Stroke Risk Factor Analysis

Team 4

Amani Chaar

Hala Chtay

Gunjan Joshi

Udbhav Srivastava

ISTM 660:601

TABLE OF CONTENTS

1 Executive Summary

This study leverages predictive analytics to identify stroke risk factors and develop models for early risk detection, aiming to mitigate the risk of stroke occurrences before it is too late. Through advanced analytical techniques, including machine learning algorithms, the study analyzes the relationships between various risk factors and their impact on stroke incidence. The analysis begins with an examination of data variables, visualizing key predictors like BMI and average glucose levels to gain insights into their distribution and potential correlations. Exploratory data analysis uncovers patterns related to stroke risk factors, laying the foundation for predictive modeling. The evaluation of predictive models highlights the strengths and limitations of each approach, focusing on metrics such as accuracy, sensitivity, and specificity. While Logistic Regression demonstrates balanced performance and interpretability, other models like K-Nearest Neighbors (KNN), Decision Trees, Random Forests, and Neural Networks exhibit varying degrees of accuracy and efficacy in predicting stroke risk. Despite challenges posed by data imbalances and model performance, the study recommends integrating the Logistic Regression model into clinical practice due to its well-rounded performance. Continuous data collection and model refinement are critical for improving predictive accuracy and informing targeted interventions to reduce the burden of stroke-related morbidity and mortality.

2 Introduction

Strokes are the second leading cause of death globally and the third leading cause of disability. These traumatic incidents pose a significant threat to public health and their frequency is on the rise, with one in four individuals facing the risk of experiencing it during their lifetime. This emphasizes the urgency of establishing more effective detection algorithms and preventive measures for early intervention.

2.1 Objectives of the Study

The study aims to leverage predictive analytics to effectively identify and assess risk factors associated with stroke while exploring the interconnection between factors. This involves the development of strong predictive models to offer patients insights for early risk detection, prompting proactive intervention to reduce the risk of stroke occurrences. By employing advanced analytical techniques to analyze the relationships among these risk factors and their impact on stroke incidence, the study will validate the performance of these predictive models using a portion

of the dataset held out for testing. The findings will then be translated into practical recommendations to be implemented in healthcare to drive targeted preventive strategies and interventions.

3 Overview of Data Variables

3.1 Variable Statistical Summary

Our dataset is comprised of 10 variables identified as potential stroke risk factors **(Figure 1)**. The dataset provides a comprehensive overview of various demographic and health-related variables. Gender distribution indicates 2994 females and 2115 males. Age spans from 0.08 to 82.00 years, with a median of 45.00 and a mean of 43.23 years. Hypertension and Heart Disease are binary variables, with 1.000 denoting the presence and 2.00 denoting the absence of these conditions. In regards to marital status, 1756 individuals are not married and 3353 are married. Work Type categorizes individuals as Children, Govt_job, Never_worked, Private, or Self-employed. Residence type shows 2513 individuals from rural areas and 2596 from urban areas. Avg Glucose Level ranges from 55.12 to 271.74, with a median of 91.88 and a mean of 106.14. BMI spans from 10.30 to 97.60, with a median of 28.10 and a mean of 28.86. Smoking Status classifies individuals as Formerly smoked, Never smoked, Smokes, and Unknown. Their interplay may amplify or mitigate stroke risk, providing us with actionable insights for prevention strategies.

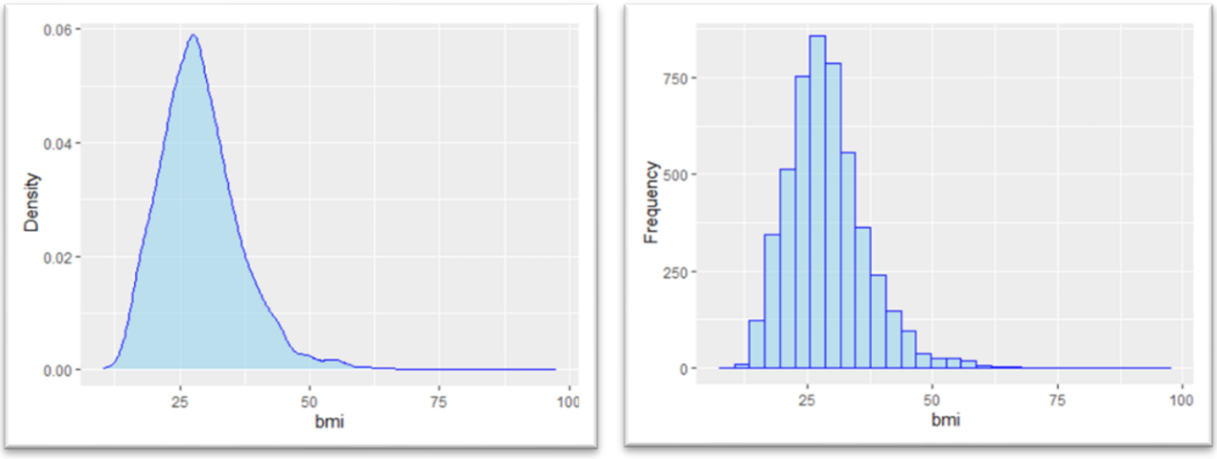
FIGURE 1: Variable Summary Statistics Table

Variable	Summary
Gender	Female: 2994, Male: 2115
Age	Min: 0.08, 1st Qu.: 25.00, Median: 45.00, Mean: 43.23, 3rd Qu.: 61.00, Max: 82.00
Hypertension	Min: 1.000, 1st Qu.: 1.000, Median: 1.000, Mean: 1.097, 3rd Qu.: 1.000, Max: 2.000
Heart Disease	Min: 1.000, 1st Qu.: 1.000, Median: 1.000, Mean: 1.054, 3rd Qu.: 1.000, Max: 2.000
Ever Married	No: 1756, Yes: 3353
Work Type	Children: 687, Govt_job: 657, Never_worked: 22, Private: 2924, Self-employed: 819
Residence Type	Rural: 2513, Urban: 2596
Avg Glucose Level	Min: 55.12, 1st Qu.: 77.24, Median: 91.88, Mean: 106.14, 3rd Qu.: 114.09, Max: 271.74
BMI	Min: 10.30, 1st Qu.: 23.80, Median: 28.10, Mean: 28.86, 3rd Qu.: 32.80, Max: 97.60
Smoking Status	Formerly smoked: 884, Never smoked: 1892, Smokes: 789, Unknown: 1544

3.2 Visualization of Data

The analysis begins with thoroughly examining Body Mass Index (BMI), a key health indicator.

FIGURE 2: Density Plot and Histogram of BMI



The density plot and histogram of BMI (**Figure 2**) offer a visual representation of its distribution across the dataset. Our findings indicate a right-skewed distribution, signaling a higher frequency of values on the lower end of the BMI scale with some outliers having extremely high values.

We further explore average glucose levels, another predictor, through a density plot (**Figure 3**). The plot reveals a notable peak within the normal range and a long tail indicating higher glucose levels in certain individuals. Such variations could signify the presence of undiagnosed prediabetes or diabetes, conditions that heighten stroke risk.

The inter-variable relationship is dissected using a scatter plot (**Figure 4**), which compares BMI against average glucose levels. This analysis uncovers potential patterns or clusters, like whether individuals with higher BMI also tend to have higher glucose levels, which could compound their stroke risk. In the observations with BMIs greater than 25, the upper end of the normal range, the data pool has greater cases of elevated average glucose levels.

FIGURE 3: Density Plot of Average Glucose Level

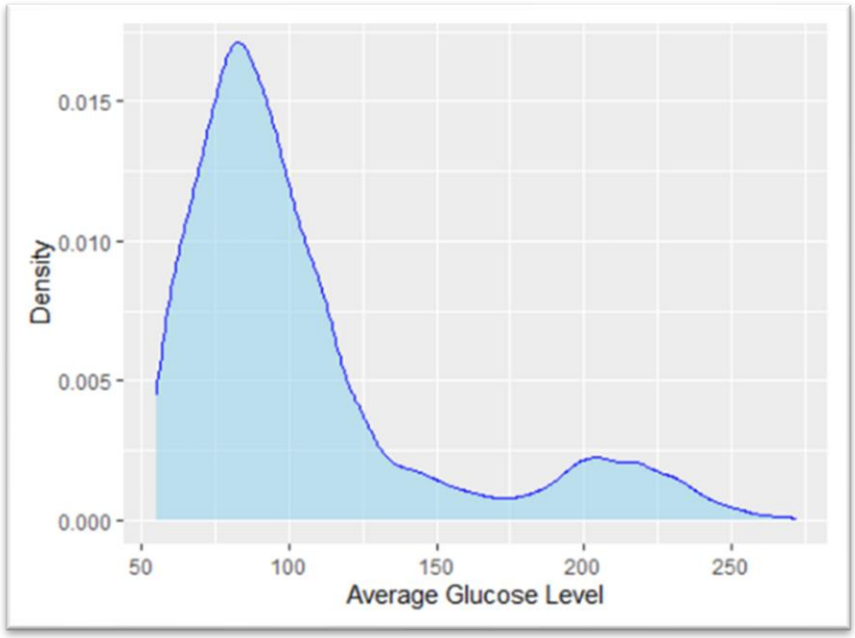
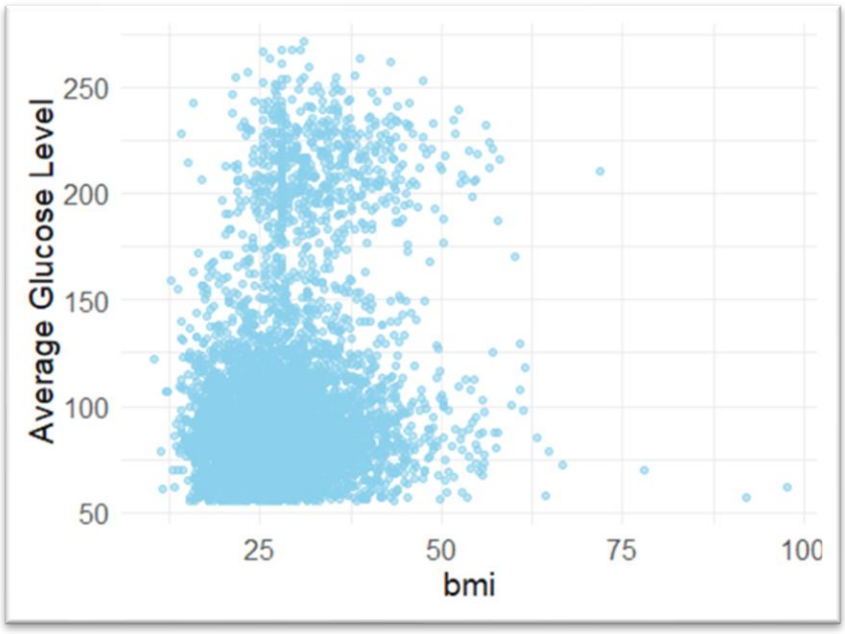


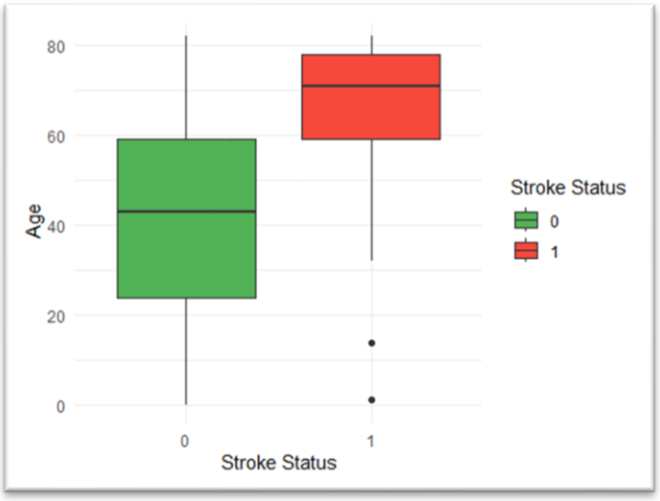
FIGURE 4: Scatter Plot of BMI vs Average Glucose Levels



Age, as depicted in the boxplot (**Figure 5**), shows its distribution by stroke status, highlighting a stark contrast in median age between those who have had a stroke (red) and those who have not (green). The median age of those who have had a stroke is significantly higher than those who

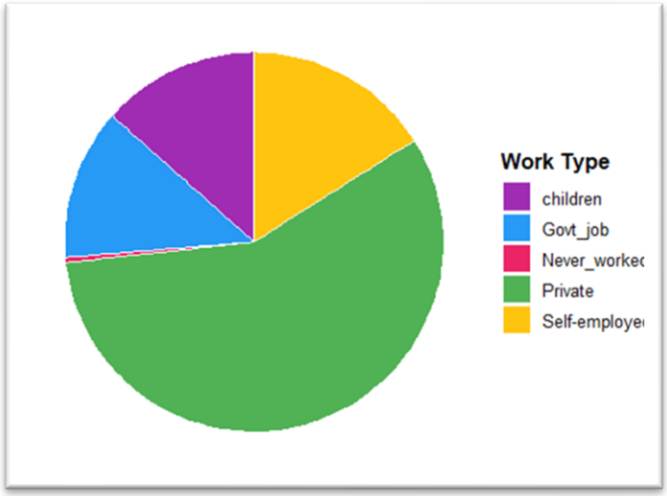
have not. This visual breakdown underscores age as a predominant risk factor and could guide age-specific prevention measures.

FIGURE 5: Boxplot of Age by Stroke Status



The distribution of work types (**Figure 6**) contextualizes the sociodemographic landscape of our data. Understanding how occupation influences lifestyle and health behaviors can illuminate indirect stroke risk factors, allowing for comprehensive risk profiling.

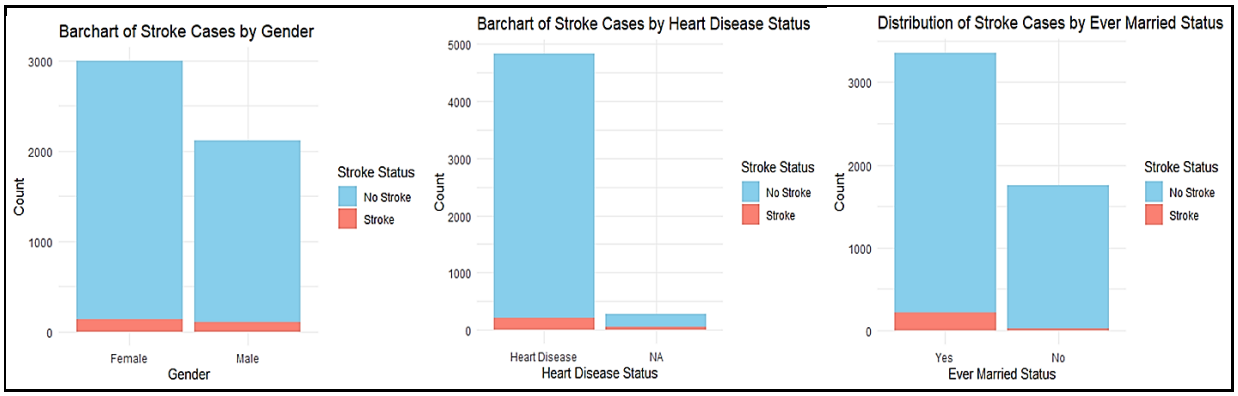
FIGURE 6: Distribution of Work Types



Our dataset analysis extends to investigating stroke incidents across key categorical variables, with a consolidated visual representation provided in **Figure 7**. This figure comprises side-by-side bar charts detailing stroke distribution by gender, heart disease status, and marital status. The gender analysis reveals a slightly higher frequency of stroke in females, which could be due

to the skewed nature of the dataset. Proportion-wise, a lower ratio of females had a stroke compared to men. While the heart disease chart affirms its strong correlation with stroke events, the data appears to be highly imbalanced as a majority of the data points have heart disease. Additionally, marital status offers an intriguing dimension, with lower stroke cases among the unmarried cohort. This could also be due to the unbalance within the data. These visual insights collectively underscore the diversity of stroke risk across demographic and health-related factors.

FIGURE 7: Bar charts Exploring Stroke Incidents Across Categorical Variables



By visualizing each variable in the dataset, we set a foundation for developing predictive models that are both accurate and reflective of real-world complexities.

3.3 Exploratory Data Analysis

Exploratory data analysis (EDA) was critical for understanding the dataset and identifying key patterns related to stroke risk factors. By exploring statistical summaries and visualizations, we gained valuable insights into the distribution of variables and potential correlations among them. Descriptive statistics helped us pinpoint trends and outliers, while graphical tools like bar charts and scatter plots provided intuitive representations of the relationships between various variables. For example, density plots highlighting BMI and average glucose levels, coupled with correlating scatter plots, offered invaluable insights into prevalent health indicators present in our dataset. Bar charts showcasing stroke incidents across demographic factors such as gender and heart disease status allowed us to recognize variations in stroke rates among different groups.

The EDA was pivotal for selecting appropriate modeling techniques and data processing steps. It indicated the need for normalization of skewed distributions and informed strategies such as stratified sampling to address observed patterns in stroke prevalence.

4 Real-World Application

4.1 Previous Applications

The evolution of machine learning in stroke risk analysis has advanced from focusing on limited risk factors like hypertension and smoking to embracing a wider array of predictors. Initially, basic statistical methods' models provided essential insights into how these well-known factors contributed to stroke risk. This foundational work was critical for the development of early risk assessment tools. As machine learning has evolved, so has its capacity to incorporate a broader spectrum of risk factors, offering a more intricate analysis and enabling more precision and accuracy.

4.2 New Application

Our study enhances stroke risk analysis by incorporating Neural Networks, an advanced machine learning technique that transcends the capabilities of traditional models. Neural Networks analyze complex data relationships, enabling a deep dive into genetic and lifestyle factors contributing to stroke risk. Together with the Synthetic Minority Over-sampling Technique (SMOTE), which corrects data imbalances, this approach offers a nuanced understanding of stroke predictors.

Addressing class imbalance in our dataset was imperative due to the highly skewed distribution of class labels. With 4860 instances classified as "no stroke" and only 249 instances labeled as "stroke," traditional machine learning algorithms could be biased towards the majority class, leading to suboptimal performance. To rectify this issue, we employed the Synthetic Minority Oversampling Technique (SMOTE), a widely used oversampling technique. SMOTE works by creating synthetic samples for the minority class, thereby balancing the class distribution and improving model generalization. By generating new instances based on the characteristics of existing minority class samples, SMOTE expands the representation of the minority class, making it more comparable in size to the majority class. This approach enhances the influence of the minority class during model training, leading to more robust predictions. In our implementation, we utilized the ovun.sample() function from the ROSE package in R to apply SMOTE to our dataset. This method effectively increased the number of minority class instances, resulting in a more balanced dataset suitable for training machine learning models.

Through this strategic use of SMOTE, we addressed the challenge of class imbalance and laid the foundation for more accurate and reliable predictions of stroke risk.

5 Data Pre-Processing

Our analysis began with a data pre-processing routine, essential for ensuring data quality and suitability for predictive modeling. The dataset, sourced from Kaggle, represents a compilation of 10 variables representing stroke risk factors from a wide demographic range, over a total of 5,109 observation points. The subsequent pre-processing steps were executed to enhance the dataset's analytical value.

5.1 Data Cleansing

Data cleansing was our initial step, involving the removal of irrelevant columns such as identifiers that offer no predictive power. We standardized categorical variables into a machine-learning-friendly format, encoding them as factors and validating that numerical variables were properly typed. Special attention was given to the 'gender' variable to rectify missing categories and to 'bmi', which presented with 201 missing entries. Given the right-skewed distribution of 'bmi', we employed median imputation to counteract the skewing effect of outliers.

5.2 Data Transformation

Data transformation is a vital step to ensure all variables have equal influence in our analysis. In our dataset, min-max scaling was used to normalize feature values to a [0, 1] range, leveling the playing field for variables of different scales. This method adjusts for scale without losing the range of values, helping to prevent models from favoring one feature over another due to size alone. By applying this technique, each variable can contribute more equally to the model's predictions, leading to more balanced and accurate results.

5.3 Data Balancing

The notable class imbalance in our dataset, evidenced by a disproportionate number of 'no stroke' instances, 4860 occurrences, compared to 'stroke' cases, 249 occurrences, necessitated the application of the SMOTE. This technique synthetically enhances the minority class, providing a balanced approach to training our predictive models. Utilizing the ovun.sample() function from the R's ROSE package allowed us to boost the underrepresented stroke cases, thus rectifying the initial bias and fostering a more equitable and effective model training environment. These pre-

processing stages were crucial in sculpting a dataset that was not only clean and normalized but also balanced, establishing a solid foundation for the deep and accurate predictive analysis that followed.

6 Predictive Modeling and Evaluation

Our models utilized the same training and testing subsets of data, with 3,832 observations, about 75% of the data, allocated to training and 1,278 observations, about 25%, left for the testing split. The decision to utilize a 75/25 training/testing split was informed by several factors, including the need to balance model complexity and generalization ability, while ensuring a sufficiently large dataset for training without compromising the robustness of our validation process. With 75% of the dataset allocated to training, we have a substantial pool of data for the model to learn from and build strong predictive capabilities. By allocating 25% of the data to the testing split, it allows for a sizable portion of data reserved for evaluating the model's performance. This split strikes a balance between providing enough data for effective model training and retaining enough for reliable performance evaluation. The imbalance in the dataset due to the scarcity of positive cases indicating stroke occurrence posed challenges; however, allocating a significant portion of data to the testing set helps mitigate potential skewness in performance metrics caused by this imbalance.

While models demonstrate promise for stroke risk prediction, the insufficient number of positive instances stresses the need for a cautious interpretation of their performance. In cases such as the one within our dataset, where the classes are imbalanced, accuracy alone can be misleading because a model can achieve high accuracy by simply predicting the majority class for all instances. This is detrimental because missing positive cases, known as false negatives, can have severe consequences - potentially leading to death or disability over the long run. While accuracy provides a general measure of a model's correctness, our recommended model will be selected by also considering other metrics such as the confusion matrix, sensitivity, and specificity.

6.1 Logistical Regression

Logistic Regression served as a benchmark in our study, providing foundational insights into stroke risk factors. This method allowed us to quantify how specific variables influence stroke risk. Logistic Regression, implemented within the Generalized Linear Model (GLM) framework using a binomial family, achieved an overall accuracy of approximately 71.4%. By applying a

classification threshold of 0.5, where probabilities greater than 0.5 are classified as stroke occurrences and those less than or equal to 0.5 are classified as non-occurrences, we converted these probabilities into predictions. This thresholding mechanism enabled the interpretation of model outputs, contributing to our understanding of stroke risk factors.

FIGURE 8: Logistical Regression Formula

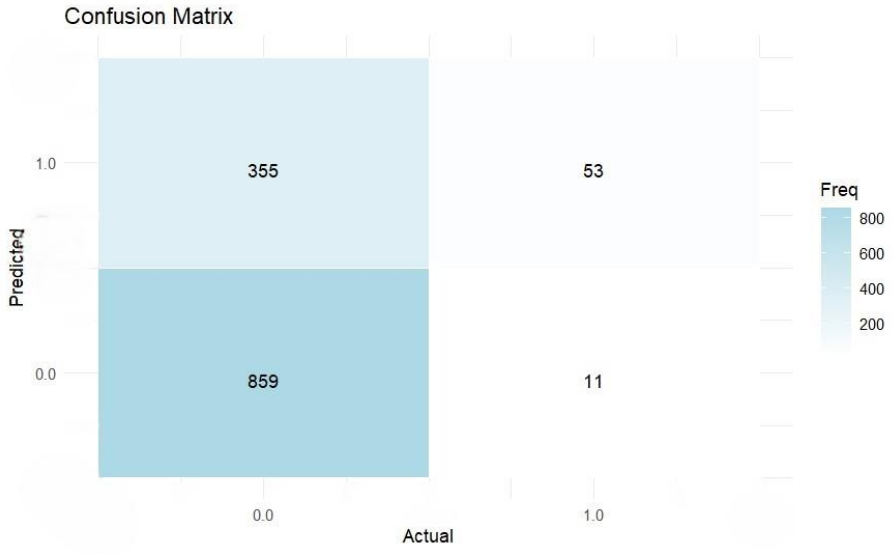
$$
\begin{aligned}
logit(p) = \ & log\left(\frac{p}{1-p}\right) \\
& = 0.94711 + 0.03677 \times Gender\ (Male) + 1.66915 \times Age \\
& + 0.19522 \times Hypertension + 0.05623 \times Heart\ Disease \\
& - 0.03082 \times Ever\ Married\ (Yes) - 0.94111 \times Work\ Type\ (Govt\ Job) \\
& - 0.87900 \times Work\ Type\ (Private) \\
& - 12.01532 \times Work\ Type\ (Never\ Worked) \\
& - 1.07536 \times Work\ Type\ (Self\ Employed) \\
& + 0.19942 \times Average\ Glucose\ Level + 0.12927 \times BMI \\
& - 0.48338 \times Smoking\ Status\ (Never\ Smoked) \\
& - 0.18427 \times Smoking\ Status\ (Unknown)
\end{aligned}
$$

This Logistic Regression formula **(Figure 8)** aims to predict the likelihood of an individual having a stroke based on several factors. The intercept (0.94711) represents the baseline probability of stroke when all other variables are zero. Gender, with a positive coefficient of 0.03677, indicates how being male slightly increases stroke risk compared to being female. Age, hypertension, and heart disease significantly increase stroke probability, with positive coefficients of 1.66915, 0.19522, and 0.05623, respectively. Marital status, with a negative coefficient of 0.03082 indicates how being married lowers stroke risk compared to never married. The type of work also plays a role, government, private, self-employed, and never worked all decrease the probability of stroke (coefficients: -0.94111, -0.87900, -1.07536, -12.01532) when compared to children. For the employment category, it is important to note that children are inherently different from adults in terms of physiological, developmental, and social factors that influence health. Using them as a baseline means that the health outcomes of all job categories, including "never worked," are assessed relative to a typically healthier, younger population. Average glucose level (0.19942) and BMI (0.12927) increase the probability of stroke risk. Smoking status, with negative coefficients of -0.48338 (never smoked) and -0.18427 (unknown status), also decreases the probability.

Before predicting the probability of a stroke using these coefficients, they must be converted into odds ratios. For each predictor variable, we exponentiated the corresponding coefficient obtained from the Logistic Regression output. This exponentiated coefficient represents how the odds of the event change for a one-unit increase in the predictor variable, holding all other variables constant. After obtaining the odds ratios, we can use them to predict the probability of the event for specific cases by plugging in the values of the predictor variables into the Logistic Regression equation.

The confusion matrix **(Figure 9)** presents a detailed breakdown of the performance of this model. The two classes represent the presence (1.0) or absence (0.0) of a stroke. The high number of true negatives, 859, indicates the model's proficiency in correctly identifying patients without strokes. However, the occurrence of false negatives, which stands at 11 instances, compared to false positives, which total 355 instances, suggests room for improvement in accurately detecting positive stroke cases. False negatives, where the model fails to identify actual strokes, are particularly concerning in medical diagnosis as they may lead to missed opportunities for timely intervention and treatment.

FIGURE 9: Logistical Regression (GLM) Confusion Matrix

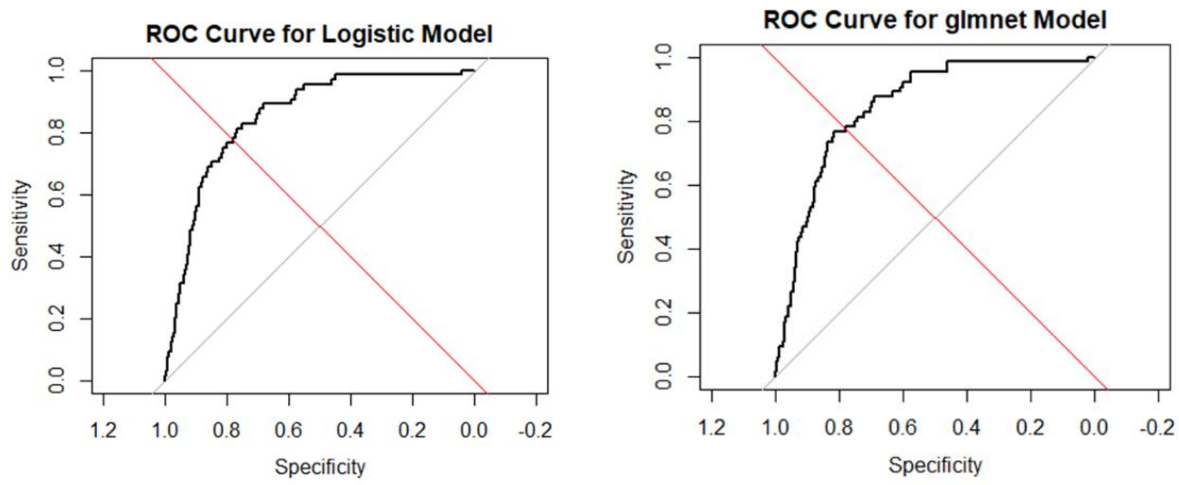


Looking beyond accuracy, sensitivity tells us how good the model is at identifying positive cases, while specificity tells us how good the model is at identifying negative cases. The sensitivity for this model is ~83%, while the specificity is ~70%, striking a good balance between the two. The classification threshold of 0.5 helps achieve a reasonable trade-off between

sensitivity and specificity, resulting in a model that performs well in identifying both positive and negative cases without skewing too heavily towards one aspect at the expense of the other.

To address the limitations of Logistic Regression, such as overfitting and poor performance with multicollinearity, we explored regularized Logistic Regression using glmnet. This approach combined both L1, known as Lasso regression, and L2, known as Ridge regression, penalties to create an elastic net model. By introducing a little bit of bias, we were able to achieve a large reduction in variance. Through 10-fold cross-validation, optimal regularization parameters were determined, further improving prediction accuracy. The elastic net approach mitigates multicollinearity effects by simultaneously considering both L1 and L2 penalties, providing more accurate predictions. Despite these enhancements, the difference in performance between Logistic Regression and glmnet was negligible. ROC analysis **(Figure 10)** demonstrated comparable performance for both models, indicating that regularization did not significantly improve predictive capability.

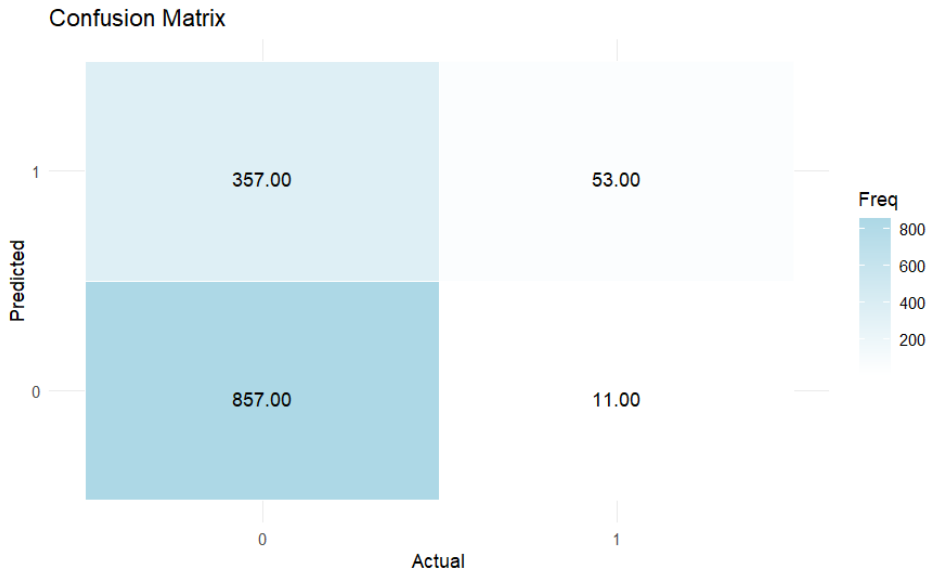
FIGURE 10: Logistical Regression ROC Curves



The glmnet model's confusion matrix **(Figure 11)** for stroke prediction reveals both strengths and areas for improvement. On the positive side, the high number of true negatives, 857, showcases the model's proficiency in accurately identifying patients without strokes. However, this is lower than the Logistic Regression model’s true negative rate of 859. The persistent issue of false negatives, 11, remains a concern as these instances represent actual stroke cases that the model failed to detect, potentially leading to missed opportunities for timely intervention and

treatment. The true positive count, 53, remained persistent signaling the same level of detection of positive stroke cases. The accuracy of glmnet was measured at 71.24%, and the sensitivity for this model is ~83%, while the specificity is ~70%, reflecting a similar performance to Logistic Regression. Sensitivity is significant as minimizing false negatives is crucial because it ensures that individuals who are truly at risk are correctly identified.

FIGURE 11: Logistical Regression (glmnet) Confusion Matrix



By integrating Logistic Regression into our predictive analysis, we gained crucial insights into stroke risk factors. Critical predictors identified through this model included age and average glucose levels, factors that significantly influence stroke risk. This finding aligns with current medical understanding and highlights the importance of these variables in preventive health screenings. Although Logistic Regression provided moderate accuracy, its strength lies in its interpretability. The model offered clear insights into the relative importance of each predictor, making it an invaluable tool for understanding complex health data. This clarity is especially beneficial in a clinical context, where understanding the weight of different risk factors aids in risk assessment and patient counseling.

6.2 K-Nearest Neighbors (KNN)

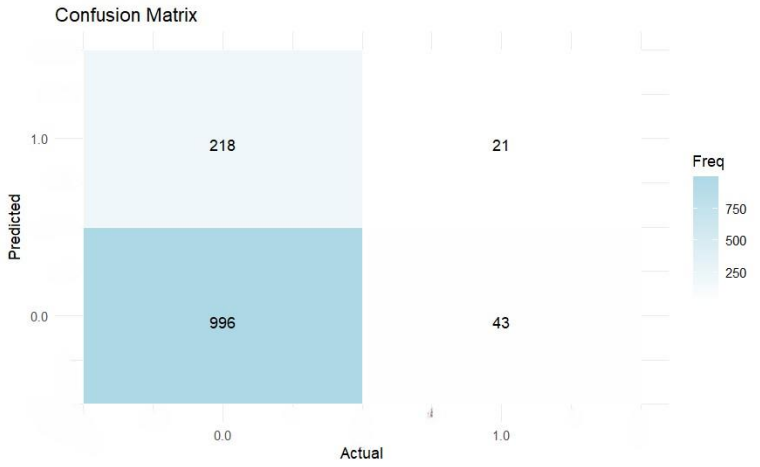
The exact nature of the relationship between predictors and response is uncertain, and it may exhibit either linear or non-linear characteristics. This uncertainty led to the exploration of KNN as a viable option. KNN makes no assumptions about the underlying data distribution, making it

a non-parametric method. This characteristic renders KNN more flexible in handling data of various types, encompassing both quantitative and qualitative variables.

In the analysis using the KNN model, the algorithm was strategically optimized to enhance its predictive accuracy for stroke risk. It is important to note that KNN is a lazy learner, meaning it does not learn a discriminative function from the training data but rather memorizes the dataset. When a prediction is required, KNN searches through the entire training dataset to find the k nearest training examples (neighbors), and the output is determined based on these nearest neighbors. The computation for the model is deferred until prediction time, hence the label "lazy." This optimization involved tuning the number of neighbors denoted as “k”, and after thorough testing, setting k=5 provided the best balance between sensitivity and specificity. Data standardization was a vital step in this process, ensuring that each variable contributed equally to the model's output. This adjustment was essential to prevent any one feature from disproportionately influencing the results, which could skew the model’s predictive accuracy.

The KNN model achieved an accuracy rate of approximately 79.5%, comparable to Logistic Regression's performance. This accuracy, derived from the confusion matrix **(Figure 12)**, signifies the model's ability to correctly classify instances into their respective categories. Among the 1278 instances evaluated, the KNN model demonstrated great efficacy in predicting stroke risk, shown by 21 true positives and 996 true negatives. True negatives indicate the model's proficiency in identifying individuals without strokes, mirroring Logistic Regression's capability in this aspect, however, the presence of 43 false negatives poses a concern.

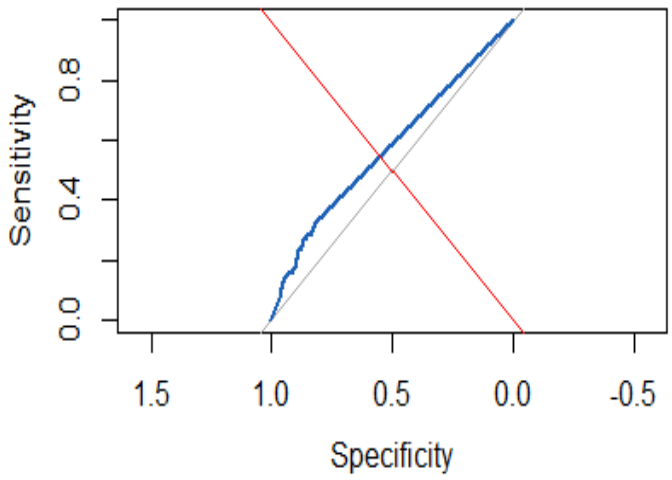
FIGURE 12: KNN Confusion Matrix



In this KNN model, the sensitivity, ~5.6%, reflects the model's ability to correctly identify individuals who are truly at risk of stroke. This low sensitivity indicates that the model misses a significant portion of true positive cases, suggesting limitations in its effectiveness at detecting those at risk. On the other hand, the specificity, ~99%, signifies the model's ability to correctly identify individuals who are not at risk of stroke. With high specificity, KNN demonstrates its capability to accurately rule out individuals who are truly negative for stroke. Despite the high specificity, the low sensitivity of the KNN model raises concerns about its reliability in identifying individuals at risk of stroke. The model's tendency to miss true positive cases could have serious consequences, as it may lead to missed opportunities for timely intervention and treatment.

While assessing the performance of the KNN algorithm through the ROC analysis, the model demonstrated limited discriminatory power. The ROC curve **(Figure 13)** plotted for the KNN model closely approached the line of no-discrimination, indicative of a random-guessing classifier, with an area under the curve (AUC) close to 0.5. This suggests that the KNN model struggled to effectively differentiate between patients at risk of stroke and those not at risk. Although the model achieved an accuracy of approximately 79.6%, as reflected by the ROC curve, its ability to correctly identify true positives against false positives was not significantly better than chance.

FIGURE 13: ROC Curve for KNN



The KNN model proved particularly insightful for understanding the proximity and similarity of risk profiles among individuals. These insights imply that while KNN models offer simplicity and ease of interpretation, they may not suffice when dealing with complex datasets where stroke indicators are subtle and non-linear.

6.3 Decision Trees

Decision Trees were essential to our predictive modeling approach due to their straightforward structure and ease of interpretation. The simplicity of Decision Trees makes them particularly appealing for clinical applications, where clear and interpretable models are essential for quick decision-making. They allow clinicians to easily trace the reasoning behind each prediction, making these models not only useful for prediction but also for educational purposes within a clinical setting.

Unlike Logistic Regression, which models the relationship between predictors and the target variable linearly, decision trees create a hierarchical structure of if-else conditions based on the input features to make predictions. The decision tree algorithm uses a splitting criterion (entropy) to determine the best feature and threshold for splitting the data at each node, minimizing entropy when creating a split. Decision trees are advantageous in capturing non-linear relationships between predictors and the target variable, as they can model complex decision boundaries, making them suitable for datasets with non-linear patterns. To avoid overfitting, we employed tree pruning techniques, simplifying the decision tree by removing branches that do not significantly improve predictive accuracy, resulting in a more generalized model that performs well on unseen data.

We fine-tuned the depth of the trees to optimize model performance and prevent overfitting, a common challenge with this model type. Our analysis determined that a tree depth of 5 was most effective. The confusion matrix **(Figure 14)** offers a granular analysis of model performance. Decision trees demonstrated proficiency in identifying individuals without strokes, with 661 true negatives. However, the occurrence of 553 false positives and 5 false negatives suggests room for enhancement in accurately predicting stroke risk. Our best model achieved an accuracy rate of approximately 58.6% and exhibited limitations compared to Logistic Regression and KNN in accurately classifying stroke cases.

FIGURE 14: Decision Tree Confusion Matrix

The model has a sensitivity of ~92% highlighting the model's ability to correctly identify individuals at risk of stroke, as true positives. Conversely, with a specificity of ~54%, the model demonstrates its capability to accurately classify individuals without stroke. The lower specificity suggests a higher tendency for misclassification of individuals without stroke as positive cases. Despite the promising sensitivity in detecting stroke risk, the relatively lower specificity raises concerns about the model's ability to accurately rule out individuals without stroke.

6.4 Random Forest

During our model development, the Random Forest algorithm stood out for its exceptional performance. This method, which utilizes an ensemble of decision trees, is recognized for its high stability and accuracy. Random Forest employs bootstrap sampling to create multiple subsets of the training data. Each tree in the forest is trained on a different bootstrap sample, involving randomly sampled instances with replacements from the original dataset. This introduces randomness in feature selection for each tree, where only a random subset of features is considered for splitting at each split in a tree, thereby reducing correlation among trees and improving generalization.

The strength of the Random Forest model lies in its ability to aggregate outputs from multiple trees, enhancing the overall prediction accuracy while significantly reducing the risk of overfitting. During prediction, the Random Forest aggregates the predictions from all individual trees. This is achieved by averaging the results of various trees, each constructed with slight

variations in the data and features used, thereby ensuring a robust model that is less prone to errors specific to any single tree.

The calibration of the model involved careful adjustments to the number and depth of trees to strike an optimal balance between complexity and performance. This careful tuning not only maximized efficiency but also ensured the model's reliability across diverse patient profiles. Given its sophisticated combination approach, the Random Forest model emerged as the top-performing algorithm in our study. Its ability to decrease variance and bias, common issues in models based on single decision trees, makes it particularly valuable for clinical applications where accurate and consistent predictions are crucial. By integrating the predictions from 100 decision trees, we achieved the highest accuracy rate of 94.0%, however further considerations are needed before we rule this out as best.

Based on the confusion matrix **(Figure 15)**, our model's performance metrics indicate that among the instances analyzed, 1202 were true negatives, showcasing the model's ability to accurately predict the absence of strokes. On the other hand, the model misclassified 12 instances as false positives, inaccurately identifying strokes that were not present. The model correctly identified 3 instances as true positives, effectively pinpointing individuals at risk of stroke, however, the 61 instances classified as false negatives are concerning, indicating instances where the model failed to detect actual stroke cases.

FIGURE 15: Random Forest Confusion Matrix

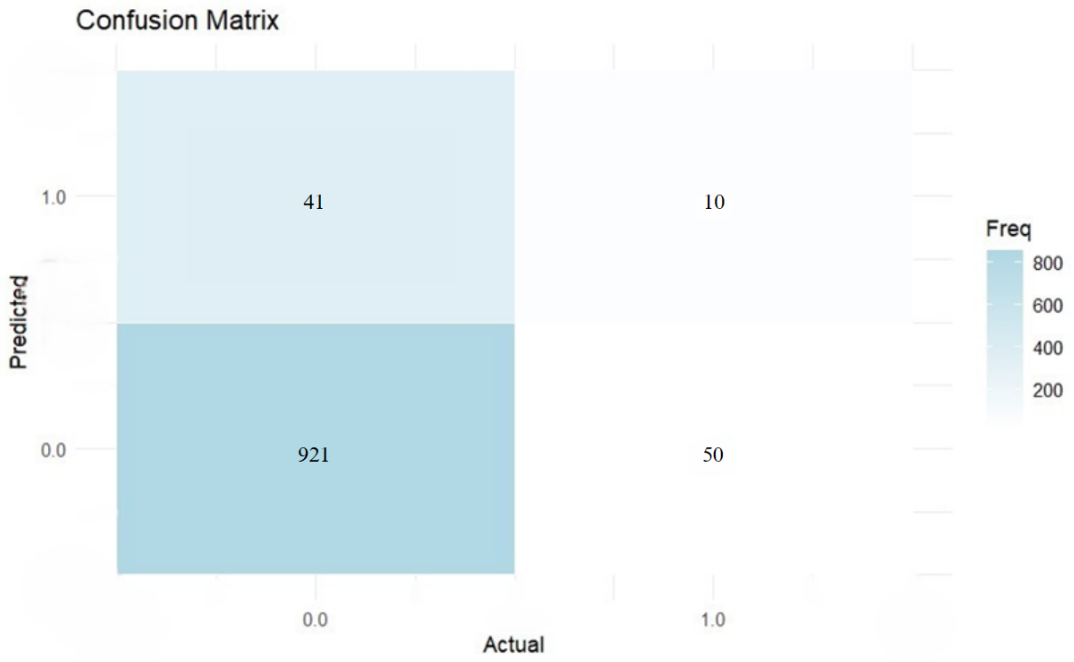
In this model, the sensitivity, ~ 4.7%, indicates the model's limited ability to accurately identify individuals at risk of stroke, suggesting a higher rate of false negatives. The specificity, ~99%, shows incredible proficiency in correctly classifying individuals without stroke. The significantly lower sensitivity implies a greater likelihood of missed stroke diagnoses, which poses major potential risks for patient health outcomes. The high specificity comes at a great cost, as the model's sensitivity is a concern that cannot be overlooked.

6.5 Neural Networks

Neural networks serve as essential components in the fields of artificial intelligence and machine learning, resembling the human brain's information processing capabilities. These networks comprise layers of nodes, or "neurons," each meticulously designed to execute specific computations. Their ability to learn from data inputs and make informed decisions renders them particularly suitable for tasks such as binary classification. They can adapt and improve over time through a process known as "training." During training, the network adjusts its weights and biases, parameters, based on the errors observed in its predictions. This iterative process involves feeding data through the network (forward propagation), evaluating prediction accuracy, and updating parameters to minimize errors (backpropagation). By creating a neural network model and setting the identification threshold to 0.3, the model achieved 91% accuracy. The decision to utilize a 0.3 threshold is due to the substantial cost of a false positive. Comparing thresholds between 0.2-0.5, we found that 0.3 results in the highest sensitivity. The sensitivity of this model is ~17%, whereas the specificity is ~96%.

The confusion matrix **(Figure 16)** shows that 921 test observations were true negatives, indicating a good ability to accurately predict the absence of strokes. On the other hand, the model misclassified 41 instances as false positives, inaccurately identifying strokes that did not occur. The model correctly identified 10 instances as true positives, identifying individuals at risk of stroke, however, the 50 instances classified as false negatives reveal that the model does not perform well in identifying true positive cases. It is important to note that this type of model was limited due to the scarcity of data points, as it historically performs better on a much larger scale.

FIGURE 16: Neural Networks Confusion Matrix



The structure of neural networks is crucial to their functionality, with each component contributing to their overall performance. A network built using sequential architecture allows data to flow seamlessly from input to output through a series of layers. These layers include an input layer, dense layers for data processing, activation functions introducing non-linearity, batch normalization for stabilization, dropout for regularization, and an output layer for classification. Each element plays a vital role in shaping the network's behavior and refining its predictions.

One key aspect that distinguishes neural networks from traditional linear models and tree-based methods is their ability to model non-linear relationships inherent in complex datasets. This nonlinearity is primarily introduced through activation functions, mathematical equations applied to the output of each neuron, which determine whether it should be activated or not. Activation functions such as ReLU (Rectified Linear Unit) and sigmoid enable neural networks to learn complex and non-linear relationships between input features and target outputs, allowing them to capture intricate patterns that may elude simpler models.

7 Conclusion

7.1 Evaluation of Results

Our evaluation focused on key metrics such as accuracy, sensitivity (recall), and specificity. The model with the highest accuracy is Random Forest at 94%; however, the cost of false negatives is significant and with the imbalance of our data, it is important to consider other metrics as accuracy tends to side with the majority class. To account for the minority class, we considered the performance metrics of sensitivity and specificity, and each model exhibited both strengths and weaknesses in predicting stroke risk. Logistic Regression, with a sensitivity of ~83% and specificity of about ~70%, demonstrates a relatively balanced performance in detecting both positive and negative cases. KNN, despite its high specificity of ~99%, struggles with sensitivity at ~5.6%, indicating a significant number of missed positive cases. Decision Trees present a high sensitivity of ~92%, suggesting effective detection of positive cases, but a lower specificity of ~54%, signifying a higher false positive rate. Random Forest shows a similar pattern to KNN, with a specificity of ~99% but a notably lower sensitivity of ~4.7%. The sensitivity of the Neural Networks model is ~17%, whereas the specificity is ~96%, revealing deficiencies in identifying true positives.

Given that the healthcare industry is dealing with individuals' quality and longevity of life, it is crucial to prioritize high sensitivity in predictive models as it is paramount where the consequences of a missed diagnosis can be severe. That being said, models with higher sensitivity, such as Logistic Regression in this case, are preferred as they perform well when identifying positive cases, even if it means sacrificing some specificity.

7.2 Recommendations and Next Steps

Out of all the models created, incorporating the Logistical Regression model into clinical practice is recommended due to its well-rounded performance. Complexity alone does not guarantee superior results as the quality and relevance of the data are paramount. Given the life-saving implications, prioritizing sensitivity in stroke prediction is essential. All models encountered difficulties due to the scarcity of positive data points indicating stroke occurrences. That being said, while Logistic Regression demonstrates strong performance in this study, ongoing efforts to collect more data points to address the data imbalances and continued model training are necessary for sustained success in stroke prediction and prevention.

8 References

1. "Stroke Facts | cdc.gov." *Centers for Disease Control and Prevention*, 6 Sept. 2023, www.cdc.gov/stroke/facts.htm.
2. Likebupt. "SMOTE - Azure Machine Learning." *Microsoft Learn*, 4 Nov. 2021, www.learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote?view=azureml-api-2.