# Predicting Stroke Susceptibility:

# Enhancing Preventive Healthcare Measures through Advanced Data Analytics

## Introduction

### Purpose

Strokes pose a critical public health challenge, with their incidences surpassing that of cancer-related fatalities annually. Identifying people who are at high risk for strokes early on is key to preventing them. Our project uses current biometric data and other factors to build predictive models that can accurately forecast who might be more susceptible to strokes. By focusing on prevention, we aim to reduce the number of strokes in the future.

### Audience

The application of stroke prediction models can impact various sectors:

Clinical/Hospitals: Integrating predictive models into electronic health records (EHR) can help clinicians identify high-risk patients early, allowing for targeted and personalized care plans, ultimately improving patient outcomes.

Insurance Companies: Utilizing these models for risk assessment can enable more accurate premium pricing and the development of preventative health programs. This approach helps in mitigating costs associated with stroke-related claims and promotes healthier lifestyles among policyholders.

Biotech Companies: By incorporating stroke prediction models into health monitoring devices, biotech firms can provide users with real-time risk assessments, facilitating timely lifestyle adjustments and medical consultations. This integration supports preventive healthcare and enhances user engagement.

These implementations demonstrate the broad applicability of stroke prediction models, highlighting their potential to improve healthcare delivery, reduce costs, and foster innovation across various industries.

# Data

## Dataset Description

The stroke dataset consists of health and demographic information for 43,400 individuals. Key features include age, hypertension, heart disease, marital status, average glucose level, BMI, smoking status, and stroke occurrence. The dataset's diversity in biometric, demographic, and lifestyle factors makes it ideal for detailed analysis and feature engineering.

## Data Preprocessing

**Our Preprocessing steps included:**

- Handling Missing Values: Imputed missing BMI values with the median and smoking status with ratios matching the original dataset.
- Encoding Categorical Variables: Converted categorical variables (gender, marital status, work type, residence type, smoking status) into dummy variables.
- Normalizing Continuous Variables: Standardized age, average glucose level, and BMI to ensure they are on a similar scale.
- Data Splitting: Divided the dataset into training and testing sets (e.g., 80-20 split) for model evaluation.

These steps ensure data quality and consistency, supporting robust predictive model development for stroke risk assessment.

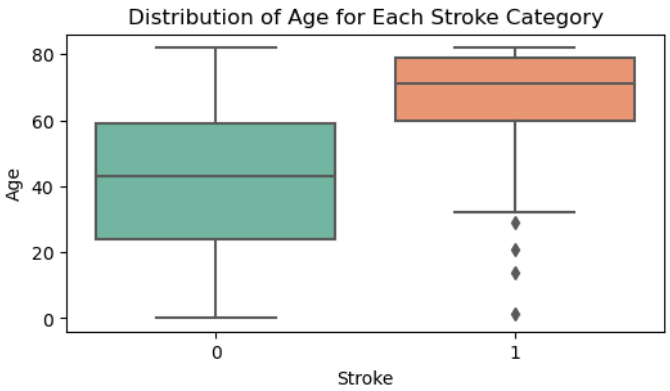
# EDA

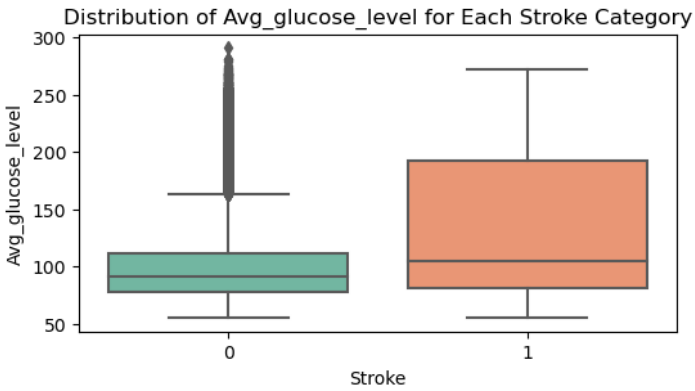
## Summary of Exploratory Data Analysis (EDA)

Our Exploratory Data Analysis (EDA) provided critical insights into the stroke dataset, particularly highlighting the significant correlation between age, average glucose levels, and stroke occurrence. These findings are crucial for understanding the key risk factors and guiding the development of effective predictive models.

**Key Findings:**

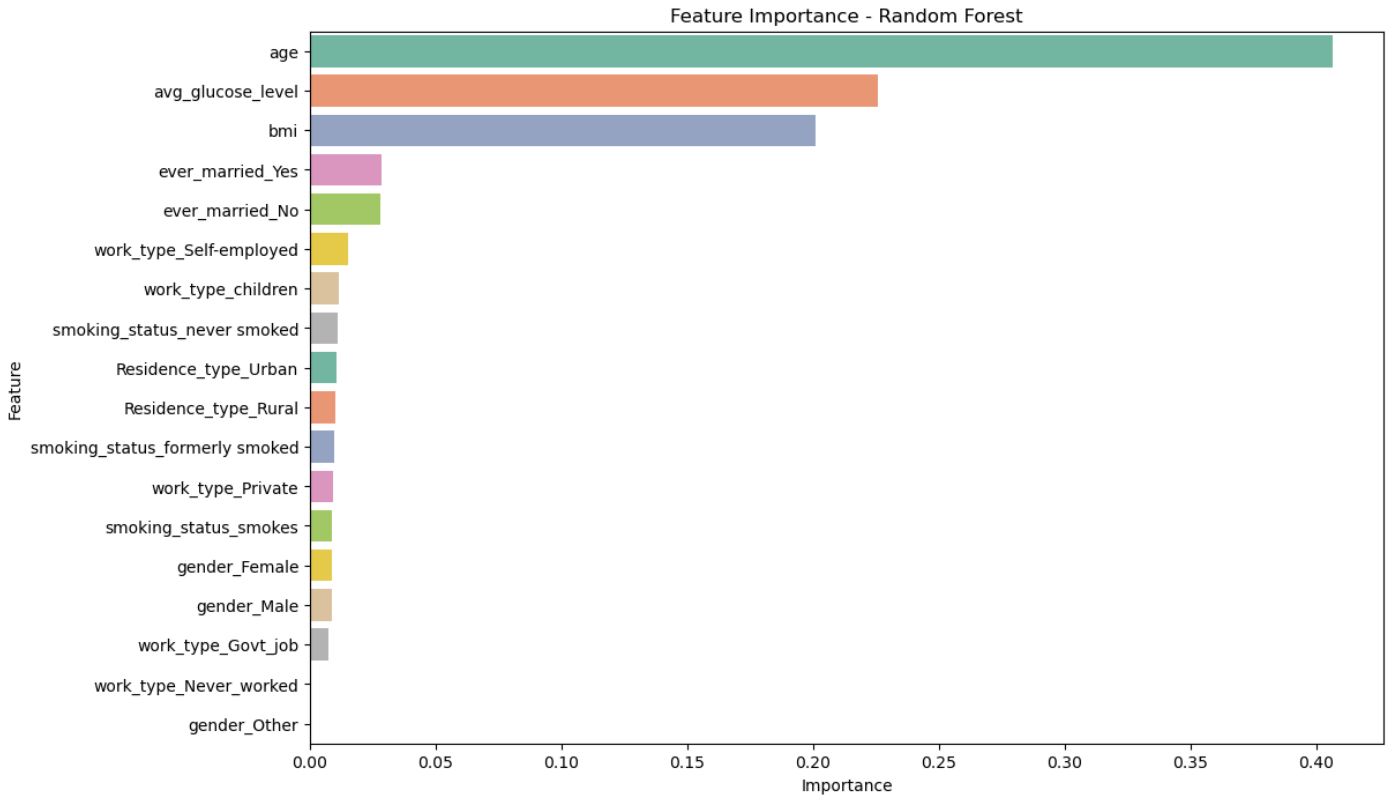
1. **Age**:
    - **Observation**: Individuals who experienced a stroke were generally older, with a median age around 70 compared to 45 for those who did not have a stroke. This is confirmed by our T-Test confirmed a significant difference in age distributions between the two groups (Statistic=32.91, P-value=1.17e-234).
    - **Importance**: Age is a well-known risk factor for many health conditions, including stroke. The strong correlation observed in our dataset reinforces the importance of age as a critical variable in predicting stroke risk.
2. **Average Glucose Level**:
    - **Observation**: Stroke patients tended to have higher average glucose levels, with a median around 150 mg/dL compared to 90 mg/dL for non-stroke individuals. Our T-Test supported this observation (Statistic=16.49, P-value=6.47e-61).
    - **Importance**: Elevated glucose levels are a critical risk factor for stroke, particularly among individuals with diabetes. The significant difference in glucose levels between stroke and non-stroke individuals underscores the importance of managing blood sugar levels as part of stroke prevention strategies.



# Modeling

Using a Random Forest model with Grid Search for hyperparameter tuning, we achieved high accuracy in predicting individuals at high risk of stroke. Key predictors identified include age, average glucose level, and BMI, which align with known risk factors for stroke. The model's robust performance and the insights gained from feature importance analysis underscore its potential for practical application in clinical settings to aid in early identification and preventive care.




Interestingly, the feature 'ever_married' shows some importance in the model, but this factor may be correlated with age rather than directly influencing stroke risk. Older individuals are more likely to be married, and their higher age alone could account for their increased stroke risk, rather than their marital status being a direct cause. Other factors like work type and smoking status also contribute to the model but to a lesser extent, and their influence may be intertwined with other demographic and health-related variables.

Overall, while some variables are highlighted in the analysis, their apparent importance could be attributed to their correlation with more directly relevant predictors such as age. This highlights the necessity of careful interpretation of feature importance in the context of known medical relationships and demographic patterns.

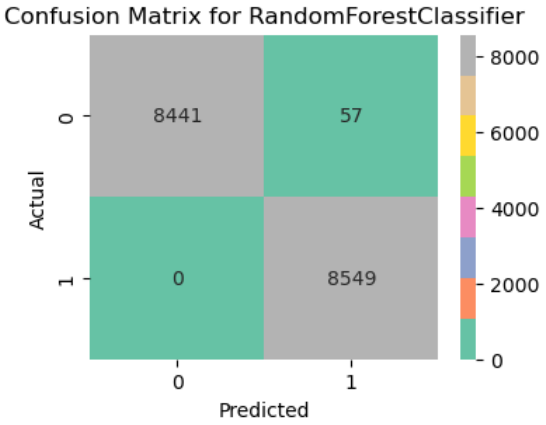
## Summary of Model Results and Their Implications for Stroke Prediction

The Random Forest model developed for stroke prediction demonstrated exceptional performance, achieving nearly perfect metrics across all evaluation parameters.

The classification report shows outstanding performance with an overall accuracy of 1.00. Key highlights include:

- **Stroke Cases ("Yes")**:
  - **Precision**: 0.9934
  - **Recall**: 1.00 (no false negatives)
- **Non-Stroke Cases ("No")**:
  - **Precision**: 1.00
  - **Recall**: 0.9933 (very low false positive rate)

These results demonstrate the model's exceptional ability to accurately predict both stroke and non-stroke cases, ensuring high reliability in identifying individuals at risk.




The confusion matrix further supports these results, showing 8441 true negatives, 57 false positives, and 8549 true positives, with no false negatives. This balanced performance is reflected in the balanced accuracy score of 0.9966, which confirms the model's robustness and reliability.

These results are particularly significant for stroke prediction because they demonstrate the model's ability to accurately and consistently identify individuals at high risk of stroke. The high precision and recall scores ensure that the model minimizes both false positives and false negatives, which are crucial in a clinical setting where accurate identification can lead to timely and potentially life-saving interventions.

Overall, the model's exceptional performance metrics indicate its potential for practical application in healthcare settings, providing a reliable tool for early identification and preventive care of individuals at high risk of stroke.

## Training vs. Test Performance Metrics

The bar chart below compares the performance metrics of the Random Forest model on training and test datasets. Key observations include:

- **Accuracy**: Both training and test sets achieve near-perfect accuracy.
- **Balanced Accuracy**: High and consistent scores across both datasets.
- **Precision**: Precision for both stroke (Yes) and non-stroke (No) categories is nearly identical in training and test sets, indicating the model's reliability.
- **Recall**: Recall scores are similarly high and consistent for both categories, confirming the model's effectiveness in identifying true positives and minimizing false negatives.

These results demonstrate that the model performs exceptionally well on both training and test data, highlighting its robustness and generalization capabilities.

# Recommendations

Based on the insights from our analysis and the strong performance of the Random Forest model, we recommend integrating this predictive model into clinical practice to enhance early identification and prevention of stroke. Specifically, healthcare providers should deploy the model within electronic health record (EHR) systems to flag high-risk individuals during routine check-ups, with a focus on monitoring older adults and individuals with elevated glucose levels. Regular tracking of key health metrics such as age, BMI, and glucose levels is crucial for timely interventions. Additionally, patient education programs should be developed to emphasize the importance of maintaining healthy glucose levels and BMI through lifestyle modifications.

Further, ongoing research should be conducted to refine the model and validate its effectiveness across diverse populations. Healthcare policies should support the integration of predictive analytics in stroke prevention programs, and resources should be allocated to ensure healthcare providers are equipped with the necessary tools and training. By implementing these recommendations, the healthcare system can improve its ability to predict and prevent strokes, ultimately leading to better patient outcomes and a reduction in the incidence and severity of strokes.

# Future Work

In the ongoing pursuit of enhancing stroke prediction, several avenues for future research and development are proposed. First, incorporating larger and more diverse datasets could improve model generalizability and accuracy across different populations. Additionally, exploring advanced modeling techniques such as ensemble learning and neural networks may further refine predictive capabilities.

Another promising area is integrating real-time data from wearable health devices, which could enable continuous monitoring and early intervention for at-risk individuals. Finally, collaboration with healthcare professionals to validate and fine-tune these models in clinical settings will be crucial for practical implementation, ensuring that predictive tools are both accurate and actionable in real-world applications.

# Conclusion

Our analysis of the stroke dataset, complemented by the development of a Random Forest predictive model, revealed significant insights into stroke risk factors. The Exploratory Data Analysis (EDA) highlighted age and average glucose levels as critical predictors, with older individuals and those with higher glucose levels showing a markedly increased risk of stroke. The model demonstrated exceptional performance, achieving near-perfect accuracy, precision, and recall on both training and test datasets. This high level of performance indicates the model's reliability and its potential for practical application in clinical settings to accurately identify individuals at high risk of stroke.

Future work should focus on further refining the model and validating its effectiveness across diverse populations to ensure its broad applicability. Additionally, integrating this predictive model into electronic health record (EHR) systems could enable healthcare providers to implement timely and targeted preventive measures, potentially reducing the incidence and severity of strokes. By leveraging these insights and tools, healthcare systems can improve patient outcomes, optimize resource allocation, and enhance the overall efficiency of stroke prevention efforts. This approach represents a significant advancement in predictive healthcare, highlighting the importance of data-driven strategies in managing public health challenges.