare professionals and patients. Clear model explanations foster better adoption in real-world clinical settings.

Practical Implications:

Patient Risk Profiling:

- The findings of our study offer practical applications for patient risk profiling. Healthcare providers can utilize the predictive models to identify individuals at higher risk of stroke, enabling targeted interventions and personalized preventive measures.

Healthcare Resource Allocation:

- Allocating healthcare resources efficiently is vital. By identifying high-risk individuals, resources can be directed toward timely interventions, potentially reducing the overall burden of stroke-related healthcare.

Thanks for listening ..

#	Name	ID
1	Nouf Alkhashan(*Leader*(	442201351
2	Sarah Alkhuraiji	442202359
3	Hoor Rammal	442201870
4	Malath AlSaif	442200413
5	Shaden AlAbdulrazaq	442200423
6	Leenah AlMashari	442200896

GitHub link :https://github.com/Noufalkhashan/SWE485.git