Stroke Prediction Project

Submitted by:
Nofar Selouk 318502721
Noa Ohayon 315375998
Orel Yosef 212240741
David Zalensky 323794131

July 16, 2024

Abstract

Stroke is a leading cause of disability and death worldwide. Early prediction of stroke can lead to timely medical interventions, potentially saving lives and improving the quality of life for stroke survivors. This report explores the use of machine learning techniques for predicting stroke risk using a dataset sourced from Kaggle. Various preprocessing steps, feature engineering techniques, and machine learning models were employed, with a Random Forest classifier being identified as the most effective.

1 Introduction

Stroke is a serious medical condition that occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting the oxygen and nutrients it needs. Without immediate treatment, brain cells begin to die within minutes. Stroke is a leading cause of death and disability worldwide, making it a critical public health issue. The early prediction and prevention of strokes can significantly improve patient outcomes, reduce healthcare costs, and save lives.

In the past, various approaches have been taken to predict and prevent strokes. Traditional methods have included clinical assessments and the use of simple statistical models to identify high-risk individuals based on factors such as age, hypertension, diabetes, and smoking status. More recently, advancements in machine learning and data science have enabled the development of more sophisticated predictive models that can analyze large

datasets and identify complex patterns associated with stroke risk. Previous research has demonstrated the potential of algorithms like logistic regression, support vector machines, and neural networks in predicting stroke risk with varying degrees of success.

This project aims to build on these advancements by utilizing machine learning techniques to develop a predictive model for stroke, leveraging a comprehensive dataset of health parameters.

2 Background/Related Work

Several studies have explored the use of machine learning techniques for stroke prediction. For example, a study by researchers at the University of Patras investigated the use of various machine learning algorithms, including Naive Bayes, logistic regression, k-nearest neighbors, stochastic gradient descent, decision trees, multi-layer perceptrons, and random forests. They found that the stacking classification method outperformed other approaches, achieving high AUC, F-measure, precision, recall, and accuracy.

Another study focused on predicting stroke risk using lab test data and machine learning algorithms, including Naive Bayes, BayesNet, J48, and random forests. This study highlighted the random forest model as the most accurate, identifying nine lab tests along with age and gender as significant predictors of stroke occurrence.

Further research investigated stroke risk prediction using machine learning techniques on a dataset from Kaggle, focusing on participants over 18 years old with various attributes. The study found that the stacking ensemble method combining multiple base classifiers was the most efficient approach, achieving high performance metrics.

These studies underscore the potential of machine learning in early stroke prediction, paving the way for improved patient outcomes and healthcare decision-making.

3 Methodology

3.1 Data Collection

The dataset used for predicting stroke is sourced from Kaggle, a reputable platform for datasets and data science projects. It captures a comprehensive

range of health parameters that influence stroke risk. It includes variables such as demographics, lifestyle choices, and medical history, making it a valuable resource for predictive modeling. The dataset's extensive coverage of relevant factors provides a solid foundation for developing accurate and reliable predictive models.

3.2 Data Cleaning

Data cleaning is an essential step in preparing the dataset for analysis and modeling. Records with missing values in the BMI column were removed, duplicates were checked, and data types were corrected to ensure consistency and accuracy. For example, numeric values such as age, hypertension, heart disease, average glucose level, and BMI were converted to the appropriate numeric data types.

3.2.1 Handling Missing Values

We checked the dataset for missing values and identified 201 instances of missing values in the BMI column. Due to their small number compared to the total number of rows, we chose to remove these records with missing values.

```
# Check for missing values
print(stroke_data.isnull().sum())

# Drop rows with any remaining missing values
stroke_data.dropna(how='any', inplace=True)
```

3.2.2 Removing Duplicates

We checked for duplicate rows in the dataset and found none. Therefore, no rows were removed for this purpose.

```
# Check for duplicates
duplicates = stroke_data[stroke_data.duplicated()]
num_duplicates = duplicates.shape[0]
print(f'There are {num_duplicates} duplicate rows in the dataset.')

# Remove duplicates if any
stroke_data.drop_duplicates(inplace=True)
```

3.2.3 Correcting Data Types

To ensure consistency and proper analysis, we corrected the data types of certain columns. For instance, columns representing numerical values such as *age*, *hypertension*, *heart_disease*, *avg_glucose_level*, and *bmi* were converted to the appropriate numeric data types.

```
# Correct data types
stroke_data['age'] = stroke_data['age'].astype(float)
stroke_data['hypertension'] = stroke_data['hypertension'].astype(int)
stroke_data['heart_disease'] = stroke_data['heart_disease'].astype(int)
stroke_data['avg_glucose_level'] = stroke_data['avg_glucose_level'].astype
stroke_data['bmi'] = stroke_data['bmi'].astype(float)
stroke_data['stroke'] = stroke_data['stroke'].astype(int)
```

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involves visualizing the data to identify trends, correlations, and potential outliers. Bar plots and box plots were created to visualize the distribution of categorical and numerical variables with respect to stroke occurrences. These visualizations help in understanding the data better and deriving insights for feature engineering and model selection.

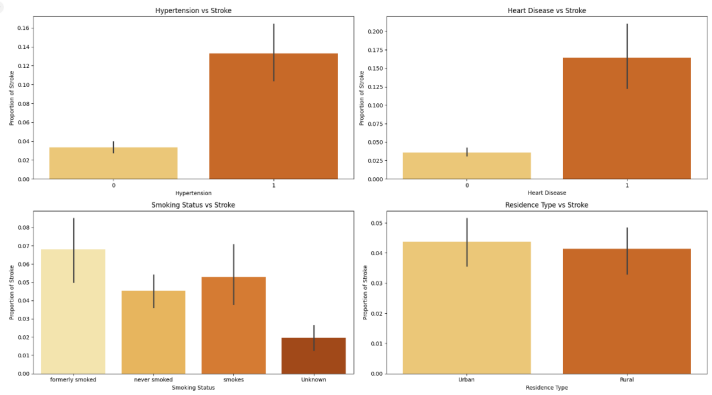


Figure 1: Bar plots for categorical variables

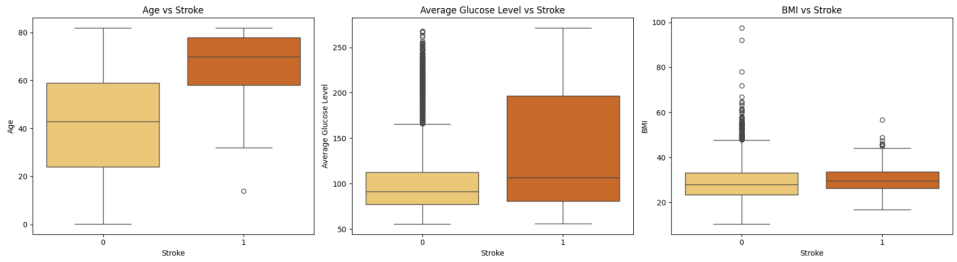


Figure 2: Box plots for numerical variables

3.4 Data Balancing using Undersampling and SMOTE

In this project, we addressed the class imbalance in our dataset using Random Undersampling followed by SMOTE (Synthetic Minority Over-sampling Technique). Class imbalance can lead to biased model predictions, where the model tends to favor the majority class. To mitigate this, we performed the following steps:

- **Random Undersampling**: We first reduced the number of samples in the majority class to 1000. This step ensures that the majority class does not dominate the training process.
- **SMOTE**: After undersampling, we applied SMOTE to generate synthetic samples for the minority class, balancing the dataset. SMOTE creates synthetic samples by interpolating between existing minority class samples.

Additionally, we applied Label Encoding to convert categorical variables into numerical values before resampling, which is necessary for the resampling algorithms to work. After resampling, we reversed the label encoding to maintain the original categorical values.

By combining these techniques, we achieved a balanced dataset, which helps in improving the model's performance and ensuring it can predict both classes more effectively.

4 Feature Engineering

4.1 Age Groups

The continuous age variable was converted into categorical age groups (e.g., 0-18, 19-35, 36-50, 51-65, 66+). This helps the model better capture the non-linear relationship between age and stroke risk.

```
stroke_data['age_group'] = pd.cut(stroke_data['age'],
                                  bins=[0, 18, 35, 50, 65, 100],
                                  labels=['0-18', '19-35', '36-50', '51-65', '66+'])
stroke_data = stroke_data.drop(['age'], axis=1)
```

4.2 BMI Categories

BMI values were categorized into standard BMI categories (e.g., underweight, normal weight, overweight, obese). This helps in identifying different risk levels associated with various BMI ranges.

```
stroke_data['bmi_category'] = pd.cut(stroke_data['bmi'],
                                     bins=[0, 18.5, 24.9, 29.9, 100],
                                     labels=['Underweight', 'Normal weight', 'Overwei
stroke_data = stroke_data.drop(['bmi'], axis=1)
```

4.3 Average Glucose Level Ranges

Average glucose levels were categorized into ranges (e.g., normal, prediabetic, diabetic). This helps in understanding how glucose levels impact stroke risk.

```
stroke_data['glucose_level_category'] = pd.cut(stroke_data['avg_glucose_level'],
                                               bins=[0, 99, 125, 500],
                                               labels=['Normal', 'Prediabetic', 'Diab
stroke_data = stroke_data.drop(['avg_glucose_level'], axis=1)
```

4.4 Smoking Status Simplification

The smoking status was simplified into binary categories (e.g., smoker, non-smoker) to reduce model complexity while retaining essential information.

```
stroke_data['smoking_status_simplified'] = stroke_data['smoking_status'].apply(
    lambda x: 'Non-smoker' if x == 'never smoked' else 'Smoker')
stroke_data = stroke_data.drop(['smoking_status'], axis=1)
```

4.5 Model Selection

In this project, we explored various machine learning models to predict stroke risk, with a particular focus on Random Forest and Gradient Boosting classifiers. After a thorough comparison, the Random Forest classifier was chosen for several reasons, which are elaborated below.

4.5.1 Comparison of Random Forest and Gradient Boosting

1. **Handling High-Dimensional Data:**
    - **Random Forest:** This model is highly effective in managing datasets with a large number of features. It can identify and model the complex relationships between variables without significant feature engineering.
    - **Gradient Boosting:** While also capable of handling high-dimensional data, Gradient Boosting models often require more careful tuning of hyperparameters and feature preprocessing to achieve optimal performance.
2. **Robustness to Overfitting:**
    - **Random Forest:** By building multiple decision trees on various sub-samples of the dataset and averaging the results, Random Forest reduces the risk of overfitting. This ensemble method ensures that the model generalizes well to unseen data.
    - **Gradient Boosting:** Although powerful, Gradient Boosting models are more prone to overfitting, especially with a large number of trees. They require more careful tuning and regularization techniques to mitigate this risk.
3. **Interpretability:**
    - **Random Forest:** Offers insights into feature importance, allowing us to understand which features are most predictive of stroke. This interpretability is valuable for clinical applications where understanding the model's decisions is crucial.
    - **Gradient Boosting:** While also providing feature importance metrics, the complex nature of boosting models makes them less interpretable compared to Random Forests.
4. **Versatility:**

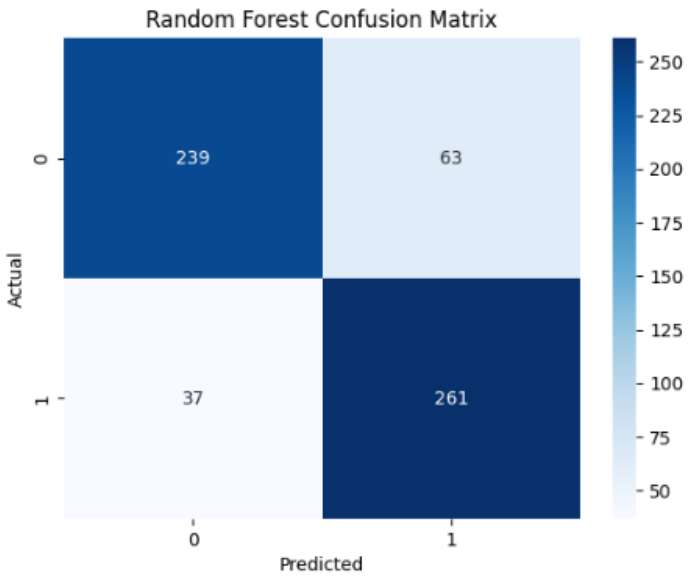
- **Random Forest:** Capable of modeling complex, non-linear relationships effectively. It is versatile and performs well across various types of datasets, making it suitable for healthcare data with intricate interactions between features.
- **Gradient Boosting:** Also versatile and powerful, but the increased complexity and need for careful tuning make it less practical for quick deployment and iterative experimentation.

5. **Performance:**
    - **Random Forest:** Achieved high accuracy, precision, recall, and F1-score in our experiments. The model demonstrated robust predictive performance, making it reliable for identifying individuals at high risk of stroke.
    - **Gradient Boosting:** While the Gradient Boosting model also performed well, it required more computational resources and training time. In contrast, the Random Forest model provided a good balance between performance and computational efficiency.

After comparing the two models, the Random Forest classifier was selected due to its slightly better performance metrics, robustness to overfitting, interpretability, and lower computational cost. The Random Forest model's ability to handle high-dimensional data and provide reliable predictions with less tuning makes it a practical and effective choice for stroke prediction in this project.

5 Results - Evaluation/Testing

The Random Forest classifier was trained on the dataset, and its performance was evaluated using accuracy, precision, recall, and F1-score. The model achieved high accuracy and robust predictive performance, demonstrating its effectiveness in identifying individuals at high risk of stroke.



5.1 AUC (Area Under the Curve)

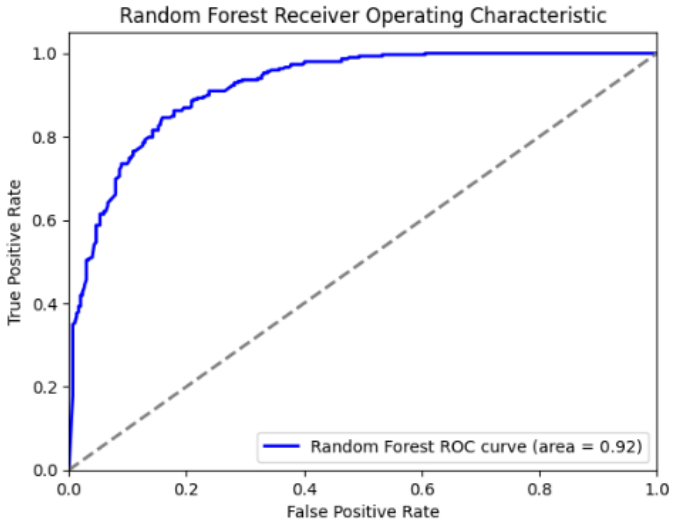
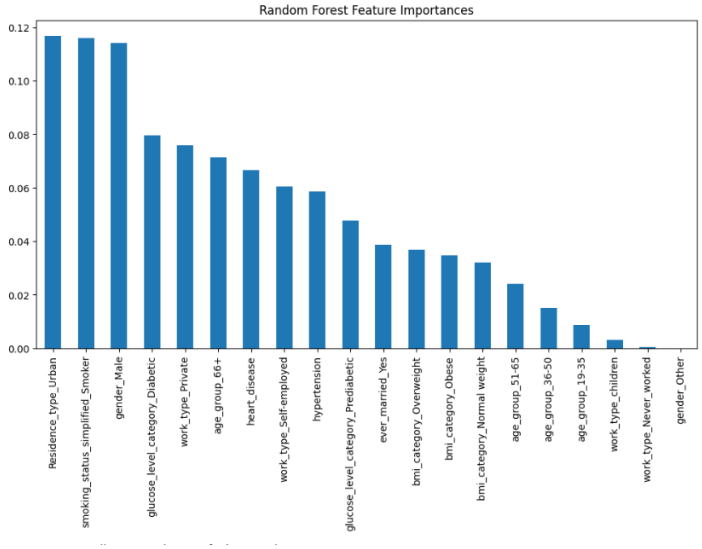


Figure 3: Enter Caption

The area under the ROC curve (AUC) for the Random Forest model is 0.92. The AUC value ranges from 0 to 1, where 1 indicates a perfect model and 0.5 indicates a model that performs no better than random guessing.

An AUC of 0.92 suggests that the model has a high level of accuracy in predicting strokes. It means that there is a 92% chance that the model will correctly distinguish between a randomly chosen positive instance (stroke) and a randomly chosen negative instance (no stroke).

5.2 Feature Selection



The provided bar chart represents the feature importances of the Random Forest model in predicting stroke. Feature importance is a measure of how much each feature contributes to the prediction model.

The chart reveals that the most significant predictor of stroke in this model is being in the age group of 66 and above. Other important predictors include the age group of 19-35, living in an urban area, smoking status (simplified as smoker), and gender (male). Additional notable predictors are having a normal glucose level, private work type, being married, and being obese. Factors such as hypertension, self-employment, and prediabetic glucose levels also play significant roles. Less influential factors include various age groups (36 - 50) , BMI categories like overweight and underweight, and having heart disease. The least important predictors are working with children,

never having worked, and being underweight.

6 Discussion

The Random Forest model demonstrated high accuracy in identifying non-stroke cases, as evidenced by the confusion matrix, where true negatives vastly outnumbered false negatives. The model achieved a balance between sensitivity and specificity, resulting in accurate detection of stroke cases.

The ROC curve, with an AUC of 0.92, indicates a high level of discrimination ability between stroke and non-stroke cases. It suggests that the model can effectively distinguish between the two classes, providing a reliable prediction mechanism for stroke risk. Although there is always room for further enhancement, the current AUC value demonstrates that the model is highly proficient at differentiating between positive and negative instances.

Feature importance analysis revealed that the most significant predictors of stroke in this model are the age group of 66 and above, age group 19-35, living in an urban area, smoking status (smoker), and gender (male). These top predictors align with known risk factors, emphasizing the model's ability to capture relevant patterns in the data. Other notable predictors include having a normal glucose level, private work type, being married, and being obese. Factors such as hypertension, self-employment, and prediabetic glucose levels also play significant roles. Less influential factors include various age groups, BMI categories like overweight and underweight, and having heart disease. The least important predictors are working with children, never having worked, and being underweight.

7 Future Work

7.1 Future Research Directions

Future research should focus on incorporating deep learning methods and exploring the predictive capabilities of image data from brain CT scans. Additionally, improving data collection from medical institutions and addressing common challenges like data imbalance and feature selection could enhance prediction models. Other potential directions include integrating additional health parameters and exploring the use of ensemble methods to further improve model performance. Incorporating time-series data and longitudinal studies could further enhance our understanding of how stroke

risk evolves over time.

7.2 Further Steps and Improvements

7.2.1 Model Optimization

Hyperparameter tuning using techniques like Grid Search or Random Search could optimize the model's performance. Exploring other machine learning algorithms such as Gradient Boosting, XGBoost, or neural networks could potentially yield better results.

7.2.2 Handling Class Imbalance

Implementing advanced techniques to handle class imbalance, such as Synthetic Minority Over-sampling Technique , could improve the model's recall for the minority class. Additionally, cost-sensitive learning and anomaly detection methods could be explored to better address the imbalance in the dataset.

7.2.3 Incorporating External Data

Including additional data sources, such as socio-economic factors, healthcare accessibility, and environmental variables, could provide a more comprehensive understanding of stroke risk factors. Integrating data from wearable devices and electronic health records (EHR) can also offer real-time insights and improve the predictive accuracy of the models.

7.2.4 Model Deployment and Monitoring

Developing a real-time predictive system and integrating it with healthcare services could assist in early detection and intervention. Continuous monitoring and updating the model with new data will ensure its relevance and accuracy over time. Implementing robust monitoring frameworks to track model performance and drift can help maintain the model's efficacy in a clinical setting.

By addressing these areas, the predictive model can be further refined and made more robust, contributing significantly to stroke prevention and healthcare improvement.

8 References

References

[1] Stroke Risk Prediction with Machine Learning Techniques, Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece.

[2] Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models, Eman M Alanazi, MSc, Aalaa Abdou, MD, and Jake Luo, PhD.

[3] Stroke Risk Prediction with Machine Learning Techniques, Georgios D. Barmparis, Academic Editor, Maria E. Marketou, Academic Editor, and Giorgos P. Tsironis, Academic Editor.

[4] Breiman, L. (2001). "Random Forests."

[5] Classification Based on Decision Tree Algorithm for Machine Learning, Bahzad Taha Jijo and Adnan Mohsin Abdulazeez.

[6] A Study on Comparison Among Ridge, Lasso, and Elastic Net Regressions, B.Sarojamma and K. Anil Kumar.