Course Project#3

Title: Stroke Risk Prediction

Business Problem:

Stroke is a leading cause of mortality and long-term disability worldwide, significantly burdening healthcare systems and impacting millions of lives. According to the World Health Organization (WHO), stroke is the second leading cause of death globally, responsible for approximately 11% of all deaths annually. Early identification of individuals at risk of stroke can enable timely intervention, reducing fatalities and improving long-term outcomes. Traditional methods for assessing stroke risk rely on statistical models that incorporate well-known factors such as hypertension, diabetes, and high cholesterol. However, these models often fail to capture complex interactions between multiple health indicators.

Machine learning provides a more sophisticated approach by leveraging large datasets to uncover hidden patterns in stroke risk factors. This project aims to develop a predictive model that can assess an individual's likelihood of experiencing a stroke based on medical history, lifestyle factors, and demographic attributes. By improving stroke prediction accuracy, this model can serve as a valuable tool for healthcare professionals and individuals to take proactive measures for prevention.

Problem Statement

Despite medical advancements, stroke continues to be a major health challenge, with many cases occurring unexpectedly, often leaving little time for emergency intervention. Given that patients who receive medical attention within the first few hours of a stroke experience significantly

better recovery outcomes, early risk assessment is critical. However, current predictive approaches often lack the precision required to identify at-risk individuals effectively.

This project seeks to bridge this gap by using machine learning techniques to analyze multiple health indicators and predict the likelihood of stroke occurrence. The model will use a dataset containing key medical and demographic variables to identify correlations and develop a reliable risk assessment system. The ultimate goal is to contribute to better preventive healthcare strategies by offering an early warning system that can guide lifestyle modifications, medical screenings, and timely interventions.

Background/History

Stroke is a medical emergency that occurs when blood flow to the brain is interrupted, leading to brain cell damage or death. There are two primary types of stroke: ischemic strokes, which occur due to blockages in blood vessels, and hemorrhagic strokes, which result from blood vessel rupture. Ischemic strokes account for approximately 87% of all strokes, making them the most common type.

Globally, nearly 15 million people experience a stroke each year. Among them, one-third succumb to the condition, while another third suffer long-term disability. In the United States alone, approximately 800,000 people have a stroke annually, equating to one stroke every 40 seconds and a stroke-related fatality every 3.5 minutes. Around 25% of these individuals have previously experienced a stroke, highlighting the increased risk of recurrence.

Several known risk factors contribute to stroke occurrence. Cardiovascular conditions such as hypertension, high cholesterol, obesity, and diabetes are major contributors. Age is also a

significant factor, with older individuals facing a higher risk. Additionally, certain demographic and geographical factors influence stroke risk, with research indicating that Black individuals are nearly twice as likely to suffer from a stroke as White individuals. Furthermore, individuals living in the Southern United States have a higher mortality risk from stroke than those residing in other regions.

Early detection and intervention are crucial in stroke prevention. The Centers for Disease Control and Prevention (CDC) reports that individuals who receive emergency medical care within three hours of stroke symptoms have better recovery outcomes and reduced long-term disability. By predicting stroke risks early, healthcare providers and individuals can take preventive measures, including lifestyle changes, medication management, and routine medical check-ups.

Dataset Explanation

The dataset used in this project was sourced from Kaggle and includes patient records containing various medical and demographic features relevant to stroke prediction. Two datasets were considered—one with approximately 5,000 records and another with over 40,000 records. The larger dataset was chosen due to its greater sample size, which enhances statistical significance despite being highly imbalanced, with stroke cases accounting for only about 2% of the total records.

The dataset consists of 12 columns, including:

- **Target Variable**: ‘stroke’ (binary 0/1 indicating whether a stroke occurred).
- **Demographic Features**: Gender, age, and residence type (Urban/Rural).
- **Medical History**: Hypertension, heart disease, and prior stroke occurrences.

- **Lifestyle Factors**: Smoking status and work type.
- **Health Indicators**: BMI and average glucose level.

Given the dataset's imbalance, several resampling techniques were considered to ensure accurate stroke prediction.

- Data Distribution:

When analyzing the dataset's distributions, one of the most striking observations was the age distribution, which exhibited significant spikes at both the lower and upper extremes. Initially, the presence of very young individuals, including infants under one year old, seemed erroneous. However, upon closer inspection, the decimals in the dataset appeared to represent fractions of a year, suggesting intentional inclusion. At the upper age range, an unusual overrepresentation of individuals aged 80-82 years was observed, while no records existed for individuals older than 82, raising questions about potential rounding or data processing methods.

(See Figure 1: Histograms)

Outliers were also assessed in key health-related features such as average glucose level ('avg_glucose_level') and BMI. While both variables fell within expected biological ranges, they exhibited noticeable skewness, with glucose levels displaying a slightly bimodal distribution. This skewness could impact model performance, necessitating appropriate transformations.

One challenge encountered when analyzing the dataset was the high class imbalance, making it difficult to observe trends in stroke occurrences within histograms. Since stroke cases constituted only a small fraction of the dataset, their distribution appeared as a nearly flat line at the bottom

of the visualizations. To better understand the relationship between different features and stroke risk, oversampling of stroke cases was conducted to create more balanced histograms.

(See Figure 2: Histograms (Balanced))

This approach revealed several interesting trends. Gender and residence type showed minimal distinction between stroke and non-stroke groups, suggesting they might not be strong predictors. However, certain health-related attributes—such as age, hypertension, heart disease, and average glucose level (often linked with diabetes)—displayed clear differences between stroke and non-stroke groups. Additionally, while work type was not initially expected to influence stroke risk, individuals who were self-employed exhibited a higher stroke prevalence compared to other employment categories. Furthermore, smoking status showed former smokers had a higher association with strokes, whereas current smoking status did not significantly affect stroke likelihood.

These findings highlight the importance of feature selection in building an accurate predictive model. While some categorical variables may require further scrutiny to assess their predictive power, health-related features such as hypertension, heart disease, and glucose levels appear to be strong indicators of stroke risk.

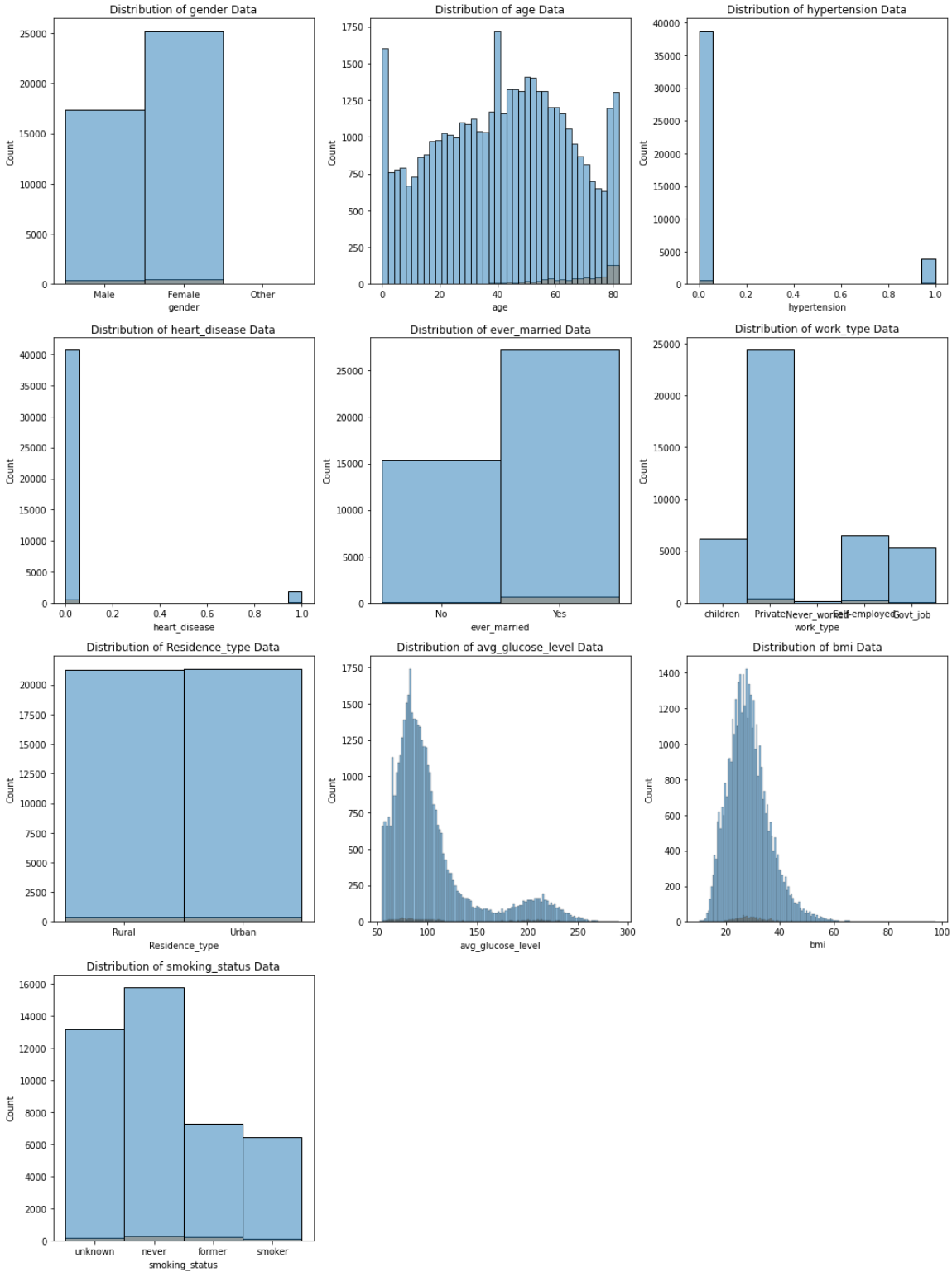

Figure 1 Histograms

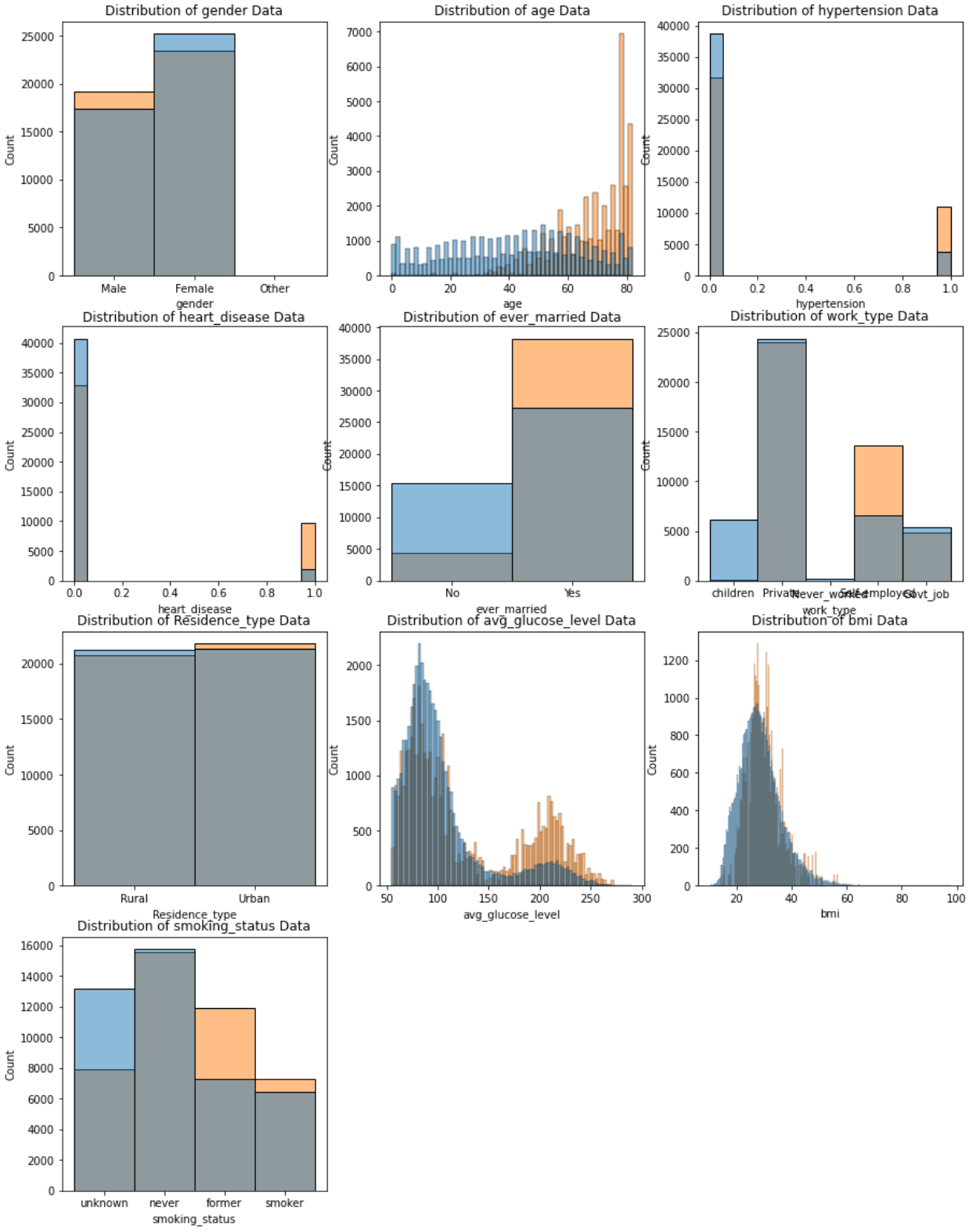

Figure 2 Histograms (Balanced)

- Handling of Null Values:

Missing values are an important aspect to consider when preparing data for machine learning models. In this dataset, the most prominent issue with missing data was in the smoking status variable, where nearly one-third of the values were missing. Given that this feature is categorical, a simple and effective solution was to create a new category labeled 'Unknown', rather than attempting to impute values based on other characteristics. Interestingly, after this modification, the 'Unknown' category became the second-largest group in the smoking status variable, raising questions about potential inconsistencies or gaps in data collection.

The BMI variable was the only numerical feature with missing values, accounting for a little over 3% of the dataset. Given the relatively small proportion of missing data, several imputation strategies were considered to fill in the gaps. Initially, a logistic regression-based approach was explored to predict missing BMI values using other available health-related attributes. However, upon comparison, datasets with BMI imputed using the mean and median values were also evaluated for performance. While logistic regression-based imputation provided reasonable estimates, it introduced unnecessary complexity. The best-performing and most efficient approach was median imputation, as it avoided extreme values and maintained the original data distribution more effectively.

- Handling Categorical Variables:

To prepare the dataset for machine learning models, categorical variables were encoded appropriately to ensure numerical representation. The 'gender', 'ever_married', and 'Residence_type' features were converted into binary representations (1s and 0s) to maintain

simplicity and reduce feature dimensionality. Meanwhile, categorical features with multiple unique values, such as 'work_type' and 'smoking_status', were transformed using one-hot encoding, creating separate binary columns for each category.

An interesting observation emerged when analyzing the impact of the newly created 'unknown' smoking category, which was introduced to handle missing values. It was found that this group was disproportionately associated with non-stroke cases. Various approaches were considered, including reducing the variable to a simple smoker/non-smoker classification or restructuring it as a one-vs-all former-smoker category, but ultimately, the decision was made to retain the original encoding structure.

Similarly, the 'work_type' variable posed a challenge, as certain categories appeared to have little correlation with stroke occurrences. A potential modification considered was converting it into a binary self-employed vs. other classification, given that self-employed individuals showed a higher incidence of stroke. However, after further evaluation, the original multi-class representation was preserved, allowing the model to capture more nuanced relationships within the dataset.

- Train/Test Split:

After completing the imputation of missing values and encoding categorical variables, the dataset was divided into training and testing sets to evaluate model performance. While best practices recommend performing the train-test split before imputation to prevent data leakage, the impact of imputing missing BMI values using the median was considered minimal, and thus, the split was conducted afterward.

However, to maintain the integrity of the test set, splitting was performed before applying transformations and scaling. This precaution ensured that no statistical information from the test set influenced the training data, preserving the model's ability to generalize effectively when encountering unseen data.

- Transformation:

To address skewness in the dataset, Box-Cox transformations were applied to the glucose levels, BMI, and age variables. These transformations are particularly useful for stabilizing variance and normalizing the data, which is important for many machine learning models. Since none of the variables had values at or below zero, no additional rescaling was required prior to applying the Box-Cox transformation.

(See Figure 3: Box-Cox Transformation)

The Box-Cox transformation converts the data into a range of approximately -2.5 to 2.5, but since the other binary variables (such as gender and residence type) were already coded as 0/1, it was necessary to scale all features to a -1 to 1 range. This uniform scaling was applied to ensure consistency across all features, which is preferred in certain machine learning models and should not negatively affect the performance of other models.

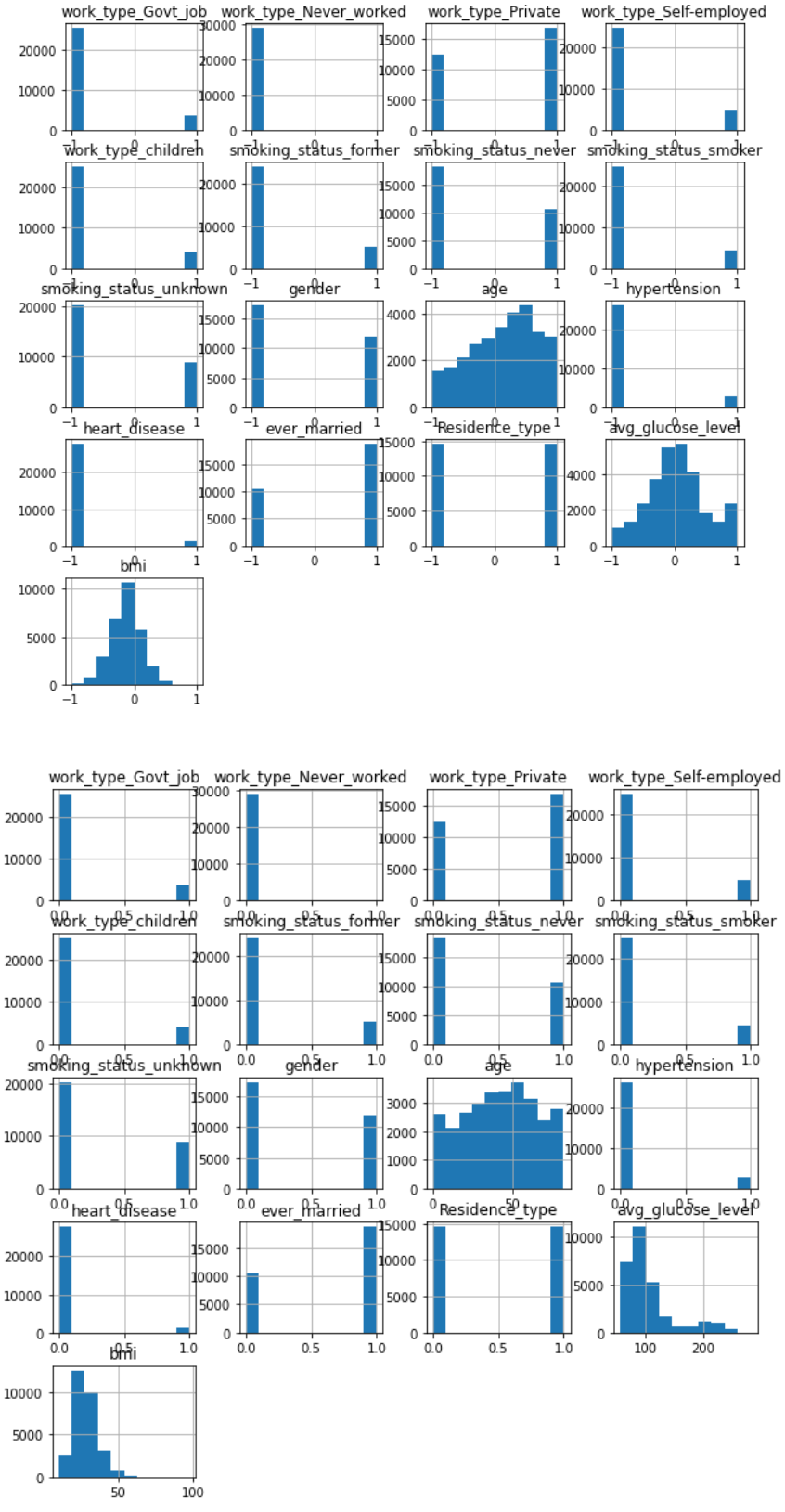

Figure 3 Box-Cox Transformation

Methodology

Developing a reliable predictive model for stroke risk required careful consideration of multiple factors, including the dataset's severe class imbalance, the selection of appropriate machine learning algorithms, and the choice of evaluation metrics. The methodology focused on optimizing data preprocessing, feature selection, balancing techniques, model selection, and hyperparameter tuning to ensure the best possible performance.

- Feature Selection

Given that the dataset contained only 10 predictor variables, no major feature reduction techniques were applied. Instead, all features were retained to maximize the available information and avoid potential loss of valuable predictors. While some categorical features, such as work_type, showed limited impact on stroke prediction, they were preserved in case they contributed to model performance when combined with other features.

- Balancing the Dataset

One of the most significant challenges in building an effective stroke prediction model was the extreme imbalance in the dataset, with stroke cases comprising only about 2% of the total records. Initial model evaluations without balancing resulted in an inflated accuracy of nearly 98%, misleadingly indicating strong model performance when, in reality, the model was heavily biased toward predicting the majority class (non-stroke cases).

To address this issue, multiple balancing techniques were tested. Initially, oversampling of the minority class (stroke cases) to match the number of non-stroke cases was attempted. However, this approach did not yield significant improvements in model performance. Next, Synthetic

Minority Over-sampling Technique (SMOTE) was applied, which artificially generates synthetic samples for the minority class. While SMOTE resulted in some improvement, it nearly doubled the dataset size without adding substantial predictive value.

A refined balancing approach was then implemented, where SMOTE was used to upsample the minority class to 10% of the majority class, followed by downsampling the majority class to twice the size of the minority class. This technique maintained a reasonable dataset size while improving model performance. Ultimately, after testing various approaches, the best results were achieved without explicit resampling, instead using the balanced class weight option available in classification algorithms to adjust for class imbalance dynamically.

- Algorithm Selection

To determine the most effective predictive model, multiple classification algorithms were tested, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines (SVM), and Neural Networks. Initial model selection was based on Accuracy, Precision, Recall, and F1-score, with the top five models achieving over 80% accuracy.

However, given the severe imbalance in the dataset (~98% non-stroke cases), accuracy alone was not a reliable evaluation metric. A false positive (incorrectly predicting stroke) is less harmful than a false negative (failing to predict stroke when it occurs). To address this, Recall was initially considered as the primary metric. However, optimizing for recall alone led to excessive false positives, which could result in unnecessary medical interventions.

Further research led to the adoption of Matthews Correlation Coefficient (MCC), a robust metric for imbalanced binary classification problems. MCC provides a balanced measure of performance

across all four confusion matrix categories (true positives, true negatives, false positives, and false negatives). Additionally, ROC AUC (Receiver Operating Characteristic - Area Under the Curve) was used to assess the trade-off between sensitivity and specificity.

(See Figure 4: Metrics for Classification Models)

	Model	Accuracy	F1_Score	Precision	Recall	MCC	AUC
0	LogReg_lbfgs	74.03	0.1015	0.0541	0.8140	0.1655	0.7765
1	LogReg_liblinear	74.03	0.1015	0.0541	0.8140	0.1655	0.7765
2	LinearSVC	76.10	0.1047	0.0561	0.7752	0.1649	0.7680
3	SGD	78.80	0.1087	0.0588	0.7171	0.1627	0.7532
4	SVC	75.75	0.0987	0.0529	0.7364	0.1517	0.7472
5	LGBM	87.50	0.1234	0.0706	0.4884	0.1492	0.6852
6	LDA	97.19	0.0987	0.1170	0.0853	0.0858	0.5367
7	GaussianNB	22.50	0.0441	0.0226	0.9922	0.0667	0.6016
8	DecisionTree	96.73	0.0824	0.0833	0.0814	0.0657	0.5325
9	XGBC	98.11	0.0145	0.1176	0.0078	0.0258	0.5033
10	BernoulliNB	97.54	0.0330	0.0566	0.0233	0.0251	0.5081
11	KNN	98.10	0.0073	0.0625	0.0039	0.0112	0.5014
12	MLP	98.20	0.0000	0.0000	0.0000	0.0000	0.5000
13	AdaBoost	98.19	0.0000	0.0000	0.0000	-0.0011	0.5000
14	RandomForest	98.16	0.0000	0.0000	0.0000	-0.0025	0.4998
15	GBC	98.14	0.0000	0.0000	0.0000	-0.0032	0.4997

Figure 4 Metrics for Classification Models

- Hyperparameter Tuning

After selecting the top-performing models, hyperparameter tuning was conducted to optimize their predictive accuracy. The primary method used was GridSearchCV, which systematically tests multiple parameter combinations to find the best-performing configuration. Due to the computational intensity of tuning deep models, RandomizedSearchCV was used for more complex algorithms, such as Multi-layer Perceptron (MLP) Neural Networks.

After refining evaluation metrics, models with MCC > 0.15 were prioritized, with a focus on those that also ranked highly in Recall and AUC scores. The best-performing models were Logistic Regression, Support Vector Classifier (SVC), and Linear SVC. Additionally, Stochastic Gradient Descent (SGD) was later included due to its competitive performance in AUC and Recall metrics, and its potential for improvement with further tuning.

Further experimentation revealed that hyperparameter tuning could be optimized using multiple evaluation metrics simultaneously. As a result, MCC, Recall, and AUC were all incorporated into the scoring function during GridSearchCV optimization. The tuning process yielded minor improvements for SVC-based models but significantly enhanced the performance of SGD.

To further improve prediction accuracy, ensemble methods were explored, including a VotingClassifier, which combines predictions from multiple models by assigning weights based on performance scores. Early testing of this approach showed notable improvements in recall, but further refinement was needed. Given more time, additional ensemble techniques, such as stacking or boosting, could further enhance model robustness.

Analysis:

Initial model evaluations demonstrated high accuracy rates exceeding 95%, but further inspection of the confusion matrices revealed a persistent issue—severe class imbalance was masking actual model performance. While accuracy appeared strong, the model struggled to correctly identify stroke cases, with positive stroke instances making up only 251 observations out of more than 14,000 records. As a result, even the highest-accuracy models failed to capture a substantial proportion of true stroke occurrences, highlighting the limitations of relying solely on accuracy as a performance metric.

Given the critical nature of stroke prediction, it became evident that minimizing false negatives (missed stroke cases) was more important than minimizing false positives (incorrectly predicting stroke for non-stroke individuals). In a real-world healthcare setting, a false negative could mean failing to warn a high-risk individual, leading to severe health consequences, whereas a false positive—though not ideal—would at most prompt further medical evaluation.

To address this, the evaluation criteria were adjusted to focus on metrics that better reflect the model's ability to detect stroke cases accurately. Instead of relying solely on accuracy, the final model comparisons prioritized Matthews Correlation Coefficient (MCC), Area Under the Curve (AUC), and Recall. MCC provided a balanced evaluation of both true positive and false positive rates, making it particularly suitable for highly imbalanced datasets. AUC helped assess the trade-off between sensitivity (true positive rate) and specificity (true negative rate), while Recall ensured that as many stroke cases as possible were correctly identified.

By refining evaluation metrics and prioritizing stroke case detection, the final analysis focused on achieving a model that, while not perfect, maximized early stroke identification—a crucial factor in real-world medical decision-making.

Ethical Problem:

Ethical considerations are crucial in developing a stroke prediction model, particularly in ensuring data privacy and compliance with regulations like GDPR and HIPAA. While the dataset is anonymized, future iterations using real patient data must follow strict security protocols to protect sensitive information. Additionally, bias in model predictions must be carefully addressed, as certain risk factors like race and socioeconomic status are not included in the dataset. The absence of such variables can lead to biased outcomes, potentially impacting healthcare recommendations for different demographic groups. Ensuring fairness through balanced data representation and unbiased modeling techniques is essential for ethical AI deployment in healthcare.

Responsible decision-making is another key aspect, as the model should serve as a supportive tool rather than a definitive diagnostic system. Healthcare professionals must validate its predictions to ensure accuracy and clinical relevance. The model's insights should complement, not replace, medical expertise, reducing the risk of misinterpretation. By prioritizing data privacy, fairness, and responsible use, this project aims to develop an ethical and reliable stroke risk prediction model that contributes meaningfully to preventive healthcare.

Challenges/Issues:

One of the primary challenges encountered in this project was handling missing values, particularly in the BMI and smoking status variables. Initially, logistic regression was explored as an imputation method, but it did not yield the expected improvements in model performance. While it was a valuable learning experience, it ultimately proved to be unnecessarily complex compared to simpler techniques like median imputation, which provided better results with less computational overhead.

Another major challenge was balancing the dataset, as stroke cases made up only about 2% of the total data. Initially, oversampling the minority class to match the number of non-stroke observations was attempted, but this approach resulted in overfitting, where the model became too specialized in recognizing stroke cases from the training data and struggled to generalize effectively.

To address this, SMOTE (Synthetic Minority Over-sampling Technique) was implemented, which generates synthetic stroke cases to balance the dataset. While SMOTE provided some improvements, it significantly increased the dataset size without necessarily adding meaningful new information. Additional strategies, such as combining SMOTE with undersampling of the majority class, were tested to create a more balanced dataset while maintaining a reasonable size.

Ultimately, the best results were achieved without explicit resampling and instead by using class weighting within the classification algorithms. Assigning higher misclassification penalties to the

minority class allowed the models to account for stroke cases more effectively without overfitting.

Assumptions

One of the key assumptions in this project is that the dataset, which was sourced from Kaggle, was collected from a legitimate and representative source. However, since no official documentation or metadata was available detailing how the data was gathered, the true origin and collection methodology remain unknown. This introduces some uncertainty regarding its real-world applicability and generalizability.

The decision to use the larger dataset over the smaller one was based on the assumption that more data generally leads to better model performance. However, without detailed background information, there is a possibility that the larger dataset was derived from the smaller one or that certain preprocessing steps had already been applied before publication. Evidence of data processing can be seen in some histogram distributions, particularly age, where spikes appear at the upper and lower extremes, possibly indicating rounding adjustments or missing value imputations. If data cleaning or augmentation had already been performed before availability, it could introduce biases that impact model training.

Despite these uncertainties, the dataset was assumed to be sufficiently diverse and realistic to train a meaningful stroke prediction model. However, in a real-world deployment scenario, verifying the data's source and ensuring its authenticity would be critical for ethical and clinical applications.

Limitations:

One of the primary limitations of the dataset is that it represents a snapshot of patient health data at a single point in time. Many individuals labeled as "no stroke" may still experience a stroke later in life, but this information is not reflected in the dataset. If the stroke column only indicates past occurrences, rather than tracking patients over time, the model is limited to short-term stroke prediction rather than long-term risk assessment. A more robust dataset would include longitudinal patient records to track whether individuals eventually experience a stroke, improving the accuracy and reliability of predictions.

Another significant limitation is the absence of certain key risk factors. While the dataset includes important predictors like age, hypertension, heart disease, glucose levels, and BMI, it lacks other well-documented stroke risk factors such as family history, geographical location, race, and previous stroke occurrences. The omission of these features may lead to suboptimal model performance, as certain high-risk groups might not be accurately identified.

Perhaps the most impactful limitation is the severe class imbalance, with stroke cases making up only about 2% of the dataset. This imbalance makes it challenging for machine learning models to learn meaningful patterns associated with stroke occurrences. Despite various resampling techniques and class weighting strategies, a dataset with a higher proportion of stroke cases would significantly improve the model's ability to generalize. Addressing these limitations in future datasets could enhance the accuracy and clinical usefulness of stroke prediction models.

Conclusion:

The findings from this project suggest that stroke prediction using machine learning is feasible, provided that the dataset contains sufficient high-quality data and an adequate number of positive stroke cases. While the model demonstrated promising results, its effectiveness was limited by the severe class imbalance and the absence of certain key risk factors. With the given dataset, false negatives remained a challenge, as the model struggled to capture all stroke cases without significantly increasing false positives. Striking the right balance between sensitivity (recall) and specificity (precision) remains critical, as an overly aggressive model could misclassify too many non-stroke cases, reducing its practical utility.

Given the complexity of stroke risk factors, an alternative approach could involve expanding the dataset to include a wider range of potential predictors, including genetic predisposition, lifestyle habits, environmental factors, and previous stroke history. A more comprehensive feature selection process could help identify which variables contribute most significantly to stroke prediction, enabling the development of a more accurate and clinically relevant model. Future iterations of this project could focus on enhancing data quality, incorporating longitudinal patient records, and refining model evaluation techniques to improve overall predictive performance and support more effective stroke prevention strategies.

Future Work

The potential applications of stroke prediction models extend beyond individual risk assessment, offering valuable insights into broader stroke risk factors that could inform preventive healthcare strategies worldwide. Future work in this area should focus on enhancing model accuracy and

generalizability by incorporating a more diverse and representative dataset, including longitudinal patient records that track stroke occurrences over time. This would allow for a more precise distinction between short-term and long-term stroke risks, improving the reliability of predictions.

Additionally, integrating a wider range of risk factors, such as family history, socioeconomic background, physical activity levels, and dietary habits, could significantly strengthen the model's predictive capabilities. Future iterations of this project could also explore the deployment of real-time stroke risk assessment tools, such as mobile health applications or clinical decision-support systems, enabling individuals and healthcare providers to take proactive preventive measures. By simplifying stroke risk assessment through machine learning, this research has the potential to contribute to early detection, reduce stroke-related mortality, and enhance overall public health outcomes.

Questions an audience may ask

1) What additional features would improve the dataset?

   To enhance the predictive capability of the model, incorporating more health-related variables such as physical activity levels, blood pressure, cholesterol levels, and genetic predisposition would be beneficial. Additionally, including socioeconomic and demographic factors like race, ethnicity, and family history of stroke could provide deeper insights into high-risk populations.

2) Can a model built using U.S. data be generalized globally?

To some extent, yes. Many stroke risk factors, such as hypertension, diabetes, and high cholesterol, are universally relevant. However, regional differences in diet, lifestyle, healthcare access, and genetic predisposition may influence stroke risks. Studies have shown that individuals living in the Southern U.S. face a higher risk of stroke-related mortality, indicating that geographic and environmental factors play a role. More research is needed to determine whether these trends apply globally or are region-specific.

3) What other methods did you consider for imputing missing values?

For BMI imputation, I experimented with logistic regression, mean imputation, and median imputation, ultimately finding that median imputation produced the most stable results. While predictive models could offer alternative imputation techniques, they add complexity and computational overhead. Another viable approach could be randomly selecting BMI values from the dataset based on a similar distribution, which would help maintain a natural-looking histogram

4) Why did you decide to keep individuals under 30 in the model?

Although there were very few stroke cases in the under-30 age group, I chose to retain them to ensure that the model accounted for any potential early-onset stroke cases. Removing younger individuals could lead to an unnecessary loss of information that might be valuable for identifying rare but significant stroke occurrences.

like injuries, referee decisions, and psychological elements of the game.

5) How did you handle the 'Other' gender category?

Ultimately, I removed the 11 cases categorized as "Other" in the dataset. While this decision may not be ideal in a real-world setting where inclusivity is critical, it was made to simplify the model and prevent complications arising from an extremely small sample size. In practical applications, a more nuanced approach should be considered.

6) What train-test split ratio did you use?

I used a 67-33% split, allocating 33% of the data for testing. However, this choice was largely arbitrary—I could have just as easily opted for a 75-25% or another ratio. The main priority was ensuring that the test set was large enough to evaluate model generalizability without compromising training performance.

7) Why did you choose your specific method for handling class imbalance?

After testing several resampling techniques, I found that leaving the dataset imbalanced while using model class weighting provided the best results. Models with built-in class weight adjustments were better suited for handling the imbalance, whereas models that lacked class weighting options performed poorly across all metrics except accuracy.

8) What evaluation metrics did you use?

Initially, I relied on accuracy, but I quickly realized that it was not appropriate for a highly imbalanced dataset. To prioritize identifying true positives, I incorporated more reliable metrics, including Matthews Correlation Coefficient (MCC), Area Under the Curve (AUC), and Recall. These metrics provided a more balanced evaluation of model performance, particularly in distinguishing stroke cases.

9) Why did you use the Box-Cox transformation?

Box-Cox transformation is an effective and straightforward method for normalizing skewed data. I frequently use it because it is quick to implement, computationally efficient, and works well across different distributions. It ensures that features with non-normal distributions are transformed into a format that improves model performance.

10) What was your biggest challenge in model selection?

The biggest limitation was time. With more time, I would have explored additional model architectures, fine-tuned hyperparameters more extensively, and experimented with ensemble techniques. Given the constraints, I focused on selecting a model that performed well without excessive optimization, though additional refinements could further enhance predictive performance.

References:

1. Amal, L. (2020, October 26). *Heart stroke*. Kaggle. https://www.kaggle.com/datasets/lirilkumaramal/heart-stroke
2. Brownlee, J. (2021, May 7). *How to develop a weighted average ensemble with Python*. Machine Learning Mastery. https://machinelearningmastery.com/weighted-average-ensemble-with-python/
3. Brownlee, J. (2020, August 26). *Train-test split for evaluating machine learning algorithms*. Machine Learning Mastery. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
4. Chicco, D., & Jurman, G. (2020, January 2). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation*. *BMC Genomics*. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7
5. Eddie_4072. (2022, June 20). *Dealing with missing values in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/
6. Emergency Nutrition Network. (2002, January 4). *The limits of human starvation*. *Field Exchange 15*. https://www.ennonline.net/fex/15/limits
7. Emon, M. U., Islam, M. S., Hossain, M. S., Sadi, M. S., & Saha, S. (2020). *Performance analysis of machine learning approaches in stroke prediction*. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. https://doi.org/10.1109/iceca49313.2020.9297525

8. Centers for Disease Control and Prevention. (2022, April 12). *Know your risk for stroke*. CDC. https://www.cdc.gov/stroke/risk_factors.htm

9. Kudva, Y. C. (2020, October 30). *Diabetes*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451

10. Nik. (2022, February 23). *One-hot encoding in scikit-learn with OneHotEncoder*. Datagy. https://datagy.io/sklearn-one-hot-encode/

11. Centers for Disease Control and Prevention. (2022, April 5). *Stroke facts*. CDC. https://www.cdc.gov/stroke/facts.htm

12. World Health Organization. (n.d.). *Stroke, cerebrovascular accident*. WHO Regional Office for the Eastern Mediterranean. http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

13. World Health Organization. (2020, December 9). *The top 10 causes of death*. WHO. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

14. NHS Choices. (2019, July 15). *What is the body mass index (BMI)?* NHS. https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/

15. Zach. (2021, September 9). *How to calculate AUC (area under curve) in Python*. Statology. https://www.statology.org/auc-in-python/