Kritika Ghimire 22068011 3.docx

Islington College,Nepal

Document Details

Submission ID

trn:oid:::3618:79776829

Submission Date

Jan 22, 2025, 12:33 AM GMT+5:45

Download Date

Jan 22, 2025, 12:42 AM GMT+5:45

File Name

Kritika Ghimire 22068011 3.docx

File Size

44.6 KB

43 Pages

6,894 Words

37,646 Characters

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

114 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

12%	Internet sources
6%	Publications
16%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

114 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

12%	Internet sources
6%	Publications
16%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.frontiersin.org	2%
2	Submitted works	islingtoncollege on 2025-01-16	<1%
3	Internet	www.peeref.com	<1%
4	Submitted works	University of College Cork on 2024-02-15	<1%
5	Submitted works	University of North Texas on 2024-03-08	<1%
6	Internet	ouci.dntb.gov.ua	<1%
7	Internet	jieee.a2zjournals.com	<1%
8	Submitted works	George Bush High School on 2023-08-30	<1%
9	Submitted works	Georgia Institute of Technology Main Campus on 2024-06-02	<1%
10	Submitted works	University of Surrey on 2025-01-02	<1%

11	Submitted works	Coventry University on 2023-02-17	<1%
12	Internet	journalofbigdata.springeropen.com	<1%
13	Internet	pubmed.ncbi.nlm.nih.gov	<1%
14	Submitted works	Liverpool John Moores University on 2023-01-15	<1%
15	Submitted works	Liverpool John Moores University on 2021-08-16	<1%
16	Submitted works	islingtoncollege on 2025-01-16	<1%
17	Submitted works	Taylor’s Education Group on 2023-10-15	<1%
18	Submitted works	The Robert Gordon University on 2024-03-17	<1%
19	Submitted works	islingtoncollege on 2025-01-16	<1%
20	Internet	aiforsocialgood.ca	<1%
21	Internet	www.jmdet.com	<1%
22	Submitted works	University of Portsmouth on 2025-01-15	<1%
23	Internet	www.ssgmce.ac.in	<1%
24	Submitted works	University of Portsmouth on 2025-01-15	<1%

25	Submitted works	islingtoncollege on 2025-01-16	<1%
26	Submitted works	Westford School of Management on 2024-07-05	<1%
27	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2024-10-17	<1%
28	Submitted works	islingtoncollege on 2025-01-16	<1%
29	Internet	www.mdpi.com	<1%
30	Submitted works	De Montfort University on 2021-09-03	<1%
31	Submitted works	University of North Texas on 2023-11-16	<1%
32	Submitted works	Bournemouth University on 2023-08-24	<1%
33	Publication	Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain,...	<1%
34	Internet	eprints.uthm.edu.my	<1%
35	Submitted works	AlHussein Technical University on 2025-01-20	<1%
36	Submitted works	Iowa State University on 2024-05-03	<1%
37	Submitted works	Stourbridge College on 2019-02-05	<1%
38	Submitted works	University of Wales Institute, Cardiff on 2023-11-12	<1%

39 Internet

www.igi-global.com <1%

40 Internet

www.ijraset.com <1%

41 Publication

Shah, Uzair. "Mind Reading! Decoding Imagined Speech From Brain Signals", Ham... <1%

42 Internet

eprints.nottingham.ac.uk <1%

43 Internet

fastercapital.com <1%

44 Internet

medium.com <1%

45 Internet

www.geeksforgeeks.org <1%

46 Internet

tehqeeqat.org <1%

47 Internet

4spepublications.onlinelibrary.wiley.com <1%

48 Submitted works

University of Portsmouth on 2025-01-15 <1%

49 Submitted works

University of Portsmouth on 2025-01-15 <1%

50 Submitted works

Gitam University on 2024-12-26 <1%

51 Publication

Gourav Bathla, Sanoj Kumar, Harish Garg, Deepika Saini. "Artificial Intelligence in... <1%

52 Internet

dokumen.pub <1%

53	Internet	
ijercse.com		<1%
54	Internet	
researchinventy.com		<1%
55	Internet	
www.researchgate.net		<1%
56	Submitted works	
City University on 2022-05-01		<1%
57	Submitted works	
Softwarica College Of IT & E-Commerce on 2022-02-14		<1%
58	Submitted works	
Southampton Solent University on 2021-09-07		<1%
59	Submitted works	
Universiteit Hasselt on 2024-08-20		<1%
60	Submitted works	
University of Bedfordshire on 2024-12-19		<1%
61	Internet	
qspace.qu.edu.qa		<1%
62	Internet	
www.irjet.net		<1%
63	Internet	
www2.mdpi.com		<1%
64	Submitted works	
De Montfort University on 2023-09-01		<1%
65	Submitted works	
University of Wales, Lampeter on 2024-01-15		<1%
66	Submitted works	
Brunel University on 2023-09-13		<1%

67 Submitted works	
Coventry University on 2021-04-01	<1%
68 Submitted works	
University of Salford on 2023-09-28	<1%
69 Submitted works	
University of Westminster on 2025-01-09	<1%

INTRODUCTION

Topic and Concept Visualization

The topic that is chosen is for the stroke disease. Stroke is considered to affect many people every year and is a major cause of disability also. A large stroke can cause in death, some strokes cause restricted physical abilities, weakness or paralysis of limbs and many more. We can find some many data analyses of this disease called stroke. This project contains supervised learning especially a classification model to help predict if a person is suffering from a stroke based on the data. Machine learning models have described remarkable accuracy in analysis, diagnosing, guiding medical treatment, and predicting patients' conditions.

The early process of stroke risk factors can help in prevention and the dataset provided helps in predicting or classifying patients having a stroke or no stroke. Classification algorithms are widely used in healthcare for tasks like disease diagnosis, prediction of risks factors in disease, patient conditions etc. This system will also assist healthcare professionals in identifying high-risk patients and also helps in early preventative measures. This project's goal is to predict whether a person is at a risk of this disease. The dataset look up with the demographic factor like age, gender, residence type. Health sector like hypertension, heart disease, glucose levels, BMI and lifestyle factors such as marital status, smoking status, and work type.

This project follows the classification algorithm. There 2 types classification algorithm they are supervised and unsupervised classification. Supervised classification is used when the algorithm is trained on a dataset which containing input features and the categories that classifies new data based on the learning patterns. Supervised classification contains algorithms like Logistic regression, K-NN, random forest etc. which are being used in the project. Unsupervised classification which is also known to be Clustering does not uses predefined labels. This classification uses algorithm groups similar data points together based on pattern and similarities without the knowledge of categories. Common unsupervised algorithms is K-means Clustering etc.

AI Association with the Concept

29 This project is focusing on including different algorithms of Artificial Intelligence and
machine learning to detect the risk of stroke. Machine Learning classification algorithm
40 is used, algorithms like Logistic Regression, K-Nearest Neighbors (KNN), and Random
Forest. This algorithms are applied to explore the effectiveness in predicting the risk of
stroke. This algorithms are most likely to provide accuracy in train, test, and predict so
these are being used. These algorithms also provides understanding about the
importance of different factors that is influencing stroke. Here, is the brief intro for the
59 AI algorithm that are being used in this project.

Logistic Regression

2 Logistic Regression is considered to be a machine learning algorithm which is mainly used for classification tasks. It helps in predicting the probability of a instance belonging to a certain class. By handling categorical variables, it makes the process of implementation, interpretation, and training more simple, which makes it a easy and widely used method.

(KNN) K-Nearest Neighbor

K nearest Neighbour is a machine learning algorithm which is working for both classification and regression tasks. In KNN, "K" represents the number of nearest neighbors considered for making predictions. The algorithm is assuming that similar data points are close to each other, which makes it particularly effective for classification problems. KNN is versatile and has applications in fields like data preprocessing, finance, and healthcare.

Random Forest

Random Forest is a supervised machine learning algorithm used for both classification and regression. It is operated by building multiple decision trees, combining their outputs to make more accurate predictions. The more trees in the forest, there is better accuracy and the lower the risk of overfitting the data. Random Forest is efficient, requiring less training time, and it can maintain good performance even with missing data, making it reliable for complex datasets.

Problem domain

Stroke is a critical health condition that is affecting in peoples life among the world. This is considered to be one of the cause of people death by not getting access to early time diagnosis and care. There are key factors influencing the risk of stroke which also includes demographic attributes like age, gender, and residence type, also medical history like hypertension, heart disease, and average glucose levels, and lifestyle of a individual like their marital status, smoking status, and work type. This project scales the problem domain by focusing on early detection and prevention, which provides opportunity to take preventive measures and avoid the risks factors mostly in the country or state with low and medium income. These low and medium developed countries are more in the risks factors. The dataset that will be used provides the model which can help in capturing the risks factors making applicable to the many people and healthcare systems.

Motivation

Stroke is a dangerous health condition people suffer from. Each person have their own individual dreams and aims because of this condition and no chance of early detection they cannot fulfill all the aims dreams they are thinking of. This project is done to help a little by early detection of risks. This is one step to finding a way for people by implementing algorithms which will reduce the burden of the patient. These algorithms helps by analyzing data which can help in revealing the hidden insights about the risks factors of stroke. This research aims to make a impactful fight against the stroke.

Aim and Objectives

AIM

34 The aim of this project is to develop a system that is capable of predicting or detecting the risk of stroke based on the data which are provided by the patient. This project is willing to help the healthcare providers and patient a system that can easily predict stroke and help them in adapting preventive measures as fast as possible.

Objectives

To collect the datasets which contains data and information that are relevant to the stroke risks prediction.

To implement different algorithms like KNN, random forest, logistic regression for the tasks of classification.

To create a user-friendly model that can be useful for healthcare for the practical use.

To provide the vision of importance of various stroke risk factors and their influence on the prediction.

To improve the outcomes for individuals at risks of stroke.

BACKGROUND

Much research was conducted for the selection of the topic. Watching different YouTube videos, searching for the datasets that could be suitable for the prediction. With the help of different research, teachers consultation, and friend's help the stroke prediction was finalized as a topic for this given coursework. The main goal of this stroke disease predication is to identify the individuals at high risk of stroke early, and medications could help them to improve their health.

Research Done on Chosen topic and problem domain

In this world, over 12 million people will have stroke this year and half of them which is 6.5 million will die because of stroke. This disease called stroke increases significantly with the age. The rates of stroke is growing the fastest mainly in the countries whose people income is low and middle, where healthcare providers finds it challenging to provide proper care needed for effective prevention and treatment of stroke.

Artificial Intelligence (AI) and Machine learning (ML) techniques are very powerful tools for analyzing large datasets to predict the stroke risks more accurately. Many supervised model or classification algorithms like logistic Regression, Random forest, and K-nearest neighbors (KNN) to predict the Stroke risk for a person. By analyzing the patterns in datasets, these algorithm can classify or predict patients is at risk of stroke or not.

There are lots of challenges and complexity of risks factors while predicting the models. Some of the problem many include in the incomplete patient datasets. Some patients datasets may got missing or the value might be mistake which makes it challenging to generate true or reliable predictions. People may doubt on the predictions made by AI if they are accurate and fair as sometimes it may predict wrong. The models should be easy to understand so that doctors can analyze and trust the results.

Description of the Selected Domain

This stroke disease belongs to the health sectors. Health sectors are very essential component of the society which are responsible for the well-being goodness of the overall people and the country. Health sector should have all the technologies that are required to treat, and prevent the illness and to promote overall health. Health sector plays a significant role in the identification of risk of an individuals and providing timely preventive measures. In the health sector medical professionals, doctors, nurses, and other members work together to monitor a patient's health and tools like stroke predictions can be a great tool to predict and detect he risks of a individual as health care sector is responsible for determining which patients are at a high risk and what preventive measures should they be provided. Machine learning algorithms like random forest, knn, logistic regression have potential to analyze complex data and predict the risk and likelihood of any medical conditions. By utilizing these tools healthcare sector can be in benefits and they can provide more accurate and timely information for risks and non-risks factors in any diseases.

Advantages of working around the problem domain/ algorithm

There are several advantages of working in stroke prediction and algorithm.

Helps in predicting high-risk individuals early and it allows patient to use preventive measures.

These algorithms helps in finding more detailed information through the data that has been provided or collected.

This allows healthcare providers to create more treatments plans based on the patients conditions.

The technology that can be scaled for helping people all over the world especially in those areas where healthcare is limited.

Chances of mistakes of human in prediction of the risks are reduced, which helps in making health care more reliable.

Considered Drawbacks Revolving Around the Problem Domain/Algorithm Implementations

There are some drawbacks related to this problem domain or algorithm. Some of them are:

There can be issues related to data qualities. If the data that being used for the training model is incomplete or wrong, the prediction can be inaccurate which can be problematic.

Sometimes AI models can be so complex that can make medical professionals hard to understand why a prediction was made. This can cause trust issues.

This also may concern about the privacy issues while handling sensitive information. It can make patients concern about their privacy.

There can be a risk that AI cannot make the decisions that are right and useful with human judgement.

These models need to be update at the certain period of time with fresh and accurate data with proper monitoring which may be tiring.

Without the proper monitoring of the system their predictions may become outdated.

Designed Dataset

The dataset for stroke prediction created for capturing essential details about a person’s health, lifestyle, and demographic factors that contribute to their risk of having of having a stroke. Datasets contains personal information including age, gender, and residence type. This personal information could be beneficial to correctly predict so these data are included. There are the health factors such as hypertension, heart disease, average glucose level, BMI, which also helps in the accuracy prediction. Lifestyle details including smoking status, work type, marital status to find out how healthy lifestyle a individual have. This datasets targeted outcome is to show whether a person has experienced a stroke. The dataset is designed to be user friendly and reasonable for the analysis. By using this dataset, we can discover the things that can help in predicting stroke risks early. This ca help in better prevention strategies and timely medical prevention for those risks. This holds the potential to make a good impact in prediction of stroke and prevention methods.

Details and Background of Dataset

The dataset that is chosen for the prediction of stroke contains important about a person’s health, lifestyle, and demographic, making easy for analyzing stroke risks. The dataset is sourced from a reliable platform. The dataset includes reliable information useful for the prediction process. The datasets includes all the major factors associated with stroke risk like age, health conditions, lifestyle etc. it provides clear data that can be used to train machine learning models for predicting risks. The dataset includes variety of people who have and haven’t experienced a stroke, which helps in creating accurate models. Some data have missing values which were filled with averages to ensure no information was lost. The dataset is designed to build machine learning models that predicts if someone is likely to experience stroke.

Figure 1 Datasets

Some information on Exploratory Data Analysis (EDA)

Exploratory data is understanding and preparing the dataset before building a model. EDA process takes place with a work flow that consists of:

Understanding the datasets: Firstly, should understand the structure of a dataset. After understanding those structures then can clean the datasets. Without understanding further action will be difficult to take.

Handling all the Missing values or data cleaning: There is missing values in BMI column which are filled with median, ensuring there are no gaps.

Data preprocessing: In this data preprocessing numerical data which has missing value was adjusted. For the categorical value s method called OneHotEncoder was used. These steps helps in making data cleaner, more organized, and ready for the model to process. This ensures the algorithm treats all the inputs and give accurate results.

26 Data Splitting: The data is splited into training and testing sets.

Analytical Review of Existing systems on the Problem Domain

Some of the existing work in the Problem domain that have used AI techniques to address stroke prediction are listed below:

Artificial Intelligence and Acute Stroke Imaging

Authors: JE Soun , DS Chow , M Nagamine, RS Takhtawala , CG Filippi, W Yu , PD Chang

Stroke prediction has been a major focus area in research due to its significant impact on health worldwide. AI, particularly in acute stroke imaging, has been progressing rapidly. According to JE Soun and team, AI techniques like Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Neural Networks are being widely used for stroke imaging and early detection. These models typically 20 depends on a limited number of adjustable parameters to optimize their performance. AI techniques like Random Forest (RF), Support Vector Machines (SVM), and Neural Networks analyzing medical images to identify early signs of a stroke. Models like KNN and RF optimizes performance by fine-tuning parameters, leading to reliable and repeatable results. By using limited parameters, these models provide a hope for improving machine learning techniques in stroke detection (JE Soun, 2024).

The Most efficient machine learning algorithms in stroke prediction: A systematic review

Authors: Farkhondeh Asadi , Milad Rahimi, Amir Hossein Daeechini, Atefeh Paghe

As mentioned by Farkhondeh Asadi and colleagues, stroke remains one of the leading causes of death globally, causing severe complications and reducing life quality. Random Forest (RF) has been pointed out as one of the best machine learning algorithms for stroke prediction. Other algorithms like SVM, Stacking, XGBOOST, DSGD, COX & GBT, ANN, Naive Bayes (NB), and RXLM have also shown efficiency. There has been a remarkable increase in the use of ML for stroke prediction in recent years, leading to significant improvements in the accuracy of these models. These machine learning algorithms have made stroke predictions more accurate. This means doctors can identify people at higher risk of having a stroke earlier, leading to faster treatments and better chances of preventing a stroke from occurring. The study mentions several different algorithms, like SVM, XGBOOST, and Neural Networks. This variety gives researchers more options, so they can choose the one that works best for the specific data they have. With more accurate predictions, healthcare providers can make better decisions about who needs immediate care, lifestyle changes, or further testing (Asadi, 2024).

Artificial intelligence in ischemic stroke images

6 Authors: Ying Liu Zhongjian Wen Yiren Wang Yuxin Zhong

6 This paper reviews the progress in using Artificial Intelligence (AI) for ischemic stroke imaging, highlighting key challenges and future research directions. It focuses on how AI is being applied to tasks like automatically identifying infarcted areas, detecting large vessel blockages, predicting stroke outcomes, assessing the risk of hemorrhagic transformation, forecasting the likelihood of recurring strokes, and grading collateral circulation. The study shows that Machine Learning (ML) and Deep Learning (DL) have great potential to improve diagnostic accuracy, speed up disease detection, and predict how a stroke might progress or respond to treatment. However, challenges remain in using these technologies clinically, such as limited data availability, difficulty in understanding how models work, and the need for real-time updates and monitoring. AI is also used to assess the risk of more severe issues like hemorrhagic transformation where a stroke turns into a bleeding event or the chance of future strokes. AI is being used to automatically detect key areas in ischemic stroke images, such as effected areas where brain tissue is damaged and large vessel blockages.

This helps doctors identify the severity and exact location of the stroke more quickly, which is crucial for providing timely treatment. AI helps predict how a stroke might respond to different treatments (Liu, 2024).

The predictive performance of artificial intelligence on the outcome of stroke: a systematic review and meta-analysis

Authors: Yujia Yang Li Tang Yiting Deng Xuzi Li Anling Luo Zhao Zhang Li He Muke Zhou

This study aimed to assess the accuracy of artificial intelligence (AI) models in predicting the prognosis of stroke. Artificial intelligence (AI) can be defined as the ability of computers or other machines to demonstrate intelligent behavior, like human being. ML techniques utilize various methods for automated data analysis, including logistic regression (LR), random forests (RF), support vector machines (SVM), and classification trees, which allow combining features (data characteristics) with flexible decision boundaries in a non-linear manner. Acute stroke ranks among the leading causes of morbidity and mortality worldwide, and it can be divided into ischemic stroke and hemorrhagic stroke. In addition, predicting the outcome of a stroke often depends on the experience of the physician clinically, but it is difficult for inexperienced young physicians to judge the prognosis. In clinical, patients are most concerned about their clinical outcomes. ML predictive models which are image-based feature recognition and segmentation and have greatly facilitated the rapid diagnosis of stroke, but stroke prognosis depends on a large number of patient-specific and clinical factors, so accurate prognostic prediction models remain challenging (yang, 2023).

Selected System/Algorithms with Problem Domain

For this research, we’ve selected three algorithms: Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN). These algorithms are popular choices in healthcare, especially for tasks that involve in reviewing patient data, such as predicting the risk of stroke. These algorithms are best for analyzing and predicting the datasets. Here, logistic regression is simpler yet powerful method for binary classification. KNN is effective for making predictions based on how similar data point is to others which can be useful in identifying patterns in patients health data. Random forest seems to have high accuracy as it creates multiple decision trees.

Summarized Review and Analysis

The growing use of AI and machine learning in predicting stroke risk, shows how algorithms like Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN) have become so much important tools in the healthcare sector and many other sectors too. There are many challenges also that should be identify and make more reliable of AI in further future in the healthcare sectors. Due to these tools and algorithms, studies, predictions, detections are being improved and helping healthcare providers identify high-risk people.

This AI system have great potential for stroke prediction and prevention but more research and development are needed to be done for the existing challenges. Additionally, these technologies help in reviewing stroke outcomes, identifying risk factors, and improving diagnostic accuracy. While these advancements show great potential, there are challenges such as incomplete patient data, some inabilities of AI models, and the need for integration. But the use of ML and AI in healthcare continues to advancement, leading to improvements in prediction accuracy and ultimately better stroke prevention and treatment. The algorithms used in this research aim to solve these challenges by analyzing patient data to predict stroke risk, assisting healthcare professionals in making timely and informed decisions.

System Architecture

The system architecture for stroke prediction starts with Data Collection, where patient details like age, BMI, and medical history are gathered. The data is cleaned, encoded, and scaled in the Data Preprocessing stage. Next, in the Model Training phase, algorithms like Logistic Regression or Random Forest analyze the data to learn patterns. The trained model is then used in the Prediction Layer to assess stroke risk based on new inputs. Finally, the User Interface allows healthcare professionals to input data and view predictions in a simple, user-friendly dashboard.

Figure 2 System Architecture

SOLUTIONS

It is considered that stroke is the second leading cause of death in the world, responsible for many deaths every year and leading the cause of disability among the patients. AI has played an significant role in transforming the care of stroke patient with the help of introducing many AI predictions platforms where physicians can detect patients with stroke more efficiently and accurately. By identifying high-risk patients early, these tools enables healthcare professionals to take preventive measures, improving the chances of preventing strokes or minimizing their impact. These AI solutions analyze medical data, like patient histories, brain scans, and lifestyle information, to predict chances of a stroke. This early detection helps doctors to act quickly to prevent strokes.

Proposed approach to solve the problem

To solve the problem of stroke prediction, plan to use machine learning algorithms that analyze patient data and identify the risk factors for stroke. The main goal is to develop a system that can predict strokes accurately and help doctors take early action. The first step involves collecting a dataset with important information like age, gender, health history, lifestyle habits, and other medical details. After that, the data will be cleaned and processed to make sure it is ready for analysis. Once the data is prepared, we will use algorithms like Random Forest, Logistic Regression, and K-

7 Nearest Neighbors (KNN) to train a model that can predict whether a person is at risk of stroke.

This approach focuses on creating a user-friendly system where data are provided and by the help of those data prediction can be instant. This helps in fast and early diagnosis and better management of the condition. By this predictions can be more accurate which helps in saving lives and reducing impact of strokes. By gathering all the data and doing all the steps we can finally find the results where the problem is solved. This approach aims to help all the health care sectors by improving patient health and quality of life.

Details of Algorithms

There are many algorithms that can be used while predicting this stroke disease. 47 Some of the AI Algorithm that are being used in this project are Logistic regression, Random Forest, and K-nearest Neighbors.

K-Nearest Neighbors

12 The K-Nearest Neighbors(KNN) algorithm operates on the principle of likelihood of 12 similarity. KNN is also categorized as a lazy learning algorithm that formulates predictions by analyzing the data structure in the data structure in real-time. KNN is the 51 simple and easy algorithm to understand. It doesn't make assumptions about the data like assuming that it follows a specific pattern. It can handle both types of problem in

classification and regression. KNN can be slow if the dataset is very large because it has to calculate distances for every points. It is important to choose the right value for k is important and too small and too large file can be issues. In the healthcare, KNN is to predict diseases by comparing patient symptoms with others. KNN calculates how far away the other data points are from the one that is being tried to predict. In KNN, once the distance is calculated, it picks the k closest ones.

44

Random Forest

Random forest is the machine learning algorithm that is mainly used for predictions or either to classify data into groups or to predict numbers. It is so much accurate then other because it combines the knowledge of many trees. It can handle datasets with missing values better than any other algorithms. It works very well for both classification and regression problems. It can show which features in the data are most important for making predictions. Random forest uses multiple trees so it may take a bit more longer to train and predict. It's harder to understand why it makes certain predictions compared to simpler models like a single decision tree. It can use a lot of memory when the dataset is large. Random forest is used in many sectors such as health sector, finance,ecommerce and so on. In, Healthcare predicting patient outcomes or diseases based on symptoms. In, finance detecting frauds and in e-commerce recommending products to users by analysing past data.

45

Logistic Regression

Logistic regression is a simple yet powerful machine learning algorithm that is mostly used for classification problems not likely for linear regression, which predicts the 53 continuous values. Logistic regression predicts categorical outcomes. logistic regression also works by applying a mathematical function called logistic function. It is easy to understand and also easy to implement. It performs well when the data contains linear relationship between input features and the output. It gives probabilities that can be helpful for the further analysis. If there are too many features, it can overfit the data. Logistic regression can be affected by extreme values in the dataset. Logistic regression is also used in healthcare, finance, marketing and soon. It estimates the chance of an event that is happening and then it makes a decision based on the probabilities.

Pseudocode and Mathematical Implementation

Pseudocode of logistic regression

```
START
Step 1: Import required libraries
   Import pandas, numpy, matplotlib, scikit-learn libraries


Step 2: Load the dataset
   Load dataset into a dataframe using pandas (pd.read_csv)


Step 3: Preprocessing of dataset
   - Handle missing values in the 'bmi' column by filling with the median
   - Normalize/scale numerical features using StandardScaler (age, hypertension, heart disease, avg_glucose_level, bmi)
   - Encode categorical features using OneHotEncoder (gender, ever_married,
```

work_type, Residence_type, smoking_status)

Step 4: Split the dataset

31 - Split the dataset into training and testing sets (70% for training, 30% for testing) using train_test_split

18 - Define X (features) and y (target variable: stroke)

2 Step 5: Train the model

- Initialize Logistic Regression classifier with a maximum of 1000 iterations

2 - Fit the model using the training data

Step 6: Generate predictions

- Use the trained model to predict outcomes on the testing set

2 Step 7: Evaluate the model

39 - Calculate accuracy score, classification report (precision, recall, F1-score), and confusion matrix to evaluate the model's performance

Step 8: Visualize and generate results

33 - Print accuracy, classification report, and confusion matrix

- Optionally, create visualizations (e.g., ROC curve) for better understanding of the results

END

Pseudocode of KNN algorithm

```
START
  Step 1: Import Required Libraries
    Import pandas, scikit-learn (train_test_split, KNeighborsClassifier, etc.), matplotlib

  Step 2: Load the Dataset
    Load the dataset from the CSV file into a pandas DataFrame

  Step 3: Handle Missing Values
    Replace missing values in the 'bmi' column with the median of that column

  Step 4: Define Features and Target
    Define the feature columns (X) and the target variable (y)

  Step 5: Identify Categorical and Numeric Columns
    Define which columns are categorical and which are numeric

  Step 6: Preprocess Data
    Apply transformations:
      - Scale numeric columns using StandardScaler
```

18 - Encode categorical columns using OneHotEncoder

5 Step 7: Split Data into Training and Test Sets

Split data into training set (70%) and test set (30%) using train_test_split

Step 8: Create KNN Model

Initialize a KNN classifier with k=5 neighbors

Step 9: Build a Pipeline

Combine the preprocessing steps and classifier into a pipeline

Step 10: Train the Model

9 Train the KNN model using the training data (X_train, y_train)

Step 11: Make Predictions

Use the trained model to make predictions on the test data (X_test)

17 Step 12: Evaluate the Model

Calculate the accuracy of the model using accuracy_score

Generate a classification report using classification_report

Generate a confusion matrix using confusion_matrix

5 Step 13: Calculate ROC and AUC

```
    Calculate the ROC curve and AUC score

  Step 14: Plot ROC Curve
    Plot the ROC curve using matplotlib, and display the plot

  Step 15: Display Results
    Print accuracy, classification report, confusion matrix, and AUC score

END
```

Pseudocode of Random Forest

```
START

  Step 1: Import Required Libraries
    Import pandas, scikit-learn (train_test_split, RandomForestClassifier, etc.), matplotlib

  Step 2: Load the Dataset
    Load the dataset from the CSV file into a pandas DataFrame

  Step 3: Handle Missing Values
    Replace missing values in the 'bmi' column with the median value of that column
```

Step 4: Define Features and Target

Define the feature columns (X) and the target variable (y)

Step 5: Identify Categorical and Numeric Columns

Specify which columns are categorical and which are numeric

Step 6: Preprocess Data

Apply preprocessing steps:

- Standardize numeric columns using StandardScaler 18
- Encode categorical columns using OneHotEncoder (dropping the first category)

Step 7: Split Data into Training and Test Sets 38

Divide the dataset into training (70%) and test (30%) sets using train_test_split, ensuring stratified sampling

Step 8: Create Random Forest Model

Initialize the Random Forest classifier with: 65

- Random seed for reproducibility
- Number of trees in the forest set to 100 63

Step 9: Build a Pipeline

Combine preprocessing steps and the Random Forest model into a pipeline

14 Step 10: Train the Model

9 Fit the pipeline to the training data (X_train, y_train)

Step 11: Make Predictions

25 Use the trained model to make predictions on the test data (X_test)

Step 12: Evaluate the Model

Calculate evaluation metrics:

- Accuracy using accuracy_score

56 - Classification report using classification_report

- Confusion matrix using confusion_matrix

Step 13: Calculate Probabilities and ROC Curve

48 Get probabilities for the positive class (stroke=1) using predict_proba

42 Compute the ROC curve (False Positive Rate vs. True Positive Rate)

Calculate the Area Under the Curve (AUC) using roc_auc_score

5 Step 14: Plot ROC Curve

Plot the ROC curve using matplotlib, with:

32 A blue line representing the ROC curve

-A dashed gray line representing a random classifier

Step 15: Display Results

```
    Print:
        Model accuracy
        Classification report
        Confusion matrix
        AUC-ROC score

END
```

Flowchart

Figure 3 Flow chart of KNN algorithm

Figure 4 Flowchart of Logistic Regression

Figure 5 Flow chart of Random Forest

Data Description

Data Preparation

Data Import

Figure 6 Data importing

Load Datasets

Figure 7 loading datasets

Handle missing values

Figure 8 Handling missing values

Data structure info

Figure 9 Info of data structure

Feature Engineering

Figure 10 defining feature

Identify Categorical and Numerical Columns

Figure 11 identifying categorical and numerical columns

Data Pre-processing

Figure 12 Data Pre-Processing

Logistic Regression

Split into train and test

Figure 13 split into train and test

Create pipeline and train model

Figure 14 Creating pipeline and train model

Making predictions

Figure 15 Making prediction

Evaluate Model

Figure 16 Evaluation of model

Data Visualization

Figure 17 Confusion Matrix data visualization

Figure 18 AUC-ROC curve

Random Forest

Split into train and test

Figure 19 Splitting into train and tests

Create pipeline and train Model

Figure 20 Creating pipeline and model for random forest

Evaluate and predict the model

Figure 21 Evaluation and prediction of model

Data Visualization

Figure 22 Data Visualization Confusion matrix

Figure 23 AUC_ROC data visualization

K-NN

Split into train and test sets

Figure 24 Split into train and tests

Evaluate pipeline and train model

Figure 25 Evaluate pipeline and train model

Make predictions and evaluate model

Figure 26 Making predictions and evaluating the model

Data Visualization

Figure 27 confusion matrix data visualization

Figure 28 AUC_ROC data visualization

3.4.5. Data Visualization

Figure 29 count of gender in bar graph

Figure 30 Histogram of age

Figure 31 Histogram of average glucose level

Figure 32 Bargraph of Work type

Figure 33 Histogram of hypertension

Data Procedure

1. Understanding the Problem

First, need to figure out exactly what we want to solve. In this case, the goal is to predict whether someone is at risk of having a stroke based on their health and lifestyle information, like their age, medical history, and lifestyle choices.

2. Gathering the Data

Next, need data to work with. For stroke prediction, this means collecting information about people’s health, such as their age, gender, if they have conditions like high blood pressure, their lifestyle habits like smoking, and other important health details. This data can come from hospitals, health studies, or even public health records.

3. Cleaning and Preparing the Data

Once we have the data, we need to get it ready for analysis. Some information might be missing in the dataset (like someone's BMI). To deal with this, we can fill in the blanks using the average or median value from the rest of the data. Some data is non-numeric, like gender or marital status. We need to convert these into numbers so the machine learning model can understand them. This is where techniques like One-Hot Encoding come in handy.

4. Picking the Important Features

Not all information in the dataset might be equally important for predicting a stroke. So, we need to decide which features matter most. For example, age, blood pressure, and glucose levels might be more important than someone's work type. We select the most important features to make our model smarter.

5. Choosing the Right Model

Now, we pick the machine learning model to use. For stroke prediction, we could choose models like:

Logistic Regression: A simple model that works well when we want to predict.

Random Forest: A more complex model that can handle a lot of data and can help us figure out which features are most important.

K-Nearest Neighbors (KNN): This model looks at similar people in the data to make predictions.

6. Training the Model

With the right model chosen, we then train it. This means we feed it the data so it can learn from the patterns in it. The more data it sees, the better it gets at making predictions.

32 7. Testing the Model

After the model has been trained, we need to see how well it performs. We do this by testing it on new data that it hasn't seen before. We can check how accurate the predictions are and see if there are areas where the model could improve.

8. Using the Model to Make Predictions

14 Finally, once the model is trained and tested, we use it to make predictions. So, if we have new patient data, the model can help doctors figure out if that person might be at

15 risk of having a stroke.

This is the basic process where understanding the problem, gathering the data, cleaning, picking a good model, training it, testing it, and then using it to make predictions.

Comparison Between Algorithms

After comparing all of the algorithms that are being used, Logistic regression and K-NN

61 both gave the accuracy of 0.9517 and random forest gave accuracy of 0.9511 which is

slightly lower than the other two algorithms. Here, we can find out logistic regression and K-NN is better than random forest but there is little difference in the accuracy which means all of these models are performing equally and well. Each algorithms have there own beneficial in using as the performance might have there own performance.

Algorithm

Accuracy

Advantage

Best based on Accuracy

Logistic Regression

0.9517

This algorithm is simple fast, and easy to use.

Highest Accuracy

K-Nearest Neighbours (K-NN)

0.9517

There is no assumption about data distribution and works well in small datasets.

Highest Accuracy

Random

Forest

0.9511

This Handles complex data and also work well with missing values.

Slightly lower than highest accuracy.

Table 1 Comparison table

Between logistic regression and K-NN logistic regression might be the better choice because it is easier, faster to interpret and it also works well for the medical data. Random forest can also be a good choice but in the case of this project it got slightly lower accuracy then other two algorithms.

Development Platforms

The development platform I used for this particular coursework is google Colab. Google Colab is an online platform that allows to write and run Python code right in your web browser. It's like an interactive notebook where can write code, add explanations, and show graphs all in one place. The best part is that, don't need to install anything just have to sign in with your Google account. Google Colab is a handy tool for anyone working on data science or machine learning. Colab gives you access to powerful hardware, like GPUs and TPUs, which are great for training machine learning models. Since it's linked to Google Drive, can easily share your work with others. Multiple people can work on the same project at the same time, just like collaborate on a Google Doc.

Figure 34 google colab

Language Used

In this project, the programming language used in Google Colab is Python, which is one of the most popular languages for data science, machine learning, and general-purpose programming. Python works really well with other programming languages and tools. Python is the language of choice in Google Colab because it’s easy to use, incredibly powerful, and well-supported by a large community.

Figure 35 Python image

CONCLUSION

Analysis of the Work Done

This system which is focused to predict risk of stroke in patients. This uses many 69 algorithms like KNN, logistic regression, and random forest. Stroke prediction focuses on careful data preparation and preprocessing to make ensure of the dataset is clean and reliable. Sometimes there might be some missing data, outliers and imbalanced classes which are managed properly. Datasets also include useful information like age, hypertension, glucose level, BMI, heart diseases, smoking status, etc which helped in 33 the process of predicting the medical condition of that patient who might be suffering from stroke or not.

The random forest algorithm helps in identifying diseases trends and risks by handling large datasets and enhance accuracy of the model. KNN algorithm helps in assuming 21 the similarity between the new data and available data and put new data into the category of most similar available categories. KNN is also simple to implement.

30 Logistic regression helps in observing predicting the categorical dependent variable 23 using a given set of independent variables. Logistic regression is used to classify the observations using different types of data and can easily determine the most effective variables used for classification. The implementation of modern AI tools like Python libraries like Pandas, NumPy, and Scikit-learn streamlined the data analysis, model building, and visualization processes.

Despite the challenges, the project successfully demonstrated the application of AI in predicting stroke risks, emphasizing the potential of machine learning to address critical healthcare issues. This research lays a solid foundation for future innovations in stroke prediction and prevention.

How the Solutions addresses the real-world problem of stroke

The AI based stroke prediction offers a solution that is practical to the critical issue by helping in identifying people who are at risk of experiencing a stroke. This AI prediction helps in early detection and prevention of this critical disease. This stroke prediction system save lives by helping hospitals and healthcare providers catch the risk warning signs of stroke more early. Here is how the solution addresses the real-world problem:

Supporting the Healthcare Professionals

This predicting system gives doctors a reliable data about their patients whether they are in risks or not. Accurate predictions can help or support health professionals by providing a layer of information that helps in guiding treatment and diagnosis plans, and improving patients health.

Reduced Healthcare cost

With the help of early detection it helps in preventing stroke and its high cost during treatments, medications, and long-trem care of stroke. This offers cost-effective way to manage and reduce stroke- related healthcare expenses.

Early detection and prevention

By analyzing all the datasets, this system predicts who is likely to cause stroke and who are at highly risks . This also allows hospitals, healthcare professionals to provide with appropriate measures such as prescribing suitable medications, advising healthy lifestyle or its changes, regular check-ups and many preventive measures.

Future Work

In this world AI has a lot more capacity that can make happen anything in any field. Stroke prediction mechanism can be more valuable tool in the coming future. As, this is very much needed tool in present too. To make the prediction more accurate and applicable, the system can use more amount of datasets with more information from various people with different backgrounds and medical history. The datasets also can include more health related details such as cholesterol levels, diet, healthy eating habits, exercise habits ad many more so can the system gets more easy to predict the health conditions. For more further success in future, can develop an easy user interface and features which can make the system applicable for both patients and healthcare to detect their health conditions.

Working and partnering with hospitals, clinics, and any other health providers to test the systemin real-world will also help in improving its functionality and encourage for more use among patients and health care providers. In the future adding more functionality like regular updates to both patients and health care providers about the health conditions, integrating this systems in mobile app, and ensuring all advancement of AI technology could be more effective and suitable. This stroke prediction system can be considered as a best tool that provides early prevention and detection of the risk and give a significant role in the management of the stroke risks.