

**Technical Report: Machine Learning – Based Predictive Model for Stroke Risk Assessment: Development, Evaluation and Clinical Implications**

**Authors: Utkarsh Khatri (Sunway)**

**Date: 26/June/2024**

**Wordcount: 1630 (excluding python)**

**Page count: 19**

**Confidential: YES – DEPARTMENT ONLY**

**Student ID: 24128440**

Table of Contents

[EXECUTIVE SUMMARY 3](#_Toc170455292)

[DOMAIN DESCRIPTION 4](#_Toc170455293)

[PROBLEM DEFINITION 4](#_Toc170455294)

[LITERATURE REVIEW 5](#_Toc170455295)

[i. EFFECTIVENESS OF ENSEMBLE METHODS 5](#_Toc170455296)

[ii. IMPORTANCE OF FEATURE SELECTION 6](#_Toc170455297)

[iii. ADVANCEMENTS IN MACHINE LEARNING TECHNIQUES 6](#_Toc170455298)

[iv. CLINICAL APPLICATIONS AND IMPLICATIONS 6](#_Toc170455299)

[DATA SET DESCRIPTION 7](#_Toc170455300)

[DATASET PRE-PROCESSING 8](#_Toc170455301)

[DATA VISUALIZATION USING MATPLOTLIB AND SEABORN 8](#_Toc170455302)

[EXPERIMENTS 13](#_Toc170455303)

[EXPERIMENTAL DESIGN 13](#_Toc170455304)

[ANALYSIS OF RESULT 14](#_Toc170455305)

[1. PREDICTING STROKE WITH RANDOM FOREST CLASSIFIER 14](#_Toc170455306)

[2. DATA VISUALIZATION AND INSIGHTS 14](#_Toc170455307)

[CONCLUSIONS 18](#_Toc170455308)

[LIMITATIONS AND FUTURE WORK 18](#_Toc170455309)

[REFERENCES 19](#_Toc170455310)

EXECUTIVE SUMMARY

The study presents the evaluation of a ML model designed to predict the chances of stroke among individuals using different health attributes. A Random Forest Classifier was used due to its robustness in handling data sets with diverse nature and their ability to provide interpretable results. Data preprocessing included addressing class imbalance through SMOTE, making sure balanced and accurate predictions. Visualizations offered insights into data distribution and their relationships. The model demonstrated high accuracy and reliability, as shown by detailed performance metrics and a confusion matrix. The studey concludes that machine learning, particularly Random Forest, can significantly improve prediction of stroke and its prevention. The integration of such models into clinical practice promises to support proactive healthcare intereventions, ultimately improving patient outcomes.

DOMAIN DESCRIPTION

One of the leading cause of mortality and disabilty globally, Stroke necessitates accurate predictive models to identify peoples at risk based on demographic, medical and lifestyle parameters [(Akpa et al., 2021).](#b1) Although deep learning has transformed stroke diagnosis, predicting outcomes remains challenging due to complex patient factors, successful machine learning application hinges on data quality and suitablity of algorithms, essential for reliable clinical improvements [(Samuel, 1959).](#b3) In the present context, rapid clinical decision making is being helped by machine leaning technology however, oversight from clinical professionals is still needed to address some parts that may not be accounted for in an automated algorithm.

PROBLEM DEFINITION

Conventional stroke risk models don’t provide an accurate risk assesment towards all patients. However, Artificial Intelligence(AI) algorithms demonstrates more efficient performance compared to conventional models [(Jamthikar et al., 2020).](#b2) The model utilizes supervised learning techniques in order to classify individuals into different stroke risk categories, supporting healthcare professionals in active management and personalized care. After analyzing medical, lifestyle and demographic factors, the model seeks to provide accurate and early diagnosis of high-risk individuals. The project brings attention on the importance of quality of data and selection of algorithm to ensure model’s relaibility. Therefore, the main goal is to integrate this predictive tool into clinical practice.

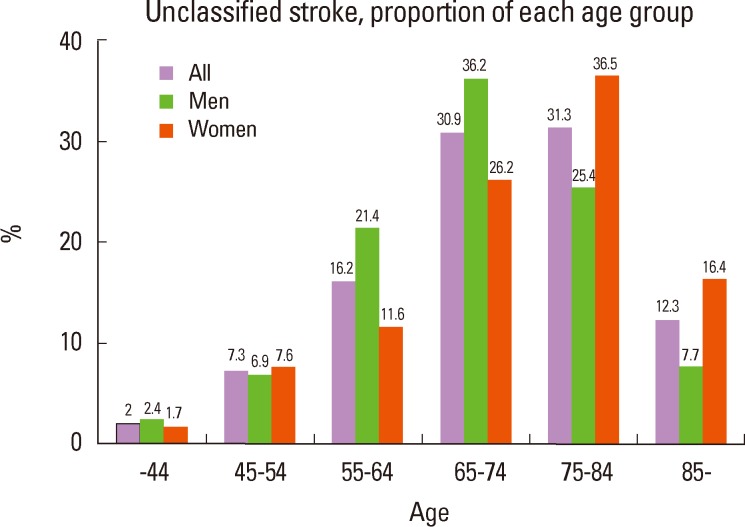


Fig 1: Risk of stroke according to age

LITERATURE REVIEW

## EFFECTIVENESS OF ENSEMBLE METHODS

Analyzing the vast amount of data from sources such as patients, reports, hospitals and IoT-based smart medical devices requires advanced data analytic tools from the machine learning (ML) domain. Although, existing ML techniques have provided acceptable outcomes, they lack in accuray. To address this issue, an ensemble learning strategy which combines the result Edge Detection Instance Preference (EDIP) and Extreme Gradient Boosting (XGboost) can be [implemented (Gao et al., 2018).](#b4) The proposed ensemble learning method significantly improves accuracy as compared to the existing methods .

## IMPORTANCE OF FEATURE SELECTION

Feature selection in data analysis is one of the key point that allows to remove different sorts of machine learning and data mining issues. Building simple and more understandable classifier models for enhancing data mining and processing performance is one of the primary objective of a feature selection method. The research identified that the combination of  (Tree Based Method using Random Forest, TBM-RF, decision tree classifier, DTC) provides accuracy higher than 85%, F1-score higher than 88% is better than  the KNN (K- Nearest Neighbour) and NB ( naïve Bayesian classifier )using the Chi-Square, RFE (Recursive Feature Elimination) and TBM-RF methods [(Chourib et al., 2022)](#b5). Additionally, focusing on important metrics like age, hypertension and lifestyle factors, feature selection helps in reducing model complexity and prevent overfitting.

## ADVANCEMENTS IN MACHINE LEARNING TECHNIQUES

Advancements in machine learning techniques recently, have shown great potential to be used to detect and intervene strokes and other diseases. Although, skeptism regarding it’s adoption in practical life exists, the use of these methods are increasing at a rapid pace. The use of deep learning , neural networks and other advanced algorithms to dissect large and complex datasets into practical and insightful information has made it possible to use large datasets obtained from hospitals, patients etc. and build reliable stroke prediction systems [(Samek et al., 2021).](#b6) These techniques have the potential to crack hidden patterns and corelation that conventional statistics based methods might miss making it more reliable and accurate.

## CLINICAL APPLICATIONS AND IMPLICATIONS

Machine learning and large scale data are key factors in developing an accurate diagnosis model for stroke and integrating ML into clinical side holds notable potential for stroke management. However, ethical considerations like patient privacy, transparency in decision making, bias mitigation and health guidelines needs to be properly addressed while building such models. Any model filled with non-transparency about how the user data is collected and bias filled decisions will result in improper diagonisis and put the patient’s health at risk [(Lorenzini et al., 2022)](#b7). Therefore, for the successful adoption of ML in this field gaining the trust of the healthcare professionals and patients is very essential.

DATA SET DESCRIPTION

The dataset used for the study was downloaded through kaggle. A wide range of medical datasets ensuring access to high-quality data for making predictive models was availabe in the dataset. It included records from patients with a lot of different data attributes necessary to ensure proper stroke analysis. The patients who took part in the dataset are made anonymous in order to ensure their privacy and maintained to promote the data’s integrity and accuracy.

A screenshot of a table

Description automatically generated

Fig 2: Dataset Sample

The dataset's extensive information enables the development of accurate predictive models and the identification of anomalies, thereby improving stroke management for medical personnels.

A screenshot of a graph

Description automatically generated

Fig 3: Dataset Description

DATASET PRE-PROCESSING

Before analysis and modeling, the dataset underwent required steps for data preparation. Firstly, the dataset was imported using pandas library then, categorical attributes like gender, age, hypertension etc. were converted to numerical equivalent for analysis. Lastly, the dataset was divided into testing and traning part in order to correctly asses the performance of the model.

A close-up of a computer screen

Description automatically generated

Fig 4: Splitting the dataset into traning and testing

DATA VISUALIZATION USING MATPLOTLIB AND SEABORN

Data visualization is mandatory to gain insights from the provided dataset as it provides clear and intuitive representation of data distribution as well as the relationship between the variables. Several visualizations were carried out to define different aspects of the dataset:

* **Stroke Distribution:** A bar plot is used to show the distribution of stroke occurrence in the dataset.  
  A graph with a bar and a number of bars

  Description automatically generated with medium confidence

Fig 5: Stroke Distribution

* **Age vs BMI Scatter Plot:** A scatter plot was used to show the relationship between age and BMI which highlights the differences between individuals that have had a stroke and those who don’t.

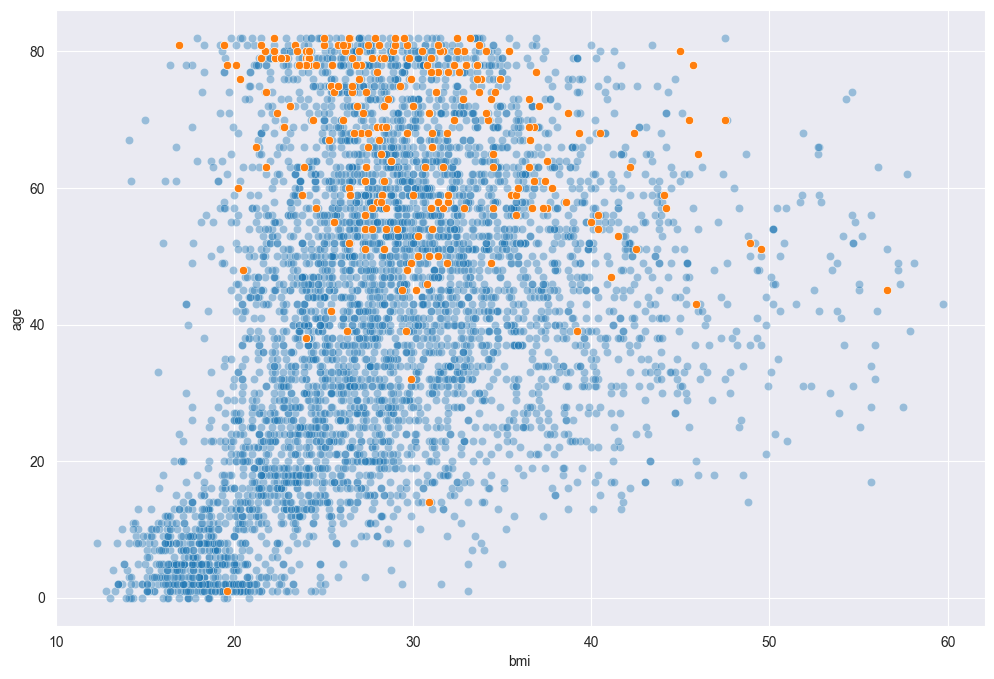


Fig 6: Age vs BMI Scatterplot

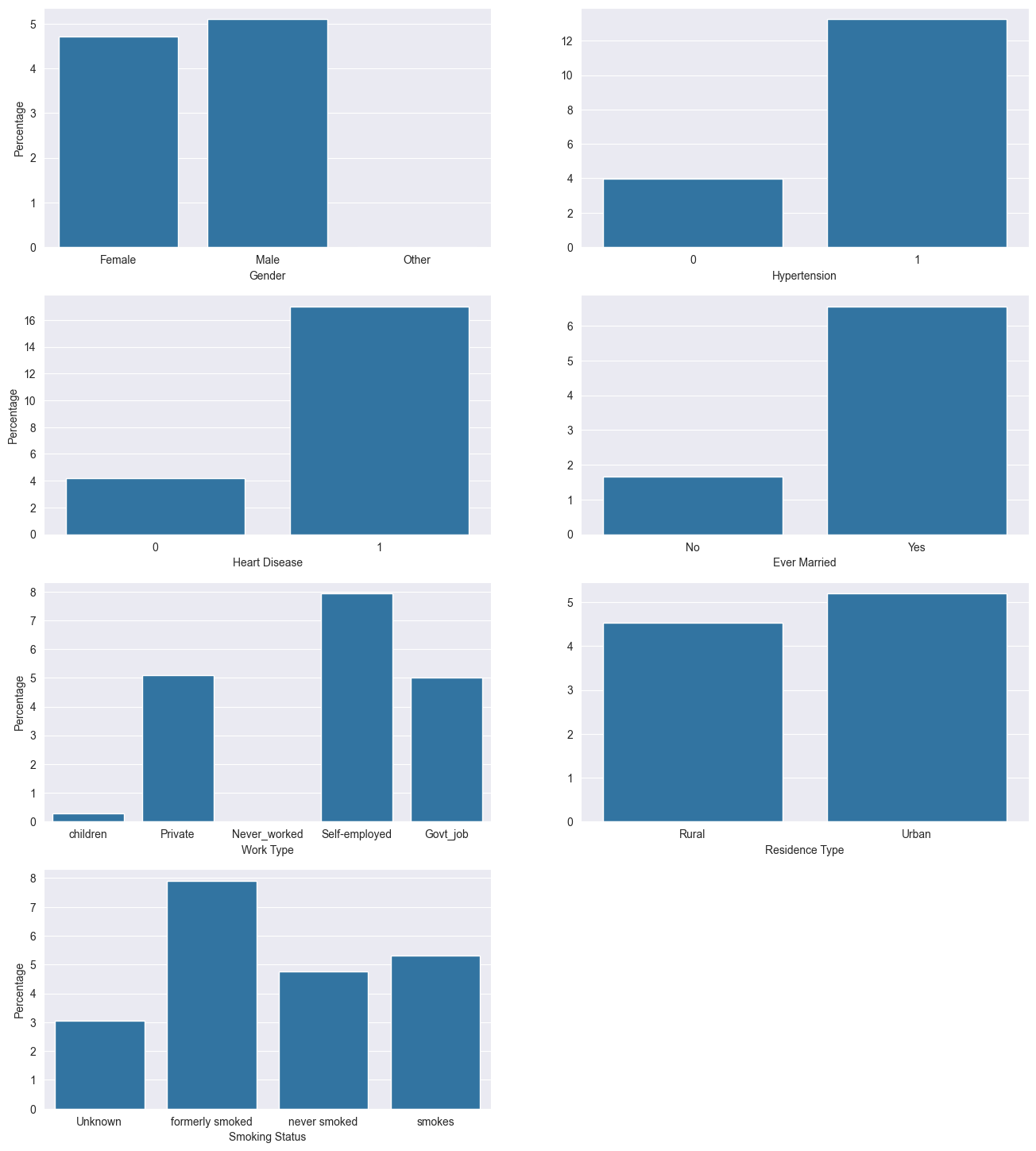
* **Stroke Percentage by Category:** Bar plots were created to showcase the percentage of stroke occurences in different categories.****

Fig 7: Stroke percentage by category

* **Dataset Balance Before and After SMOTE:** Bar plots that shows the distribution of the target variable (stroke) before and after applying the SMOTE techniques to handle the class imbalance were created.

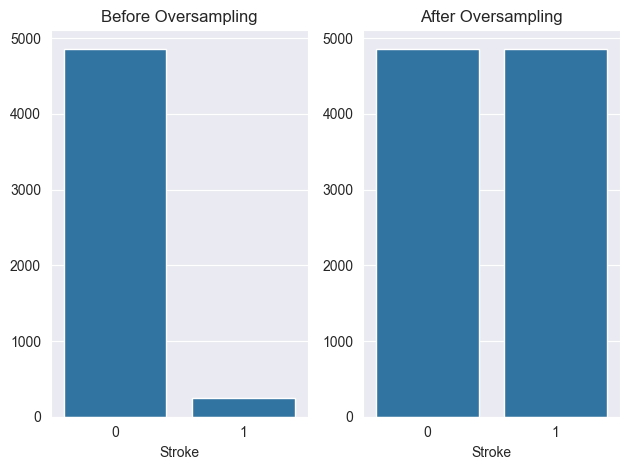


Fig 8: Dataset balance before and after SMOTE

* **Confusion Matrix:** A confusion matrix visualizing the performance of machine learning model on test data, showing the true positive, false positive, true negative and false negative rates was generated.

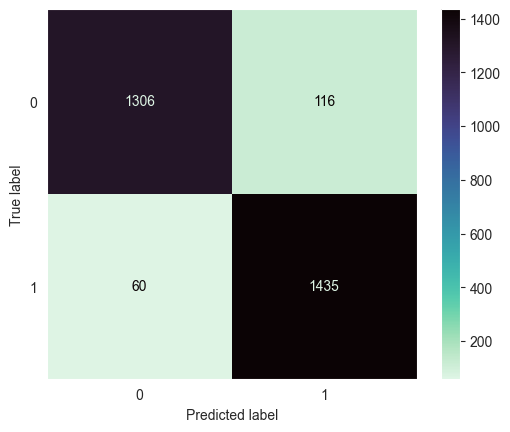


Fig 9: Confusion Matrix

EXPERIMENTS

## EXPERIMENTAL DESIGN

* **Objectives:** The main goal of this project is to determine the likelihood of stroke in individuals by utilizing a Random Forest Classifer and to analyze the dataset for drawing out patterns and insights correlating with stroke occurrence.
* **Methods:** 
  + **Random Forest Classifier for Stroke Prediction:** 
    - **Justification:** Random forest is chosen due to its robustness, ability to deal with both numerical and categorical data and its effectiveness in avoiding overfitting. It aggregates the result from multiple decision trees to improve prediction accuracy.
    - **Code Snippet:**

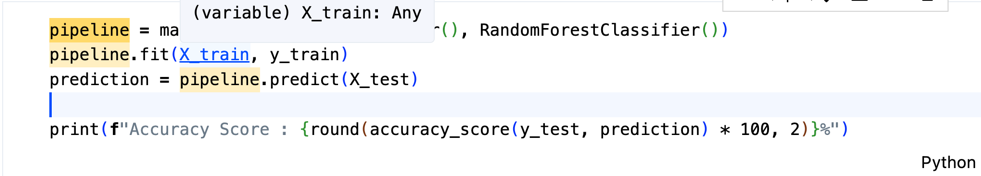
****

Fig 10: Random Forest classifier

* + **SMOTE for Handling Imbalanced Data:**
    - **Justification:** The SMOTE is selected to inscribe the class imbalance in the dataset, which is important for improving the model's ability to predict the minority class (stroke occurrences).
    - **Code Snippet:**

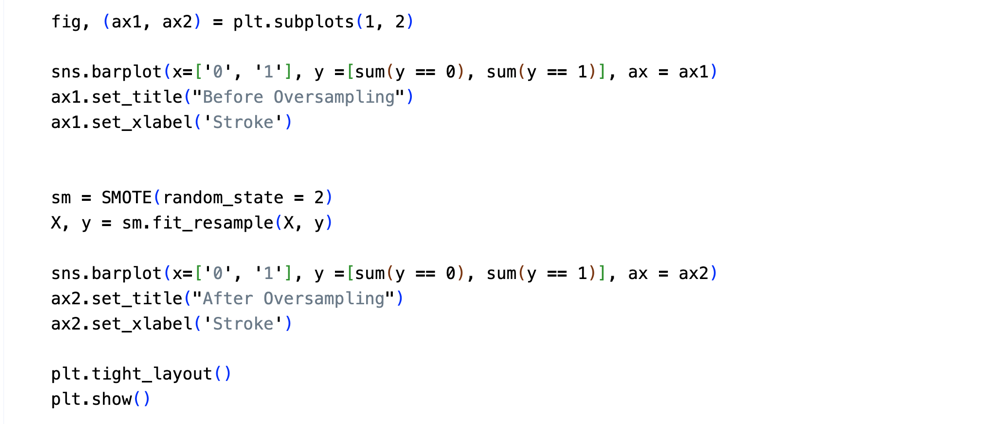
**

Fig 11: SMOTE code

ANALYSIS OF RESULT

## PREDICTING STROKE WITH RANDOM FOREST CLASSIFIER

**MODEL PERFORMANCE:**

* **Accuracy Score:**The accuracy score is calculated to evaluate the proportion of correct predictions made by the model. The accuracy of my model is (93.97 %).
* **Classification Report:**It provides comprehensive metrics such as precision, recall, and F1-score for each class.
* **Confusion Matrix:**Visual representation of the performance of the classification model.

**INTERPRETATION:**

A screenshot of a graph

Description automatically generated

Fig 12: Prediction value and precision

## DATA VISUALIZATION AND INSIGHTS

**STROKE DISTRIBUTION:**

A barplot shows the distribution of stroke occurences, giving a glimpse of the class imbalance in the dataset.

**A comparison of a graph

Description automatically generated**

Fig 13: Before and after oversampling

**AGE VS BMI SCATTER PLOT:**

A scatter plot reveals the relationship between age and BMI, highlighting differences between stroke and non-stroke individuals.

A blue and orange dots

Description automatically generated

Fig 14: BMI vs AGE

**STROKE PERCENTAGE BY CATEGORY:**

Bar plots illustrate the percentage of stroke occurrences across various categorical features such as marital status.

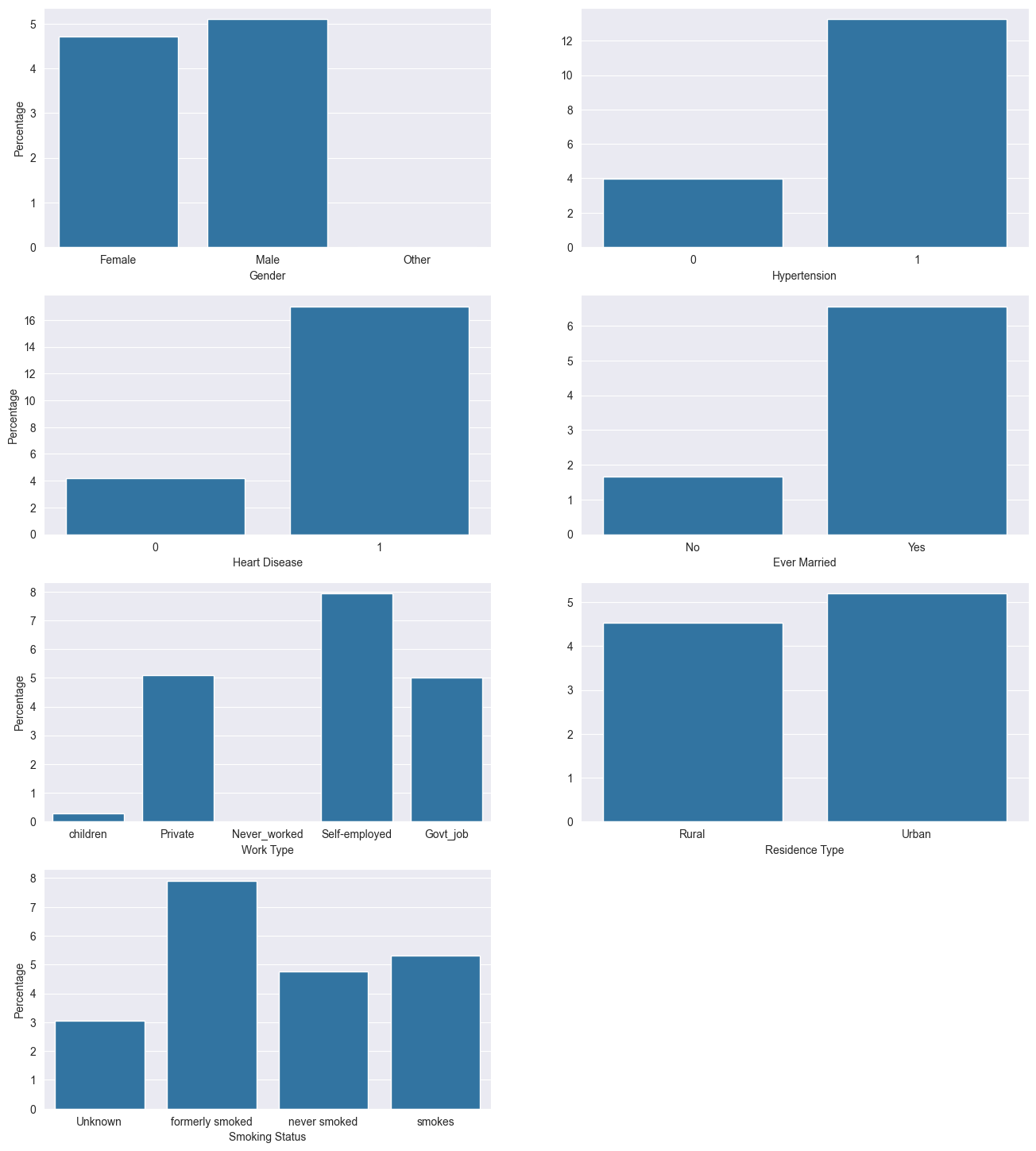


Fig 15: Stroke Percentage By Category

CONCLUSIONS

**PREDICTIVE MODELLING:**

The Random Forest classifier effectively predicts stroke likelihood, achieving high accuracy and robust performance metrics. This model's strength lies in its ability to handle diverse data types and provide interpretable results. However, further enhancements can be made by incorporating more advanced techniques and additional data features.

**DATA VISUALIZATIONS:**

Visualizations provide valuable insights into the dataset, revealing patterns and relationships among variables. These graphical representations aid in understanding the distribution of stroke occurrences and the impact of various factors on stroke risk.

LIMITATIONS AND FUTURE WORK

The scope of this study is limited by the size and diversity of the dataset and the choice of a single model. Potential biases in the dataset and the limitations of the Random Forest model may affect the generalizability of the findings.

Future work should explore more advanced models, such as neural networks or ensemble methods, to improve predictive accuracy. Longitudinal studies and the inclusion of time-series analysis would provide deeper insights into stroke risk factors and trends over time.

REFERENCES

Akpa, O. *et al.* (2021) ‘A novel afrocentric stroke risk assessment score: Models from the Siren Study’, *Journal of Stroke and Cerebrovascular Diseases*, 30(10), p. 106003. doi:10.1016/j.jstrokecerebrovasdis.2021.106003.

Samuel, A.L. (1959) ‘Some studies in machine learning using the game of Checkers’, *IBM Journal of Research and Development*, 3(3), pp. 210–229. doi:10.1147/rd.33.0210.

Jamthikar, A.D. *et al.* (2020) ‘Artificial Intelligence Framework for predictive cardiovascular and stroke risk assessment models: A narrative review of integrated approaches using carotid ultrasound’, *Computers in Biology and Medicine*, 126, p. 104043. doi:10.1016/j.compbiomed.2020.104043.

Gao, S., Dai, J. and Shi, H. (2018) ‘Discernibility matrix-based ensemble learning’, *2018 24th International Conference on Pattern Recognition (ICPR)* [Preprint]. doi:10.1109/icpr.2018.8545243.

Chourib, I. *et al.* (2022) ‘Stroke treatment prediction using features selection methods and machine learning classifiers’, *IRBM*, 43(6), pp. 678–686. doi:10.1016/j.irbm.2022.02.002.

Samek, W. *et al.* (2021) ‘Explaining Deep Neural Networks and beyond: A review of methods and applications’, *Proceedings of the IEEE*, 109(3), pp. 247–278. doi:10.1109/jproc.2021.3060483.

Lorenzini, G. *et al.* (2022) ‘Machine learning applications in healthcare and the role of informed consent: Ethical and practical considerations’, *Clinical Ethics*, 18(4), pp. 451–456. doi:10.1177/14777509221094476.