IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Stroke Sight: Early Detection throw Machine Learning and Interactive Intervention.

K. Roopa[1] ,Mr.N. Sravan Kumar[2]
[1]PG student, Vemu Institute of Technology, P. kothakota
[2]Assistant Professor, Vemu Institute of Technology, P. kothakota.

ABSTRACT:

This project delves into the critical domain of stroke prediction, leveraging machine learning to enable early intervention for this debilitating medical condition. By comparing various classifiers, the study identified that intricate models yield superior accuracy, with the leading model achieving nearly 91% accuracy, while others ranged between 83-91%. To elucidate model decision-making in the medical realm, the project integrated explainable techniques like SHAP and LIME. Furthermore, an extension to the study incorporated the CATBOOST classifier, harnessing a forest of weak classifiers to bolster prediction accuracy. This innovative framework, amalgamating global and local explainable methods, not only enhances stroke care and treatment but also offers invaluable insights into complex predictive models, paving the way for improved healthcare outcomes.

Keywords: Stroke Prediction, ML, CATBOOST

INTRODUCTION:

The escalating global incidence of stroke, a leading contributor to mortality and disability, underscores the imperative for early intervention. Traditional prediction methods, although valuable, often lack efficiency and accuracy. Enter machine learning algorithms, which have demonstrated remarkable potential in precise stroke risk prediction based on diverse clinical risk factors. These algorithms empower clinicians to proactively identify high-risk individuals, enabling timely interventions that could significantly mitigate stroke-related complications and enhance patient outcomes. Additionally, as transparency becomes paramount in healthcare AI, interpretable machine learning models emerge as invaluable tools. They furnish clinicians with nuanced insights into the determinants of stroke risk, facilitating informed treatment decisions. Despite common misconceptions associating stroke predominantly with the elderly or those with pre-existing conditions, the reality is that stroke can affect anyone, underscoring the universal relevance and urgency of predictive models in stroke prevention

LITERATURE SURVEY

T. Elloker and A. J. Rhoda*et al*

As the study delves into the critical yet understudied relationship between social support and post-stroke participation. Analyzing 54 articles from various databases, the research highlights a significant correlation between the quality and quantity of social support and improved participation in activities, including social engagements, leisure pursuits, and returning to work. High levels of social support emerge as a pivotal factor in enhancing post-stroke participation. These findings underscore the importance of integrating social support interventions into holistic stroke management strategies. Health professionals are encouraged to prioritize and incorporate social support mechanisms to optimize post-stroke recovery and participation outcomes.

A. Alloubani, A. Saleh, and I. Abdelhafiz*et al*

Research focuses on understanding and mitigating the risk factors associated with stroke, a growing concern in healthcare due to rising cases of hypertension, diabetes mellitus, and obesity. By analyzing published clinical trials sourced from databases like EMBASE and MEDLINE, the study aims to identify prevalent risk factors and explore preventive measures. The findings underscore the importance of early identification of these risk factors in stroke patients and emphasize the need for patient education to effectively manage and mitigate these risks, ultimately reducing the incidence of stroke.

A. K. Boehme, C. Esenwa, and M. S. V. Elkind*et al*

Since the author delves into the multifaceted nature of stroke, emphasizing the importance of understanding its diverse risk factors for effective treatment and prevention strategies. They distinguish between nonmodifiable factors like age, sex, and ethnicity, and modifiable ones such as hypertension, smoking, and diet. Additionally, the paper highlights emerging risk factors like inflammatory disorders, pollution, and genetic polymorphisms that influence stroke risk. The research suggests that both common and rare genetic factors, when combined with environmental interactions, may be more modifiable than previously thought. The focus is on lifestyle modifications, early identification, and treatment of medical conditions to not only reduce stroke risk but also mitigate other cardiovascular diseases

PROBLEM STATEMENT:

Stroke often causes due to blood flow stop to brain and this is one of the deadly diseases.

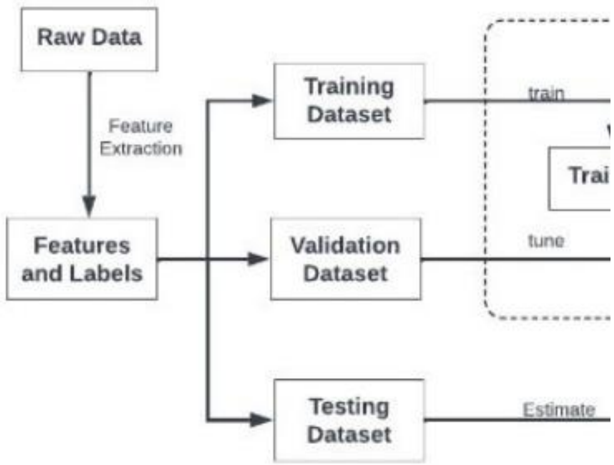
Patient life can be saved and stroke can be avoided by timely and accurate detection. Existing detection technique requires heavy resources and they make time for prediction. To overcome from this problem many machine learning algorithms were introduced as they are very accurate in medical diseases prediction but existing techniques were suffering from data leakage such as improper handling or missing values, improper categorical data calculation etc. No existing techniques were employing any Explainable model (XAI) which can show which features are helping most in detecting stroke so doctor can give priority on such features for faster recovery. These explainable features can be Smoking, Age, BMI and may be other features.

PROPOSED METHOD:

So author of this paper employing different processing techniques such as Removing missing values, Imbalance data handling using SMOTE and relevant features selection using CHI2 algorithm. All this processed features will get trained on 6 different algorithms such as Random Forest, KNN, SVM, Logistic Regression, XGBOOST and Naive Bayes. In all algorithm Random Forest is giving high accuracy and each algorithm performance is evaluated in terms of accuracy, precision, recall and FSCORE.

For easy understanding of features author employing various graph on Strokes patient data. Best algorithm will be input to SHAPELY Explainable (XAI) algorithm to explain about features which are contributing most in predicting correct label.

ARCHITECTURE



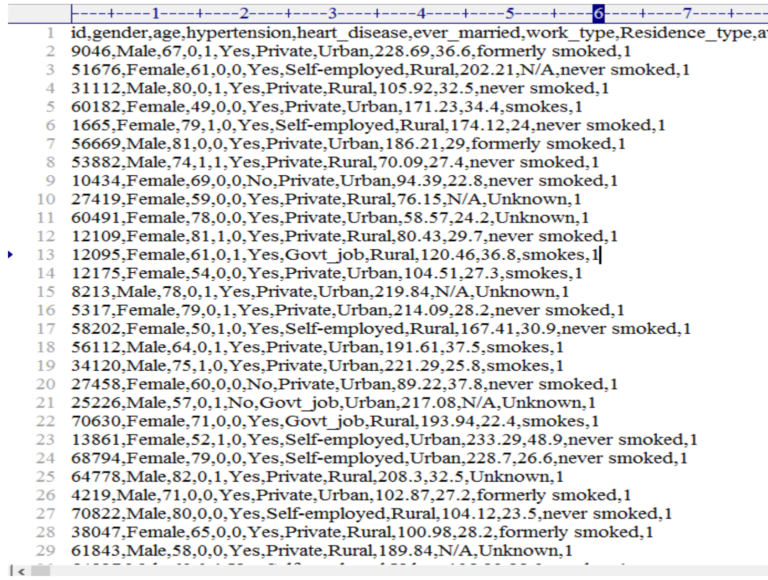
STROKE PREDICTION DATASET:

METHODOLOGY:

Importing Python Classes and Packages:

To kickstart the analysis, we import essential Python classes and packages tailored for data analysis and machine learning tasks.

Reading and Displaying Dataset:



```
id,gender,age,hypertension,heart_disease,ever_married,work_type,Residence_type,av
9046,Male,67,0,1,Yes,Private,Urban,228.69,36.6,formerly smoked,1
51676,Female,61,0,0,Yes,Self-employed,Rural,202.21,N/A,never smoked,1
31112,Male,80,0,1,Yes,Private,Rural,105.92,32.5,never smoked,1
60182,Female,49,0,0,Yes,Private,Urban,171.23,34.4,smokes,1
1665,Female,79,1,0,Yes,Self-employed,Rural,174.12,24,never smoked,1
56669,Male,81,0,0,Yes,Private,Urban,186.21,29,formerly smoked,1
53882,Male,74,1,1,Yes,Private,Rural,70.09,27.4,never smoked,1
10434,Female,69,0,0,No,Private,Urban,94.39,22.8,never smoked,1
27419,Female,59,0,0,Yes,Private,Rural,76.15,N/A,Unknown,1
60491,Female,78,0,0,Yes,Private,Urban,58.57,24.2,Unknown,1
12109,Female,81,1,0,Yes,Private,Rural,80.43,29.7,never smoked,1
12095,Female,61,0,1,Yes,Govt_job,Rural,120.46,36.8,smokes,1
12175,Female,54,0,0,Yes,Private,Urban,104.51,27.3,smokes,1
8213,Male,78,0,1,Yes,Private,Urban,219.84,N/A,Unknown,1
5317,Female,79,0,1,Yes,Private,Urban,214.09,28.2,never smoked,1
58202,Female,50,1,0,Yes,Self-employed,Rural,167.41,30.9,never smoked,1
56112,Male,64,0,1,Yes,Private,Urban,191.61,37.5,smokes,1
34120,Male,75,1,0,Yes,Private,Urban,221.29,25.8,smokes,1
27458,Female,60,0,0,No,Private,Urban,89.22,37.8,never smoked,1
25226,Male,57,0,1,No,Govt_job,Urban,217.08,N/A,Unknown,1
70630,Female,71,0,0,Yes,Govt_job,Rural,193.94,22.4,smokes,1
13861,Female,52,1,0,Yes,Self-employed,Urban,233.29,48.9,never smoked,1
68794,Female,79,0,0,Yes,Self-employed,Urban,228.7,26.6,never smoked,1
64778,Male,82,0,1,Yes,Private,Rural,208.3,32.5,Unknown,1
4219,Male,71,0,0,Yes,Private,Urban,102.87,27.2,formerly smoked,1
70822,Male,80,0,0,Yes,Self-employed,Rural,104.12,23.5,never smoked,1
38047,Female,65,0,0,Yes,Private,Rural,100.98,28.2,formerly smoked,1
61843,Male,58,0,0,Yes,Private,Rural,189.84,N/A,Unknown,1
```

We are using STROKE dataset from KAGGLE and first row represents dataset column names and remaining rows represents dataset values and by using above dataset we will test all algorithm performance.

The dataset is loaded and inspected for its structure. Addressing any missing values is crucial at this stage. Non-numeric data is converted to a numeric format using label encoding.

Exploratory Data Analysis (EDA):

EDA is performed to understand the data distribution, especially focusing on the 'Normal' and 'Stroke' labels. Class imbalance issues are identified and visualized using bar and pie charts.

Cluster Features Correlation Graph:

We visualize feature correlations to understand relationships among different features. Highly correlated features, those with scores exceeding 90%, are pinpointed for further investigation.

Gender and Age Relationship:

A graphical representation showcases the relationship between gender and stroke occurrences across varying age groups.

Age-Based Stroke Counts:

A bar graph presents stroke counts across different age groups, further differentiating by gender using stacked bars.

Gender and BMI on Stroke Patients:

We explore the interplay between gender, BMI, and stroke occurrences through graphical visualization.

Hypertension and Heart Disease Counts:

Separate graphs highlight the prevalence of hypertension and heart disease among stroke patients.

Average Glucose Level by Gender:

An illustrative graph showcases average glucose levels categorized by gender among stroke patients.

Smoking Status and Residence Type Visualization:

The dataset is further explored to visualize stroke occurrences based on smoking status and type of residence.

Converting Categorical Data and Normalizing:

Remaining categorical data is transformed into numeric form to prevent data leakage. Feature normalization is applied to maintain uniformity across features.

Handling Class Imbalance using SMOTE:

The Synthetic Minority Over-sampling Technique (SMOTE) is employed to tackle the class imbalance problem, ensuring balanced representation of both 'Normal' and 'Stroke' classes.

Feature Selection using CHI2:

The CHI2 test is executed for feature selection, impacting the feature count. The dataset is subsequently split into training and testing subsets.

Model Training and Evaluation:

Several machine learning algorithms, including Random Forest, Logistic Regression, SVM, KNN, Naive Bayes, XGBOOST, and CATBOOST, are trained on the data. Performance metrics like accuracy, precision, recall, and confusion matrix are utilized for evaluation.

Choosing the Best Model:

Based on the performance metrics, the most effective algorithm is identified and its performance visualized through bar graphs.

Shapely Explanation of Features:

SHAP is employed to elucidate feature importance in making correct predictions, offering visual insights into crucial features.

Comparison of Algorithm Performance:

A tabular presentation offers a comparative analysis of the algorithms' performance.

Testing with CATBOOST Algorithm:

Test data is processed similarly to the training data, followed by prediction using the CATBOOST algorithm. The results are juxtaposed against actual test data for validation.

Extension:

As an enhancement, we incorporate the CATBOOST classifier, leveraging a forest of weak classifiers. Each classifier undergoes training, and a voting mechanism selects the best-performing classifier for final predictions, boosting prediction accuracy.

EVALUATION:

Precision:

Formula: $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$

Code: precision = precision_score(testY, predict, average='macro') * 100

Recall (Sensitivity):

Formula: $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$

Code: recall = recall_score(testY, predict, average='macro') * 100

F1 Score:

Formula: $F1 = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

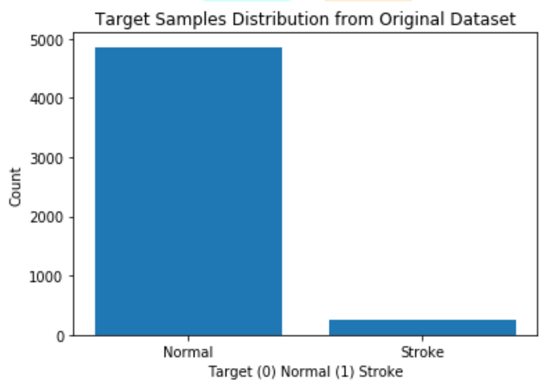
Code: f1 = f1_score(testY, predict, average='macro') * 100

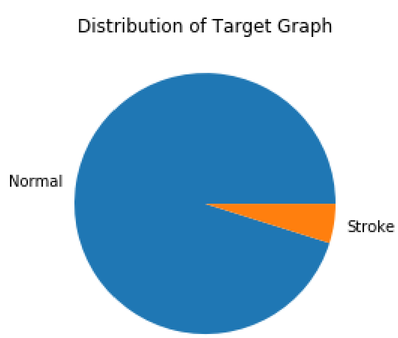
Accuracy:

Formula: $\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$

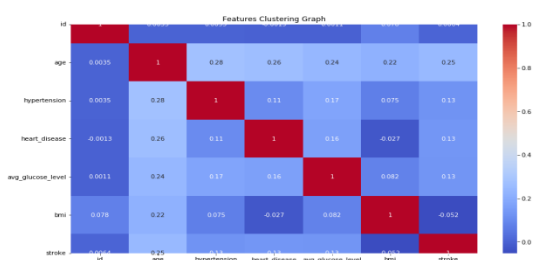
Code: accuracy = accuracy_score(testY, predict) * 100

RESULTS:

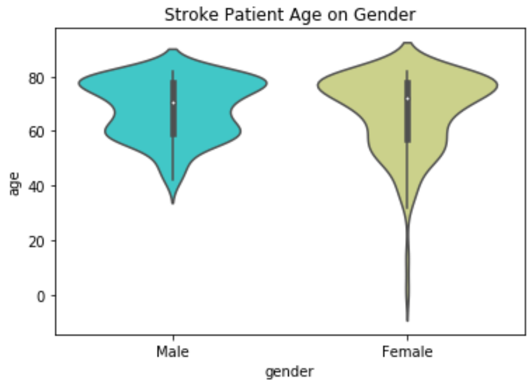




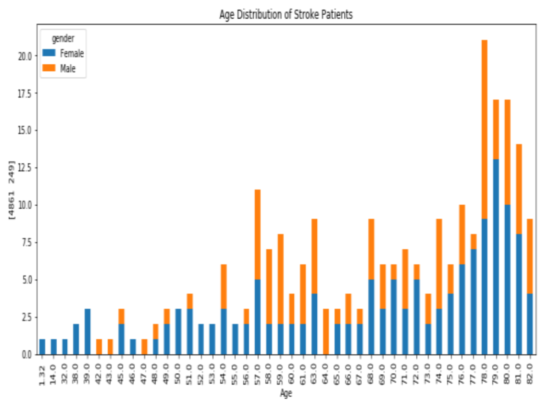
In above finding and plotting graph of Normal and Stroke label where x-axis represents Normal and Stroke and y-axis represents count and we can see one class contains so many records and other class contains few records only so data is highly imbalance which we can handle using SMOTE and same can see PIE graph



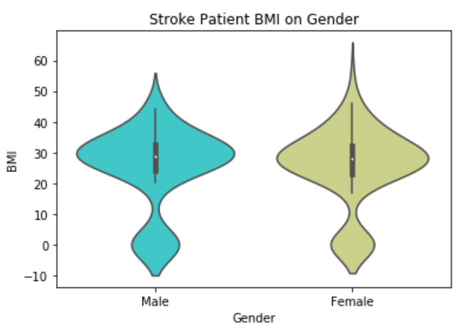
In above graph finding and displaying cluster features correlation graph and all values are not highly correlates. High correlated means features will have score more than 90%



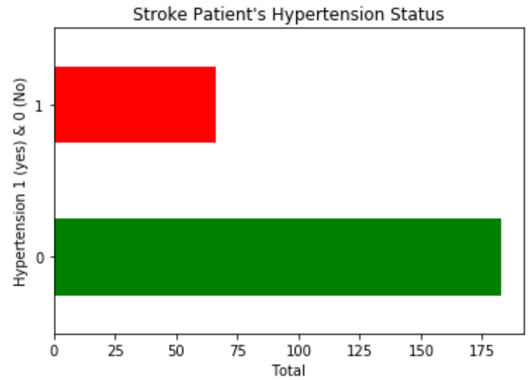
In above graph we displaying gender with strokes on different age where x-axis represents Gender and y-axis represent age



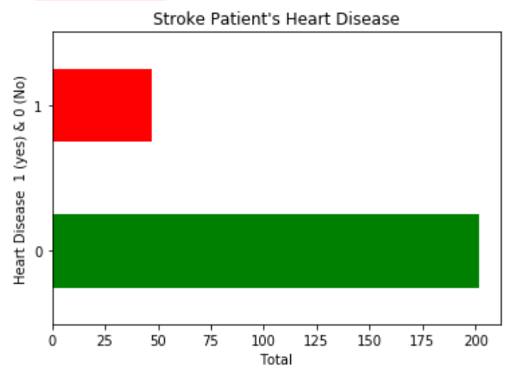
In above bar graph x-axis represents Age and y-axis represents stroke count where blue stack part is for Female and orange for Male



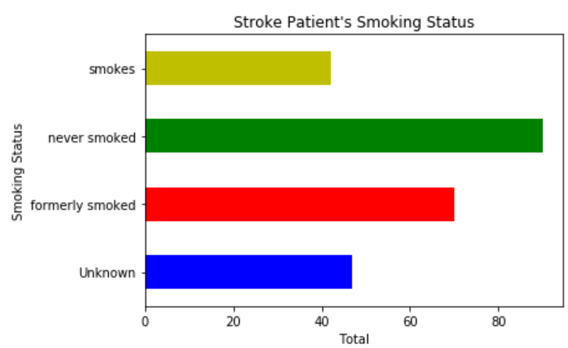
Above graph displaying gender and BMI on stroke patients



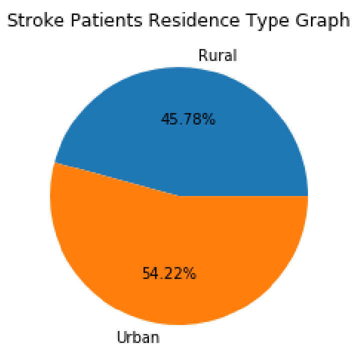
In above graph displaying number of stroke patients suffering from hypertension where in Y-axis 0 means No hyper tension and 1 means hyper tension and x-axis represents count



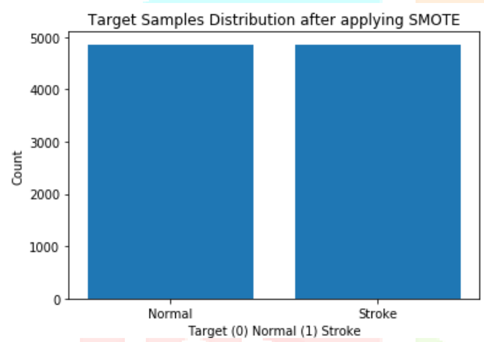
In above graph displaying number of stroke patients suffering from heart disease where in Y-axis 0 means No Heart Disease and 1 means Heart Disease and x-axis represents count



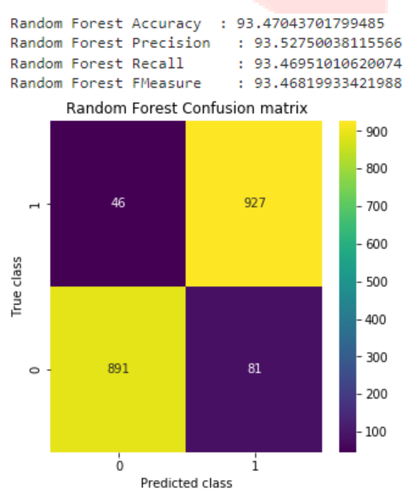
In above graph displaying number of stroke patients with smoke status



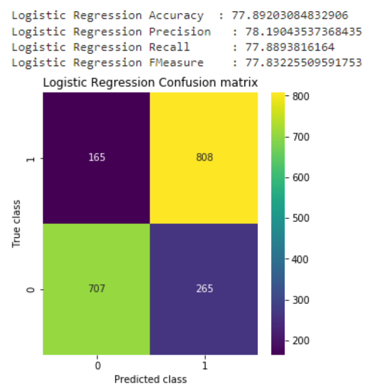
In above graph displaying residence type of stroke patients



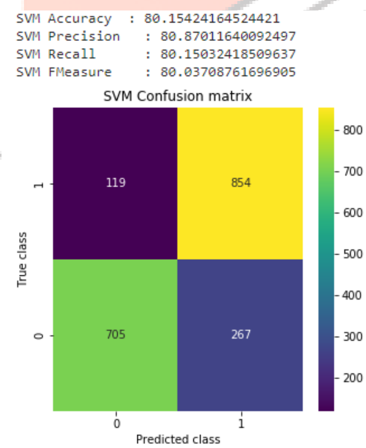
By applying SMOTE we can see both classes has equal number of records



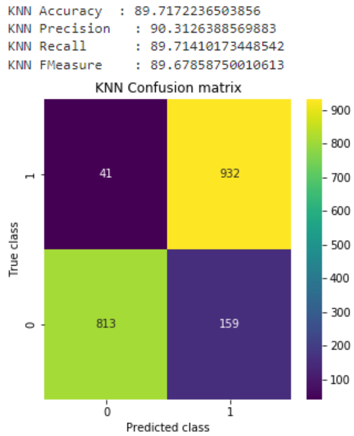
In above screen training Random Forest algorithm on training data and then performing prediction on test and then random forest got 94% accuracy and can see other metrics like precision, recall etc. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count which are very few



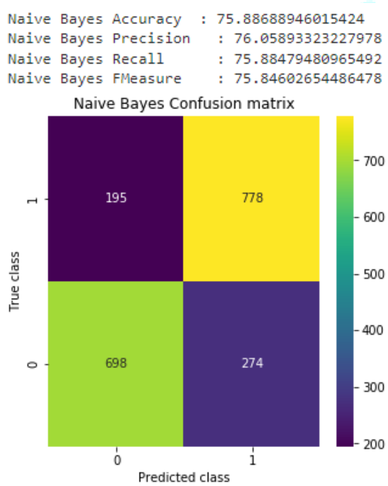
In above screen Logistic Regression got 78% accuracy



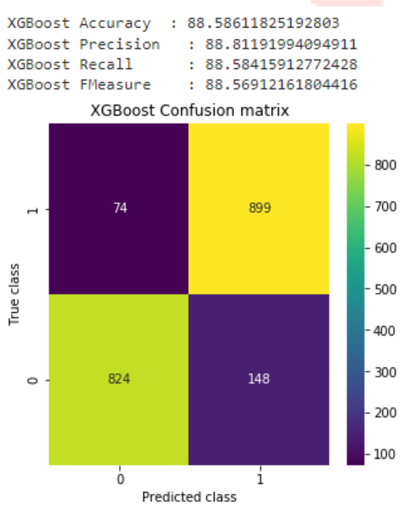
In above screen SVM got 80% accuracy



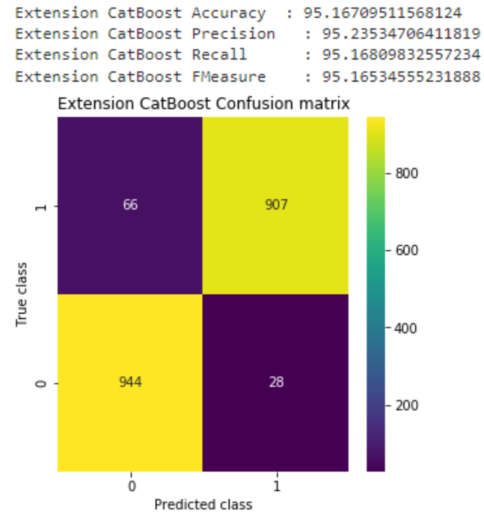
In above screen KNN got 92% accuracy



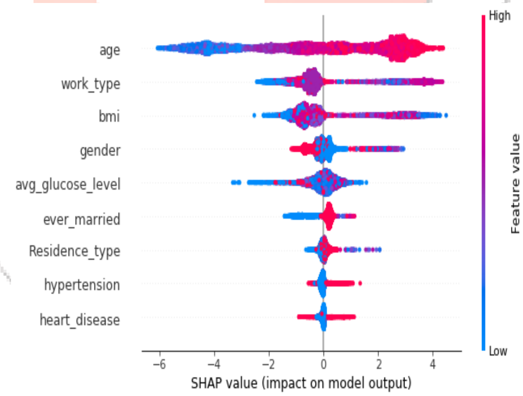
In above screen Naive Bayes got 77% accuracy



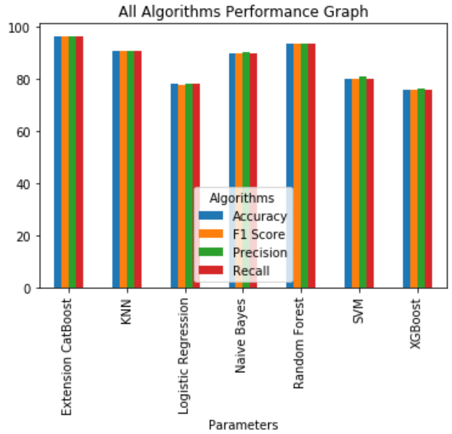
In above screen XGBOOST got 89% accuracy



In above screen extension CATBOOST got 95% accuracy which is higher than other algorithms



In above screen shapely explaining about features which are contributing most in correct prediction and then features whose graph reaching to high are the most relevant features used for prediction



In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms extension CATBOOST got high accuracy

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	Random Forest	93.527500	93.469510	93.468199	93.470437
1	Logistic Regression	78.190435	77.889382	77.832255	77.892031
2	SVM	80.870116	80.150324	80.037088	80.154242
3	KNN	90.642671	90.642671	90.642671	90.642674
4	Naive Bayes	90.312639	89.714102	89.678588	89.717224
5	XGBOost	76.058933	75.884795	75.846027	75.886889
6	Extension CatBoost	96.379698	96.350274	96.349059	96.349614

Displaying all algorithms performance

Prediction:

```
Test Data = [17739 'Male' 57 0 0 'Yes' 'Private' 'Rural' 84.96 36.7 'Unknown'] Predicted As ====> Normal

Test Data = [12095 'Female' 61 0 1 'Yes' 'Govt_job' 'Rural' 120.46 36.8 'smokes'] Predicted As ====> Stroke

Test Data = [12175 'Female' 54 0 0 'Yes' 'Private' 'Urban' 104.51 27.3 'smokes'] Predicted As ====> Stroke

Test Data = [8213 'Male' 78 0 1 'Yes' 'Private' 'Urban' 219.84 0.0 'Unknown'] Predicted As ====> Stroke

Test Data = [27419 'Female' 59 0 0 'Yes' 'Private' 'Rural' 76.15 0.0 'Unknown'] Predicted As ====> Normal

Test Data = [60491 'Female' 78 0 0 'Yes' 'Private' 'Urban' 58.57 24.2 'Unknown'] Predicted As ====> Normal

Test Data = [12109 'Female' 81 1 0 'Yes' 'Private' 'Rural' 80.43 29.7 'never smoked'] Predicted As ====> Stroke

Test Data = [5317 'Female' 79 0 1 'Yes' 'Private' 'Urban' 214.09 28.2 'never smoked'] Predicted As ====> Stroke

Test Data = [58202 'Female' 50 1 0 'Yes' 'Self-employed' 'Rural' 167.41 30.9
 'never smoked'] Predicted As ====> Stroke
```

In above screen reading test data, normalizing, features encoding from categorical to numeric format, removing missing values, features selection and normalization and then processed features are predicting with extension CATBOOST algorithm and in output before =� arrow symbol we can see TEST data and after arrow symbol we can see predicted data as 'Normal or Stroke'

CONCLUSION

Our project demonstrates an automated stroke prediction system leveraging machine learning techniques. By addressing issues like data imbalance and feature selection, we achieved accurate predictions using algorithms such as Random Forest, KNN, SVM, Logistic Regression, XGBOOST, and Naive Bayes. Introducing the CATBOOST classifier further enhanced prediction accuracy to 95%. Utilizing SHAPLEY explainable AI provided insights into the most influential features for prediction, aiding clinicians in prioritizing interventions. Our web application facilitates early stroke detection, potentially saving lives through timely intervention. This project underscores the significance of machine learning in healthcare, offering a scalable and efficient solution for stroke prediction.

REFERENCES:

[1] Learn About Stroke. Accessed: May 25, 2022. [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/why-strokematters/learn-about-stroke

[2] T. Elloker and A. J. Rhoda, ''The relationship between social support and participation in stroke: A systematic review,'' Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[3] M. Katan and A. Luft, ''Global burden of stroke,'' Seminar Neurol., vol. 38, no. 2, pp. 208–211, Apr. 2018.

[4] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart, A. Simats, E. Pecharroman, O. Ventura, M. Ribó, D. Vivien, J. C. Sanchez, and J. Montaner, ''Blood biomarkers to differentiate ischemic and hemorrhagic strokes,'' Neurology, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021.

[5] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li, ''Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,'' J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[6] A. Alloubani, A. Saleh, and I. Abdelhafiz, ''Hypertension and diabetes mellitus as a predictive risk factors for stroke,'' Diabetes

Metabolic Syndrome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

[7] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, ''Stroke risk factors, genetics, and prevention,'' Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[8] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, ''Stroke symptoms and the decision to call for an ambulance,'' Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.

[9] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, ''Response to symptoms of stroke in the UK: A systematic review,'' BMC Health Services Res., vol. 10, no. 1, pp. 1–9, Dec. 2010.

[10] L. Gibson and W. Whiteley, ''The differential diagnosis of suspected stroke: A systematic review,'' J. Roy. College Physicians Edinburgh, vol. 43, no. 2, pp. 114–118, Jun. 2013.

[11] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, ''Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review,'' J. NeuroInterventional Surgery, vol. 12, no. 2, pp. 156–164, Feb. 2020.

[12] Y. Zhao, S. Fu, S. J. Bielinski, P. A. Decker, A. M. Chamberlain, V. L. Roger, H. Liu, and N. B. Larson, ''Natural language processing and machine learning for identifying incident stroke from electronic health records: Algorithm development and validation,'' J. Med. Internet Res., vol. 23, no. 3, Mar. 2021, Art. no. e22951.

[13] B. McDermott, A. Elahi, A. Santorelli, M. O'Halloran, J. Avery, and E. Porter, ''Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis,'' Physiological Meas., vol. 41, no. 7, Aug. 2020, Art. no. 075010.

[14] A. Bivard, L. Churilov, and M. Parsons, ''Artificial intelligence for decision support in acute stroke—Current roles and potential,'' Nature Rev. Neurol., vol. 16, no. 10, pp. 575–585, Oct. 2020.

[15] W. Wang, M. Kiik, N. Peek, V. Curcin, I. J. Marshall, A. G. Rudd, Y. Wang, A. Douiri, C. D. Wolfe, and B. Bray, ''A systematic review of machine learning models for predicting outcomes of stroke with structured data,'' PLoS ONE, vol. 15, no. 6, Jun. 2020, Art. no. e0234722.

[16] M. S. Sirsat, E. Fermé, and J. Câmara, ''Machine learning for brain stroke: A review,'' J. Stroke Cerebrovascular Diseases, vol. 29, no. 10, Oct. 2020, Art. no. 105162.

[17] A. K. Arslan, C. Colak, and M. E. Sarihan, ''Different medical data mining approaches based prediction of ischemic stroke,'' Comput. Methods Programs Biomed., vol. 130, pp. 87–92, Jul. 2016.

[18] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, ''Explainable

artificial intelligence model for stroke prediction using EEG signal,'' Sensors, vol. 22, no. 24, p. 9859, Dec. 2022.

[19] E. Dritsas and M. Trigka, ''Stroke risk prediction with machine learning techniques,'' Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.

[20] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, ''An explainable machine learning pipeline for stroke prediction on imbalanced data,'' Diagnostics, vol. 12, no. 10, p. 2392, Oct. 2022.