Analysis of Stroke Prediction with Dataset using Machine Learning Classification Algorithm

Dr. NAVANEETHA KRISHNAN M, M.E. Ph.D., Head of the Department,

Department of Computer Science and Engineering

Mr. K. SETHURAMAN B.E, Student of Computer Science Engineering

Mr. M. YOKESH, B.E, Student of Computer science and Engineering

St. Joseph College of Engineering, Sriperumbudur, Chennai.

Abstract

Stroke is a blood clot or bleeds in the brain, which can make permanent damage that has an effect on mobility, cognition, sight or communication. Stroke is considered as medicalurgent situation and can cause long-term neurological damage, complications and often death. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humansin ability to learn things.

Machine learning is actively being used today, perhaps in many more places than one would expect. A variety of machine learning techniques are employedin the health care industry to aid in diagnosing and early detection of illnesses. Several elements that lead to stroke are considered in the current investigation. First, we're lookinginto the characteristics of those who are more likely to suffer from a stroke than others. The dataset is gathered and multiple classification algorithms are used to predict the occurrence of a stroke shortly.

The algorithms are namely K-Nearest Neighbors, ExtremeGradient Boosting and Light Gradient Boosting Machine. By applying these three algorithms, accuracy obtained were 70%, 80% and 83%. The best accuracy was shown inExtreme Gradient Boosting which is of 83 percent. Finally, various preventative steps suchas quitting smoking, avoiding alcohol, and other factors are recommended to reduce the risk of having a stroke.

Key Terms: BMI - Body Mass Index, KNN – K Nearest Neighbor, LightGBM – Light Gradient Boosting Machine, ML – Machine Learning, SKLearn – Sci-Kit Learn, SMOTE – Synthetic Minority Oversampling Technique.

Introduction

Machine learning is the modern science of finding patterns and makingpredictions from data based on work in multivariate statistics, data mining, pattern recognition, and advanced/predictive analytics. Stroke denies an individual's oxygen and supplements, which results in the death of dead cells when stroke occurs. It's not onlyvery expensive for the medical treatments and a permanent disability but can at lastprompt demise. By and large, Data Mining assumes an imperative part in the forecast ofillnesses in the medical care industry. A significant subject of AI in medication is utilizedin this project.

A machine learning model would take patient's information and proposea bunch of suit expectations. The framework can remove concealed information from achronicled clinical data set and can anticipate patients with infection and utilize theclinical profiles like Age, blood pressure, Glucose, and so forth it can foresee theprobability of patients getting an illness. Grouping calculations are utilized with numberof properties for expectation of illness.

The clinical record additionally comprises hisclinical history of illnesses and strokes he has had a stroke before too and we take all thatdata and train the machine dependent on various models, for example, Decision tree,SVM, Logistic regression, and so on. To address the issue of deals expectation of thingsdependent on client's future requests in various Big Marts across different areas diverse Machine Learning algorithms like XGBoost, LightGBM and K-Nearest Neighbours. Hence, they were the best suited model for stroke prediction and can feasibly be used byphysicians to predict stroke in real world.

Literature Survey

The paper "Stroke Risk Prediction Model based on Demographic Data." By Teerapat Kansadub,Sotarat Thammaboosade kiattisin in 2015 , provide the development of model for prediction based on the demographic data of the patients. This study aim to compare accuracy, false positive (FP), false negative (FN), and area under ROC Curve (AUC) resulted from three methods among Decision tree, Naïve Bayes, and Neural Network and then converted to rule. The best rule is selected for the benefits of population who have risk in stroke.

The paper "Prediction of Stroke using Data Mining Classification Techniques." By Ohoud Almadani,Riyad Alshammari in 2018, have considered Several assessments and prediction models, Decision Tree, Naive Bayes and Neural Network, showed acceptable accuracy in identifying stroke-prone patients. This project hence helps to predict the stroke risk using prediction model and provide personalized warning and the lifestyle correction message through

a web application. By doing so, it urges medical users to strengthen the motivation of health management and induce changes in their health behaviors.

The paper" Prediction of Stroke Using Machine Learning." By Kunder Akash Mahesh, Srikanth S, Shashank H N in 2020, helps to predict the stroke risk using data mining classification techniques.

The main objectives of this research are twofold: i) Use data mining techniques to predict patient at risk of developing stroke; and ii) Find the patient with who has higher chances to develop stroke.

System Design

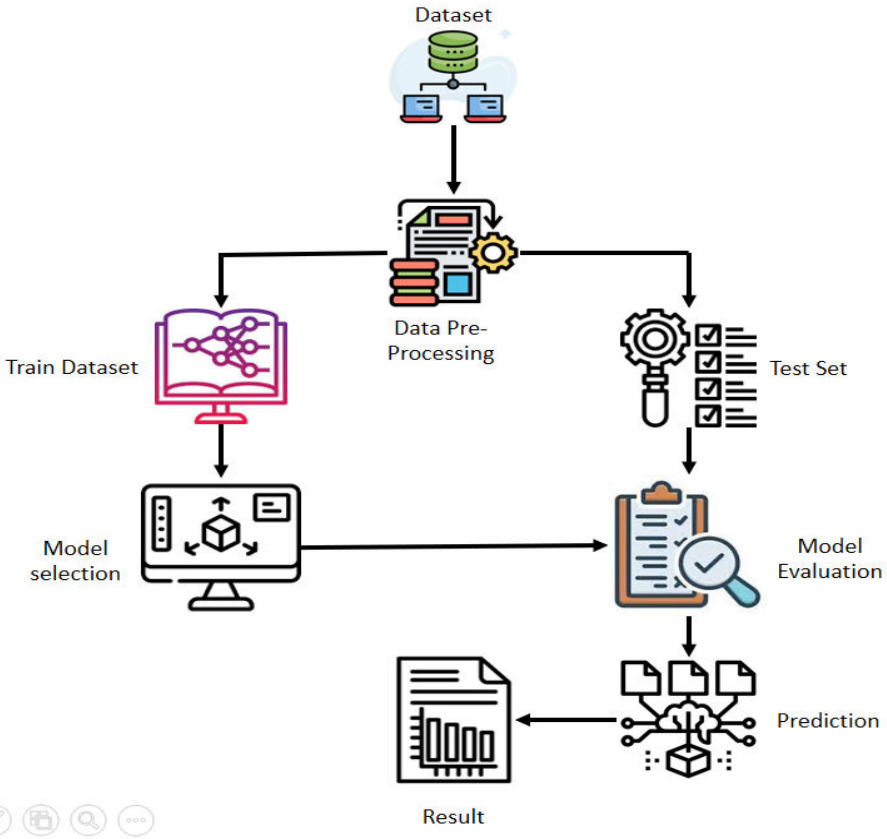
The dataset is the collection of data that is used to train and evaluate the machine learning model. It typically includes features such as age, gender, medical history, lifestyle factors, and other relevant information that may contribute to the risk of stroke.

Data preprocessing involves preparing the dataset for use in the machine learning model. This may include cleaning the data, removing outliers and errors, filling in missing values, and scaling or normalizing the data.

The training dataset is a subset of the overall dataset that is used to train the machine learning model. The model learns from the patterns in the training dataset to make accurate predictions.

The test dataset is a separate subset of the overall dataset that is used to evaluate the performance of the machine learning model. It is used to measure the accuracy of the model's predictions on

new data that it has not seen before.



Model selection involves choosing the appropriate machine learning algorithm for the task. Commonly used algorithms for stroke prediction include logistic regression, decision trees, random forests, and support vector machines.

Model evaluation involves testing the performance of the machine learning model using the test dataset. Common metrics used to evaluate the model include accuracy, precision, recall, and the area under the receiver operating characteristic curve.

Once the machine learning model has been trained and evaluated, it can be used to make predictions on new data. In the case of stroke prediction, the model can be used to predict the risk of stroke for a given patient based on their demographic and medical information.

The result of the machine learning model is the prediction of stroke risk for a given patient. This information can be used by clinicians to make more informed decisions about patient care, such as recommending lifestyle changes or prescribing medication to reduce the risk of stroke.

IMPLEMENTATION

Importing Library Files:

```
import pandas as pd
import numpy as n
import matplotlib.pyplot as p
```

Loading Dataset and Checking missing values:

```
df=pd.read_csv('/content/sample_data/healthcare-datasetstrokedata.csv')
df.head()
df.info() df.isnull().sum()
df['bmi'].value_counts()
```

Filling missing values:

```
df.describe()
df['bmi'].describe()
df['bmi'].fillna(df['bmi'].mean())
df['bmi'].fillna(df['bmi'].min(),inplace=True)
df.isnull().sum() df.drop('id',axis=1,inplace=True)
```

Outlier Removal Checking:

```
import seaborn as sns

p.rcParams['Figureure.Figuresize']=(40,10)

df.plot(kind='box')  p.show()
```

Normalization of Data:

```
from sklearn.preprocessing

import LabelEncoderenc=LabelEncoder()

enc.fit_transform(df['gender'])

gender=enc.fit_transform(df['gender'])

smoking_status=enc.fit_transform(df['smoking_status'])

work_type=enc.fit_transform(df['work_type'])
Residence_type=enc.fit_transform(df['Residence_type'])
ever_married=enc.fit_transform(df['ever_married']

 df['gender']=gender df['smoking_status']=smoking_status

df['work_type']=work_type df['Residence_type']=Residence_type
df['ever_married']=ever_married df.info()
```

Training and Testing the Data:

```
X=df.drop('stroke',axis=1).values

Y=df.stroke.values   from sklearn.model_selection

import train_test_split X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,rando
m_state=101)
```

SMOTE Technique

```
from imblearn.over_sampling
import SMOTE sm= SMOTE(random_state = 2)
X_train_res, Y_train_res = sm.fit_resample(X_train, Y_train.ravel())
```

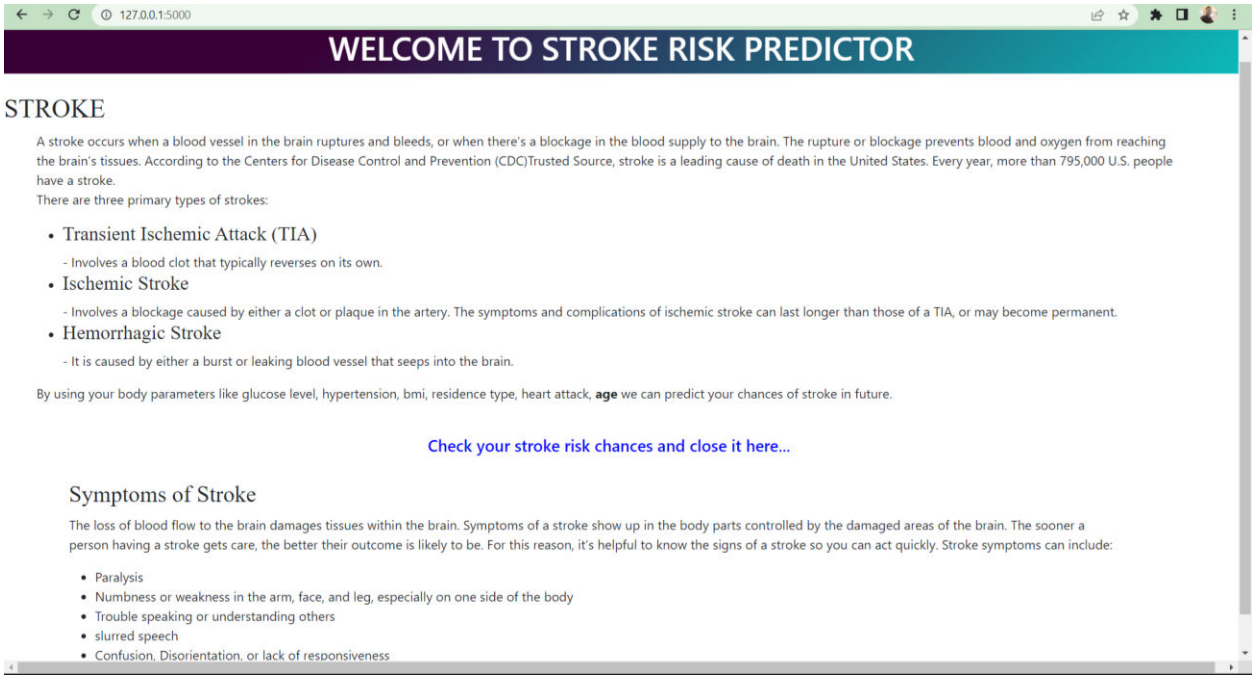
Algorithm selection

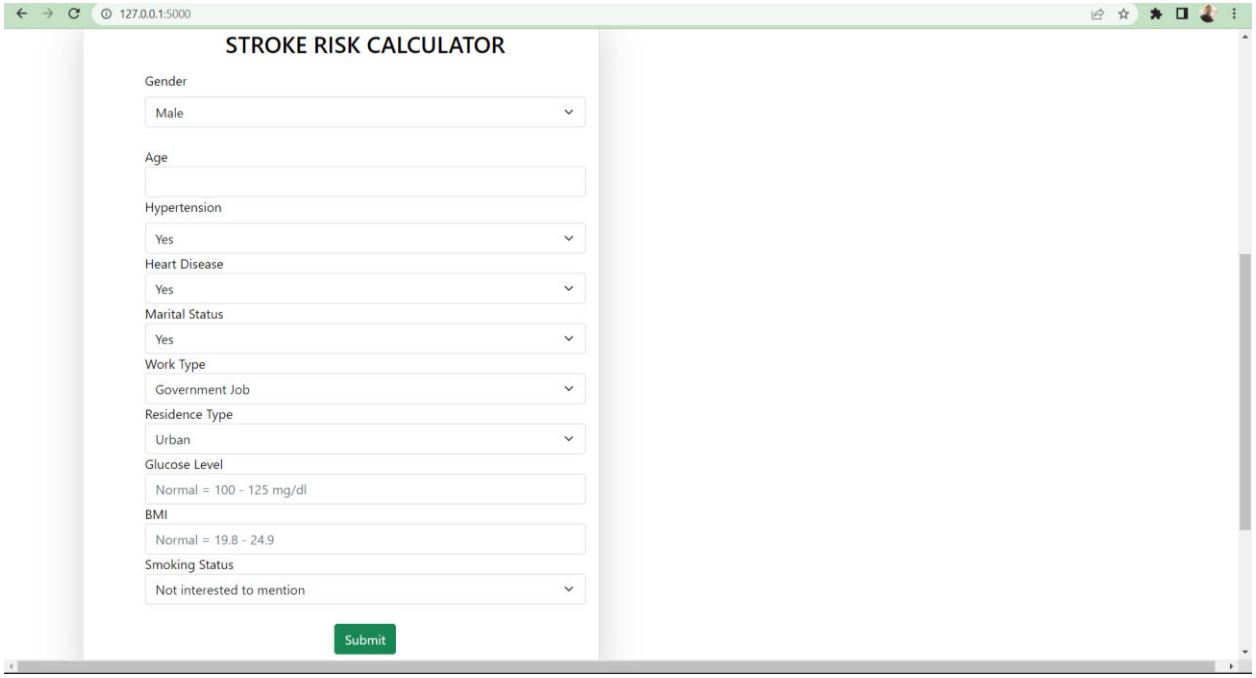
```
from sklearn.neighbors import KNeighborsClassifier
from numpy.ma.core import sqrt n = sqrt(n)
neigh = KNeighborsClassifier(n_neighbors=70) neigh.fit(train_x,train_y) y_pred =
neigh.predict(test_x) acc_knn = accuracy_score(y_pred,test_y)*100
import lightgbm as ltb model = ltb.LGBMClassifier() model.fit(train_x_val,train_y) predicted_y
=      model.predict(test_x_val)
acc_lgbm = accuracy_score(predicted_y, test_y)*100
from xgboost import XGBClassifier
xgc=XGBClassifier(objective='binary:logistic',n_estimators=100000,max_
depth=5,learning_rate=0.001,n_jobs=-1) xgc.fit(train_x_val,train_y) predicted_val =
xgc.predict(test_x_val) acc_xgc = accuracy_score(predicted_val, test_y)*100
import joblib
filename = 'xgbooster-new-version-model-joblib-file.sav'joblib.dump(xgc, open(filename,'wb'))
import matplotlib.pyplot as plt importseaborn as sns
y=[acc_knn,acc_lgbm,acc_xgc] x=['accuracy_knn','accuracy_lightbgm','accuracy_xgc']
plt.Figureure(Figuresize=(10, 8))
plots = sns.barplot(x, y)for bar in plots.patches: plots.annotate(format(bar.get
height(), '.2f'),
```

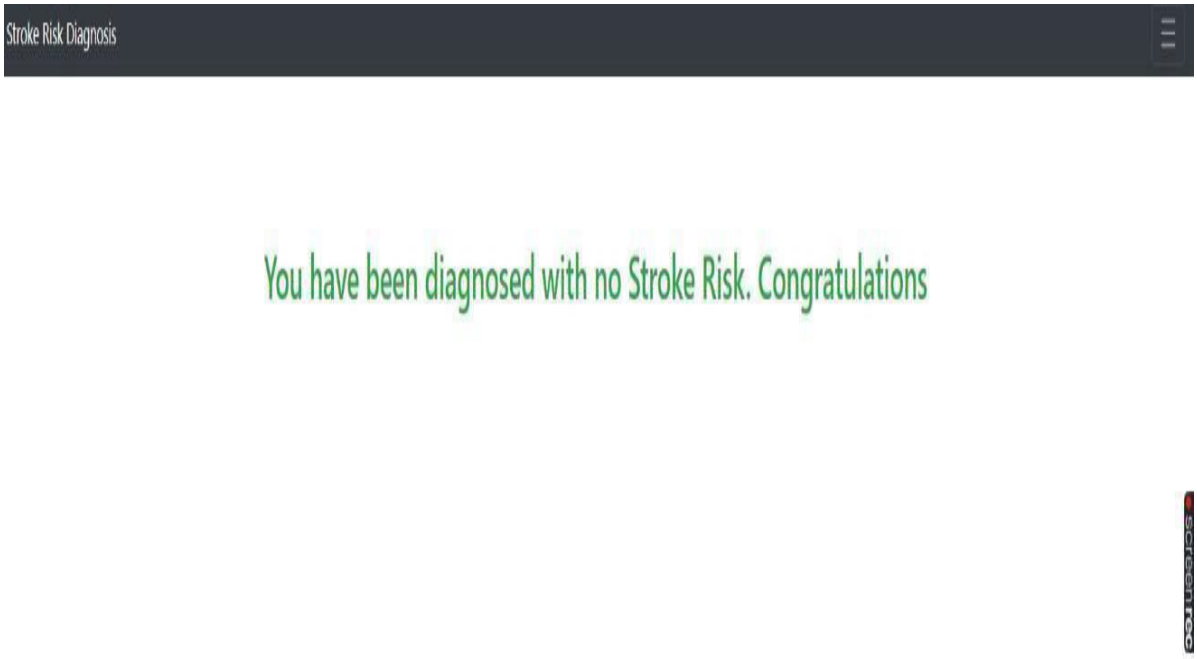
(bar.get_x() + bar.get_width() / 2, bar.get_height()), ha='center', va='center', size=15, xytext=(0, 8),textcoords='offset points')plt.xlabel("Algorithm", size=14) plt.ylabel("Accuracy", size=14) plt.title("Algorithm Accuracy")

plt.show() plt.saveFigure("Algorithm.jpeg")

SNAPSHOTS






Stroke Risk Diagnosis
You have been diagnosed with no Stroke Risk. Congratulations

You have been diagnosed with Stroke Risk
Based on your body condition the result is here. Please consult a Doctor.
List of Best hospitals for stroke.
1. Apollo Hospital, Greams Road, Chennai
2. Nanavati Hospital, Mumbai
3. Indraprastha Apollo Hospitals, Delhi
4. Manipal Hospital, HAL Road
5. Fortis Hospital, Bannerghatta Road, Bangalore

CONCLUSION

Upon the observation from the data processing, the glucose level acts as an major factor for the future stroke risk. Glucose level is directly proportional to the stroke risk. Therefore maintaining average range of glucose level is much more important. Nowadays, every human needs to know how to handle the work and life pressure, since work pressure also leads to increase the risk of stroke in a peak. To lead a peaceful life, every person should need to manage a good and even a better health condition in both physical and mental way.

FUTURE ENHANCEMENTS

This project helps to predict the stroke risk using prediction model in older people and for people who are addicted to the risk factors are mentioned in the project. In future the sameproject can be extended to give the update in stroke risk percentage using the output of current project. This project can also be used to find the stroke probabilities in young people and underage by collecting respective risk factor information's and doctor consulting.

REFERENCES:

[1] Michael Regnier, " Focus on stroke: Predicting and preventing stroke ".

[2] A.Sudha, P.Gayathiri, N.Jaishankar, " Effective analysis and predictive model ofstroke disease using classification ".

[3] Ohoud Almadani, Riyad Alshammari, " Prediction of Stroke using Data Mining Classification Techniques ". In: International Journal of Advanced Computer Science and Applications (IJACSA) (2018).

[4] Kansadub, T., Thammaboosadee, S., Kiattisin, S., Jalayondeja, " Stroke risk prediction model based on demographic data ". In: 8th Biomedical Engineering International Conference (BMEiCON) IEEE (2015), 2013.

[5] Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy, " Predictionof Brain Stroke Severity Using Machine Learning ". In: International Information and Engineering Technology Association (2020).

[6] Fahd Saleh Alotaibi: "Implementation of Machine Learning Model to Predict Heart Failure Disease ". In: International Journal of Advanced Computer Scienceand Applications (IJACSA) (2019).

[7] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, Honglak Lee:An Integrated Machine Learning Approach to Stroke Prediction. In:Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (2010).

[8] Dataset named ‘Stroke Prediction Dataset’ from Kaggle.

[9] Shanthi, D., Sahoo, G., Saravanan, N.: Designing an artificial neural network model for the prediction of thrombo-embolic stroke. Int. J. Biometric Bioinform.(IJBB) (2009).

[10] Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach,A. K.: DRFS: Detecting Risk Factor of Stroke Disease from Social Media UsingMachine Learning Techniques. In Springer (2020).

AUTHOR 1

Dr.M.Navaneethakrishnan M.E., PhD is a Head of the Department in the Department of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai, Tamil Nadu. He has completed his Ph.D, in Cyber Security - Computer Science and Engineering in 2017 from Manonmaniam Sundaranar University (MSU) Tirunelveli, Tamilnadu. He has done his M.E, CSE in Anna University Chennai in the year 2008. Dr.M.Navaneethakrishnan has 15 years of teaching experience and has 58 publications in International Journals and Conferences. His research interests include network security, Computer Networks, data science and Machine Learning. He is an active member of ISTE, CSI, IEANG and IEI

AUTHOR 2

Mr. K. Sethuraman B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops, Seminars in Python, Machine Learning. I got placed in Reputed Companies like Click Solutions, Q Spider and some respected companies.

AUTHOR 3

Mr. M. Yokesh B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops and Seminars in the area of Python and Machine Learning.