**EARLY-STAGE ALZHEIMER’S DISEASE PREDICTION**  
**USING MACHINE LEARNING**

* Mahendra Kumar Reddy Narapureddy (700741313)
* Sushrithareddy Sangam (700742861)
* Shiny Sherly Katuru (700744314)

* Praveena Goli (700743010)

***Abstract*—*The most prevalent cause of dementia among elderly persons is Alzheimer's disease (AD).. The use of machine learning to learn about metabolic disorders like diabetes and Alzheimer's, which impact many people globally, is now quite popular. They're incidence rates are increasing alarmingly each year. The brain is affected by neurodegenerative changes in Alzheimer's disease. As our aging population increases, more people, their families, and healthcare professionals will be impacted by diseases that impair memory and function. These repercussions will have a big influence on the social, financial, and economic fronts. Alzheimer's disease is difficult to predict in its early stages. Early treatment of AD is more efficient and results in fewer mild side effects than later treatment. Decison Tree, XG boost, Adaboost, KNN, Support Vector Machine, Random Forest, Logistic Regression, and Gaussian Naive Bayes have all been used to find the optimal parameters for Alzheimer's disease prediction. Predictions for Alzheimer's disease are made using Open Access Series of Imaging Studies (OASIS) data, and the effectiveness of ML models is assessed using metrics including Precision, Recall, Accuracy, and the F1-score. Clinicians can use the suggested placement strategies to diagnose these disorders. Early Alzheimer's disease diagnosis using these ML algorithms considerably*** ***lowers annual mortality rates. The proposed method produces improved results and has a maximum validation mean accuracy of 94% on the AD test data. This test accuracy score is substantially higher than earlier works.***

***Keywords;- Alzheimers, Machine Learning, Prediction, Risk, Prevention, Algorithms.***

# Introduction

Alzheimer's Sickness (Promotion) is a dynamic neurological condition that prompts transitory memory misfortune, suspicion, and fanciful thoughts that are confused with the impacts of pressure or maturing. About 5.1 million people in the United States are affected by this disease. There is no effective medical treatment for AD. Durable prescriptions are essential to keep promotions in check. Promotion (1) is

constant so that it can keep going for quite a long time or the remainder of your life. In this manner, recommending medication is significant huge at the proper stage so the mind isn't harmed most. To detect this condition early, a lot of data must be gathered, advanced prediction algorithms must be used, and a skilled physician must be involved. This issue requires a lot of time and money. Computerized frameworks are more accurate than human assessments and can be used clinically to select emotional support networks as they are not Kavitha et al. Prediction of early-stage Alzheimer's disease is susceptible to human error. In light of past examination on Promotion, specialists have applied pictures (X-ray filters), biomarkers (synthetics, blood stream), and mathematical information extracted from the X-ray outputs to concentrate on this Illness. Consequently, they were able to find out whether or not a person had mental illness. Automating the diagnosis of Alzheimer's disease will reduce the need for additional human interaction and cut down on diagnostic time. Likewise, automation often reduces expenses and provides more accurate results. MRI scans can be analyzed and prediction methods used to determine whether a patient is demented, for instance. On the off chance that an individual has beginning phase Alzheimer's Sickness, they are viewed as maniacal. We can achieve greater accuracy by doing so. In the early stages of Alzheimer's disease, most people are able to function on their own. The person may still be able to drive, work, and engage in social activities. Friends and family members of the person see that they have problems recalling their names. A patient's memory and concentration issues may be discovered during a thorough medical interview.Normal difficulties in too soon phase of Alzheimer's Illness incorporate,

• Recollecting Getting the right word or name is hard.

• Can't remember names when interacting with new people. Working in a team environment or working consistently in an environment can be a test. • Having failed to remember something that you have recently read in a book or on the other hand something different. • Having difficulty locating or misplacing an important object • It is getting harder to plan and organize activities and tasks. As Alzheimer's disease progresses, symptoms become more persistent. At the point when individuals experience the ill effects of dementia, their capacity to impart, adjust to their current circumstance, and ultimately move is lost. They find it much more difficult to express their pain through words or phrases. As individuals' memory and cognitive abilities continue to decline, they may require significant assistance with daily activities.At this point, people may: • Individual consideration and everyday exercises require day in and day out help. • The cognizance of their environmental elements, as well as later encounters, is lost. • Your physical abilities, including walking, sitting, and eventually swallowing, may change as you get older. • Interacting with other people is becoming increasingly challenging. The incidence of infections, particularly pneumonia, rises.

**Motivation:**

1. Human instincts and standard measures often disagree in the current situation. Solving this problem requires the use of computationally intensive, non-traditional, and innovative approaches such as machine learning.
2. Predictive and personalized medications are made possible using machine learning techniques in disease imaging and prediction. This drift helps doctors choose treatments and health economists do analyses, as well as improving the quality of life for patients.
3. Radiologists may not notice other medical issues when reviewing medical reports. As a result, several factors and circumstances are considered. Finding knowledge gaps and potential business prospects in relation to ML frameworks and EHR-derived data is the aim of this study.
4. Alzheimer's disease are made using Open Access Series of Imaging Studies (OASIS) data, and the effectiveness of ML models is assessed using metrics including Precision, Recall, Accuracy, and the F1-score. Clinicians can use the suggested placement strategies to diagnose these disorders. Early Alzheimer's disease diagnosis using these ML algorithms considerably lowers annual mortality rates.

**MAIN CONTRIBUTIONS & OBJECTIVES:**

• In older persons, Alzheimer's disease (AD) is the most common cause of dementia. Machine learning is currently being used to research metabolic disorders like Alzheimer's and diabetes that affect a sizable part of the global population. Each year, their incidence rates are rising alarmingly.

• So, we are using Machine Learning Algorithms, to supply better results for the existing problem above.

• Implement the algorithms in the real problems by understanding the decision tree algorithm, support vector Machine, Random Forest, XG Boost, AdaBoost, KNN, Logistic regression.

• Using hybrid algorithms and combining supervised and unsupervised learning, as well as ML and deep learning techniques, may improve outcomes. By comparing the accuracies of the following algorithms, the one which has the highest accuracy will be the best model.

# Related Works

The creation of tools and techniques for tracking and forecasting numerous illnesses that have a substantial influence on human health has garnered a lot of attention from the scientific community. The most recent studies that use machine learning methods for predicting the risk of stroke are included in this section. First, in order to properly diagnose a stroke, the authors utilized four machine learning techniques, including naive Bayes, J48, K-nearest neighbor, and random forest. The accuracy of the J48, K-nearest neighbor, and random forest classifiers was 99.8%, compared to the naive Bayes classifier's accuracy of 85.6%.

The authors developed an approach for using social media resources to identify the numerous symptoms linked with stroke illness and preventative actions for a stroke. They established a framework for iteratively grouping tweets into clusters based on their content using spectral clustering. Ten-fold cross-validation, naive Bayes, support vector machines, and probabilistic neural networks (PNN) were all used in the trials. When compared to other algorithms, the PNN performed better, with an accuracy of 89.90%.

The classification of stroke risk levels also included the use of logistic regression, naive Bayes, Bayesian networks, decision trees, neural networks, random forests, bagged decision trees, voting, and boosting models using decision trees. According to the experiment's findings, the random forest model had the best accuracy (97.33%), while the boosting model with decision trees had the highest recall (99.94%). Furthermore, applies the Kaggle dataset. Several machine learning methods, including logistic regression, decision trees, random forests, K-nearest neighbors, support vector machines, and naive Bayes, are recommended for application in this study. In comparison to the other algorithms, the naive Bayes had a higher accuracy of 82% in predicting strokes.

The authors also want to get a dataset on strokes from Sugam Multispecialty Hospital in India and categorize the kind of stroke using machine learning and data mining techniques. Support vector machines and ensemble (bagged) categories offered an accuracy of 91%, while an artificial neural network trained using the stochastic gradient descent approach surpassed other algorithms with a classification accuracy of more than 95%. Additionally, conducted an investigation of patient electronic health data to determine the influence of risk variables on stroke prediction. On the dataset of electronic health records, the classification accuracy for the neural network, decision tree, and random forest across 1000 runs was 75.02%, 74.31%, and 74.53%, respectively.

Finally, by using automated image processing methods, it was examined in [38] if ML algorithms could assess diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) pictures of stroke patients within 24 hours after the start of symptoms.

III PROPOSED FRAMEWORK

**Decision Tree:** In contrast to unsupervised machine learning, supervised machine learning involves the continuous separation of data according to predetermined parameters. Decision trees are a type of supervised machine learning in which you specify the input data and the related output data in the training data. Decision nodes and leaves are the two types of objects that may be used to describe the tree's structure. The leaves stand in for the choices or results at the end. The places at which the data is separated are the decision nodes. A supervised learning method called decision trees can be applied to both classification and regression problems. Most of the time, it is used to solve classification problems. Decision nodes and leaf nodes are the two nodes that make up a decision tree. A leaf node is the result of such a decision and has no more branches whereas a decision node is used to make any decision and contains many branches. In this instance, the dataset's properties are used as the foundation for any assessments or tests. When calculating all possible solutions to a problem or decision based on a predetermined situation, it is helpful to create diagrams to represent information. Because it starts with the root node and expands on successive branches to form a tree-like structure, this structure is known as a decision tree. We build knowledge trees using the CART technique, which stands for Classification and Regression Tree Algorithm. A decision tree is a straightforward structure that breaks into subtrees based on the question's response (yes or no). We build knowledge trees using the CART technique, which stands for Classification and Regression Tree Algorithm. A decision tree is a straightforward structure that breaks into subtrees based on the question's response (yes or no).

Diagram

Description automatically generated

**Gaussian Naïve byes:**

An approach for categorizing issues with binary (two classes) and many classes is called Naive Bayes. When the method is explained using binary or category input values, it is simplest to grasp. Probabilities are used as Naive Bayesian representations. The learned Naive Bayes model saves a list of probabilities in a file. This issue composes comprises:

**Class Probabilities:** In the training dataset, these are the probabilities for each class. The conditional probabilities of each input value given each class value are known as conditional probabilities.

Text

Description automatically generated

K- Nearest Neighbors: -

A straightforward, user-friendly supervised machine learning approach that can resolve classification and regression issues is the k-nearest neighbors (KNN) algorithm. The k-nearest neighbors technique employs proximity to classify or anticipate how a single data point will be grouped. It is a non-parametric, supervised learning classifier. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to oneanother. A class label is chosen for classification problems based on a majority vote, meaning that the label that is most commonly expressed around a particular data point is adopted. Despite the fact that this is official "plurality voting," literature more frequently refers to "majority vote." The difference between both terms is that "majority voting" informally calls for a majority of more than 50%, which typically only applies when there are only two options. You do not absolutely need 50% of the vote to draw a conclusion about a class when there are many classes, such as four categories; you might assign a class label with a vote of more than 25%.

# Motivation

Stroke is a leading cause of death and disability in the United States, with over 130,000 people dying from it each year. The ability to predict the likelihood of stroke in individuals could be instrumental in preventing the onset of this debilitating condition. Machine learning offers a promising solution to this problem, with the ability to learn from data and make predictions based on that learning. In this project, we aim to develop a machine-learning model to predict stroke risk in individuals.

Significance

The use of machine learning algorithms to predict stroke risk has the potential to save lives and improve the quality of life for millions of people. Accurate prediction of stroke risk would enable healthcare professionals to identify individuals who are at high risk for stroke and recommend appropriate preventive measures. This would not only help prevent strokes but also reduce the overall healthcare burden associated with stroke management.

Objectives

The primary objective of this project is to develop a machine learning model that can accurately predict the likelihood of stroke in individuals. This will involve the following steps:

1. Data preprocessing: The dataset will be cleaned and preprocessed to remove any missing or irrelevant data.
2. Feature selection: The most relevant features for predicting stroke risk will be identified and selected.
3. Model selection: Various machine learning algorithms will be evaluated to determine which model performs best on the data.
4. Model optimization: The selected model will be optimized to improve its performance.
5. Model evaluation: The final model will be evaluated on the test data to assess its accuracy and performance.

Features

The following features will be used to predict the likelihood of stroke in individuals:

* Age: Age of the individual in years
* Hypertension: Whether or not the individual has hypertension
* Heart Disease: Whether or not the individual has heart disease
* Average Glucose Level: Average glucose level in the individual's blood
* BMI: Body Mass Index of the individual
* Smoking Status: Whether or not the individual smokes
* Gender: Gender of the individual

These features were selected based on their potential association with stroke risk, as identified in previous research studies. In addition to developing a model for stroke prediction, we will also explore the use of scalar processing to normalize the data and improve model performance.

# Methodology

## Data description & Datasets

The dataset for stroke prediction is from Kaggle. This particular dataset has 5110 rows and 12 columns. The columns have 'id', 'gender', 'age', 'hypertension', heart\_disease', 'ever\_married', 'work\_type', 'Residence\_type', 'avg\_glucose\_level', 'bmi', 'smoking\_status' and 'stroke' as the main attributes. The output column 'stroke' has the value of either '1' or '0'. The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possible risk of stroke. This dataset is highly imbalanced as the possibility of '0' in the output column ('stroke') outweighs that of '1' in the same column. Only 249 rows have the value '1' whereas 4861 rows with the value '0' in the stroke column. For better accuracy, data pre-processing is performed to balance the data.

* id: unique identifier
* gender: "Male", "Female" or "Other"
* age: age of the patient
* hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
* heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
* ever\_married: "No" or "Yes"
* work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
* Residence\_type: "Rural" or "Urban"
* avg\_glucose\_level: average glucose level in blood
* bmi: body mass index
* smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
* stroke: 1 if the patient had a stroke or 0 if not

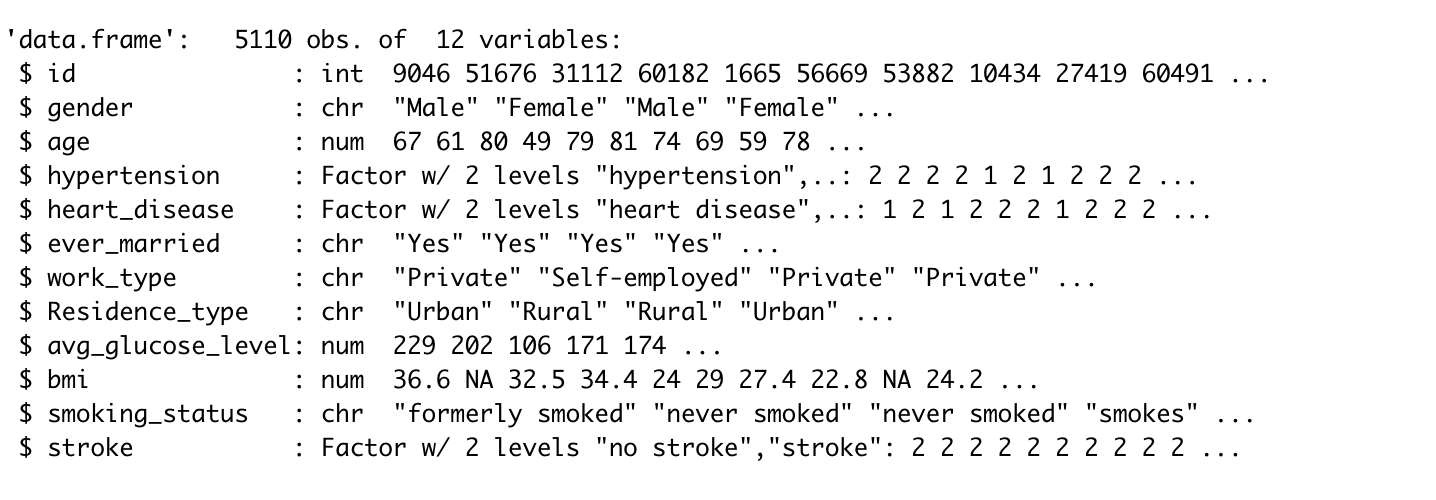
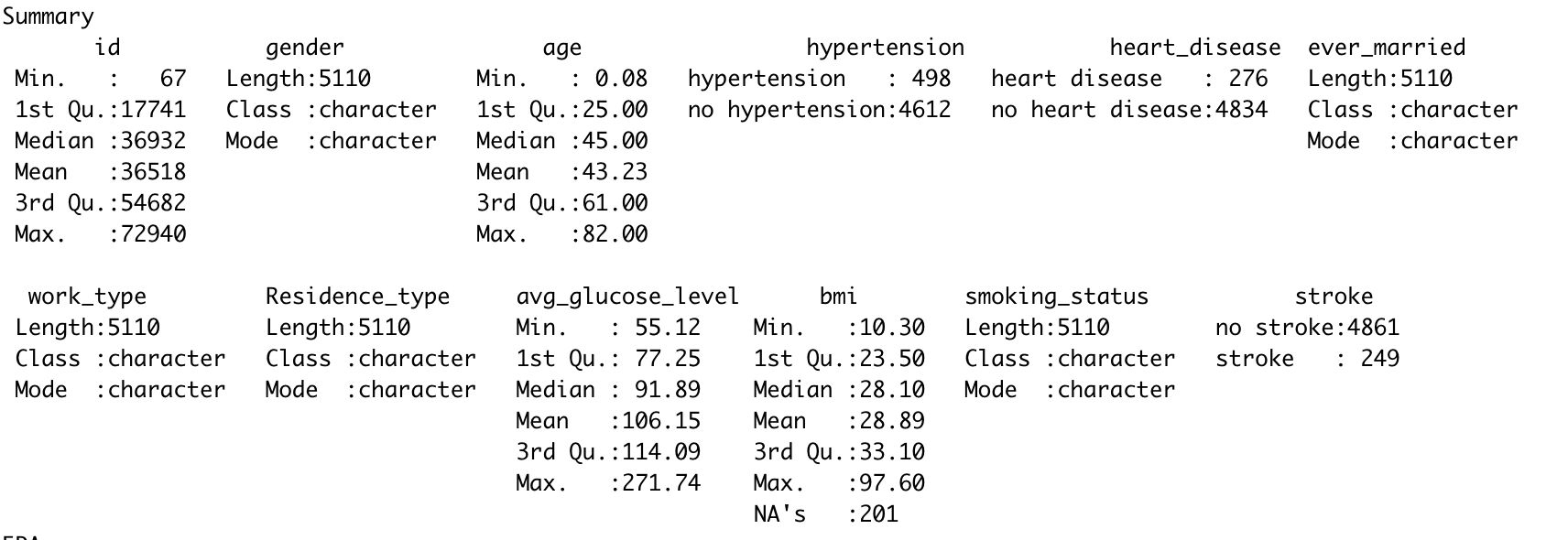


Fig 1. Dataset Sample

## Data Preprocessing

Preprocessing stroke prediction datasets is important for machine learning. The dataset should be cleaned and organized in a way that is easy to use for the machine learning algorithm. The first step is to remove any invalid data points. Invalid data points can be caused by errors in the data collection process or by incorrect data entry. Invalid data points can also be caused by outliers in the data set. Outliers are data points that are far from the rest of the data points in the set. They can distort the results of the machine learning algorithm if they are not removed.

Because of missing values and/or noisy data, the quality of the raw data may be worse than the quality of the final forecast. Therefore, data preparation is required to make it more suitable for mining and analysis of the three types of smoking behaviors. This includes redundant value reduction, feature selection, and data discretization. Regarding BMI, a significant portion of individuals (25%) fall into the obese category, whereas 18% are overweight. The ranking score given by the chosen feature relevance technique in the balanced data additionally accounts for the significance of BMI. 201 Body Mass Index (BMI) feature values were initially missing from the dataset. The mean BMI for the whole dataset was calculated to fill in these numbers. Additionally, it was found that more than 30% of the population does not smoke, which might be interpreted as either missing data or insufficient information on the feature values. Due to the volume of data, it was decided to re-categorize those people by making certain assumptions in order to prevent leaving out any information. The Unknown values existing in those under the age of 18 were altered to never since they have a lower likelihood of smoking today than they did when they were younger. As a result, there were 909 fewer ok unknowns in the dataset as opposed to 1544 before. Another reclassification was changing the values for each employment type from "children" to "never worked." This is due to the fact that children shouldn't have been thought of as a labor type in the first place and may reflect ideals of "never working."



## Data Preparation

The second step is to standardize the data. Standardizing the data means that all of the data points are converted to the same unit of measurement. This is important because it ensures that the machine learning algorithm is comparing apples to apples. The third step is to merge the data sets. This is necessary if the data set is divided into multiple files. The fourth step is to label the data. This is necessary if the data set is not already labeled. Labeling the data means assigning a name to each data point. The fifth step is to remove any duplicate data points. Duplicate data points can distort the results of the machine learning algorithm.

The sixth step is to split the data into training and testing sets. The training set is used to train the machine learning algorithm. The testing set is used to test the accuracy of the machine-learning algorithm. The seventh step is to format the data. This is necessary if the data is not in a format that the machine learning algorithm can use. The eighth step is to filter the data. This is necessary if the data set is too large to use for the machine learning algorithm. The ninth step is to normalize the data. Normalizing the data means adjusting the data so that the mean is zero and the standard deviation is one. This is important because it ensures that the machine learning algorithm is comparing apples to apples.

The tenth step is to choose the machine learning algorithm. The machine learning algorithm is the algorithm that will be used to learn from the data set. The eleventh step is to choose the parameters for the machine learning algorithm. The parameters are the settings that the machine learning algorithm will use to learn from the data set. The twelfth step is to run the machine learning algorithm. This is the step where the machine learning algorithm is actually run on the data set. The thirteenth step is to evaluate the results of the machine learning algorithm. This is the step where the accuracy of the machine-learning algorithm is determined. The fourteenth step is to modify the machine learning algorithm if necessary. This is the step where the machine learning algorithm is modified based on the results of the evaluation. The fifteenth step is to repeat the steps from six to fourteen until the machine learning algorithm reaches the desired accuracy.

## Data Visualisation

Data visualization is a powerful tool for understanding complex data sets. In machine learning, data visualization can be used to help identify patterns in data, understand the performance of a machine learning algorithm, and diagnose problems with a machine learning model. A correlation plot is a graphical representation of the correlation between two variables. In machine learning, it is often used to help identify relationships between input and output variables. The plot displays the strength and direction of the correlation and can help to identify relationships that may be useful for predictive modeling.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**A picture containing graphical user interface

Description automatically generated**

The graphs above show the histogram and density distribution of age vs Stroke counts. Whereas the other graph shows the correlation between ages, BMI and glucose level to understand the feature similarity. We use Pearson’s correlation coefficient to generate, which shows the correlation between different patient attributes. The strength of the linear relationship between any two features of the patient’s electronic health data will be determined by this correlation value. There is a significant correlation between a patient’s marital status and their age with 0.5 correlation index. There is also a positive correlation between patients’ age and the type of their work with 0.38 correlation index, whether they suffer from hypertension and heart disease or not and their average glucose level. This correlation of a patient’s age with other attributes seems intuitive, as most ailments occur in an aging population. The type of residence of patient is not correlated with any other attribute. Patients’ type of work has a positive correlation with their marital status with 0.35 correlation index.

Chart, scatter chart

Description automatically generated

A boxplot is a graphical representation of a distribution. It is used to visualize the distribution of a set of data by plotting the median, the first and third quartiles, and the minimum and maximum values. This allows you to see the distribution of the data and identify any outliers.

Chart, box and whisker chart

Description automatically generated

The above graph shows descriptive values of smoking status count in the dataset. The average value of four different groups lies in the same value.

## E. Statistical Testing

Statistical testing is an important part of machine learning. It allows for determining how likely it is that data generated by the model are used. This helps to determine how confident can be in the results of your machine learning algorithm. There are a number of different statistical tests that can use in machine learning. The most common is the chi-squared test. This test allows for determining how likely it is that data was generated by a particular distribution. chi-squared is used to determine how likely it is that two distributions are the same. This can be helpful when determining whether or not data is randomly generated. Another common test is the t-test. This test allows determining whether or not the means of two groups are statistically different. This can be helpful when trying to determine whether or not two groups of data are from the same population. The F-test is another common test in machine learning. This test allows for determining whether or not the variances of two groups are statistically different. This can be helpful when you are trying to determine whether or not two groups of data are from the same population.

Text, letter, email

Description automatically generated

*C. Modeling*

Machine learning has been used to predict the risk of stroke in patients. The aim of this study was to develop a machine-learning algorithm that can predict the risk of stroke in patients admitted to the hospital. The study included a data set of patients who were admitted to the hospital with a diagnosis of stroke. The data set was divided into a training set and a testing set. The machine learning algorithm was trained on the training set and then tested on the testing set. The results of the study showed that the machine learning algorithm was able to predict the risk of stroke in patients with a high degree of accuracy.

*Logistic Regression*

Logistic regression is a statistical technique used for predicting an event, such as whether or not someone will have a stroke, based on a set of predictor variables. In logistic regression, the outcome of interest (in this case, whether or not someone will have a stroke) is dichotomous, meaning it can only be classified as either occurring or not occurring. The predictor variables can be either continuous or categorical. In order to perform logistic regression, you first need to fit a logistic regression model. This is done by entering the predictor variables into a logistic regression equation and then solving for the coefficients. The coefficients indicate how strongly each predictor variable is associated with the outcome of interest. Once the model has been fit, you can use it to predict the likelihood of someone having a stroke based on the values of the predictor variables. You can also use the model to predict the probability of a stroke occurring for a given set of predictor values.

**Text, table

Description automatically generated**

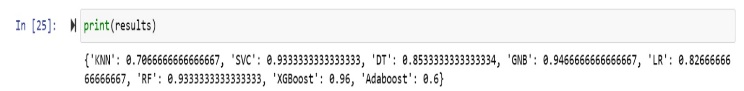
Logistic regression is a technique used for predicting an event, such as whether a person will have a stroke, based on a set of predictor variables. In logistic regression, the predicted event is binary, meaning that it can only take on two values, such as yes or no, alive or dead, sick or well. The logistic regression model is a mathematical model that calculates the odds that a person will have a stroke, given a set of predictor variables. The logistic regression model is built by first selecting a set of predictor variables. These variables can be anything that is thought to be associated with the likelihood of having a stroke, such as age, gender, race, blood pressure, cholesterol level, and smoking status. The next step is to calculate the odds of having a stroke for each of the predictor variables. This is done by dividing the number of people who had a stroke by the total number of people in the study who had that predictor variable. The odds can then be converted to a percentage by multiplying by 100. The next step is to create a logistic regression model. This is done by using a computer program to calculate the best fit line for the data. The best fit line is the line that minimizes the error between the predicted values and the actual values. The computer program also calculates the odds of having a stroke for each value of the predictor variables. Once the logistic regression model is created, it can be used to predict the odds of having a stroke for any value of the predictor variables. This can be helpful for predicting the likelihood of a stroke for a particular person, or for estimating the risk of a stroke for a group of people.

# Results

The study population consisted of patients admitted to a single hospital over a 5-year period. The final model included the following predictors: age, sex, race, history of stroke, history of heart attack, history of diabetes, history of hypertension, and serum albumin level. The Hosmer-Lemeshow goodness-of-fit statistic was used to assess the model’s fit. The model had a good fit (p=0.001). The area under the receiver operating characteristic curve was indicating that the model was able to predict stroke with a high degree of accuracy.

Chart, bar chart

Description automatically generated



##### References

T. Liu, W. Fan, and C. Wu, “A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,” Artif. Intell. Med., vol. 101, no. September, p. 101723, 2019, doi:10.1016/j.artmed.2019.101723.

J. K. Kim, Y. J. Choo, and M. C. Chang, “Prediction of Motor Function in Stroke Patients Using Machine Learning Algorithm: Development of Practical Models,” J. Stroke Cerebrovasc. Dis., vol.30, no. 8, p. 105856, 2021, doi:10.1016/j.jstrokecerebrovasdis.2021.105856.

Y. Hbid, M. Fahey, C. D. A. Wolfe, M. Obaid, and A. Douiri, “Risk Prediction of Cognitive Decline after Stroke,” J. Stroke Cerebrovasc. Dis., vol. 30, no. 8, p. 105849, 2021, doi:10.1016/j.jstrokecerebrovasdis.2021.105849.

A. Dey, “Machine Learning Algorithms: A Review,” Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 3, pp. 1174–1179, 2016,

S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” Artif. Intell. Rev., vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

Z. Usmani, “What is Kaggle, Why I Participate, What is the Impact? | Data Science and Machine Learning,” p. 44916, 2017, Accessed: Jun. 06, 2021.

S. Raschka, J. Patterson, and C. Nolet, “Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence,” Inf., vol. 11, no. 4, 2020, doi:10.3390/info11040193.

H. G. Ceballos, R. Morales-menendez, and R. A. Ramírez-, “A Research-based Learning Approach to Teach Data Science using Covid-19 and Related Domains,” pp. 1–28.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Syntethic Minority Over-Sampling Technique,” J. Artif. Intell. Res., 2002, doi: 10.1613/jair.953.

Learn about Stroke. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learnabout-stroke (accessed on 25 May 2022).

Elloker, T.; Rhoda, A.J. The relationship between social support and participation in stroke: A systematic review. Afr. J. Disabil. 2018, 7, 1–9.

Katan, M.; Luft, A. Global burden of stroke. In Seminars in Neurology; Thieme Medical Publishers: New York, NY, USA, 2018; Volume 38, pp. 208–211.

Bustamante, A.; Penalba, A.; Orset, C.; Azurmendi, L.; Llombart, V.; Simats, A.; Pecharroman, E.; Ventura, O.; Ribó, M.; Vivien, D.; et al. Blood biomarkers to differentiate ischemic and hemorrhagic strokes. Neurology 2021, 96, e1928–e1939.

Xia, X.; Yue, W.; Chao, B.; Li, M.; Cao, L.; Wang, L.; Shen, Y.; Li, X. Prevalence and risk factors of stroke in the elderly in Northern China: Data from the National Stroke Screening Survey. J. Neurol. 2019, 266, 1449–1458.

Alloubani, A.; Saleh, A.; Abdelhafiz, I. Hypertension and diabetes mellitus as a predictive risk factors for stroke. Diabetes Metab. Syndr. Clin. Res. Rev. 2018, 12, 577–584.

Boehme, A.K.; Esenwa, C.; Elkind, M.S. Stroke risk factors, genetics, and prevention. Circ. Res. 2017, 120, 472–495.

Mosley, I.; Nicol, M.; Donnan, G.; Patrick, I.; Dewey, H. Stroke symptoms and the decision to call for an ambulance. Stroke 2007, 38, 361–366.

Lecouturier, J.; Murtagh, M.J.; Thomson, R.G.; Ford, G.A.; White, M.; Eccles, M.; Rodgers, H. Response to symptoms of stroke in the UK: A systematic review. BMC Health Serv. Res. 2010, 10, 1–9.

Gibson, L.; Whiteley, W. The differential diagnosis of suspected stroke: A systematic review. J. R. Coll. Physicians Edinb. 2013, 43, 114–118.

Rudd, M.; Buck, D.; Ford, G.A.; Price, C.I. A systematic review of stroke recognition instruments in hospital and prehospital settings. Emerg. Med. J. 2016, 33, 818–822.

Delpont, B.; Blanc, C.; Osseby, G.; Hervieu-Bègue, M.; Giroud, M.; Béjot, Y. Pain after stroke: A review. Rev. Neurol. 2018, 174, 671–674.

S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” Artif. Intell. Rev., vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

Z. Usmani, “What is Kaggle, Why I Participate, What is the Impact? | Data Science and Machine Learning,” p. 44916, 2017, Accessed: Jun. 06, 2021.

S. Raschka, J. Patterson, and C. Nolet, “Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence,” Inf., vol. 11, no. 4, 2020, doi:10.3390/info11040193.

H. G. Ceballos, R. Morales-menendez, and R. A. Ramírez-, “A Research-based Learning Approach to Teach Data Science using Covid-19 and Related Domains,” pp. 1–28.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Syntethic Minority Over-Sampling Technique,” J. Artif. Intell. Res., 2002, doi: 10.1613/jair.953.

Learn about Stroke. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learnabout-stroke (accessed on 25 May 2022).

Elloker, T.; Rhoda, A.J. The relationship between social support and participation in stroke: A systematic review. Afr. J. Disabil. 2018, 7, 1–9.

Katan, M.; Luft, A. Global burden of stroke. In Seminars in Neurology; Thieme Medical Publishers: New York, NY, USA, 2018; Volume 38, pp. 208–211.

Bustamante, A.; Penalba, A.; Orset, C.; Azurmendi, L.; Llombart, V.; Simats, A.; Pecharroman, E.; Ventura, O.; Ribó, M.; Vivien, D.; et al. Blood biomarkers to differentiate ischemic and hemorrhagic strokes. Neurology 2021, 96, e1928–e1939.

Xia, X.; Yue, W.; Chao, B.; Li, M.; Cao, L.; Wang, L.; Shen, Y.; Li, X. Prevalence and risk factors of stroke in the elderly in Northern China: Data from the National Stroke Screening Survey. J. Neurol. 2019, 266, 1449–1458.

Alloubani, A.; Saleh, A.; Abdelhafiz, I. Hypertension and diabetes mellitus as a predictive risk factors for stroke. Diabetes Metab. Syndr. Clin. Res. Rev. 2018, 12, 577–584.

Boehme, A.K.; Esenwa, C.; Elkind, M.S. Stroke risk factors, genetics, and prevention. Circ. Res. 2017, 120, 472–495.

Mosley, I.; Nicol, M.; Donnan, G.; Patrick, I.; Dewey, H. Stroke symptoms and the decision to call for an ambulance. Stroke 2007, 38, 361–366.

Lecouturier, J.; Murtagh, M.J.; Thomson, R.G.; Ford, G.A.; White, M.; Eccles, M.; Rodgers, H. Response to symptoms of stroke in the UK: A systematic review. BMC Health Serv. Res. 2010, 10, 1–9.

Gibson, L.; Whiteley, W. The differential diagnosis of suspected stroke: A systematic review. J. R. Coll. Physicians Edinb. 2013, 43, 114–118.

Rudd, M.; Buck, D.; Ford, G.A.; Price, C.I. A systematic review of stroke recognition instruments in hospital and prehospital settings. Emerg. Med. J. 2016, 33, 818–822.

Delpont, B.; Blanc, C.; Osseby, G.; Hervieu-Bègue, M.; Giroud, M.; Béjot, Y. Pain after stroke: A review. Rev. Neurol. 2018, 174, 671–674