
**Title of the project: Brain stroke analysis and classification**

**Group Number: 7005**

## Group Members

| Student Name | Student ID |
|---|---|
| Kaushik Datta | 23341060 |
| Aurchi Roy | 20341010 |
| Ashakuzzaman Odree | 20301268 |
| Debabrata Bhowmick | 20301374 |

## Table of Contents

## Abstract:
This report proposes a comprehensive approach to brain stroke prediction using machine learning. Given the devastating impact of brain strokes on global health, early prediction and intervention play a pivotal role in improving patient outcomes. Herein, the potential of machine learning in stroke diagnosis, the dataset, preprocessing steps, model training, and results are discussed.

## 1. Introduction
This project is all about using machine learning to understand and classify brain strokes. We gather information about factors linked to brain strokes, gender,age, hypertension, heart disease, marital status, residence type, average glucose level, BMI, work type and smoking status  and then we use smart computer techniques to make sense of this data. We teach the computer to recognize patterns that indicate someone might be at risk of a brain stroke.
We tested different models and tried to find the model with the best accuracy to predict braind stroke and classify it.

## 2. Motivation
Brain strokes, often termed "silent killers," primarily owe this moniker to their elusive nature. Their early indicators are frequently so subtle that even well-experienced medical professionals might overlook them, leading to late diagnoses. This latency often results in irreversible neural damage, drastically reducing the quality of life for patients, or in severe cases, leading to death. It's a grim picture that has remained largely unchanged for decades, underscoring the critical need for innovation in early stroke detection methodologies.

Traditional diagnostic methods, while effective to an extent, are often constrained by the limits of human observation and the inherent variability of patient presentations. These methods rely heavily on tangible, observable symptoms, which might manifest only during the advanced stages of a stroke. Moreover, with the increasing patient-to-doctor ratio in many regions worldwide, it's becoming even more challenging to ensure timely and accurate diagnoses for all patients.

Enter the era of digitization and the subsequent data explosion. The contemporary medical landscape is characterized by vast reservoirs of health data – from electronic health records to wearable-generated biometric data. This treasure trove of information remains largely untapped and underutilized in traditional diagnostics.

Machine learning offers a transformative solution to this conundrum. By training sophisticated algorithms on these expansive datasets, it becomes possible to identify patterns and correlations that might be nearly impossible for a human to discern. This is not just about supplementing human observation but augmenting it to levels beyond our innate capabilities. With machine learning models, we can potentially predict the onset of a stroke based on intricate patterns and a myriad of variables, long before tangible symptoms manifest. Such a paradigm shift in diagnostic methods could result in timely interventions, drastically reducing stroke-induced morbidities, and saving countless lives.

Furthermore, integrating machine learning into stroke diagnostics can significantly reduce the economic burden on healthcare systems. Early interventions mean reduced hospital stays, lesser complications, and more effective treatments, translating to millions saved in healthcare costs.

In essence, the motivation behind harnessing machine learning for stroke prediction isn't just a scientific pursuit but a humanitarian one. With the potential to redefine stroke care, this endeavor aims to offer individuals a fighting chance against one of the deadliest medical adversaries known to humanity.

## 3. Dataset Description

Link: [Cerebral Stroke Prediction-Imbalanced Dataset | Kaggle]

Number of Features: 12

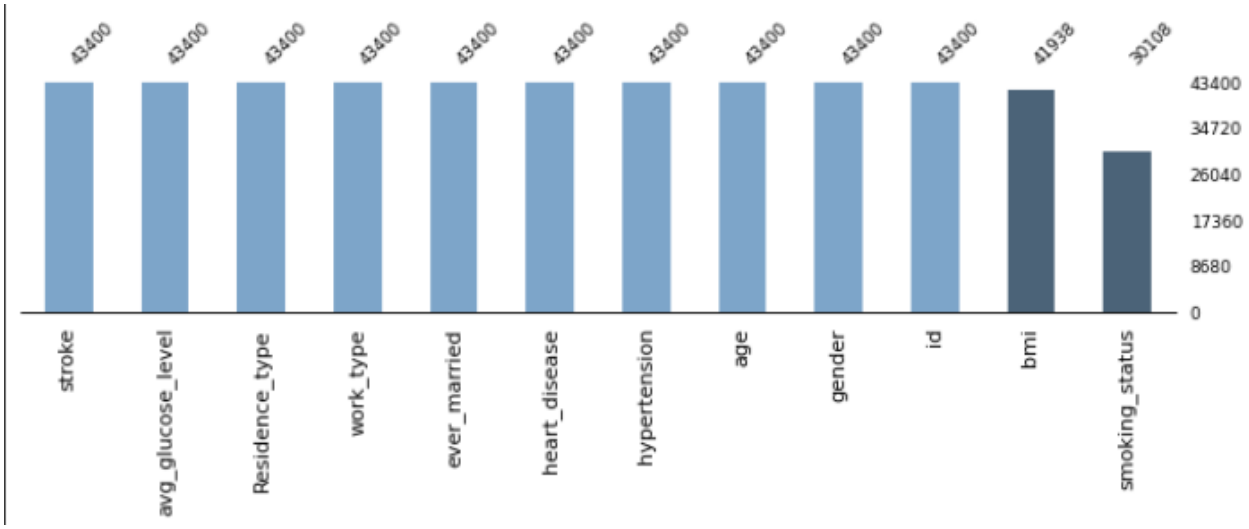
Type of class/label: Categorical and Continuous

Number of data points: 43400

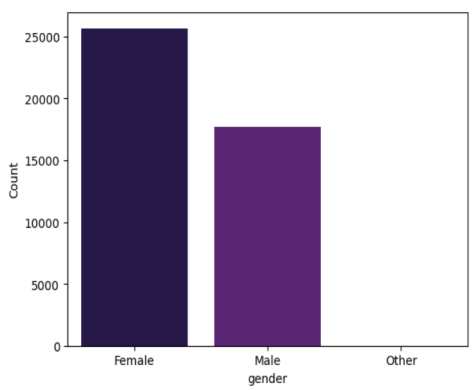
```
<class 'pandas.core.frame.DataFrame'> RangeIndex: 43400 entries, 0 to
43399 Data columns (total 12 columns):

# Column Non-Null Count Dtype
--- ------ -------------- -----
0 id 43400 non-null int64
1 gender 43400 non-null object
2 age 43400 non-null float64
```

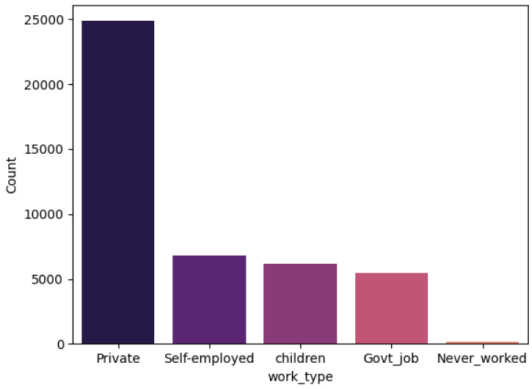
```
3 hypertension 43400 non-null int64
4 heart_disease 43400 non-null int64
5 ever_married 43400 non-null object
6 work_type 43400 non-null object
7 Residence_type 43400 non-null object
8 avg_glucose_level 43400 non-null float64
9 bmi 41938 non-null float64
10 smoking_status 30108 non-null object
11 stroke 43400 non-null int64
dtypes: float64(3), int64(4), object(5)
```

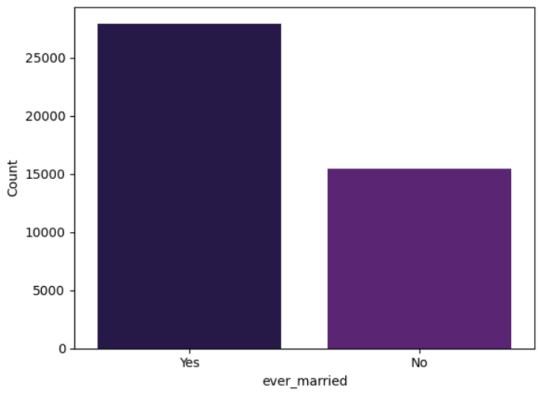
## Biasness/Balanced

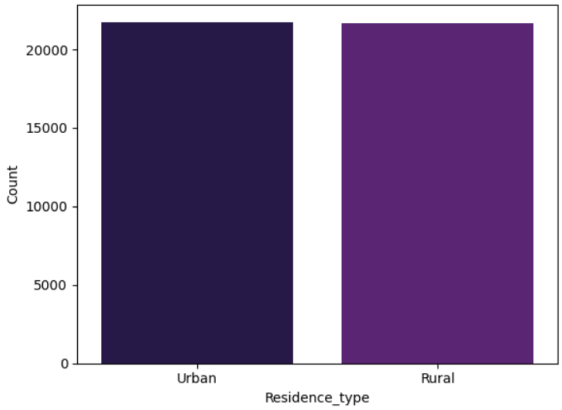


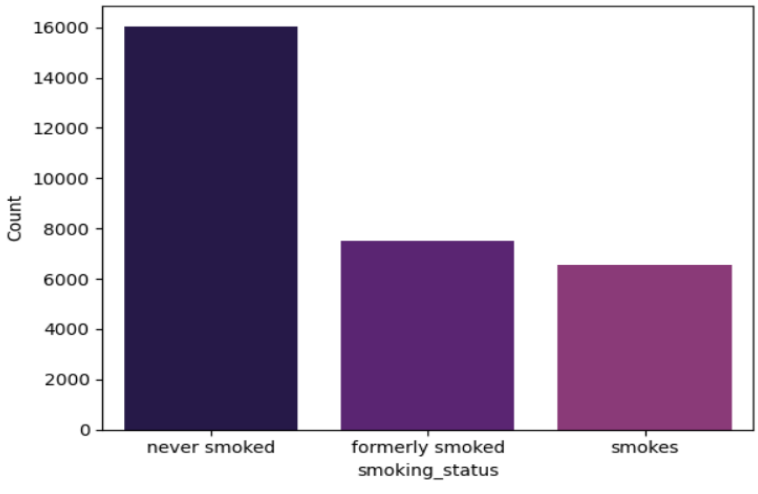

## 4. Dataset Pre-processing



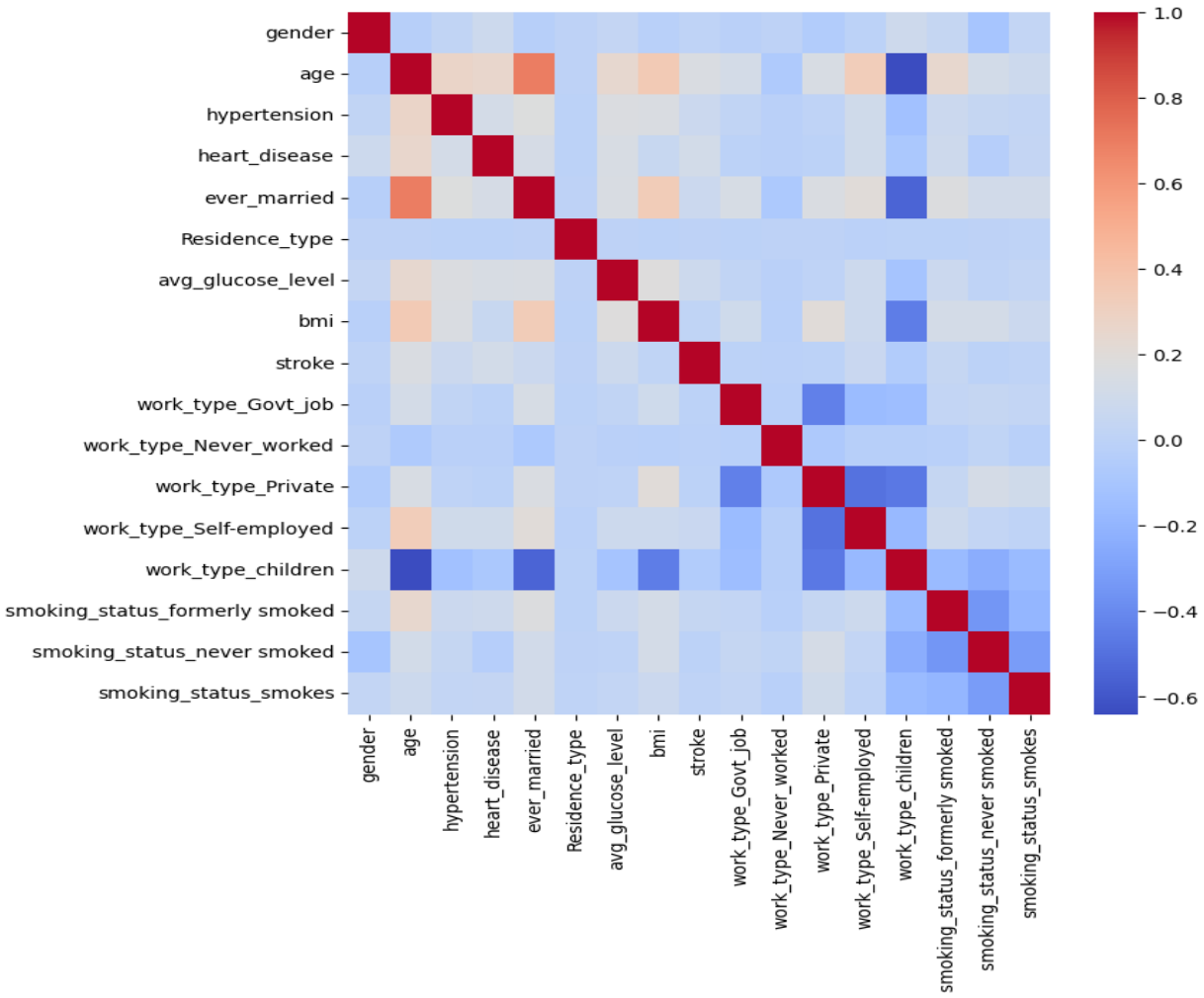




Issues such as null values and outliers were addressed:
5 features contained a total of 868 null values in BMI and 13292 null values in smoking status. Used mean imputation.

‘ID’ column was unnecessary. It was droped.
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 80-20 split was maintained - 8680 for training and 34720 for testing.

## 6. Model Training

Upon testing, the Random Forest model consistently outperformed others in predicting brain stroke from the given indicators.

| Model Name | Accuracy (%) | Error(%) |
|---|---|---|
| Logistic Regression | 90% | 10% |
| Decision Tree | 95% | 5% |
| KnnClassifier | 90% | 10% |
| Random forest | 97% | 3% |

From the table, we can see that the Random forest model showed the best performance with 97% accuracy and only 3% error. On the other hand, the worst performance was given by Logistic regression and Knn classifier 90% accuracy and the error was 10%. After that, Decision tree was mostly satisfying with 95% accuracy.

## Logistic Regression

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.91 | 0.95 | 8538 |
| 1 | 0.04 | 0.25 | 0.08 | 142 |
| accuracy | | | 0.90 | 8680 |
| macro avg | 0.52 | 0.58 | 0.51 | 8680 |
| weighted avg | 0.97 | 0.90 | 0.93 | 8680 |

Also, from this comparison we can see that Random forest has the best performance than other models.

**KNN**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.85 | 0.91 | 8538 |
| 1 | 0.04 | 0.39 | 0.07 | 142 |
| accuracy | | | 0.84 | 8680 |
| macro avg | 0.51 | 0.62 | 0.49 | 8680 |
| weighted avg | 0.97 | 0.84 | 0.90 | 8680 |

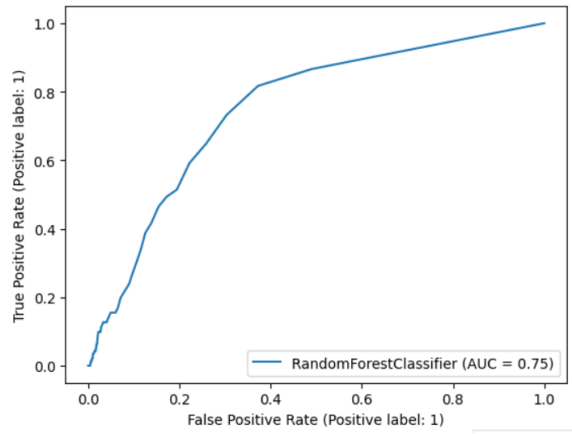
**Decision Tree**

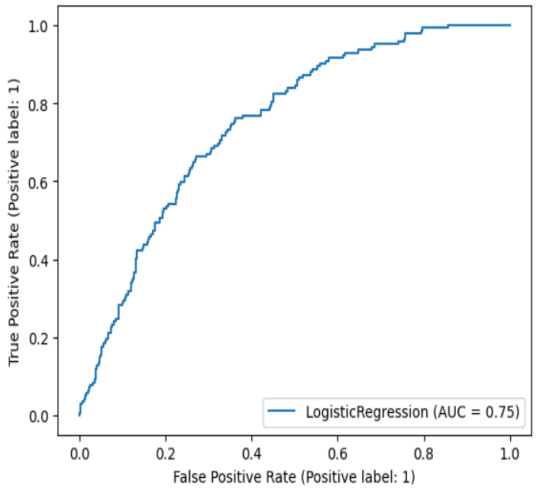
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.94 | 0.96 | 8538 |
| 1 | 0.05 | 0.20 | 0.08 | 142 |
| accuracy | | | 0.93 | 8680 |
| macro avg | 0.52 | 0.57 | 0.52 | 8680 |
| weighted avg | 0.97 | 0.93 | 0.95 | 8680 |

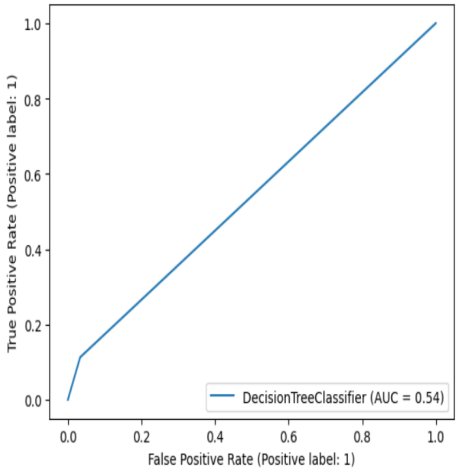
**RandomForest**

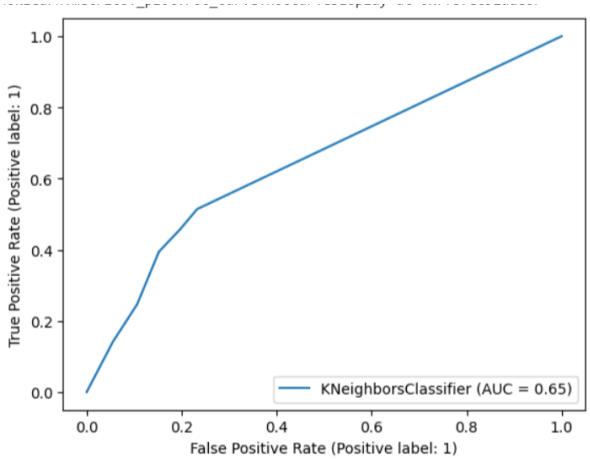
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.97 | 0.98 | 8538 |
| 1 | 0.05 | 0.11 | 0.07 | 142 |
| accuracy | | | 0.95 | 8680 |
| macro avg | 0.52 | 0.54 | 0.52 | 8680 |
| weighted avg | 0.97 | 0.95 | 0.96 | 8680 |

## 7. Model Testing




Various 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 results obtained from applying various machine learning algorithms such as Logistic Regression, Decision Tree, Kth Nearest Neighbor and Random forest, Classifier to predict brain stroke using personal key indicators are promising. With accuracy ranging from 90% to 97%, these models demonstrate good performance in predicting brain stroke.
Among the models tested, Random forest performed the best with accuracy rates of 97%. Decision tree also showed promising results with an accuracy of 95%. Although Logistic regression and Knn classifier achieved an accuracy of 90%, they may require further refinement to improve its predictive performance.

## 9. Future Extensions

The machine learning-driven approach to predicting brain strokes, while promising, has vast horizons yet unexplored. The dynamic nature of both medical science and technology demands continuous evolution. As we consider future trajectories for this endeavor, several areas emerge as prime candidates for further exploration and expansion:

Deeper Feature Analysis: The features or variables used in any predictive model are its backbone. While our current set has yielded significant results, a more profound and comprehensive analysis could uncover nuances previously overlooked. By utilizing techniques such as Recursive Feature Elimination or Feature Importance Ranking, we could prioritize variables that most influence stroke predictions. Additionally, new, emerging biometric data sources, like wearables and smart devices, could offer more granular data points, further refining our feature set.

Advanced Model Training: While traditional machine learning models like Random Forest have shown substantial promise, the rapidly evolving field of deep learning offers

even more advanced tools. Neural networks, especially convolutional and recurrent types, have the capability to recognize intricate patterns in vast datasets, potentially offering even higher prediction accuracies. Incorporating these into our model suite could provide a more holistic prediction mechanism.

Real-world Clinical Trials: A model's true test lies in its real-world applicability. The next logical step would be to collaborate with medical institutions for pilot projects, deploying our models in actual clinical settings. By evaluating its performance in real-time, we can gather invaluable feedback, making necessary adjustments to ensure the model's utility and efficiency in diverse environments.

Model Interpretability: As we tread deeper into complex algorithms, a new challenge emerges – model interpretability. For healthcare professionals to trust and adopt these tools, they must understand how these models arrive at a particular prediction. Techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can be integrated to offer clear, interpretable insights into the model's decision-making process. This not only ensures trust but also facilitates a more collaborative approach between man and machine.

Continuous Data Integration: The medical field is continually evolving, with new researches and findings emerging regularly. To ensure our models remain contemporary and relevant, there's a need for a mechanism that allows continuous integration of new data, research findings, and patient histories. This ensures the model's learning is always updated, making predictions more aligned with current medical knowledge.

Ethical Considerations: With the integration of AI and machine learning into healthcare, ethical considerations, especially concerning patient data privacy and security, become paramount. Future work must emphasize establishing robust data handling and privacy protocols, ensuring patients' rights and security are uncompromised.

In conclusion, the roadmap ahead is both challenging and exhilarating. The intersection of technology and medicine promises a future where predictive accuracy is not just a metric but a tool, continually refined, to ensure the best healthcare outcomes for individuals across the globe.