BRAC UNIVERSITY

Inspiring Excellence

Title of the project: Brain stroke analysis and classification

Group Number: 04

Group Members

Student Name	Student ID
Rejwan Shafi Wasef	23241108
Homaira Ferdousia	21301162
Istiak Islam	21301218
Tasmia Tabassum	21301495

Table of Contents

Abstract:

This paper suggests a thorough machine learning-based method for brain stroke prediction. Since brain strokes have a catastrophic effect on the health of the world, early detection and treatment are essential in increasing the lives of patients. This paper discusses the dataset, preprocessing techniques, model training, outcomes, and the perspective of machine learning in stroke diagnosis.

1. Introduction:

The goal of this project is to comprehend and categorize brain strokes using machine learning. We use advanced computer techniques to interpret the data we collect about variables associated with brain strokes, such as gender, age, hypertension, heart disease, marital status, type of residence, average glucose level, BMI, type of work, and smoking status. We train the computer to identify patterns that suggest an individual may be at risk of a brain stroke. We worked with various models in an attempt to identify the most accurate model for both classifying and determining brain damage stroke.

2. Motivation:

Commonly referred to as "silent killers," brain strokes mostly get their nickname from their intriguing nature. Even skilled medical experts may miss their early warning signs since they are sometimes so invisible, which results in delayed diagnoses. This lag frequently causes permanent brain damage, which significantly lowers the patients' standard of life or, in extreme circumstances, results in death. It's a dreadful picture that has not changed much in decades, which highlights the importance of early stroke detection methods that need to be developed.

Although they can be somewhat successful, traditional diagnostic techniques are frequently limited by the constraints of human monitoring and the inherent uncertainty of patient appearances. These techniques mainly depend on physical, observable symptoms, which may not appear until later in the progression of a stroke. Moreover, it's getting harder to guarantee prompt and precise diagnoses for every patient as the patient-to-doctor ratio rises in many parts of the world.

As we enter into the digital age and the resulting data surge. Large databases of information about health, ranging from accessible biometric data to electronic health records, define the modern medical environment. For the majority of sections, traditional diagnostics has yet to fully use this abundance of data.

This particular issue can be solved with the use of machine learning. These large datasets can be used to train advanced algorithms to find relationships and patterns that would be extremely difficult for a human to notice. This is not about improving human observation beyond what is possible for us to perceive on an inherent level, not just adding to it. By implementing machine learning models, it is possible to anticipate the development of a stroke by examining complex patterns and a wide range of factors, even before noticeable symptoms appear. A revolutionary change like this in diagnostic techniques could lead to prompt treatment options, a significant reduction in stroke-related complications, and the saving of many lives.

Furthermore, the financial strain on healthcare systems can be greatly decreased by incorporating machine learning into the diagnosis of strokes. Initial measures result in fewer hospital stays, fewer problems, and more successful treatments, which saves millions of dollars in medical expenses.

In short, there are humanitarian as well as scientific reasons for using machine learning to assume strokes. This project attempts to give people an opportunity to defend against one of the worst medical threats known to humanity, with an opportunity to transform stroke care.

3. Dataset Description:

Link: https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset/

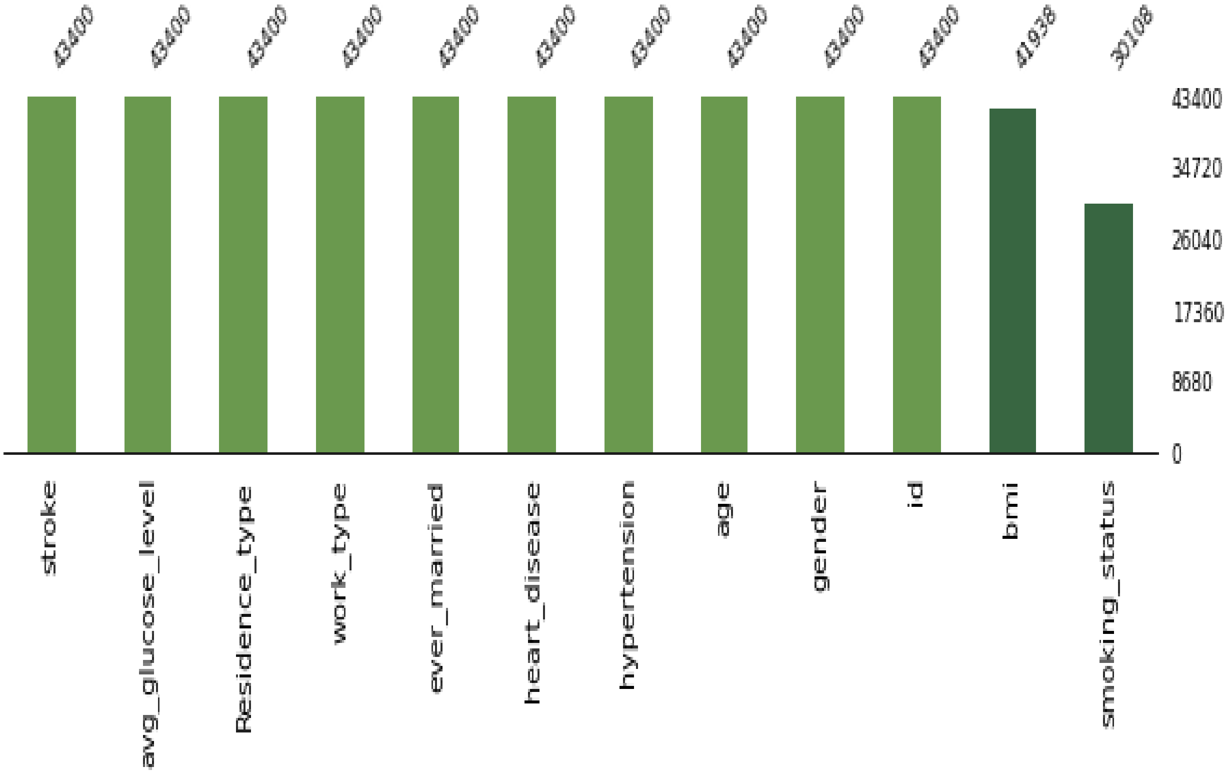
Number of Features: 12

Type of class/label: Categorical and Continuous

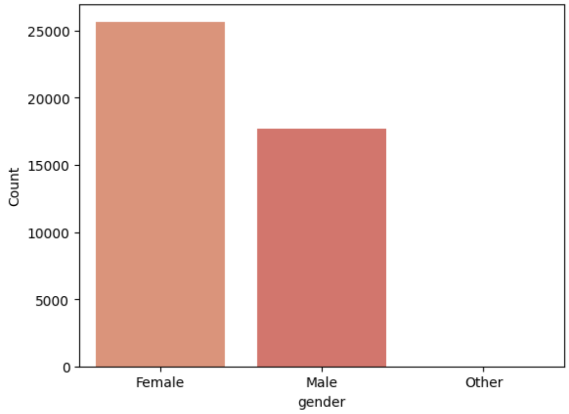
Number of data points: 43400

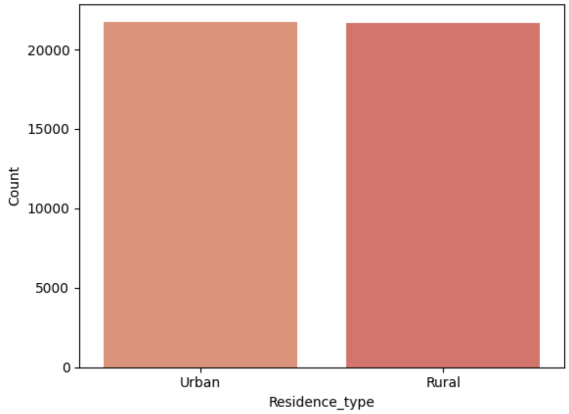
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43400 entries, 0 to 43399
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 43400 non-null  int64
 1   gender             43400 non-null  object
 2   age                43400 non-null  float64
 3   hypertension       43400 non-null  int64
 4   heart_disease      43400 non-null  int64
 5   ever_married       43400 non-null  object
 6   work_type          43400 non-null  object
 7   Residence_type     43400 non-null  object
 8   avg_glucose_level  43400 non-null  float64
 9   bmi                41938 non-null  float64
 10  smoking_status     30108 non-null  object
 11  stroke             43400 non-null  int64
dtypes: float64(3), int64(4), object(5)
memory usage: 4.0+ MB
```

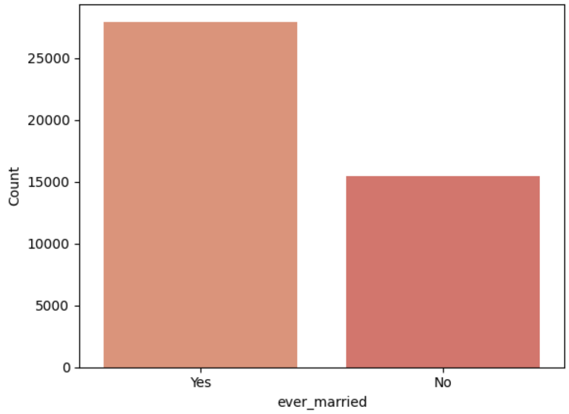
Biasness/Balanced

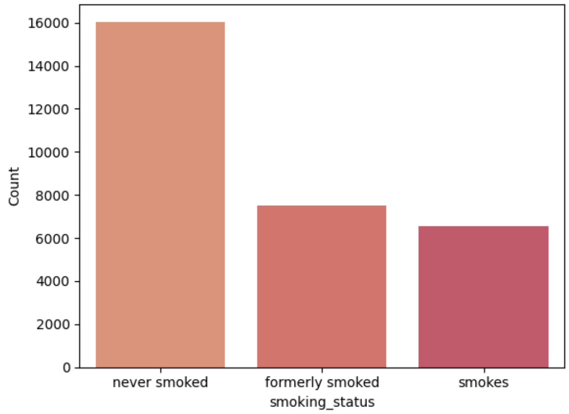

43400
43400
43400
43400
43400
43400
43400
43400
43400
43400
41938
30108
43400
34720
26040
17360
8680
0
stroke
avg_glucose_level
Residence_type
work_type
ever_married
heart_disease
hypertension
age
gender
id
bmi
smoking_status

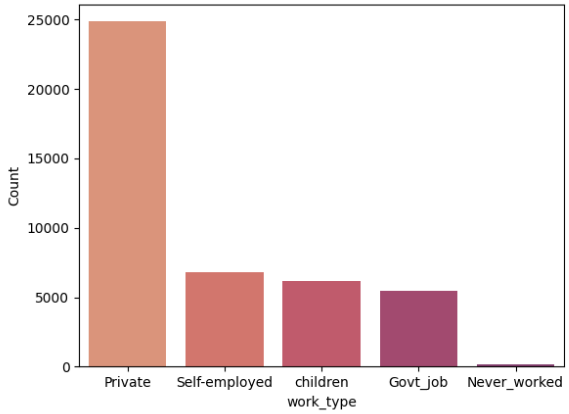
4. Dataset Pre-processing:







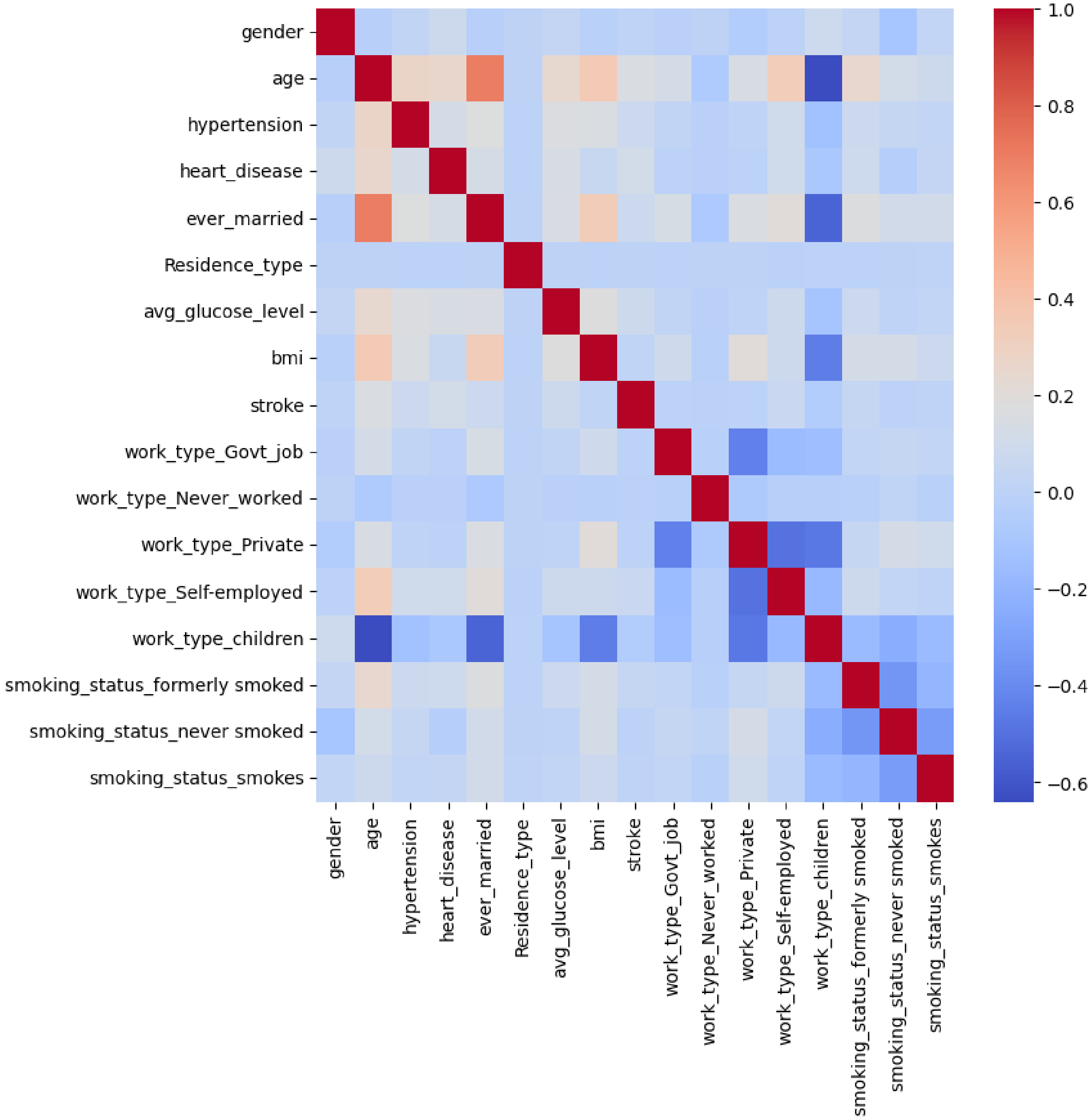




The concerns such as null values and outliers were discussed.

A total of 868 null values for BMI and 13292 null values for smoking status were present in 5 features made use of mean imputation.

‘ID’ column was unnecessary. It was dropped. Mapped ‘ever_married’ Yes into ‘1’ and No into ‘0’.Mapped ‘Residence_type’ ‘Urban’ into 1, ’Rural’ into 0. Using ‘one-hot encoding’ to convert categorical (non-numeric) variables into a numerical format.



5. Dataset Splitting:

A classic 70-30 split was maintained - 30380 for training and 13020 for testing.

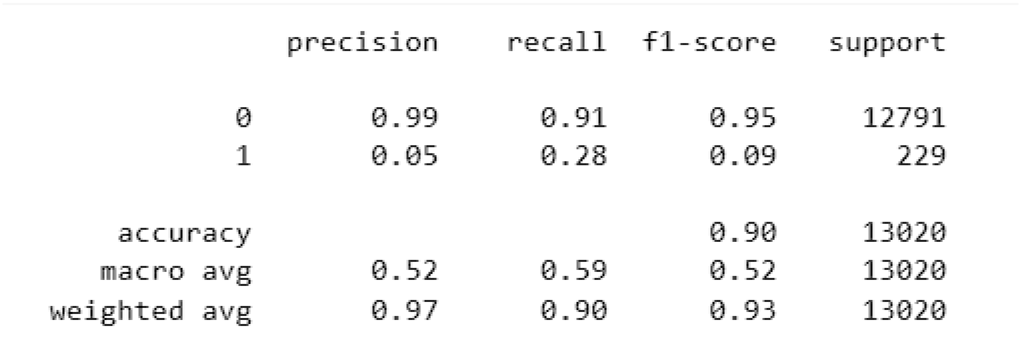
6. Model Training:

Testing revealed that the Random Forest model performed significantly better than the others at estimating brain stroke based on the presented signals.

Model Name	Accuracy(%)	Error(%)
Logistic Regression	89.68%	10.32%
Decision Tree	94.98%	5.02%
KnnClassifier	84.71%	15.29%
Random forest	96.79%	3.21%

As we can see from the table, with 96.79% accuracy and only 3.21% inaccuracy, the Random Forest model performed the best. Conversely, the Knn classifier with 10% error and 90% accuracy by logistic regression demonstrated the lowest performance. Following that, the decision tree's 95% accuracy was largely satisfactory.

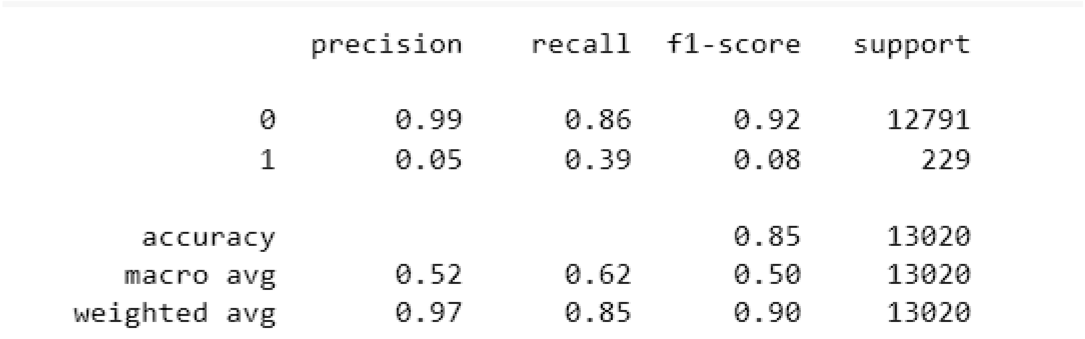
Logistic Regression



```
              precision    recall  f1-score   support

           0       0.99      0.91      0.95     12791
           1       0.05      0.28      0.09       229

    accuracy                           0.90     13020
   macro avg       0.52      0.59      0.52     13020
weighted avg       0.97      0.90      0.93     13020
```

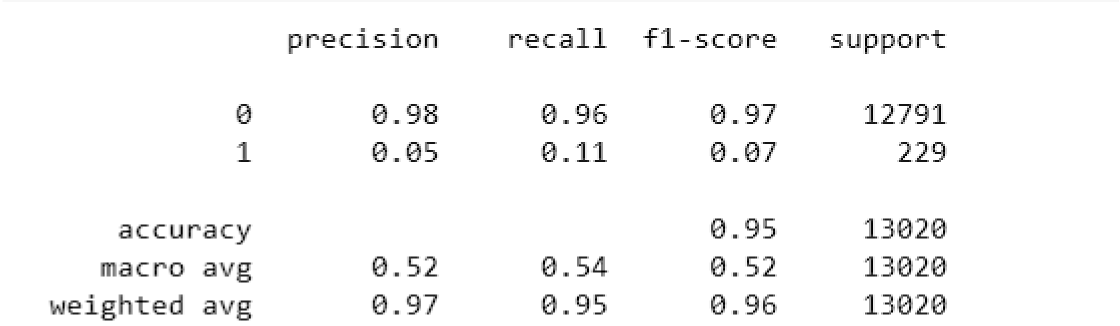
Further, the analysis shows that Random Forest executes better than the other models.

KNN



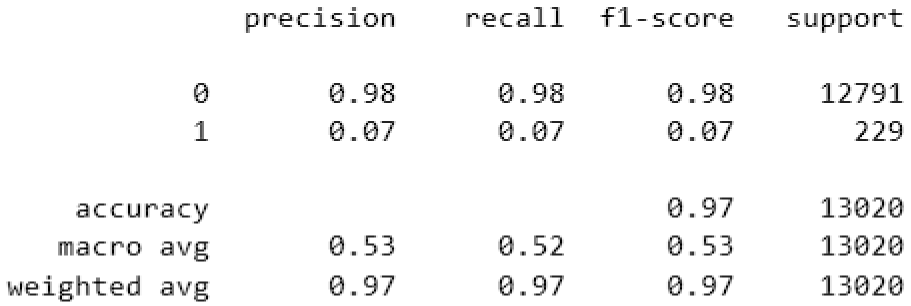
```
              precision    recall  f1-score   support

           0       0.99      0.86      0.92     12791
           1       0.05      0.39      0.08       229

    accuracy                           0.85     13020
   macro avg       0.52      0.62      0.50     13020
weighted avg       0.97      0.85      0.90     13020
```

Decision Tree



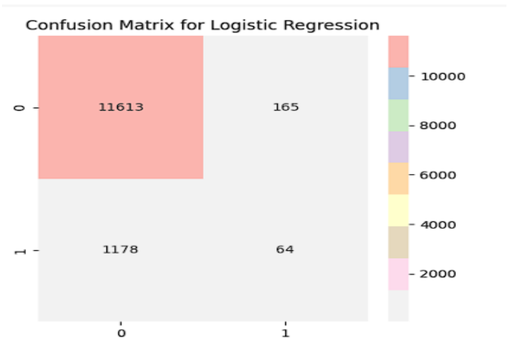
```
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     12791
           1       0.05      0.11      0.07       229

    accuracy                           0.95     13020
   macro avg       0.52      0.54      0.52     13020
weighted avg       0.97      0.95      0.96     13020
```

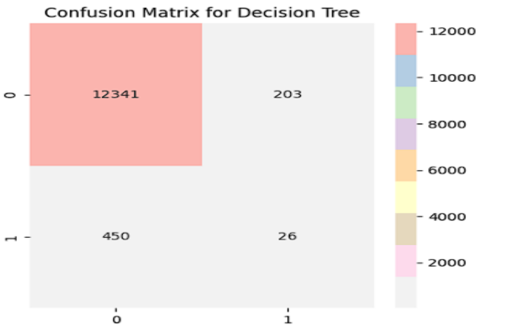
Random Forest

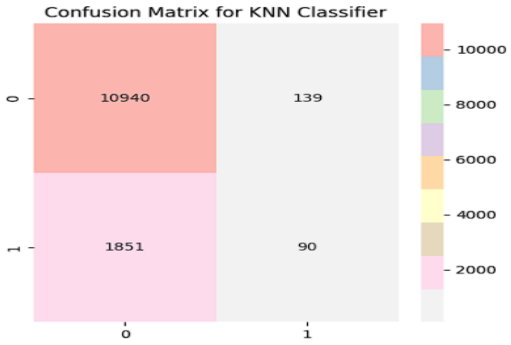


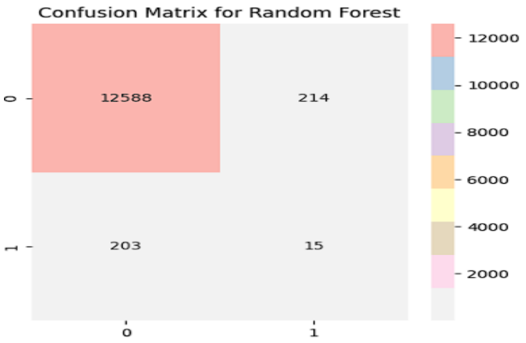
```
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     12791
           1       0.07      0.07      0.07       229

    accuracy                           0.97     13020
   macro avg       0.53      0.52      0.53     13020
weighted avg       0.97      0.97      0.97     13020
```

7. Model Testing:









Here, different machine learning algorithms were trained:
Logistic Regression: 90% Accuracy
Decision Tree: 95% Accuracy
KnnClassifier: 90% Accuracy
Random Forest: 97% Accuracy

8. Conclusion:

In conclusion, the outcomes of utilizing a wide range of machine learning techniques, including random forests, Kth Nearest Neighbour, logistic regression, and decision trees, to foresee brain stroke using individual significant signs are optimistic. These models perform well in brain stroke prediction, with an accuracy that fluctuates between 84.71% to 96.79%. With an accuracy rate of 96.79%, Random Forest outperformed the other tested models. A decision tree with a 94.98% accuracy rate also produced encouraging results. Even though the Knn classifier and logistic regression have an 84.71% and 89.68% accuracy rate respectively, they might still need to be modified for better prediction.

9. Future Extensions:

Although optimistic, the field of machine learning-based brain stroke prediction remains mostly unknown. Technology and medical science are dynamic fields that require constant advancement. While we contemplate the future directions of this project, a number of areas stand out as excellent choices for more research and development:

More in-depth feature analysis: The foundation of any forecasting model is its distinctive characteristics or parameters. While the results from our present collection are noteworthy, a deeper and more thorough examination can reveal details that were missed before. We might rank the variables that have the biggest impact on stroke predictions by using methods like Recursive Feature Elimination or Feature Importance Ranking. Furthermore, wearables and smart devices—new and developing biometric data sources—may provide more detailed data points, allowing us to improve and enrich our feature set.

Advanced Model Training: Although conventional machine learning models such as Random Forest have demonstrated significant potential, the fast developing discipline of deep learning provides further advanced methods. Convolutional neural networks and continuous neural networks in particular are capable of identifying complex patterns in large datasets, which may lead to even improved prediction accuracy levels. Including these in our model suite might result in a prediction process that is more comprehensive.

Clinical Trials in the Real World: The real test of a model is how well it applies in the actual world. Working with medical institutions on trial projects to implement our models in real clinical settings would be the next obvious step. We can obtain important feedback by monitoring its performance in real time, and we may make the required modifications to guarantee the model's effectiveness and usefulness in a variety of settings.

Model Interpretability: As we go farther into advanced algorithms, we encounter a new difficulty: model interpretability. Medical professionals need to be aware of the models' methodology in order to trust and use these resources. It is possible to incorporate methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to provide understandable and easily understandable reports of the model's decision-making process. This helps to foster a more cooperative relationship between humans and machines in addition to ensuring trust.

Continuous Data Integration: New studies and discoveries are frequently made in the medical area, which is always changing. We need a method that enables the continual integrating of new data, research findings, and patient histories to make sure our models stay current and appropriate. In doing so, the model's learning is continuously revised, improving the predictability of results in relation to present clinical knowledge.

Ethics: As AI and machine learning are incorporated into healthcare, ethical issues—particularly those related to the safety and confidentiality of patient data—become ever more significant. It is imperative that future research prioritize the development of strong data management and confidentiality regulations to safeguard patients' rights and security.

In short, the future holds thrilling as well as challenging possibilities. The future of forecasting accuracy at the junction of technology and medicine seems optimistic, with continuous improvement making it a device rather than just a statistic to guarantee the greatest possible healthcare results for people everywhere.

Project Video Recording: ***https://youtu.be/ZkGuzc3zZQI***