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. Rhodaet 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. Abdelhafizet 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. Elkindet 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:

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.

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:

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:

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

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

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 training Random Forest algorithm on training data and then
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.

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Precision</th>
<th>Recall</th>
<th>FScore</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Random Forest</td>
<td>93.527500</td>
<td>93.469510</td>
<td>93.468199</td>
</tr>
<tr>
<td>1</td>
<td>Logistic Regression</td>
<td>78.190435</td>
<td>77.889382</td>
<td>77.832255</td>
</tr>
<tr>
<td>2</td>
<td>SVM</td>
<td>80.870116</td>
<td>80.150324</td>
<td>80.037088</td>
</tr>
<tr>
<td>3</td>
<td>KNN</td>
<td>90.642671</td>
<td>90.642671</td>
<td>90.642671</td>
</tr>
<tr>
<td>4</td>
<td>Naive Bayes</td>
<td>90.312639</td>
<td>89.714102</td>
<td>89.678588</td>
</tr>
<tr>
<td>5</td>
<td>XGBoost</td>
<td>76.058933</td>
<td>75.884795</td>
<td>75.846027</td>
</tr>
<tr>
<td>6</td>
<td>Extension CatBoost</td>
<td>96.379698</td>
<td>96.350274</td>
<td>96.349059</td>
</tr>
</tbody>
</table>

Displaying all algorithms performance

Prediction:

Test Data = [17739 'Male' 57 0 1 'Yes' 'Private' 'Rural' 64.96 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 = [12176 'Female' 54 0 1 'Yes' 'Private' 'Urban' 104.51 27.3 'smokes'] Predicted As ----> Stroke
Test Data = [8213 'Male' 78 0 1 'Yes' 'Private' 'Urban' 219.04 0.0 'Unknown'] Predicted As ----> Stroke
Test Data = [27419 'Female' 59 0 1 'Yes' 'Private' 'Rural' 76.15 0.0 'Unknown'] Predicted As ----> Normal
Test Data = [60491 'Female' 78 0 1 '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 25.2 'never smoked'] Predicted As ----> Stroke
Test Data = [58202 'Female' 59 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:


