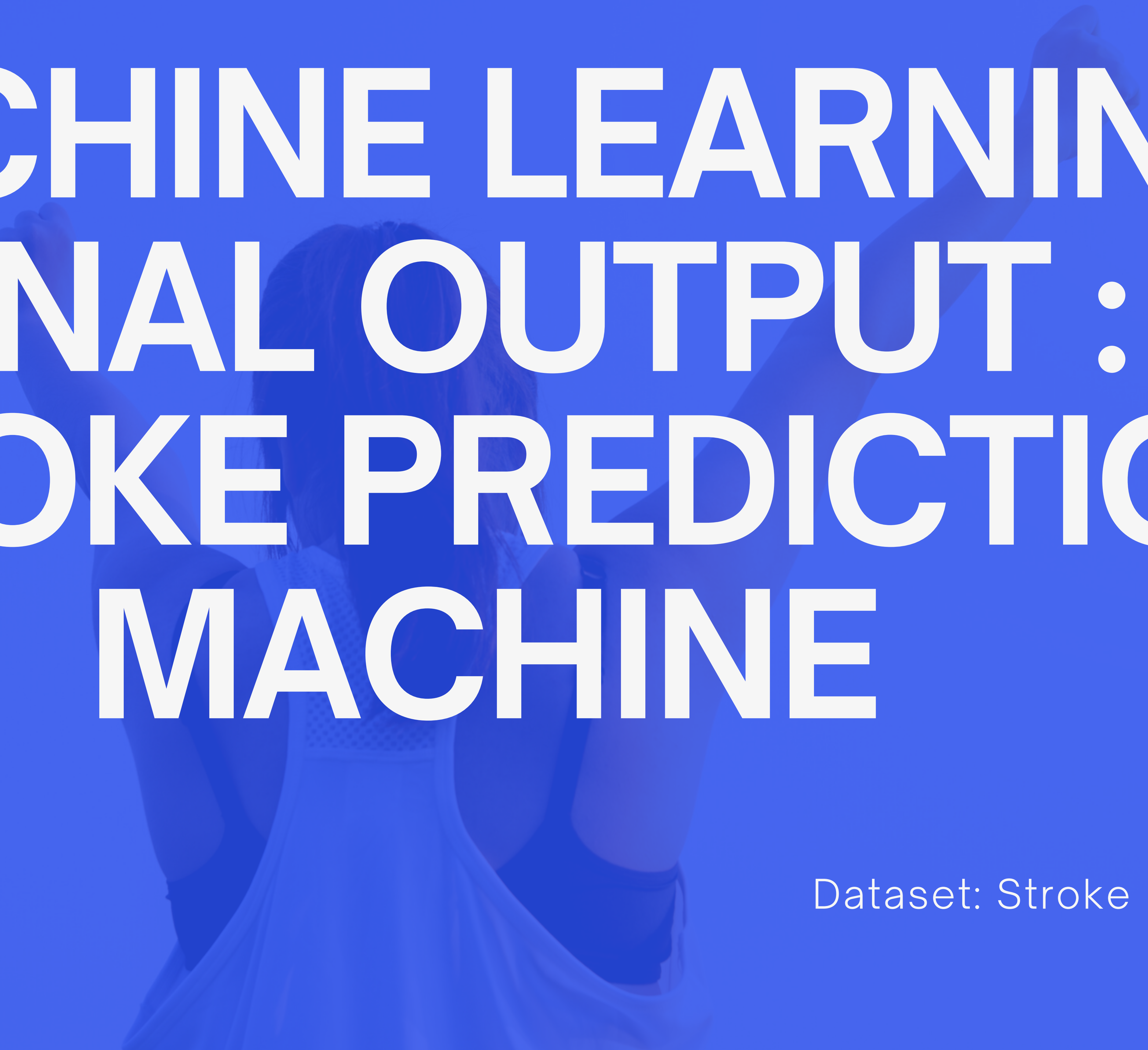
STINTSY MCO GROUP 15
CHU, CHUAN-CHEN
DELIMA JR., REYNALDO K.
JATICO II, NILO CANTIL K.
TAN, JEDWIG SIEGFRID S.



MACHINE LEARNING FINAL OUTPUT : STROKE PREDICTION MACHINE

Dataset: Stroke Prediction Dataset

Defining Stroke

Overview

A Stroke occurs when the supply of blood to the brain is interrupted or diminished, which prevents brain tissues from acquiring proper nutrients and oxygen. This results in brain cells dying in a short span of time. if not given immediate attention to, Stroke could cause life altering disabilities, and could be fatal.

Causes

The causes of stroke can be both attributed to lifestyle and medical risks including :

Lifestyle Related Risks:

- Illegal drug usage
- Being obese
- Physical inactivity
- Heavy drinking

Medical Related Risks:

- Exposure to smoking
- Having High blood pressure
- Sleep apnea
- Cardiovascular Diseases
- Diabetes
- Family History of Strokes

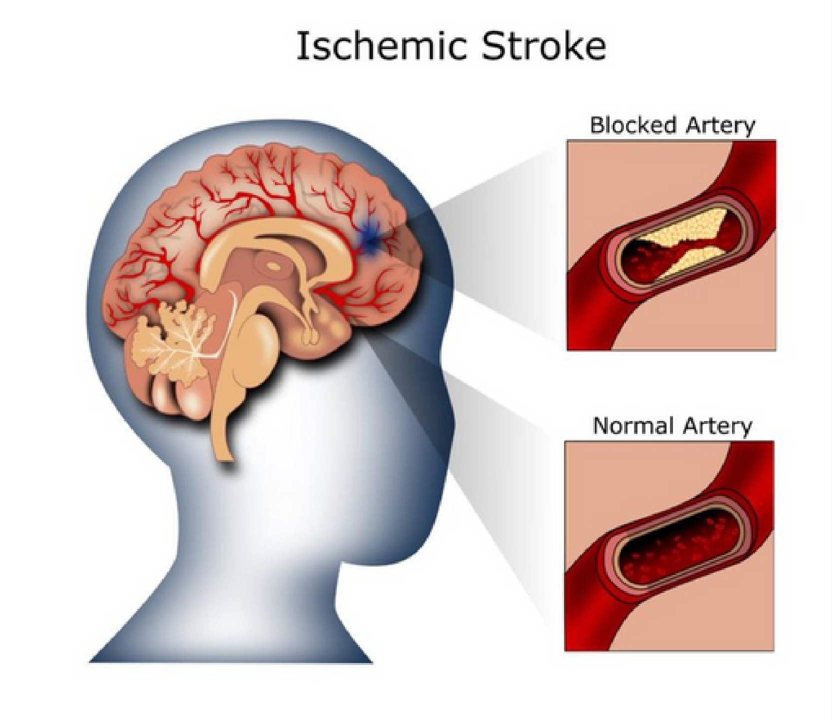
Effects

A stroke could lead to life lasting complications including paralysis, difficulty in eating or conversing, lasting pain,etc. and cause death if not given immediate response and medical treatment to.

The Two Types of Stroke

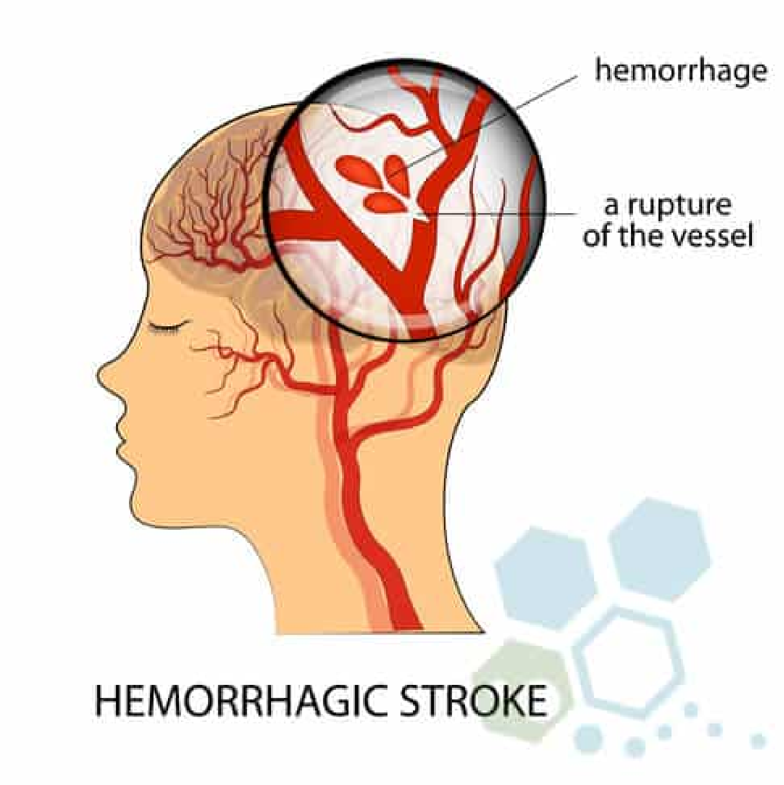
Ishcemic Stroke

This is when an artery is blocked or narrowed, resulting in a reduction in blood flow to the brain. This Results in a lost of blood flow and oxygen to the brain, which may cause brain cells to become damaged or die



hemorrhagic stroke

This is when a blood vessel to the brain is ruptured or burst, resulting in blood accumulating around the tissue in the ruptured part, causing a loss of blood and possible physical stress to the brain.



Only 38%

of the respondents knew all the major symptoms of a stroke and call 911 in a survey conducted on stroke awareness

Second Leading cause of death

According to a Study by WHO in 2016

795,000

People in the United States alone experience a stroke yearly

Every 4 minutes

someone in the United States died of Stroke in 2018

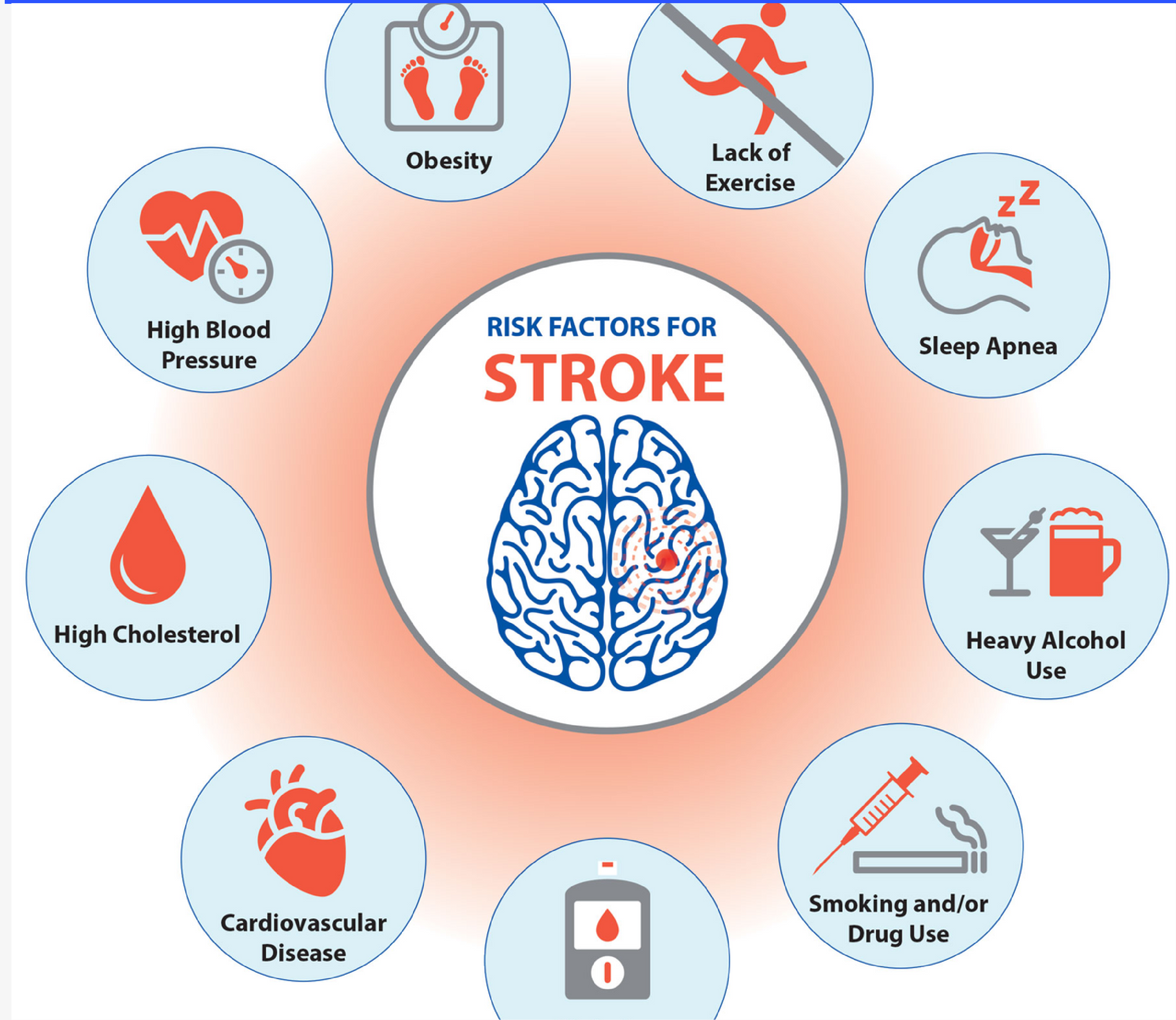
Dataset Overview

Dataset Name
Stroke Prevention Dataset

Author
Fede Soriano

Source
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Description
It features approximately 5,000 patient entries with 12 features. Each of the first 11 feature pertains to a risk factor that may lead to a stroke. The final feature is a binary value indicating whether a patient in question had a stroke or not. The origin of the patients featured in the dataset were stated as confidential in the acknowledgment of the dataset. The dataset was created to predict possible risk in having a stroke given the state of the patient in various input parameters.



Dataset Features

ID

Patient Unique Identifier

GENDER

Patient Gender Orientation

AGE

Patient Age Orientation

HYPERTENSION

Indicator if patient has hypertension

HEART DISEASE

Indicator if patient has heart disease

EVER MARRIED

Indicator if patient has ever married

WORK TYPE

Indicator of patient's work type

RESIDENCE TYPE

Indicator of patient's Resident type

AVG GLUCOSE LEVEL

Indicator of patient's average glucose level

BMI

Indicator of patient's Body Mass Index

SMOKING STATUS

Indicator if the patient has smoked

STROKE

Indicator if the patient has experienced a stroke

Methodology

PLACEHOLDER

- ## DATA CLEANING

  Some of the values in the observations needed to be modified due to outliers.

- ## FEATURE EXTRACTION

  Nine features are extracted from the dataset.

- ## HYPERPARAMETER TUNING

  Each of the machine learning algorithm will be tested to find the best result based on the given hyperparameters.

- ## MODEL TRAINING

  Six machine learning algorithms were tested and analyzed.

Data Cleaning

Exploratory Data Analysis

To further understand the dataset, initial investigation was made to find patterns, relationships between the columns and anomalies.

Dropping of features

The features, "id" and "ever_married", are dropped from the dataset.

Modifying values of observations

The values of "bmi" was modified to prevent any outliers, and one observation in gender was modified as well.

Normalization of data was also performed to the "age", "bmi" and "avg_glucose_level".

Feature Extraction

Number of features

Nine features are used to train in each model.

Description of features

There are four features (gender, age, work_type and Residence_type) that pertain to the background of the patient.

There are five features (bmi, avg_glucose_level, hypertension, heart_disease and smoking_status) that pertain to the health information of the patient.

Model Training

TRAINING AND TESTING MODEL EFFICACY

In this section of the methodology, six models were utilized by the research team, this is done in order to get a good overview of the models available to the researchers, and assess which model is best suited for the purposes of this study through evaluation through the use of Cross Validation.

Models Used

K Nearest Neighbors

Estimates which group the data belong based on the **k** number of nearest points (Subramanian, 2019).

Support Vector Machines

Analyzes the data and sorts it into one of two categories as it outputs a map of the data with margins between the two (Adankon, 2009).

K-Means

With **k** number of centroids, it allocates every data point to the nearest cluster, while keeping the centroids as small as possible (Garbade, 2018).

Naive Bayes

Utilizes Bayes' theorem, which states that the probability of an event can be adjusted as new data is introduced, to classify objects (Webb, 2009).

Logistic Regression

A statistical analysis method used to predict a data value based on prior observations of a data set (Rosencrance, 2019).

Decision Tree Classifier

Utilizes decision trees, in which data is continuously split according to a certain parameter, to classify data (Chakure, 2019).

Evaluation Specifics

Cross Validation

Cross Validation is performed on the six models to determine and compare performances.

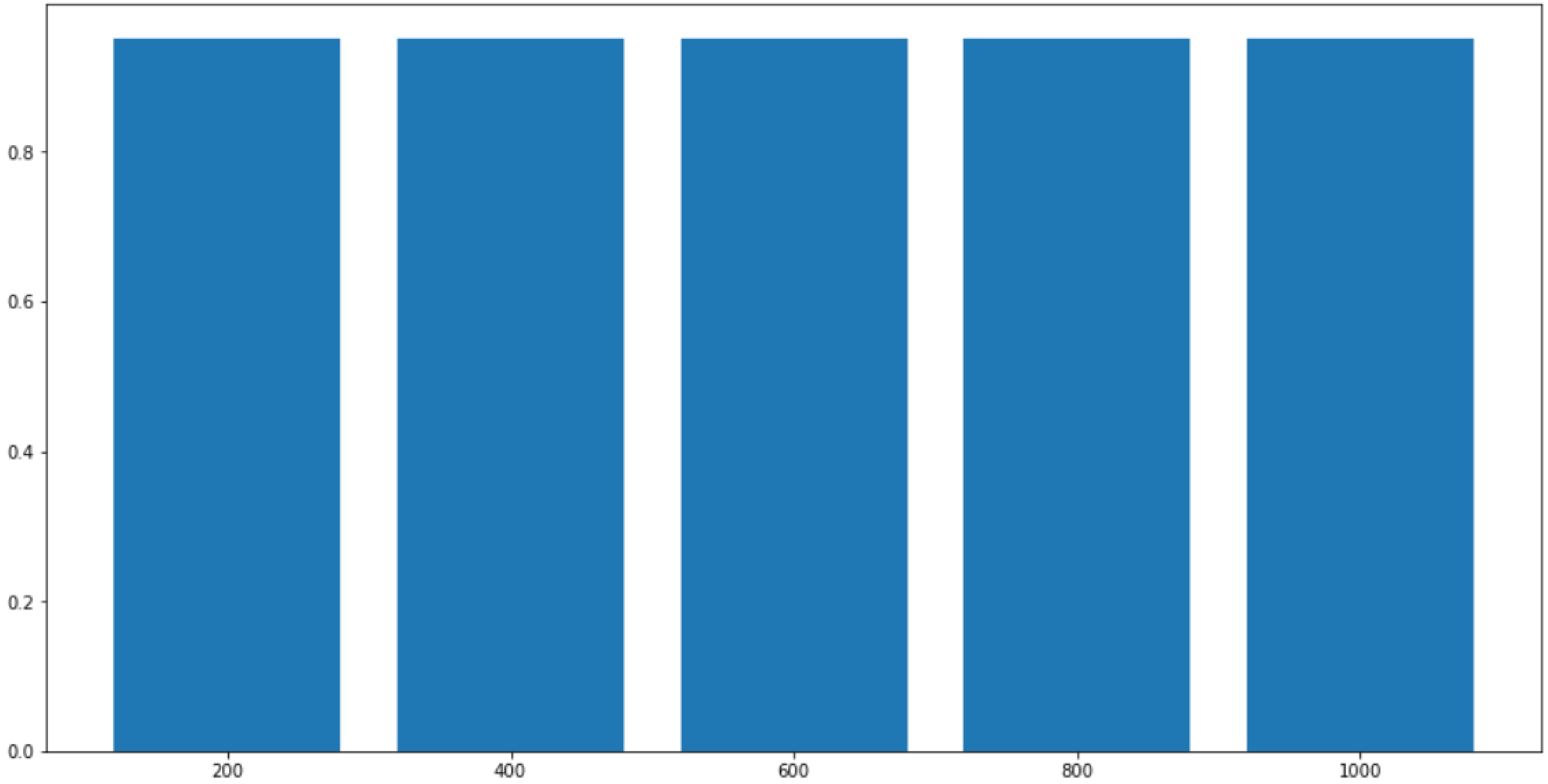
Accuracy Score

Accuracy Score is used as a metric to determine the efficacy/performance of each models

K-Fold

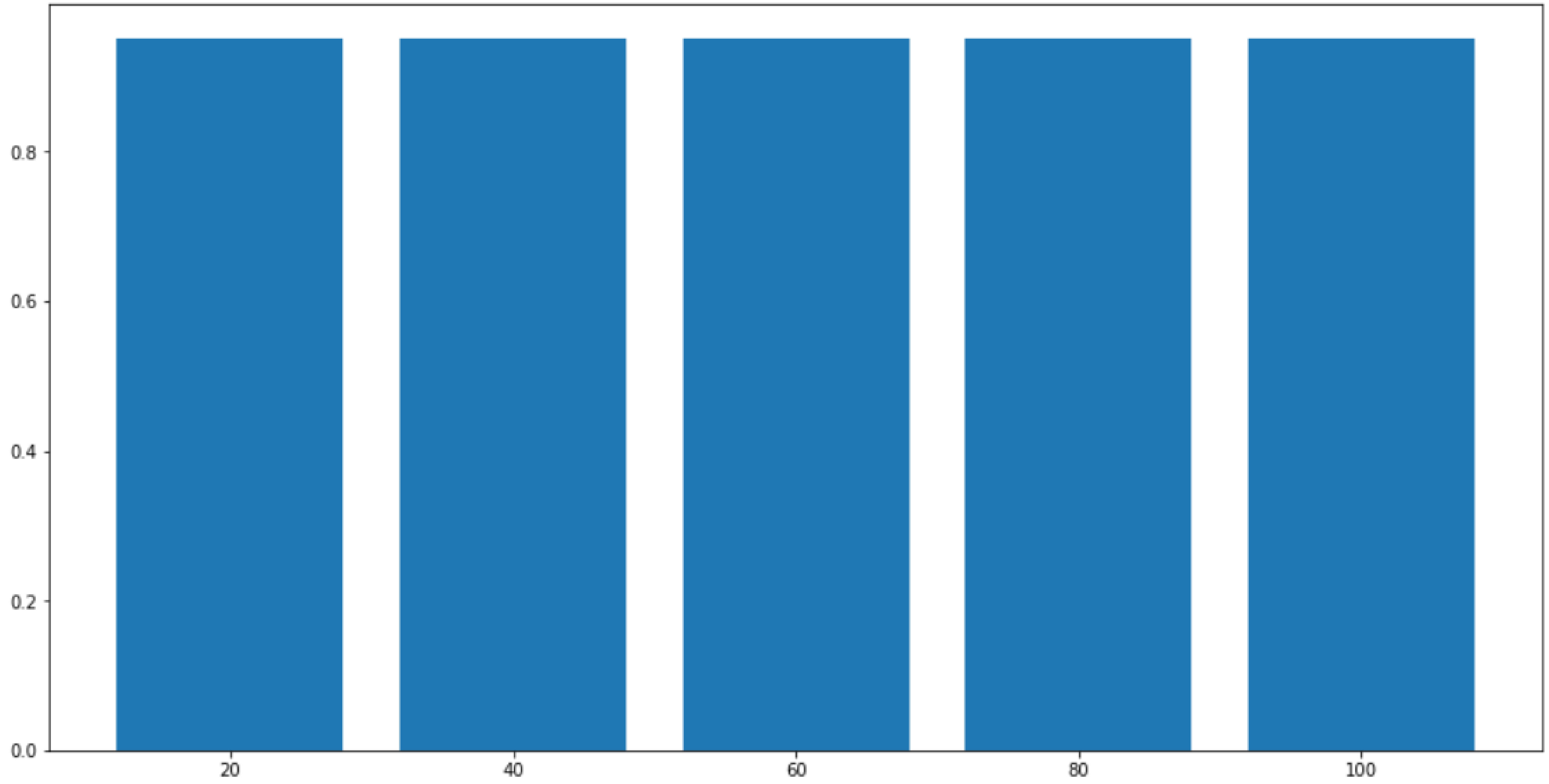
For the cross-validation splitting strategy, the number of folds in a stratified K-fold used for cross validation is 15.

Results and Analysis: Logistic Regression Model



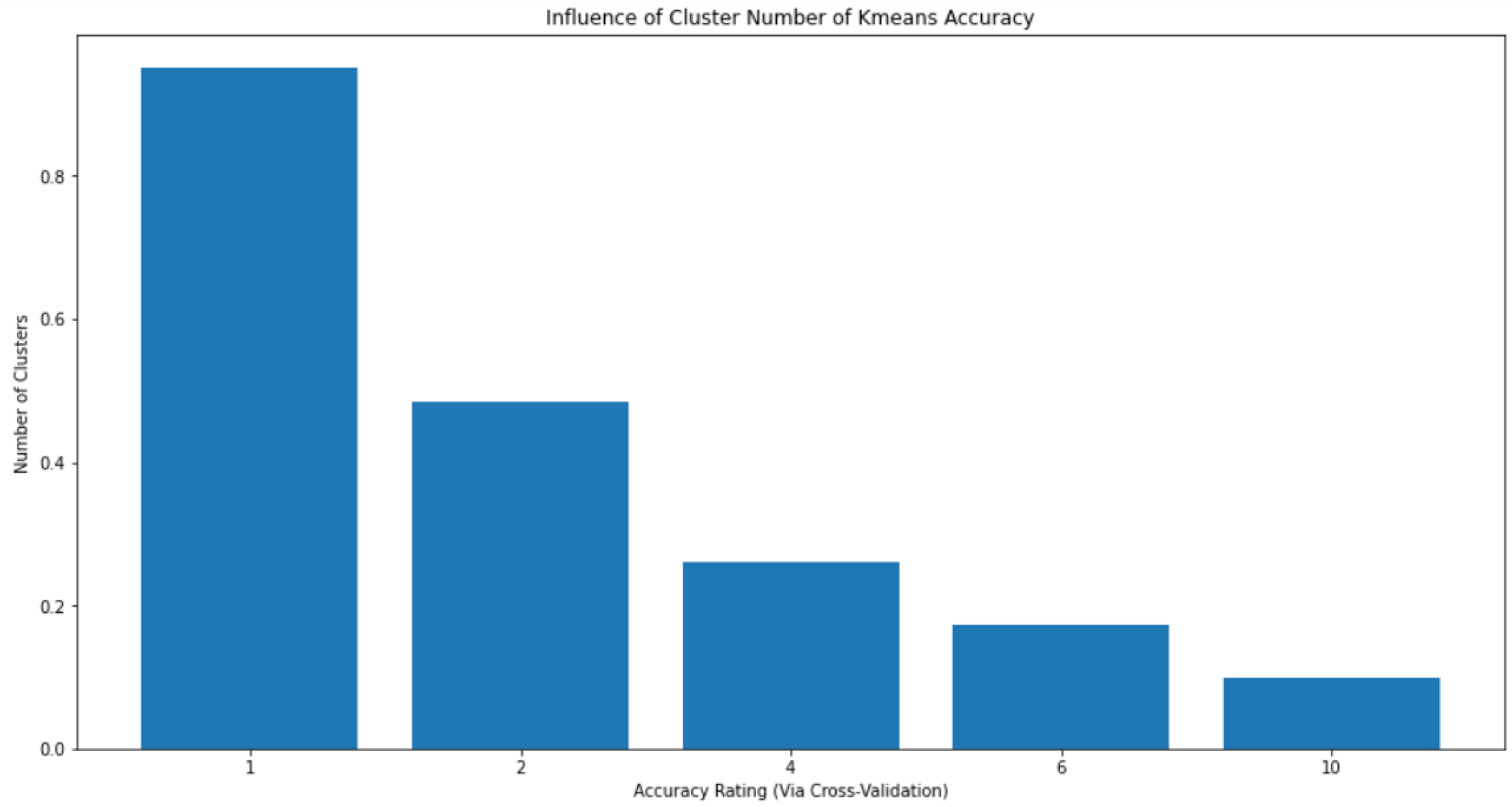
The logistic regression model has an accuracy rating of **95.5%,** with a cross evaluation rating of **95.5%**. Hyperparameter **max_iter** do not necessarily affect the overall performance of the model. Due to the stark distribution of non-stroke and stroke patients in the dataset, it result to an overfitted model.

Results and Analysis: kNN Model



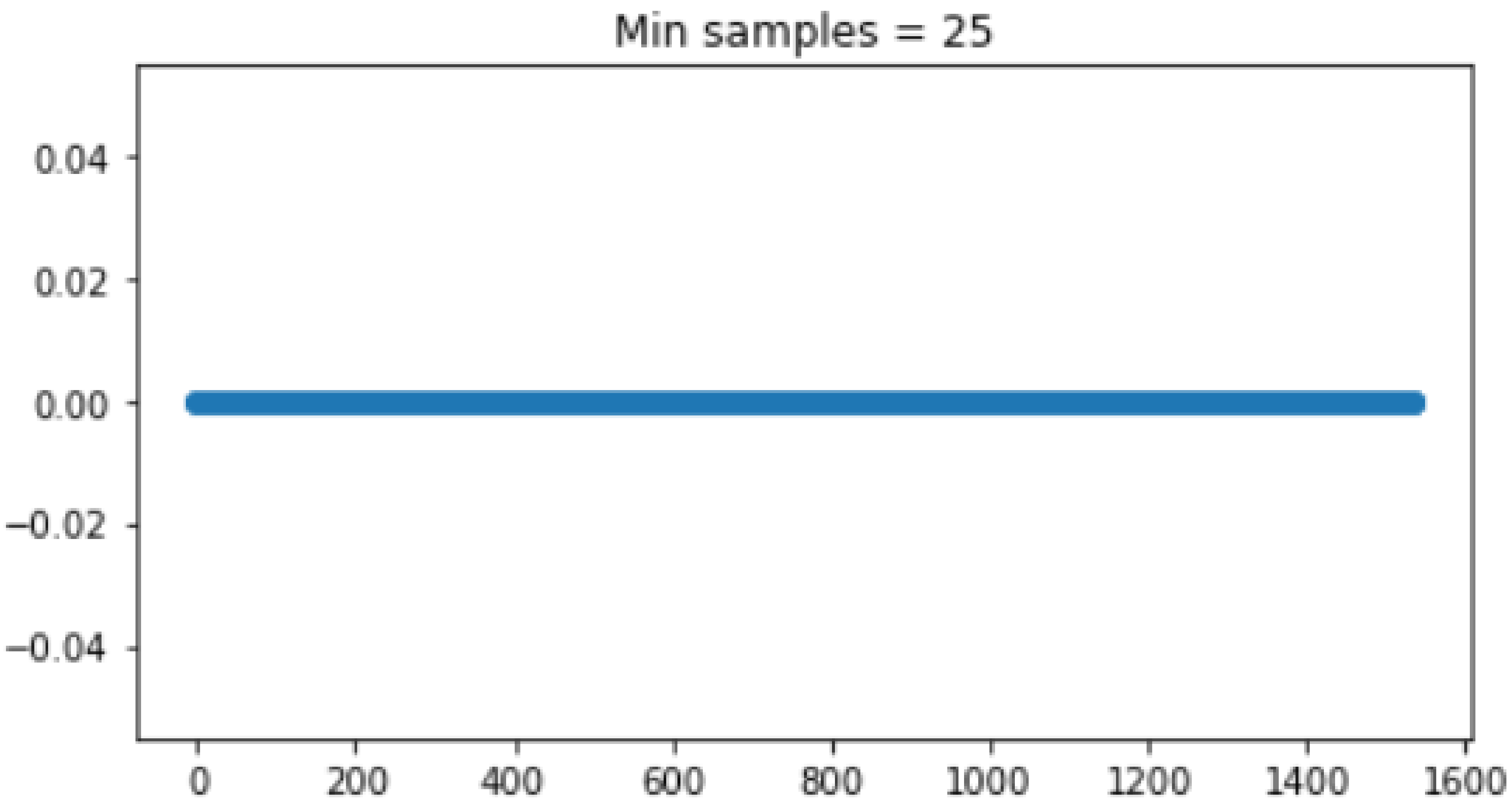
The kNN model has an accuracy rating around **95.43%**, with a cross evaluation rating of around **94.94%**. The hyperparameters **n_neighbors** and the **leaf_size** as well as the distance formula utilized only marginally increased the accuracy of the model up to a certain extent.

Results and Analysis: K-Means Model



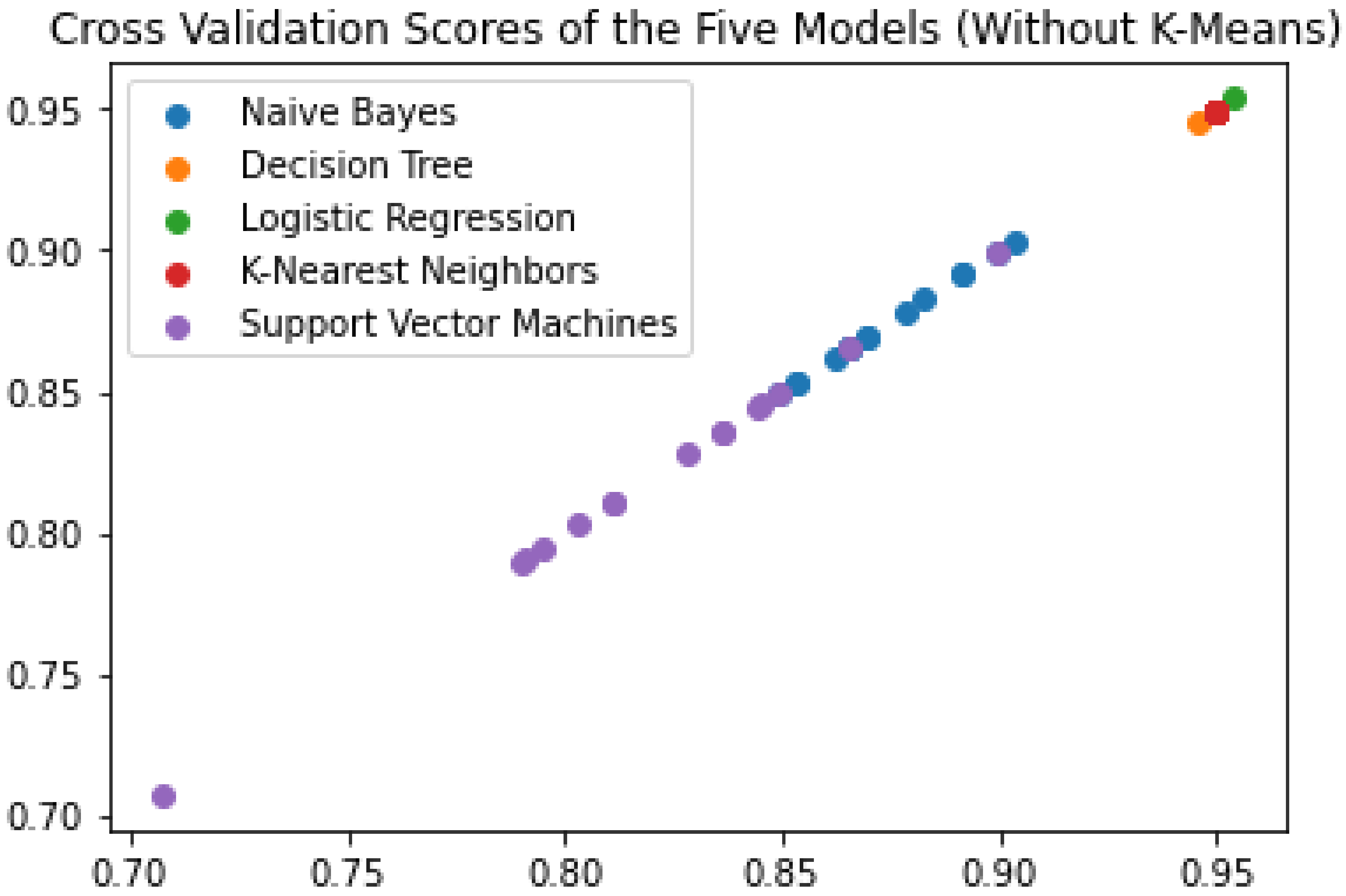
The K-Means model achieved an accuracy score of **11.4%**, and a cross-validation accuracy score of **14.1%**. Out of the six models this model performed the worst. Further testing of the **n_clusters** hyperparameter (number of clusters), shows that as the number of clusters is increased, the overall efficacy of the model severely deteriorates.

Results and Analysis: K-Means Model



The decision tree model has an accuracy rating of **95.5%**, with a cross evaluation rating of **94.97%.** The hyperparameters **max_depth** and **max_features** do not necessarily affect the model. The minimum sample and algorithm used also do not affect the efficacy of the model in a significant manner.

Results and Analysis: Naive Bayes



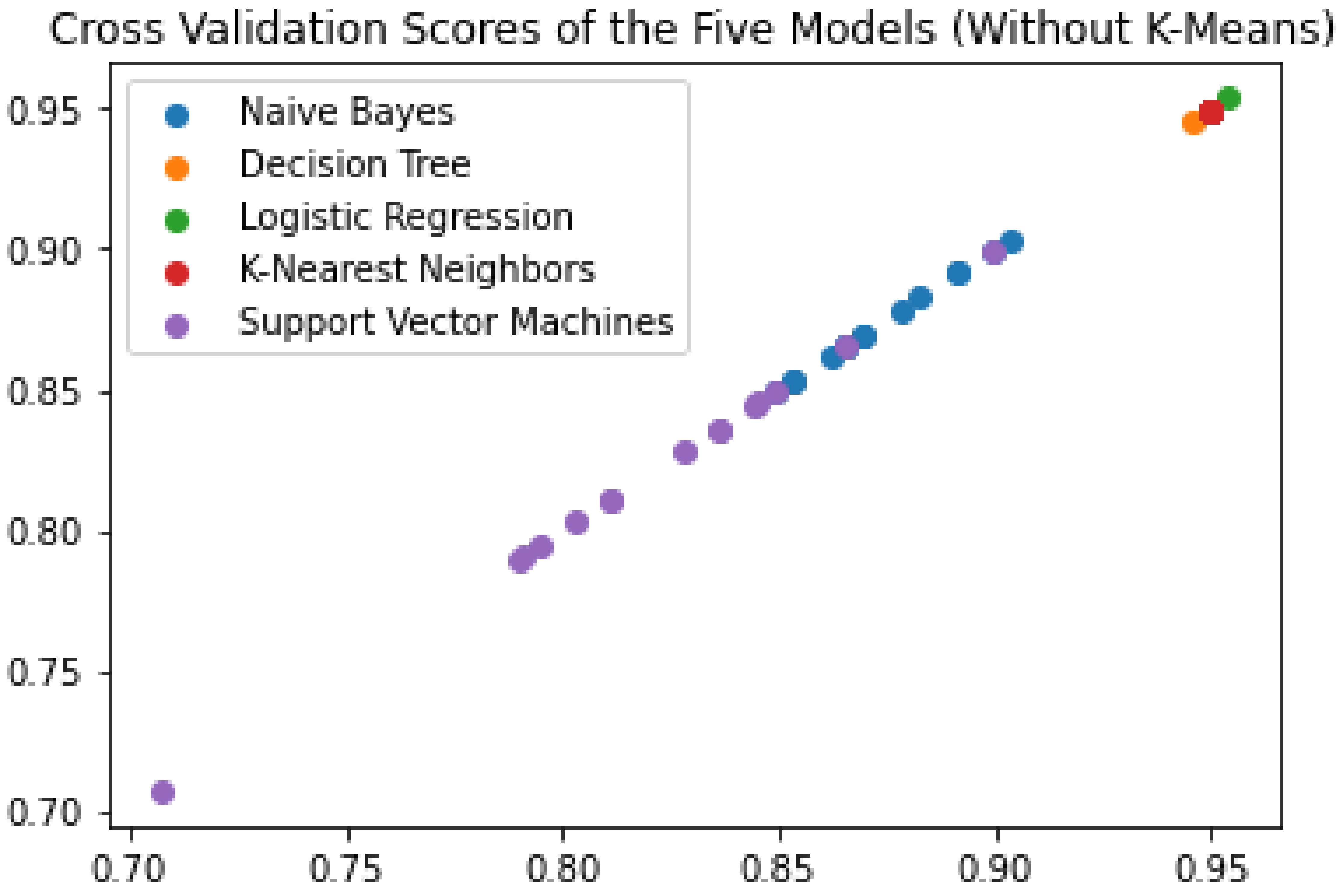
Accuracy of Naive Bayes: ~87.28%

Cross-evaluation score of Naive Bayes: ~87.34%

Naive Bayes is one of the best models since it does not overfit the data.

Naive Bayes has no hyperparameters.

Results and Analysis: Support Vector Machine



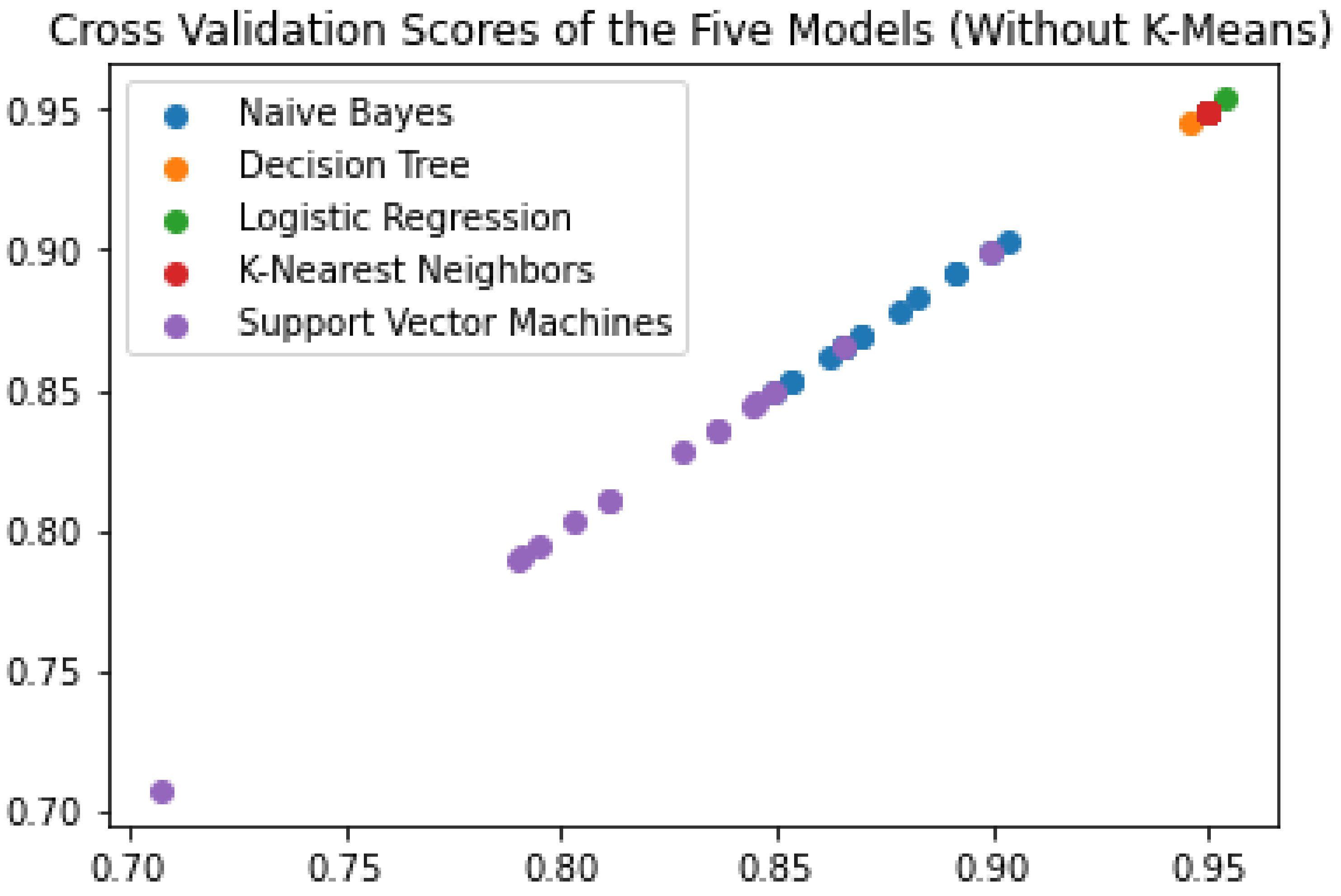
Accuracy of SVM: ~79.91%

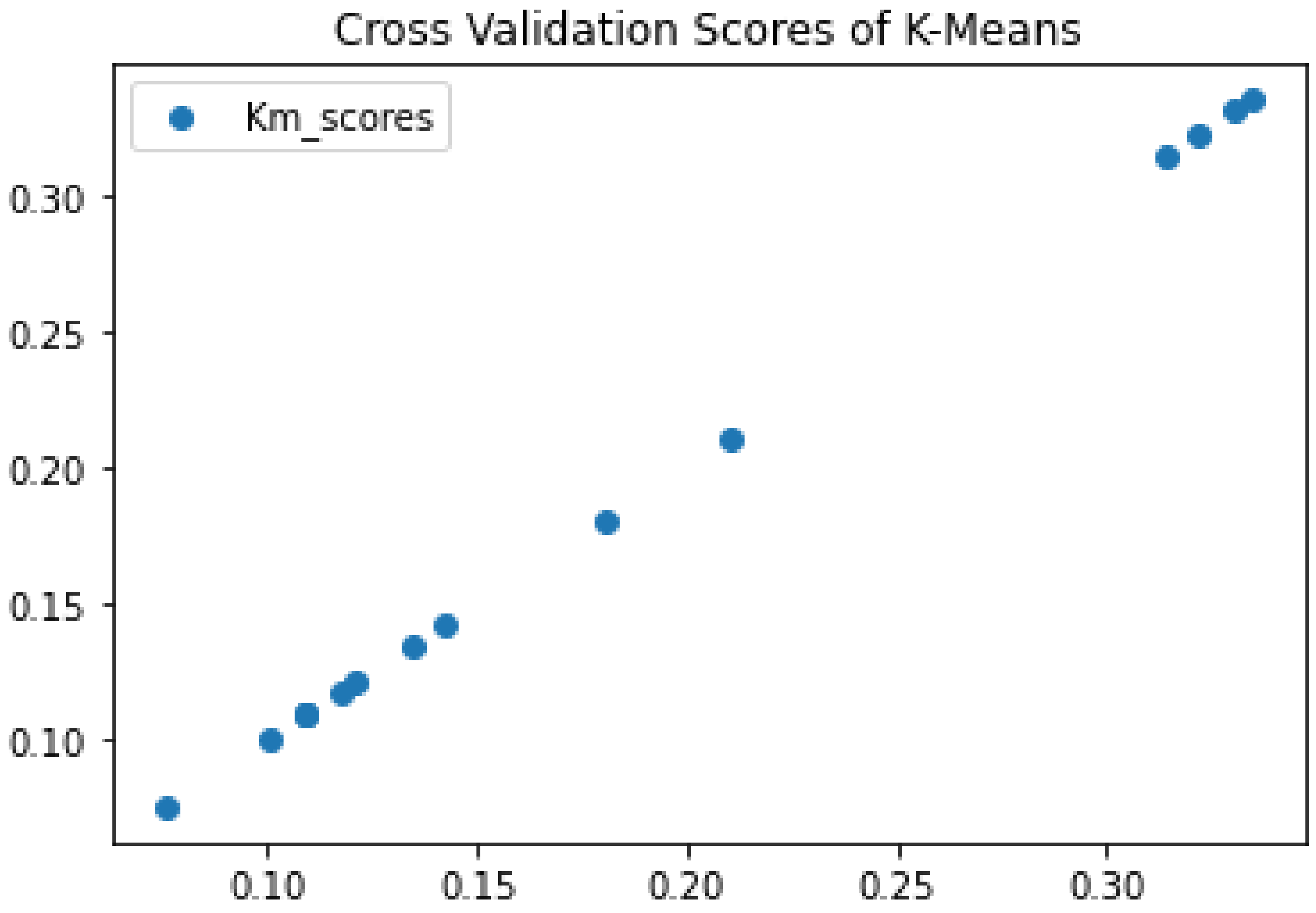
Cross-evaluation score of SVM: ~81.38%

SVM is one of the best models since it does not overfit the data.

The hyperparameter, max_iter, does not necessarily affect the performance of the model.

Results and Analysis: Comparison of Models





Conclusion and Recommendation

Conclusion

The research group was able to implement a machine that is able to determine if a certain patient is at risk of having a stroke by taking into account the medical and lifestyle risk attributed to stroke. Out of the six models presented, Logistic Regression is the best model based on the cross-validation using 15 folds.

Recommendations

Training and testing datasets with higher number of stroke patients can lead to better results and help find better hyperparameters. This can help improve the process of fitting the model.

References

Chakure, A. (2019). Decision Tree Classification. *Medium*. Retrieved from https://medium.com/swlh/decision-tree-classification-de64fc4d5aac

Garbade, M. J. (2018). Understanding K-means Clustering in Machine Learning. *Towards Data Science*. Retrieved from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Rosencrance, L. (2019). Logistic Regression. *Tech Target*. Retrieved from https://searchbusinessanalytics.techtarget.com/definition/logistic-regression

Adankon, M. M. (2009). Support Vector Machine. *Springer Link*. Retrieved from https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-73003-5_299

Subramanian, D. (2019). A Simple Introduction to K-Nearest Neighbors Algorithm. Towards Data Science. Retrieved from https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e

Webb, G. (2009). Naive Bayes. *Springer Link*. Retrieved from https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_576

.

References

CJohnson, W., Onuma, O., Owolabi, M., & Sachdev, S. (2016, September). Stroke: A global response is needed. Retrieved from https://www.who.int/bulletin/volumes/94/9/16-181636/en/

Holland, K. (2019, October 16). Stroke: Causes, symptoms, diagnosis, and treatment. Retrieved from https://www.medicalnewstoday.com/articles/7624

Stroke. (2021, February 09). Retrieved from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

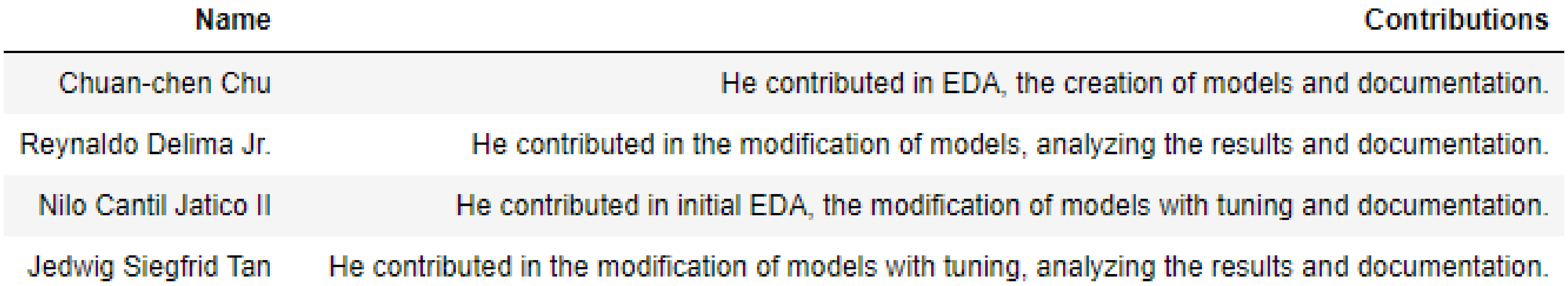
Liu, Q., Wang, X., Wang, Y., Wang, C., Zhao, X., Liu, L., Li, Z., Meng, X., Guo, L., & Wang, Y. (2018). Association between marriage and outcomes in patients with acute ischemic stroke. Journal of neurology, 265(4), 942–948. https://doi.org/10.1007/s00415-018-8793-z

Shawnita Sealy-Jefferson, Molly Roseland, Michele L. Cote, Amy Lehman, Eric A. Whitsel, Jason Booza, and Michael S. Simon.Women's Health Reports.Dec 2020.326-333.http://doi.org/10.1089/whr.2020.0034

Reeves, M. J., Bushnell, C. D., Howard, G., Gargano, J. W., Duncan, P. W., Lynch, G., Khatiwoda, A., & Lisabeth, L. (2008). Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. The Lancet. Neurology, 7(10), 915–926. https://doi.org/10.1016/S1474-4422(08)70193-5

.

Contributions



Name	Contributions
Chuan-chen Chu	He contributed in EDA, the creation of models and documentation.
Reynaldo Delima Jr.	He contributed in the modification of models, analyzing the results and documentation.
Nilo Cantil Jatico II	He contributed in initial EDA, the modification of models with tuning and documentation.
Jedwig Siegfrid Tan	He contributed in the modification of models with tuning, analyzing the results and documentation.