1. Introduction - Shravya - done
2. Lit Survey (last)
3. Procedure and Experimental Methodology
    a. Proposed Model - bhairav
    b. System Methodology - sadhavi - done (recheck block diagram)
    c. Dataset Description - Shravya - done
    d. Data Pre-processing (sarthak and shravya) - done
    e. Data Visualization -
4. Algorithms
    a. Random Forest - Eshanika
    b. Gradient Boosting - Eshanika
    c. Logistic Regression - Keegan
    d. KNN - Sadhavi - done
    e. Decision Tree - Sarthak - done
5. Result/Evaluation Matrix
    a. Feature dependence
    b. Result analysis
6. Analysis
    a. DT benefits
    b. Drawbacks
7. Conclusion
8. Future work (opt)
9. References

1. Introduction

Data is a crucial aspect of every industry and organisation. With exponential amounts of data being produced every day, data science has experienced massive growth and found applications in several industries. Historically, the healthcare sector has been an early adopter of technological advancements and benefited greatly from them. Traditionally, diagnosis and treatment of diseases solely relied on the professionals' discretion. A large amount of essential data such as stress level, heart rate, sugar level, blood pressure, brain activities need to be observed and processed to arrive at an accurate conclusion. The entire process can be time-consuming and subject to human error. Several life-threatening ailments have been known to be cured by early diagnosis or by identifying patterns discovered from extensive research. Data-driven decision-making increases the quality of healthcare services, especially in implementing preventive measures for a medical emergency. This is evident from the detailed research done on Stroke, a medical emergency that arises due to a lack of blood flow to critical organs due to a ruptured blood vessel or blockage. This prevents oxygen from reaching the organ cells, and the oxygen-starved cells begin to die rapidly. Strokes are known to occur suddenly and require immediate medical attention. They are majorly observed in the brain, and the heart and are known as silent killers since, according to the CDC, a heart stroke occurs every forty seconds in America. With cardiac diseases on the rise, it has become imperative to establish an efficient system that inculcates the various factors known to have an impact on the heart and, based on the observed patterns, predict a person's vulnerability to such life-threatening diseases. This predictive system will enable the patients to take necessary precautions and improve the probability of early diagnosis. Such a system can be established using Machine Learning, a subset of Artificial Intelligence capable of processing copious amounts of data in a short period and building a comprehensive view of the data that assists in discovering patterns and providing practical insights. Based on the data and its patterns, a model trained by Machine Learning concepts can predict the probability of occurrence of a heart stroke in a patient based on their personal and medical history. The accuracy of this prediction is a crucial factor for successful early diagnosis; hence, this proposed research aims to conduct a comprehensive analysis of the supervised learning algorithms, their accuracy and the pros and cons of using them to design an effective Heart Stroke Prediction model.

---

2. Literature Survey

---

3. Procedure and Experimental Methodology :

SYSTEM METHODOLOGY : (check for F1 score, recall score)

Different datasets were considered to begin with the implementation. Among the pre-existing datasets available online, an appropriate dataset was collected from Kaggle for model training.

After dataset collection, the next step is ***data pre-processing*** which includes handling missing values, Null values, imbalanced data, and label encoding.

The preprocessed data attributes were then visualised to check for correlations between the different features in our dataset. ***Data Visualisation*** helps in performing feature engineering and dimension reduction. This was followed by model training.

Several models were available to choose from, based on our objective: we used Random Forest Classification algorithm, Gradient Boosting, Logistic Regression, K-Nearest Neighbors algorithm and Decision Tree Classification algorithm.

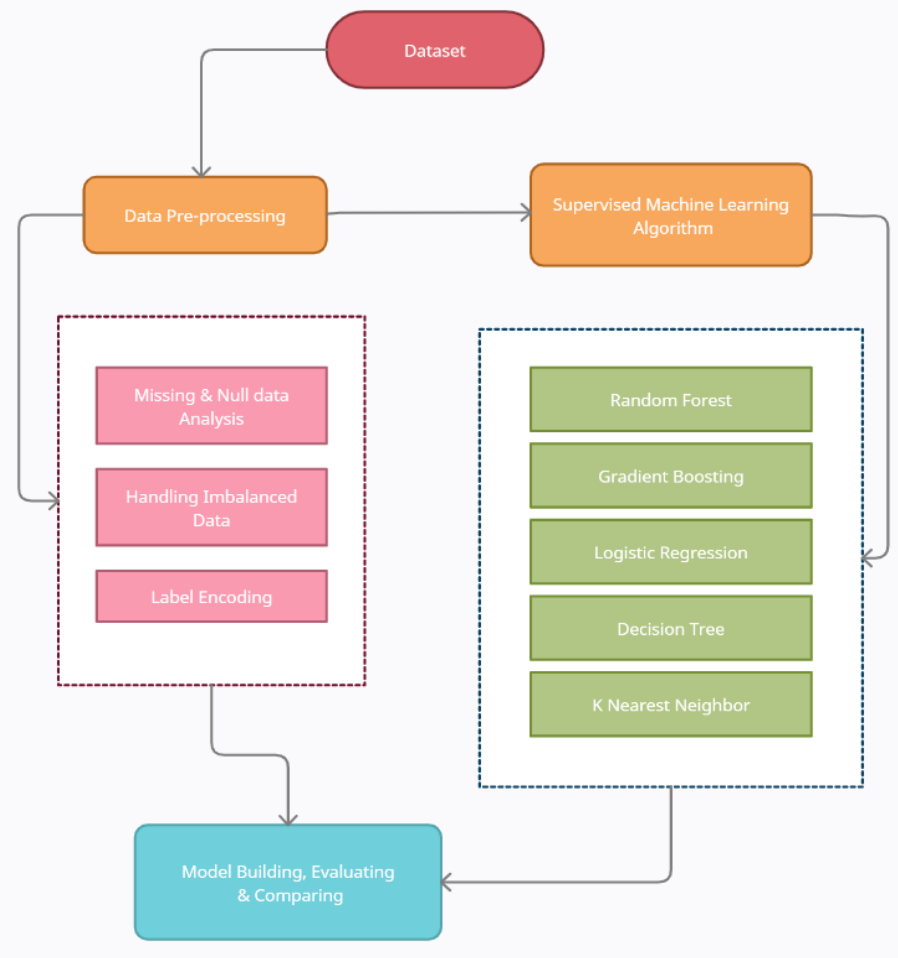

Fig 1. Block Diagram for System Methodology -

DATASET DESCRIPTION

We used the heart stroke dataset available on the Kaggle website to train the machine learning models and conduct our analysis. This dataset consists of a total of twelve attributes which have been described as follows -

1. id: It is numerical data that refers to the person's identity and counts the data present in the dataset.
2. Age: It is numerical data that refers to the person's age.
3. Gender: It is categorical data that refers to the person's gender.
4. Hypertension: It is numerical data that indicates the history of hypertension of the person.
5. Work type: It is categorical data that describes the person's employability scenario.
6. Residence type: It is categorical data that describes the person's residential area.
7. Heart disease: It is numerical data that indicates the person's prior history of heart diseases.
8. Avg glucose level: It is numerical data that indicates the person's average glucose levels.
9. Bmi: It is numerical data that indicates the person's body mass index.
10. Ever married: It is categorical data that refers to the person's marital status.
11. Smoking Status: It is categorical data that indicates the person's smoking history.
12. Stroke: It is numerical data that indicates whether a person has previously suffered from a stroke. Attribute stroke is the decision class, and the rest is the response class.

DATA PRE-PROCESSING

1. **Data Load -** In the proposed study, we use a dataset from Kaggle to train the different Machine Learning models.

2. **Exploratory Data Analysis -** We use the Pandas library in python to visualise the data into workable rows and columns and discern the patterns.

3. **Feature Engineering -** Feature engineering covers the division of the data into three categories, namely, Binary, Continuous and Categorical.
    a. **Binary:** One hot-encoding is used to convert the features Gender, Marital status and Residential area into binary (0 or 1) values.
        - **Gender:** 0 represents female, and 1 represents male
        - **Marital Status:** 0 represents unmarried, and 1 represents married.

- **Residential Area:** 0 represents rural area, 1 represents urban area.

b. **Continuous Features:** A function has been devised to divide the features of Average Glucose level, BMI and Age into constant ranges.

- **Age:** Age is divided into continuous ranges of 20 years, starting from 0-19 years, 20-39 years, 40-59 years and above 60 years.
- **Average Glucose Level:** The glucose levels have been divided into blocks of percentages with respect to the maximum value. The blocks are of 25% each, starting from below 25%, 26-50%, 51-75% and above 75%.

c. **Categorical Features:** First, we implement One-hot encoding on the previously transformed continuous features. These are then mapped into matrices of binary values, making it easier to process the data.

Note: *The **BMI** feature had 101 null values; since the data is typically distributed, the null values are replaced with the mean of the BMI as the 50% line is similar to the mean.*

---

4. Algorithms

Random Forest :

Random Forest belongs to the supervised learning technique. It can be used for Classification and Regression problems in ML. It is based on ensemble learning, which integrates several classifiers to solve a complex problem and increase the model's performance.

"Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions.

The larger the number of trees in the forest, the more accurate it is, and the problem of overfitting is avoided.

The following are some reasons why we should utilise the Random Forest algorithm:

- When compared to other algorithms, it takes less time to train.
- It predicts output with reasonable accuracy and runs rapidly even with a massive volume of data.
- When a substantial portion of the data is missing, it can still maintain accuracy.

K-Nearest Neighbor (KNN) Algorithm :
(0 plagiarism)*

K Nearest Neighbors is one of the simplest Supervised Machine Learning algorithms used for Classification. It classifies data points based on the Classification implemented on its neighbours. KNN stores all available cases and classifies new cases based on similarity measures. KNN is used when the data is labelled and noise-free and when the dataset is small because KNN is a ***Lazy Learner*** and a ***non-parametric algorithm***.

K in KNN is a parameter related to the number of nearest neighbours. It is responsible for including in the majority voting process. This algorithm is based on feature similarity. Choosing the correct value of k (parameter tuning) is vital for better accuracy.

In KNN, first, we select the number of neighbours, i.e. the value of k. Then the distance between the new point and each training point (the neighbours) is calculated using the ***distance metric***. Then the closest k data points are selected, the average of these data points is the final prediction for the new point, and our model is ready.

DECISION TREE

A decision tree is a tree-shaped diagram with leaves/nodes branched out to determine further action. Each branch represents a decision-making possibility, a reaction or an occurrence bound to happen. It is an example of a supervised learning model used for problems such as regression and classification. Regression is used when continuous values are present in the dataset, whereas classification is used with discrete data points. Decision trees are easy to use and simple to understand, interpret, and visualise.

Some advantages of Decision trees are:

1. Easier to prepare than other techniques as complex calculations are not involved.

2. Data preparation is easier since less data cleaning is required as the variables are created beforehand.