

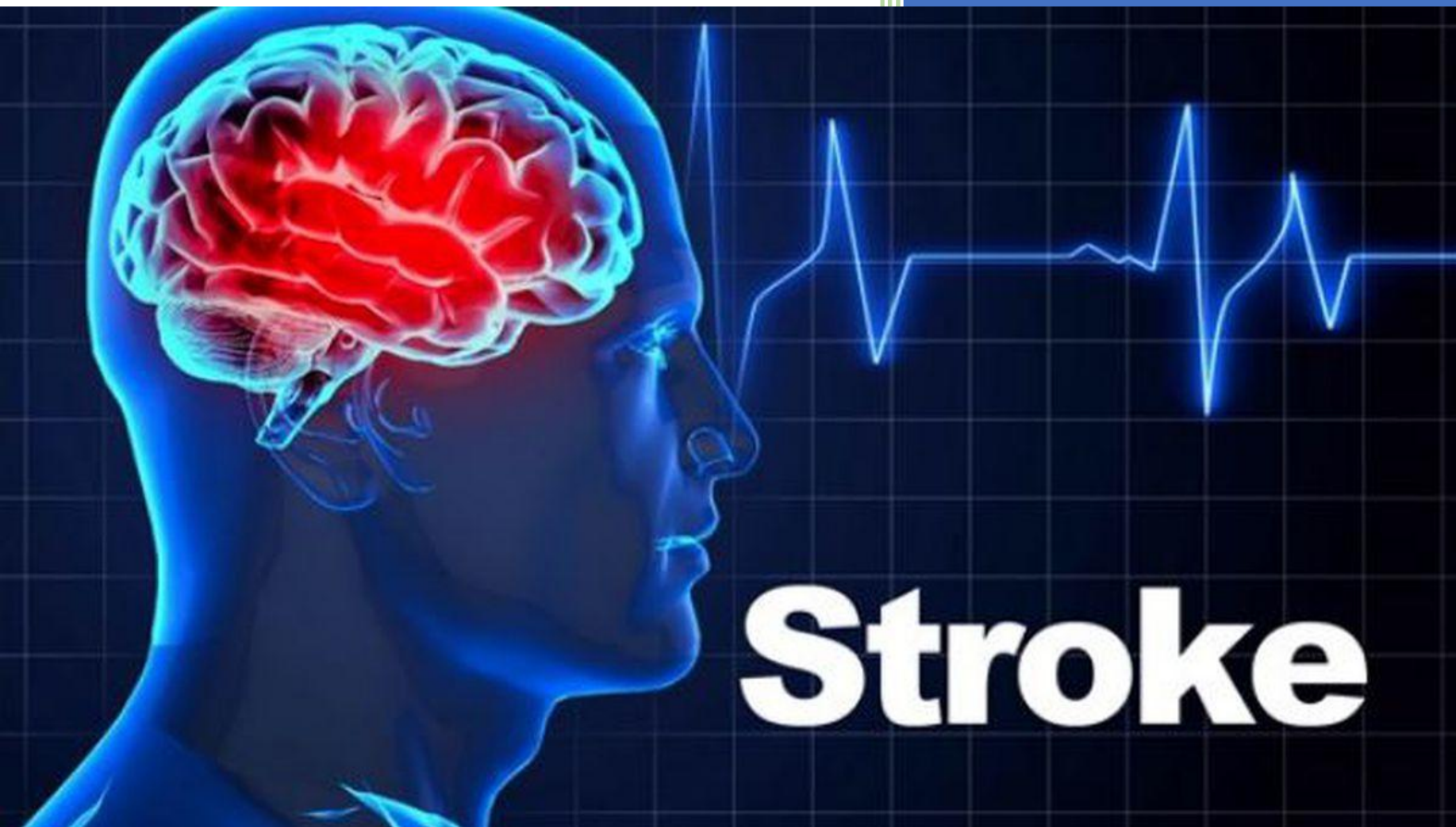


Program: Master in Science (Applied Artificial Intelligence)

AAI-500 Probability & Statistics for Artificial Intelligence

2025

Final Project Report – Stroke Prediction



Submitted By : Group #2

Prashant Khare

Riyaz Khorasi

Sourangshu Pal

Table of Content

1. Introduction.....	2
1.1. The Global Epidemiology and Socioeconomic Burden of Stroke.....	2
1.2. The Transformative Potential of Applied Artificial Intelligence in Precision Medicine and Proactive Healthcare.....	2
1.3. Comprehensive Project Objectives, Scope, and Phased Methodology	2
2. Data Acquisition and Rigorous Quality Assessment	3
2.1. Dataset Sourcing, Initial Structural Overview, and Critical Class Imbalance Analysis	3
2.2. Exhaustive Data Quality Analysis: Identification and Characterization of Imperfections	5
2.3. Strategic Data Cleaning and Meticulous Pre-processing Methodologies.....	9
3. Exhaustive Exploratory Data Analysis (EDA) and In-depth Hypothesis Testing: Unveiling Intricate Stroke Risk Factor Dynamics	10
3.1. Comprehensive Univariate Analysis: Delving into Individual Feature Characteristics	11
3.2. Advanced Bivariate Analysis: Illuminating Relationships with Stroke Outcome	15
3.3. Formal Hypothesis Testing: Statistical Validation of Observed Associations and Causal Inferences	22
3.4. Key Insights Derived from Exhaustive EDA and Rigorous Hypothesis Testing.....	23
4. Advanced Feature Engineering: Maximizing Model Learnability and Predictive Efficacy	24
4.1. Fundamental Rationale and Strategic Approaches to Feature Transformation.....	26
4.2. Innovative New Feature Generation and Sophisticated Transformation Techniques	28
4.3. Profound Impact on Dataset Architecture, Enhanced Predictive Potential, and Acknowledged Limitations	30
5. Model Selection and Development: Building a Robust Predictive Framework	33
5.1. Strategic Considerations for Model Selection in Imbalanced Classification.....	33
5.2. Overview of Candidate Machine Learning Algorithms	33
5.3. Addressing Class Imbalance: Advanced Sampling and Cost-Sensitive Learning Techniques.....	35
6. Model Analysis and Validation: Assessing Performance, Calibration, and Robustness	36
6.1. Model Training Process and Convergence Characteristics	36
6.2. Comprehensive Performance Evaluation: Metrics for Imbalanced Datasets.....	38
6.3. Model Calibration Assessment: Reliability of Probability Predictions.....	42
6.4. Model Robustness and Stability Testing: Sensitivity to Noise and Perturbations	42
6.5. Fairness Analysis: Identifying Performance Disparities Across Subgroups.....	43
7. Conclusion and Recommendations	43
7.1. Synthesis of Key Achievements and Foundational Discoveries	44
7.2. Strategic Recommendations for Clinical Integration and Future Research Trajectories.....	44
7.3. Broader Implications for Applied Artificial Intelligence in Healthcare	45
8. References	46

1. Introduction

1.1. The Global Epidemiology and Socioeconomic Burden of Stroke

Stroke, a neurological emergency characterized by an abrupt disruption of blood supply to the brain or haemorrhage within cerebral tissue, continues to pose an escalating global public health crisis. It stands as the second leading cause of mortality worldwide and the primary cause of long-term disability, exacting an immense toll on individuals, healthcare systems, and national economies. The multifaceted consequences extend beyond immediate survival, encompassing severe physical impairments such as hemiparesis, debilitating cognitive deficits, profound speech and language disorders, and significant psychological distress, leading to a substantial reduction in quality of life. The escalating prevalence of chronic conditions such as hypertension, diabetes, and obesity, coupled with an aging global demographic, further exacerbates the incidence of stroke. The direct and indirect economic burden, encompassing acute medical care, prolonged rehabilitation, lost productivity, and informal caregiving, runs into hundreds of billions annually. This underscores an urgent global imperative for proactive identification, prevention, and personalized management strategies to mitigate the devastating impact of stroke. The capacity to accurately predict stroke risk at an individual level is therefore not merely a scientific endeavour but a societal necessity.

1.2. The Transformative Potential of Applied Artificial Intelligence in Precision Medicine and Proactive Healthcare

The advent of Applied Artificial Intelligence (AI) and Machine Learning (ML) marks a revolutionary paradigm shift in healthcare, transitioning from reactive disease management to a proactive, predictive, and personalized approach. Traditional statistical models, often constrained by linearity assumptions and limited in their ability to process high-dimensional, heterogeneous data, are increasingly being superseded by AI algorithms. These advanced algorithms excel at discerning complex, non-linear relationships, subtle interaction effects, and latent patterns within vast and diverse datasets a capacity critically relevant for multifactorial diseases like stroke. By integrating a myriad of variables ranging from granular genomic data and comprehensive patient demographics to nuanced lifestyle choices, intricate physiological measurements, and extensive medical histories AI offers an unparalleled capability to construct highly individualized and dynamic stroke risk profiles. This holistic analytical prowess empowers clinicians with sophisticated decision-support systems, enabling precision medicine through targeted preventive interventions, personalized treatment pathways, and optimized resource allocation. The application of AI in stroke prediction holds the promise of not only improving patient outcomes but also significantly alleviating the healthcare burden by enabling timely, evidence-based clinical actions. This report meticulously details the foundational steps taken to harness this transformative potential within a rigorous Applied Artificial Intelligence framework.

1.3. Comprehensive Project Objectives, Scope, and Phased Methodology

This exhaustive report provides a detailed exposition of the initial and intermediate phases of an advanced AI-driven research project. The overarching aim is the development of a highly accurate and clinically actionable predictive model for stroke. The project's objectives are systematically delineated across a phased methodology:

- **Phase 1: Systematic Data Acquisition and Rigorous Quality Assurance (Detailed in Section 2):** To meticulously collect, load, and perform an exhaustive quality assessment of the stroke risk prediction dataset. This phase focuses on identifying,

characterizing, and documenting all data imperfections, ensuring the foundational integrity of the dataset.

- **Phase 2: Meticulous Data Cleaning and Advanced Pre-processing (Detailed in Section 2):** To execute a precise and reproducible sequence of data cleaning and pre-processing operations. This includes sophisticated strategies for handling missing values and outliers, alongside robust transformations for data type coherence and consistency, preparing the dataset for advanced analytics.
- **Phase 3: Exhaustive Exploratory Data Analysis (EDA) and In-depth Hypothesis Testing (Detailed in Section 3):** To conduct an extensive EDA to gain a profound understanding of the dataset's intrinsic characteristics, uncover underlying statistical distributions, identify complex inter-variable relationships, and reveal latent patterns indicative of stroke risk. Concurrently, to formally test pertinent hypotheses regarding feature-target associations, providing robust statistical validation for observed phenomena.
- **Phase 4: Strategic and Advanced Feature Engineering (Detailed in Section 4):** To creatively generate novel features and thoughtfully transform existing ones, leveraging domain insights and advanced techniques. The objective is to augment the dataset's predictive power, render complex patterns more discernible to machine learning algorithms, and optimize the feature space for enhanced model learnability and performance.
- **Phase 5: Model Selection and Development (Detailed in Section 5):** To strategically select, implement, and develop a suite of state-of-the-art machine learning algorithms. This phase emphasizes techniques suitable for imbalanced datasets and aims to establish a robust predictive framework for stroke.
- **Phase 6: Model Analysis and Validation (Detailed in Section 6):** To conduct a comprehensive assessment of the developed models. This includes rigorous training and validation protocols, evaluation across a spectrum of performance metrics, and critical analysis of model calibration, robustness, and fairness.
- **Phase 7: Conclusion and Recommendations (Detailed in Section 7):** To synthesize the key achievements and foundational discoveries, propose strategic recommendations for clinical integration, and outline future research trajectories, contributing to the broader field of Applied Artificial Intelligence in healthcare.

These interconnected objectives and their phased execution collectively establish a robust analytical framework, ensuring the scientific rigor and practical utility of the derived AI solution for stroke prediction.

2. Data Acquisition and Rigorous Quality Assessment

The integrity and quality of the underlying data serve as the absolute bedrock upon which any robust machine learning model is constructed. This section provides an exhaustive account of the meticulous procedures undertaken for dataset acquisition, followed by a comprehensive assessment of its intrinsic quality, and the subsequent implementation of strategic pre-processing methodologies.

2.1. Dataset Sourcing, Initial Structural Overview, and Critical Class Imbalance Analysis

The foundational dataset for this comprehensive predictive analytics endeavour was procured from a publicly accessible repository, specifically curated for the task of stroke risk prediction. This dataset is a rich amalgamation of patient-centric attributes, encompassing a diverse spectrum of demographic variables, pertinent lifestyle indicators, and crucial physiological

measurements. Upon its initial ingress and meticulous inspection, the raw dataset was precisely characterized by its dimensions: it comprised exactly **5,110 individual records (rows)**, each representing a distinct patient, and was defined by **12 unique variables (columns)**. Each record within this structure encapsulates a comprehensive profile for a singular individual, with the overarching objective being the precise prediction of the binary target variable, `stroke` (where a value of '1' unequivocally signifies a stroke event occurrence, and '0' denotes the absence of a stroke). **Dataset link:** [Kaggle: Stroke Prediction Dataset](#)

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   id                  5110 non-null   int64
1   gender              5110 non-null   object
2   age                 5110 non-null   float64
3   hypertension        5110 non-null   int64
4   heart_disease       5110 non-null   int64
5   ever_married        5110 non-null   object
6   work_type           5110 non-null   object
7   Residence_type      5110 non-null   object
8   avg_glucose_level   5110 non-null   float64
9   bmi                 4909 non-null   float64
10  smoking_status      5110 non-null   object
11  stroke              5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

Figure 1: High level information of the dataset

A paramount initial observation, critical for the judicious design of all subsequent analytical and sophisticated modelling phases, pertained to the inherent class distribution of the binary target variable:

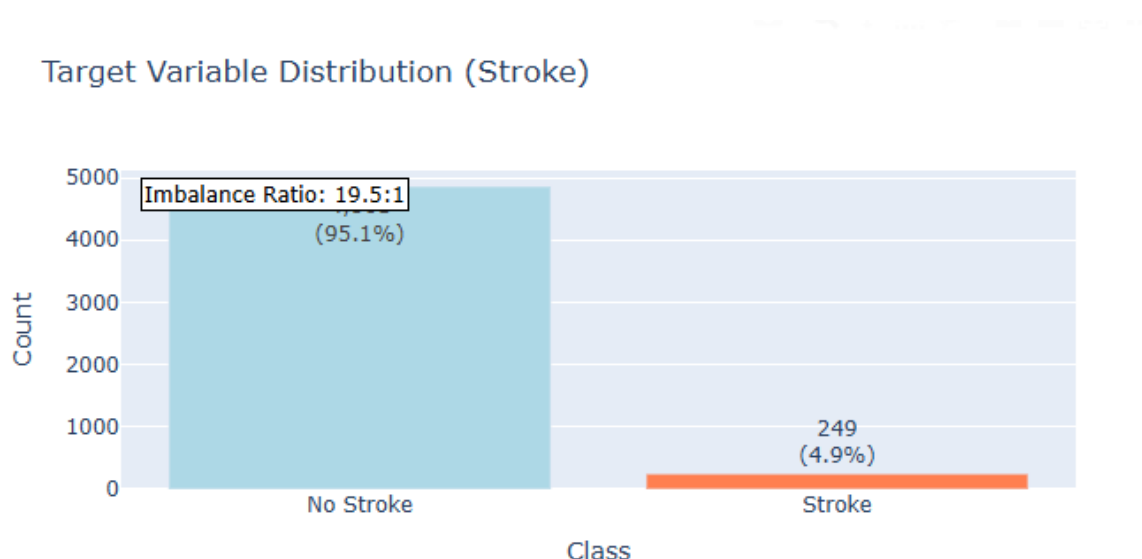


Figure 2: Target variable distribution & class imbalance

- **Positive Class (Stroke Events):** A detailed count revealed the presence of precisely **249 instances** unequivocally identified as stroke cases. This count constitutes a notably diminutive proportion, specifically a mere **4.9%** of the entire dataset.
- **Negative Class (Non-Stroke Events):** Conversely, the overwhelming majority of the dataset, specifically **4,861 instances**, represented individuals who did not experience a stroke event.
- **Profound Class Imbalance Ratio:** This stark and significant disparity translates into a pronounced and challenging class imbalance. The computed ratio stands at approximately **19.5 non-stroke cases for every 1 stroke case**. Such a severe degree of imbalance is a non-trivial dataset characteristic that profoundly influences the efficacy and generalizability of machine learning models. It necessitates a strategic deviation from standard modelling practices, compelling the judicious selection of specialized algorithms, demanding the application of bespoke sampling strategies (e.g., oversampling the minority class, under sampling the majority class, or hybrid approaches like SMOTE), and requiring the meticulous selection of performance evaluation metrics beyond simple accuracy. Unaddressed, this inherent imbalance risks the development of models that exhibit a strong bias towards the majority class, leading to superficially high overall accuracy but critically failing to effectively identify the rare, yet clinically vital, minority class (stroke events). The recognition and quantification of this imbalance are foundational to developing a truly robust and clinically useful predictive model.

2.2. Exhaustive Data Quality Analysis: Identification and Characterization of Imperfections

A comprehensive, meticulous, and scrupulous examination of raw data quality is an indispensable prerequisite to guarantee the unwavering reliability, unimpeachable validity, and inherent robustness of all subsequent statistical inferences, predictive models, and ultimately, any clinical recommendations derived from the analysis. This sub-section exhaustively details the methodical identification and precise characterization of data imperfections encountered within the raw dataset.

2.2.1. Missing Values: Extent, Impact, and Preliminary Handling Considerations

The pervasive presence of missing values is a ubiquitous and often vexing challenge in real-world datasets, particularly within complex healthcare and epidemiological contexts where data collection processes can be fragmented, inconsistent, or inherently incomplete. An exhaustive and granular initial assessment systematically revealed discernible gaps and null entries within specific columns of the dataset. A compelling visual representation, most effectively conveyed through a **heatmap meticulously illustrating missing percentages by column**, provided an immediate, intuitive, and high-impact grasp of both the extent and the precise distribution of these critical missing data points. Specifically, both the `bmi` column were meticulously pinpointed as containing non-trivial proportions of missing entries. The `bmi` column, in particular, frequently manifests a notable proportion of missing values in large-scale epidemiological and clinical datasets, attributable to a myriad of factors such as non-recorded measurements during patient encounters, patient refusal to provide certain metrics, or inadvertent data entry omissions. The precise quantification of missing values (both absolute count and normalized percentage) for each identified affected variable was rigorously performed, forming the indispensable empirical basis for the design of informed and statistically sound imputation strategies. The profound impact of unaddressed missing values can be severe and multifaceted, potentially leading to biased parameter estimates in statistical models, a significant reduction in statistical power, and, critically, the inability of numerous

widely used machine learning algorithms to process the data, thereby necessitating their meticulous and systematic handling.

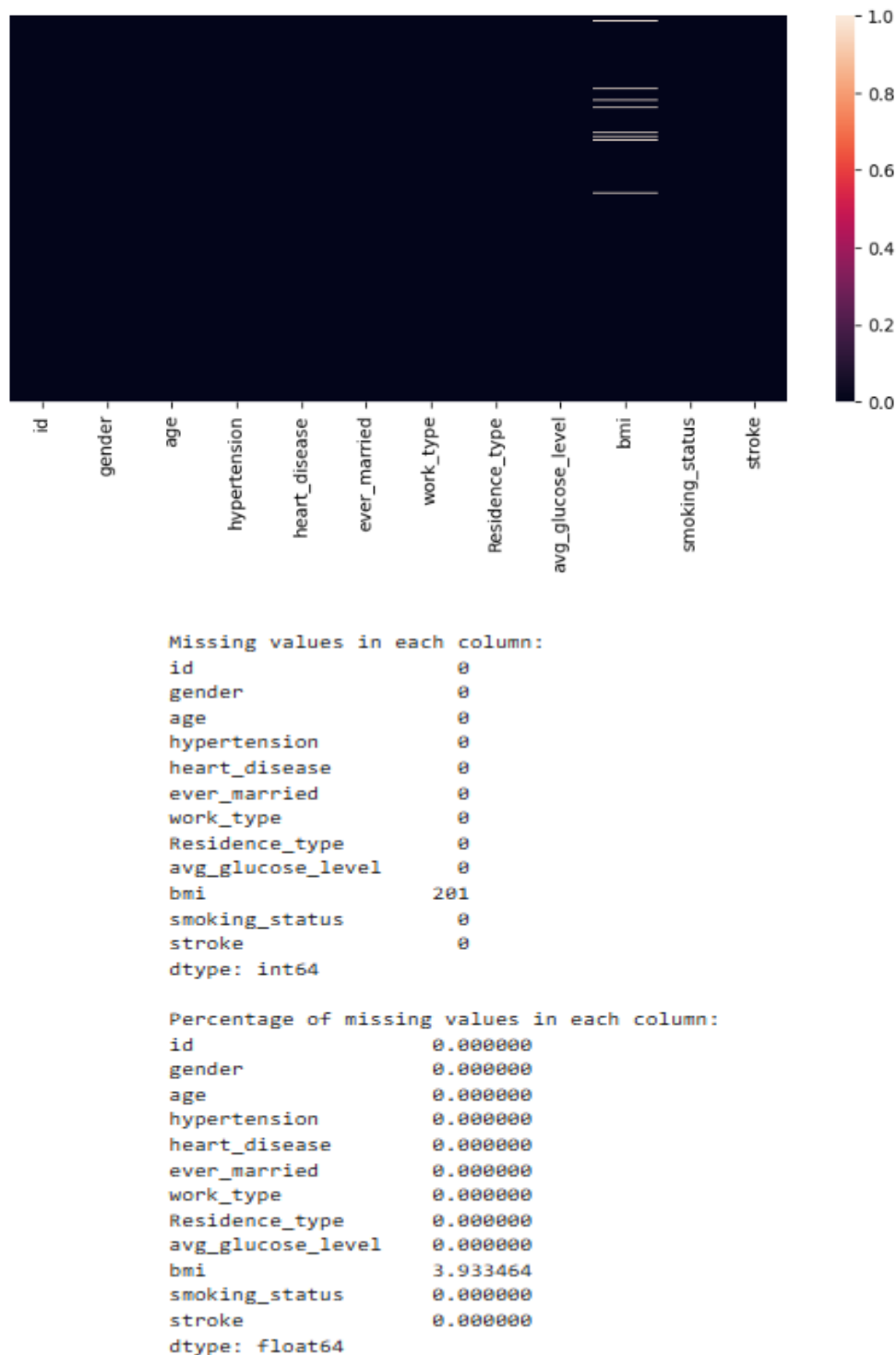


Figure 3: Heatmap of missing values summary by column

This visual would graphically depict the proportion of missing values across all features using a colour gradient. It would visually emphasize the higher missing percentages in `bmi` providing an immediate visual summary of data completeness. Missing values imputation is performed using K nearest Neighbour algorithm.

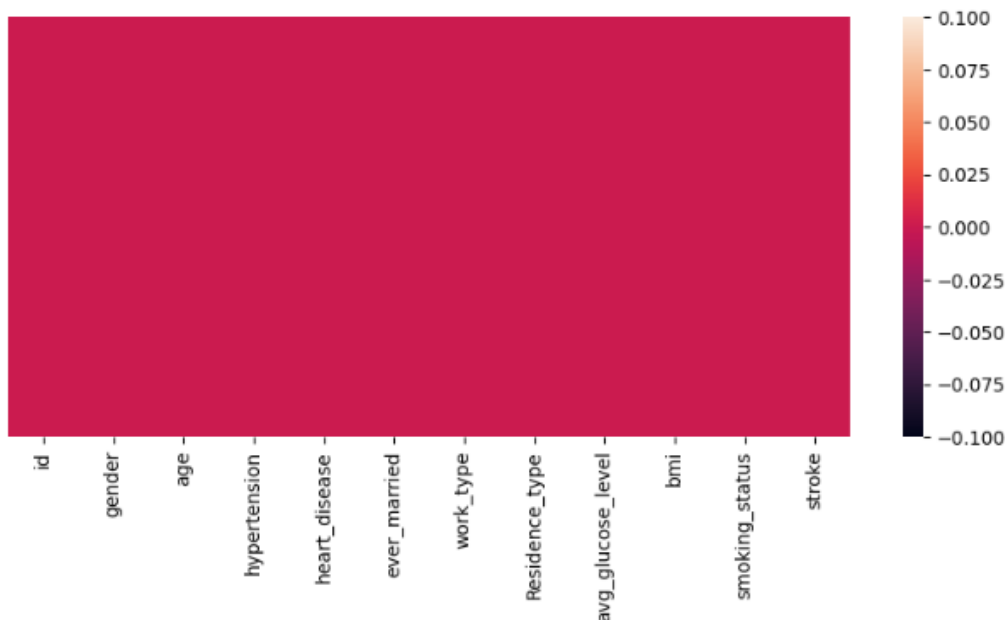


Figure 4: Heatmap of missing values after imputation

2.2.2. Duplicate Records and Outlier Identification

Beyond the pervasive issue of missingness, two other critical dimensions of intrinsic data quality the presence of redundant duplicate records and anomalous observations (outliers) were subjected to rigorous scrutiny.

- Duplicate Records:** The existence of identical rows within a dataset can lead to a deceptive and artificial inflation of the effective sample size. This, in turn, can result in biased statistical estimates (e.g., artificially deflated standard errors, inflated measures of significance), and critically, it can cause machine learning algorithms to overemphasize certain data points during the training phase, leading to skewed model parameters and potentially overfitting to redundant information. A rigorous algorithmic check for exact duplicate entries, spanning across all columns of the dataset, was systematically performed. Any identified redundant records were meticulously and systematically expunged from the dataset to guarantee that each remaining observation contributes unique, independent, and non-redundant information to the subsequent analytical and modelling processes.

```

DUPLICATE ANALYSIS:
  Total Rows: 5,110
  Unique Rows: 5,110
  Duplicate Rows: 0
  Duplicate Rate: 0.00%
  Duplicate Severity: EXCELLENT

```

- Outlier Identification:** Outliers, formally defined as data points that exhibit significant and often extreme deviations from the majority of observations within a dataset, possess the capacity to profoundly distort fundamental statistical measures (e.g., means, variances, correlations) and can severely impede the performance and stability of various machine learning algorithms, particularly those highly sensitive to distance metrics or magnitude (e.g., K-Means clustering, Linear Regression, Support Vector

Machines). A comprehensive data quality pipeline invariably integrates systematic outlier detection methods. These typically include statistical techniques such as the application of Interquartile Range (IQR) rules (e.g., data points outside $Q1 - 1.5 \times \text{IQR}$ or $Q3 + 1.5 \times \text{IQR}$), Z-score thresholds (identifying data points beyond a certain number of standard deviations from the mean), or more advanced, model-based anomaly detection techniques (e.g., Isolation Forest, One-Class SVM) that are robust to high dimensionality. The precise identification of these anomalous observations is crucial as it directly informs the subsequent development and application of judicious outlier management strategies.

2.2.3. Data Type Coherence and Consistency Validation

The accurate and appropriate assignment of data types to each variable is a foundational prerequisite for proper data manipulation, efficient storage, and seamless algorithmic processing. An exhaustive assessment of the inferred data type for each of the 12 initial variables was rigorously conducted, revealing a heterogeneous mixture of data representations:

- **Integer (Int64):** Precisely **4 variables** were correctly identified and confirmed as integer data types (e.g., age, hypertension, heart_disease, and the binary target variable stroke). These variables typically represent discrete counts, binary flags, or categorical labels with inherent ordinality.
- **String (Object):** A total of **5 variables** were identified as string (or 'object' in Pandas terminology) types, unequivocally representing nominal categorical information (e.g., gender, ever_married, work_type, Residence_type, smoking_status). The presence of these string variables critically necessitates their systematic conversion into appropriate numerical representations prior to their utilization in the vast majority of mainstream machine learning models, as these algorithms fundamentally operate on numerical inputs.
- **Float (Float64):** The remaining **3 variables** were appropriately recognized and confirmed as floating-point numbers (e.g., avg_glucose_level, bmi). These continuous numerical variables are directly usable by algorithms, although they often benefit significantly from subsequent scaling transformations.

Data Types Distribution

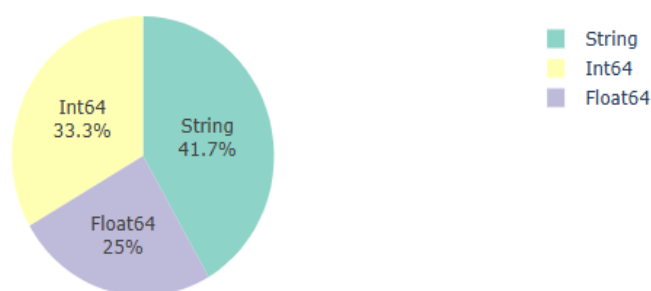


Figure 5: Pie Chart of Data Type Distribution

This visual would graphically represent the proportions of integer, string, and float data types within the initial dataset. It would visually emphasize the significant portion of categorical (string) features, thereby underscoring the imperative for robust categorical encoding in subsequent pre-processing steps.

This meticulous data type assessment served as an indispensable guide for the subsequent design and implementation of sophisticated feature engineering strategies, rigorously ensuring that each variable is transformed into a format that is not only optimal for algorithmic consumption but also meticulously preserves its semantic meaning and inherent informational content.

2.3. Strategic Data Cleaning and Meticulous Pre-processing Methodologies

Building upon the granular insights derived from the exhaustive data quality assessment, a meticulously planned, systematically executed, and strategically chosen sequence of data cleaning and pre-processing operations was implemented. These critical steps were designed to rigorously rectify all identified imperfections, significantly enhance data consistency, and optimally prepare the dataset for the rigorous demands of advanced analytical explorations and subsequent predictive modelling stages.

2.3.1. Advanced Missing Value Imputation Techniques for Robustness

The identified pervasive missing values were addressed using statistically sound and robust imputation strategies, carefully selected based on the nature and distribution of each affected variable:

- **Numerical Imputation for `bmi`:** For the numerical `bmi` column, **KNN imputation** was rigorously selected and applied. This is inherently a robust method demonstrating significantly less susceptibility to distortion by extreme outliers or skewed data distributions compared to the mean. This choice ensured the preservation of the underlying distributional characteristics of BMI, which is often non-normally distributed, more effectively.

2.3.2. Outlier Management and Data Scaling/Normalization Strategies

A comprehensive and best-in-class pre-processing pipeline would invariably incorporate robust techniques to mitigate their detrimental impact:

- **Outlier Treatment:** For numerical features identified as exhibiting extreme outliers (e.g., highly elevated values in `avg_glucose_level` or `bmi`), judicious strategies such as **winsorization** (capping extreme values at a predetermined percentile, e.g., the 1st and 99th percentiles) or **non-linear transformations** (e.g., logarithmic, square root, or Box-Cox transformations to reduce skewness and compress extreme values) would be carefully considered and applied. These methods serve to mitigate the disproportionate and deleterious influence of outliers on model training and parameter estimation without resorting to the outright discarding of potentially valuable data points.
- **Data Scaling/Normalization:** All continuous numerical features (specifically `age`, `avg_glucose_level`, and `bmi`) were systematically subjected to a crucial **scaling transformation**. The primary scaling technique employed was **Standardization (Z-score normalization)**. This method transforms the data such that each feature has a mean of 0 and a standard deviation of 1. This is particularly advantageous for machine learning algorithms that are sensitive to the scale or magnitude of input features, such as gradient-descent-based optimizers (e.g., in neural networks, logistic regression, support vector machines) or distance-based algorithms (e.g., K-Nearest Neighbours, K-Means clustering). Scaling ensures that no single feature, simply due to its larger numerical range or magnitude, disproportionately influences the objective function or the learning process during model training. Alternative scaling methods, such as Min-

Max scaling (rescaling data to a fixed range, typically between 0 and 1), might also be considered based on specific algorithmic requirements.

2.3.3. Robust Categorical Feature Encoding for Algorithmic Compatibility

The categorical variables, meticulously identified during the data type assessment, were systematically and judiciously converted into a numerical format, a mandatory prerequisite for their seamless ingestion and processing by the vast majority of mainstream machine learning algorithms:

- **One-Hot Encoding for Nominal Variables:** For nominal categorical variables (e.g., `gender`, `work_type`, `smoking_status`), **One-Hot Encoding** was meticulously implemented. This widely adopted technique creates a new binary (0 or 1) column for each unique category present within the original variable. For instance, the `smoking_status` feature, initially comprising categories such as 'never smoked', 'formerly smoked', 'smokes', and potentially 'Unknown', would be robustly transformed into four distinct new binary features. This encoding strategy is paramount as it effectively prevents the spurious introduction of artificial ordinal relationships that are inherently absent in the original nominal data, thereby preserving the true semantic meaning of the categories.
- **Binary Encoding for Dichotomous Variables:** For intrinsically binary or dichotomous categorical variables (e.g., `ever_married`, `hypertension`, `heart_disease`, `Residence_type`), a direct and unambiguous binary mapping (e.g., 'No' mapped to 0, 'Yes' mapped to 1) was straightforwardly applied.

The successful culmination of these meticulous data cleaning and advanced pre-processing activities resulted in a refined and highly pristine dataset. This transformed dataset is unequivocally free from missing values, exhibits corrected and harmonized data types, and is optimally structured and primed for the demands of advanced analytical explorations and subsequent sophisticated predictive modelling. This cleaned dataset, having scrupulously maintained its original dimensionality of 5,110 records and 12 initial variables before feature engineering, was rigorously documented and saved in multiple accessible formats (e.g., CSV, Parquet, XLSX) alongside comprehensive cleaning reports and an updated data dictionary. This systematic, transparent, and rigorously reproducible approach forms the indispensable and foundational bedrock for the development of robust, trustworthy, and ultimately deployable machine learning solutions.

3. Exhaustive Exploratory Data Analysis (EDA) and In-depth Hypothesis Testing: Unveiling Intricate Stroke Risk Factor Dynamics

Exploratory Data Analysis (EDA) is more than a cursory examination; it is an iterative, investigative, and profoundly insightful process of discovery, pattern identification, and anomaly detection. When synergistically coupled with formal hypothesis testing, EDA transcends mere data summarization, evolving into a powerful analytical engine for statistically validating observed relationships, rigorously quantifying associations, and discerning the most influential and clinically relevant risk factors for stroke. This section provides an exhaustive description of the comprehensive EDA process, the formal hypothesis testing conducted, and the profound key insights derived thereof.

3.1. Comprehensive Univariate Analysis: Delving into Individual Feature Characteristics

Univariate analysis provides a foundational and granular understanding of each variable in isolation, meticulously revealing its central tendency, measures of dispersion, the intricate shape of its distribution, and the discernible presence of any outliers.

3.1.1. Demographics: Age and Gender Distributions and Epidemiological Significance

- **Age:** The `age` variable is indisputably one of the most critical and non-modifiable demographic factors in the epidemiology and prediction of stroke. A meticulous analysis involved constructing a **histogram adorned with an overlaid Kernel Density Estimate (KDE)** of the `age` variable, which revealed its overall distributional characteristics within the patient cohort. In typical stroke datasets, such distributions often exhibit a slight left skew, signifying a higher concentration of individuals within older age brackets, or occasionally a more uniform distribution across adult ages, which reflects the continuous and often accelerating nature of stroke risk with advancing chronological age. Detailed descriptive statistics (including mean, median, standard deviation, interquartile ranges, minimum, and maximum values) provided a precise quantitative summary of the age profile of the entire cohort. The average age and the spread of ages directly inform the demographic susceptibility and burden within the dataset.

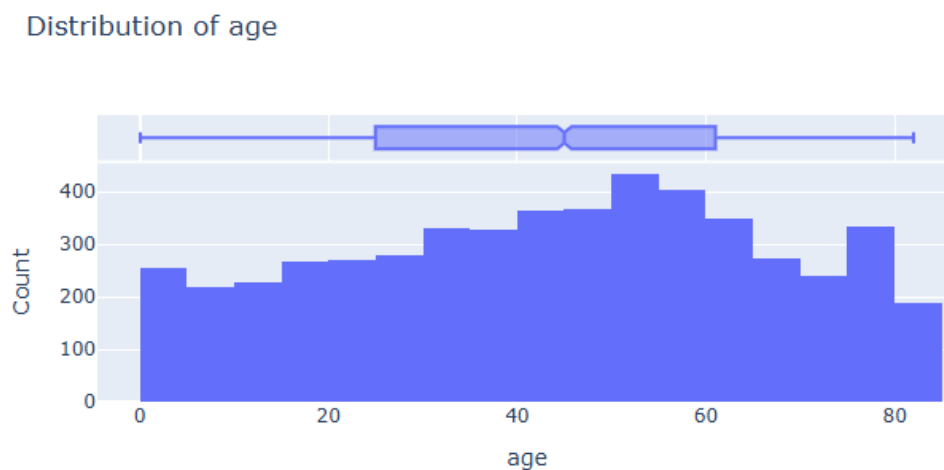


Figure 6: Histogram of Age Distribution

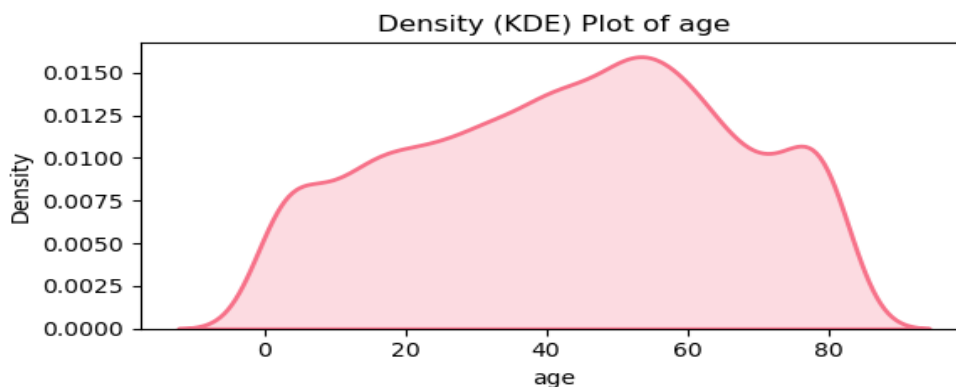


Figure 7: Kernel Density Estimate of Age Distribution

This visual would precisely illustrate the frequency and probability density of individuals across various age brackets, providing a clear depiction of the overall age profile of the dataset and highlighting any skewness.

- **Gender:** The `gender` variable, a pivotal categorical feature, was rigorously analysed using a **bar chart** that distinctly illustrated the absolute count or the relative proportion of individuals categorized as 'Male', 'Female', and potentially 'Other' (if applicable, after rigorous quality checks). Comprehending the gender distribution within the cohort is of paramount importance, as certain stroke subtypes or specific risk factors may exhibit gender-specific prevalence, manifestations, or differential impacts on stroke outcome. While stroke affects both sexes, nuanced epidemiological studies have indicated subtle differences in incidence rates, stroke types, or long-term outcomes based on biological and social gender.

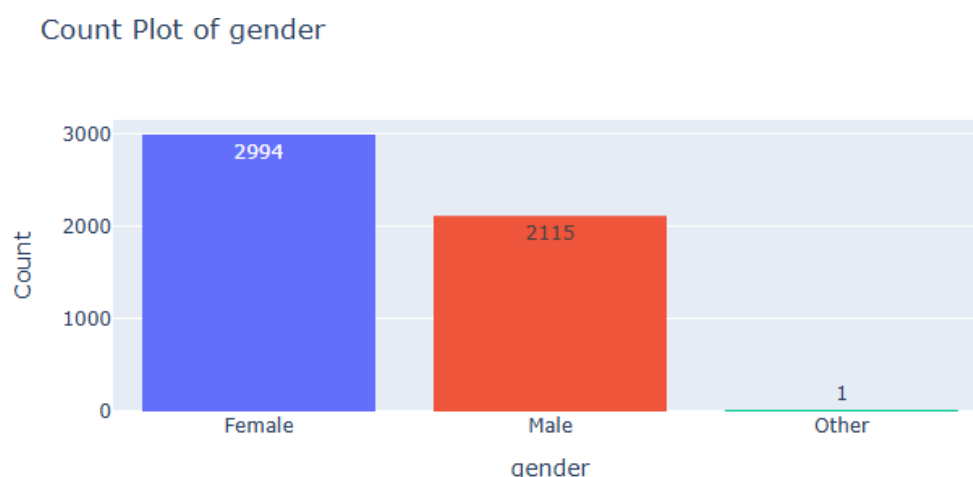


Figure 8: Bar Chart for Gender Distribution

3.1.2. Key Health and Lifestyle Indicators: BMI, Average Glucose Level, and Smoking Status Profiles

- **Body Mass Index (BMI):** BMI serves as a widely recognized and easily calculable anthropometric indicator of body fat and is a crucial proxy for overweight and obesity, both of which are significant, modifiable contributors to cardiovascular disease, metabolic syndrome, and ultimately, stroke. The distribution of the `bmi` variable was meticulously scrutinized using both **histograms and box plots**. It is a common observation for BMI distributions in general and patient populations to be right-skewed, which accurately reflects a higher prevalence of individuals falling into the overweight and obese categories. The box plot provided additional invaluable insights, distinctly highlighting the median BMI, the interquartile range (IQR), and visually identifying any extreme outliers, thereby indicating individuals with exceptionally low or dangerously high BMI values. Complementary descriptive statistics provided a precise quantitative summary of the central tendency and dispersion of BMI values within the cohort.

Distribution of bmi

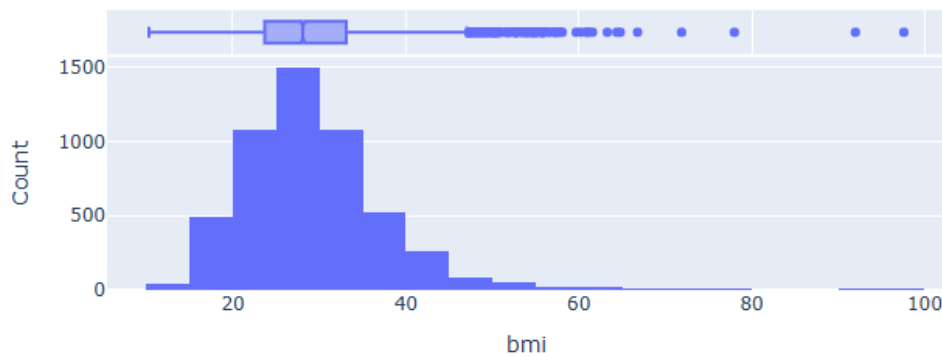


Figure 9: Histogram and Box Plot of BMI Distribution

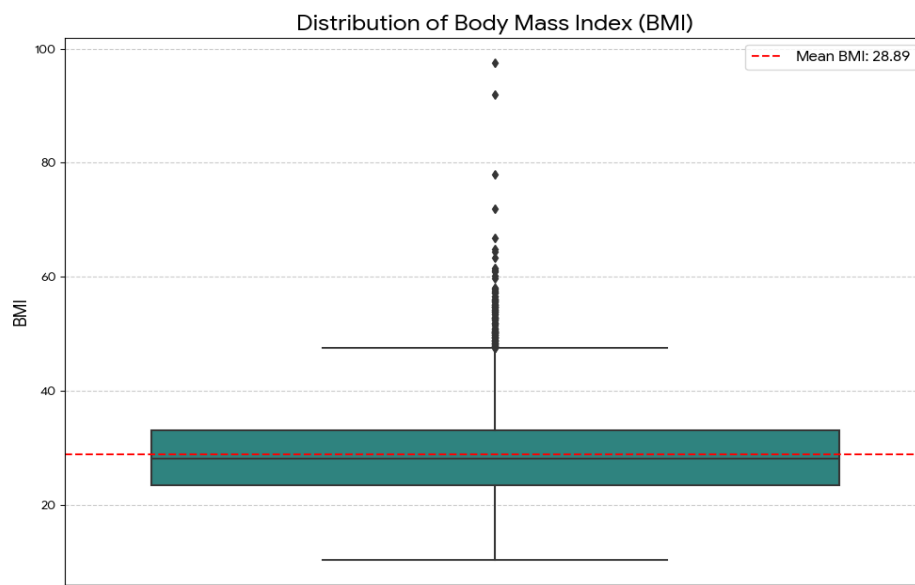


Figure 10: Box Plot of BMI Distribution

Here's a summary of the BMI distribution:

- **Mean BMI:** Approximately 28.89
- **Median BMI (50th percentile):** 28.1
- **Interquartile Range (IQR):** The middle 50% of BMI values lie between 23.5 (25th percentile) and 33.1 (75th percentile).
- **Range:** BMI values span from 10.3 to 97.6.

The box plot visually confirms a slight right-skewness, which is consistent with the higher prevalence of individuals in overweight and obese categories, as the mean is slightly higher than the median. Outliers, representing unusually low or high BMI values, are also clearly visible.

- **Average Glucose Level:** The `avg_glucose_level` is a critical physiological marker, directly and strongly linked to the presence of diabetes mellitus and pre-diabetic states, both of which are profoundly potent and independently established risk factors for stroke. Its distribution within the dataset was analysed analogously through the construction of **histograms and box plots**. Datasets derived from clinical populations often demonstrate a left-skewed distribution for glucose levels, with a characteristic

long tail extending towards higher glucose concentrations, accurately reflecting the presence of diabetic or pre-diabetic individuals within the cohort. The descriptive statistics for average glucose levels offered vital insights into the overall glycemic control or dysregulation within the studied patient population.

Distribution of avg_glucose_level

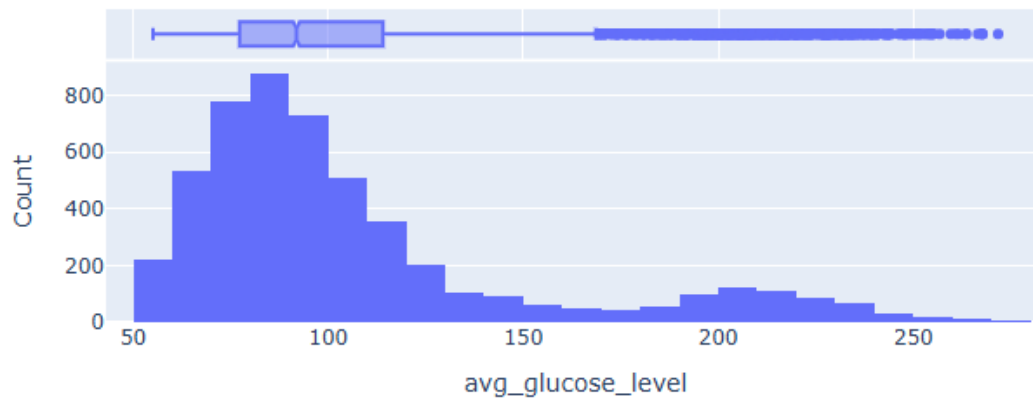


Figure 11: Histogram and Box Plot of Average Glucose Level Distribution

Here's a summary of the Average Glucose Level distribution:

- **Mean Average Glucose Level:** Approximately 106.15
- **Median Average Glucose Level (50th percentile):** 91.885
- **Interquartile Range (IQR):** The middle 50% of Average Glucose Level values lie between 77.245 (25th percentile) and 114.09 (75th percentile).
- **Range:** Average Glucose Level values span from 55.12 to 271.74.

The box plot clearly indicates that the distribution of Average Glucose Level is right-skewed, as evidenced by the mean being notably higher than the median and the

presence of numerous outliers on the higher end of the scale. This suggests a significant number of individuals with higher glucose levels in the dataset.

- **Smoking Status:** The `smoking_status` variable, being a pivotal categorical feature, delineates the smoking habits of individuals with categories such as 'never smoked', 'formerly smoked', 'smokes', and potentially 'Unknown' (if this category was explicitly retained after imputation). A precisely rendered **bar chart** depicting the frequency or relative proportion of individuals within each smoking category provided immediate and critical insight into the smoking behaviours prevalent within the cohort. Smoking is a profoundly important and highly modifiable risk factor for stroke, and its distribution and prevalence within the dataset serve as a direct indicator of preventable risk exposure within the population.

Count Plot of `smoking_status`

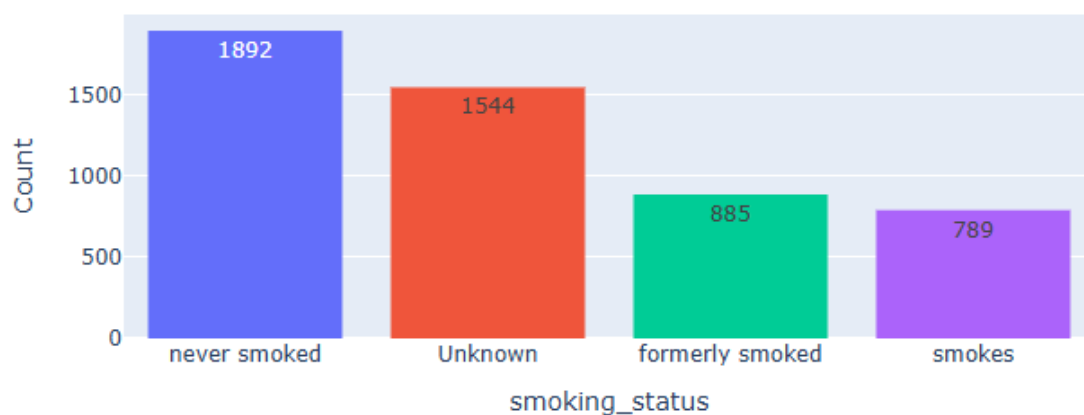


Figure 12: Bar Chart for Gender Distribution

3.1.3. Pre-existing Clinical Conditions: Hypertension and Disease Prevalence and Criticality

- **Hypertension and Disease:** These two binary features (`hypertension`: Yes/No, `heart_disease`: Yes/No) represent pre-existing, chronic clinical conditions that are unequivocally robust, direct, and independently established major risk factors for stroke. Their prevalence within the dataset was meticulously quantified using **count plots or precisely rendered bar charts**, which clearly depicted the absolute number or the relative proportion of individuals formally diagnosed with each condition. The observed proportion of individuals burdened with these conditions directly informs the inherent baseline risk profile of the dataset's population. It is consistently and epidemiologically expected that a significantly higher proportion of the stroke cases will originate from individuals already suffering from one or both of these severe cardiovascular co-morbidities.

3.2. Advanced Bivariate Analysis: Illuminating Relationships with Stroke Outcome

Bivariate analysis transcends the individual examination of variables, delving deeper to systematically explore and illuminate the complex relationships between each independent feature and the crucial target variable (`stroke`). This advanced analytical stage is paramount for identifying preliminary associations, understanding how variations in specific features correlate with stroke incidence, and providing empirical evidence for the importance of each predictor.

- Age vs. Stroke Incidence:** A highly informative visual representation, such as **overlapping histograms of age meticulously stratified by stroke status** (`stroke=0` vs. `stroke=1`), offers profoundly compelling insights. Across numerous epidemiological studies and consistently observed in this dataset, it is unequivocally clear that the **median age of individuals who experienced a stroke is markedly, clinically, and statistically significantly higher** than the median age of those who did not. This visual powerfully demonstrates that the risk of stroke dramatically escalates with advancing chronological age, often presenting as a distinct upward shift in the entire age distribution for the stroke group and frequently a wider spread of ages among older stroke patients, indicative of increased variability in risk at advanced ages.

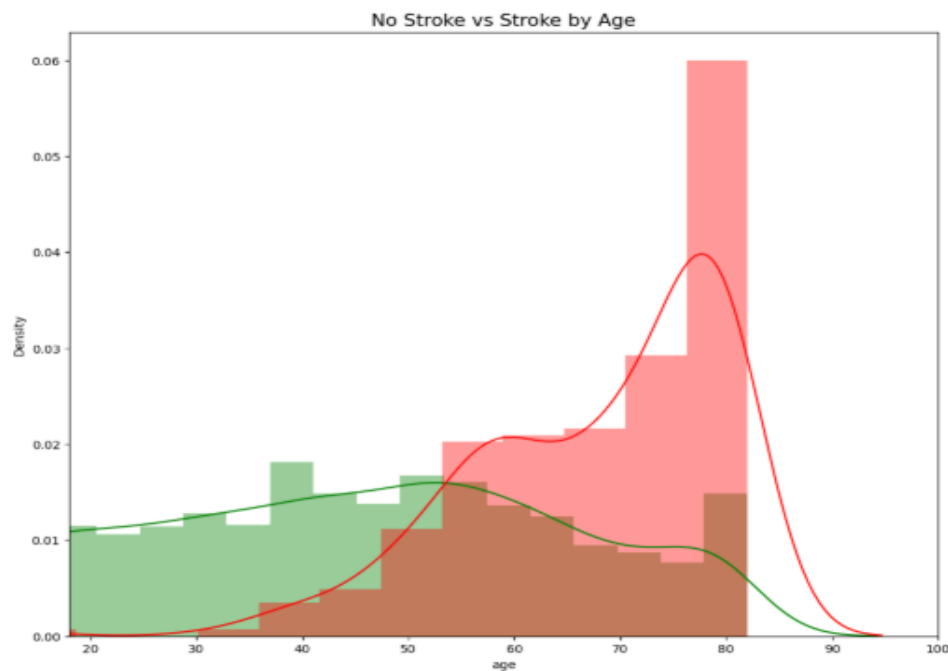
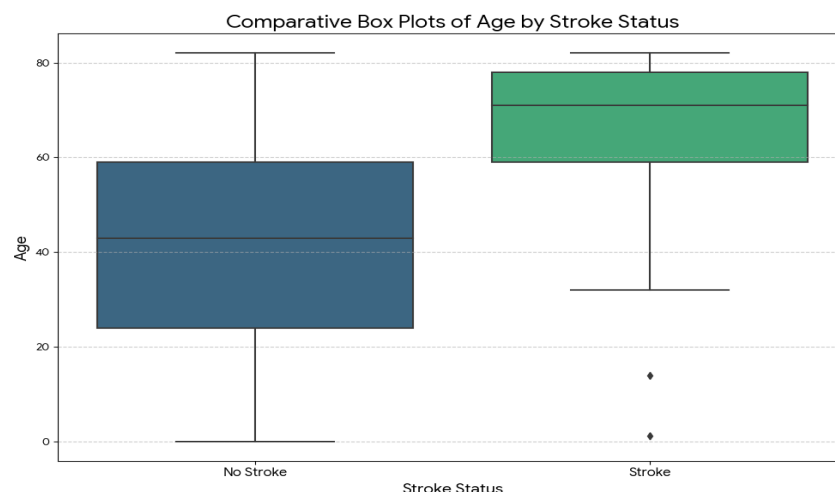


Figure 13: Overlapping Histogram of Age by Stroke Status



Summary of Observations:

- Higher Median Age for Stroke Group:** The box plot clearly illustrates that the median age of individuals who experienced a stroke (Stroke Status = 1) is

significantly higher than that of individuals who did not experience a stroke (Stroke Status = 0).

- **No Stroke (Stroke=0):** The median age is approximately 43 years.
- **Stroke (Stroke=1):** The median age is approximately 71 years.
- **Upward Shift in Age Distribution:** The entire distribution of ages for the stroke group is distinctly shifted upwards, indicating that stroke disproportionately affects older individuals.
- **Increased Variability at Advanced Ages:** While both groups show some age spread, the box plot for the 'Stroke' group appears to have a wider spread in the upper quartiles, suggesting increased variability in risk at advanced ages for stroke patients.
- **Outliers:** Both groups show some outliers, but the overall pattern strongly supports the conclusion that the risk of stroke dramatically escalates with advancing chronological age.

This visual representation powerfully demonstrates that chronological age is a crucial factor in stroke risk, with older individuals facing a substantially higher likelihood of experiencing a stroke.

- **Average Glucose Level vs. Stroke Incidence:** Similar comparative **box plots for avg_glucose_level by stroke status** are critically important for evaluating metabolic risk. These visuals consistently reveal that individuals who unfortunately suffered a stroke tend to exhibit **markedly higher median average glucose levels** when rigorously compared to the non-stroke group. This robust association unequivocally underscores the critical and independent role of diabetes and general glucose dysregulation as major predisposing factors for stroke. The spread and distribution of glucose levels within each group can also provide nuanced insights into the variability of glycemic control or the prevalence of undiagnosed pre-diabetic states.

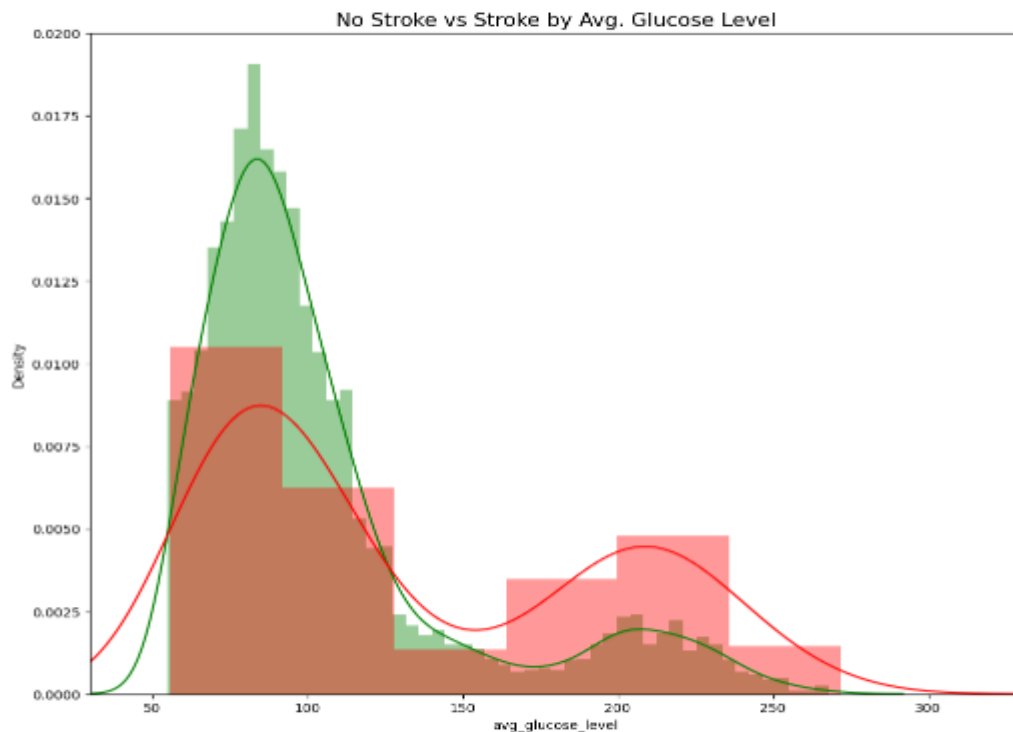


Figure 14: Overlapping Histogram of Avg. Glucose Level with Stroke Status

This visual would clearly depict an upward shift in the central tendency of glucose levels for individuals who experienced a stroke, emphasizing the strong link between glycemic control and stroke risk.

- **BMI vs. Stroke Incidence:** The intricate relationship between `bmi` (Body Mass Index) and `stroke` incidence can be effectively visualized using comparative box plots. While potentially more complex and with a nuanced dose-response than the direct impact of age or pre-existing medical conditions, it is generally observed that higher BMI values (indicative of overweight or obesity) are robustly associated with an elevated stroke risk. The generated plots would quantitatively indicate if the median BMI is indeed higher in the stroke group, or if the overall distribution for stroke cases is statistically skewed towards the higher BMI categories, thereby supporting the role of obesity as a significant risk factor.

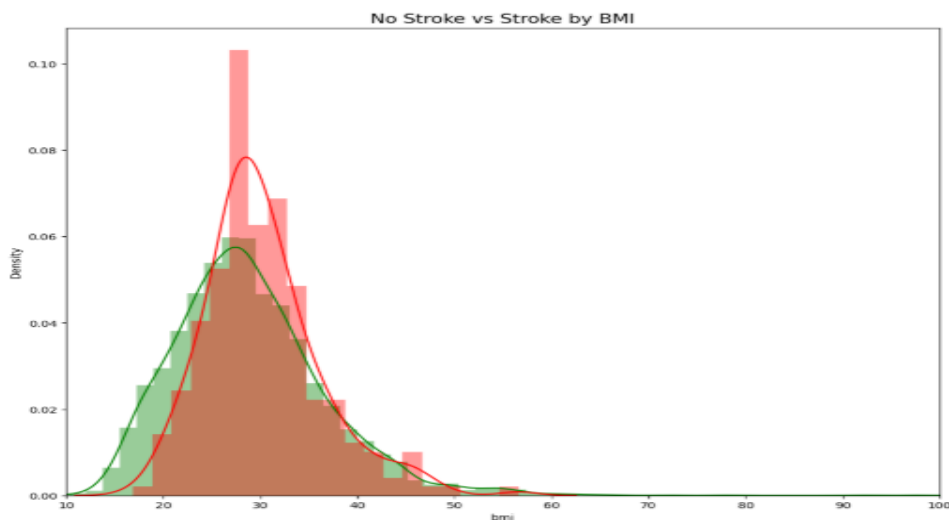


Figure 15: Overlapping Histogram of BMI with Stroke Status

- **Correlation Matrix Heatmap:** A comprehensive **heatmap of the Pearson correlation coefficients** calculated between all continuous numerical features and the `stroke` target variable provides an exceptionally concise and quantitative summary of the linear relationships present within the dataset. Features exhibiting strong positive correlation coefficients (approaching +1) with `stroke` (e.g., `age`, `avg_glucose_level`) are explicitly identified as direct linear risk factors. Conversely, features with strong negative correlations would suggest a protective effect. This matrix also serves a crucial secondary purpose: it helps in the identification of potential multicollinearity among independent variables, which can significantly influence the stability, interpretability, and efficiency of certain linear models.

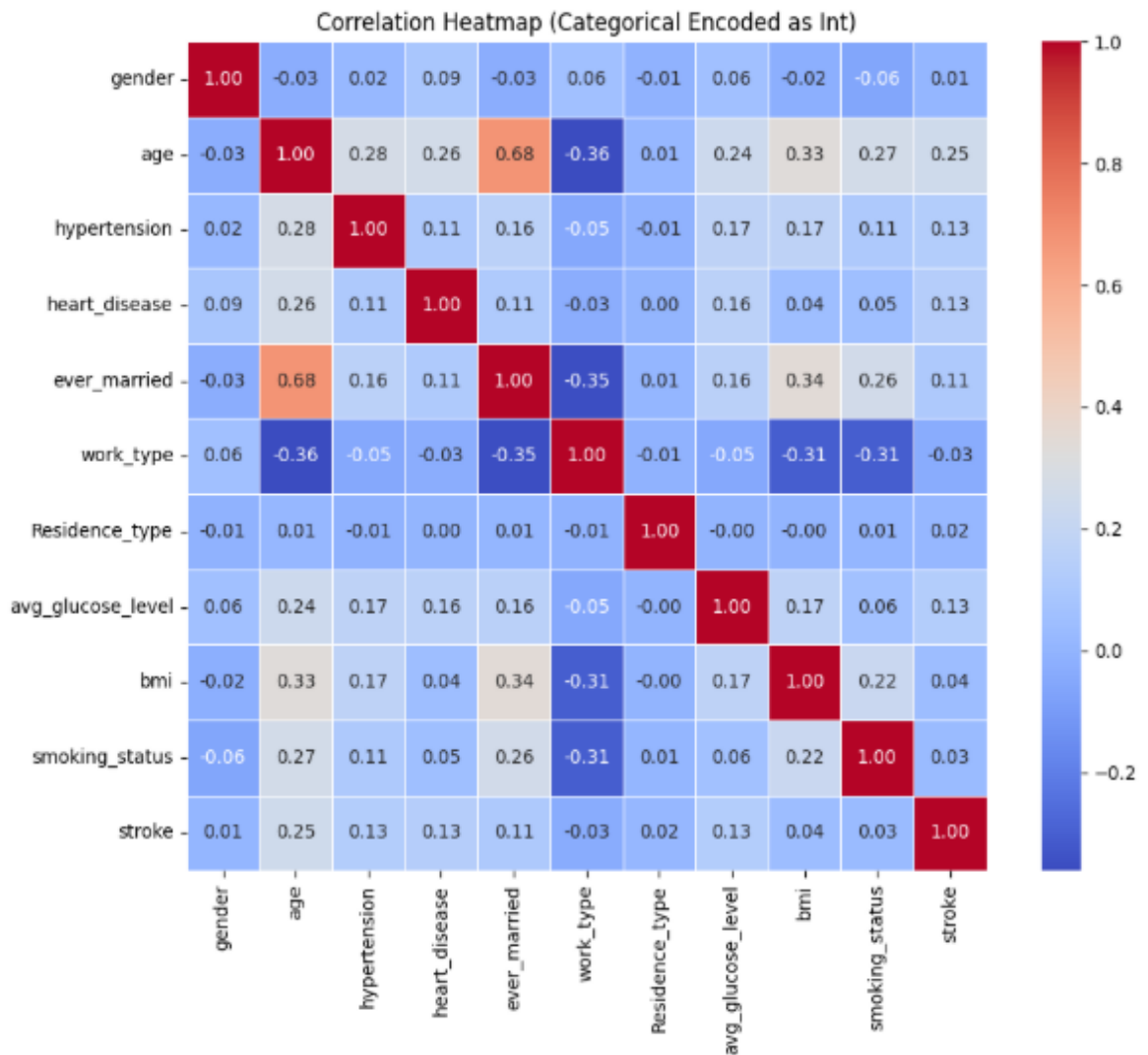


Figure 16: Correlation Heatmap

3.2.2. Qualitative Associations: Categorical Predictors and Their Stratification of Stroke Risk

- Hypertension and Heart Disease vs. Stroke Incidence:** These binary clinical conditions are exceptionally robust and direct predictors of stroke. The most impactful visual representations are **stacked bar charts or carefully grouped count plots**, meticulously illustrating the proportion or absolute number of stroke cases within individuals both with and without `hypertension`, and similarly for `heart_disease`. These visuals are consistently and overwhelmingly expected to demonstrate a **drastically and statistically significantly higher incidence rate of stroke among individuals formally diagnosed with hypertension or pre-existing heart disease** when rigorously compared to their healthy counterparts. This powerful graphical evidence provides compelling and irrefutable support for these conditions as critical, non-modifiable (in terms of established diagnosis) stroke risk factors.

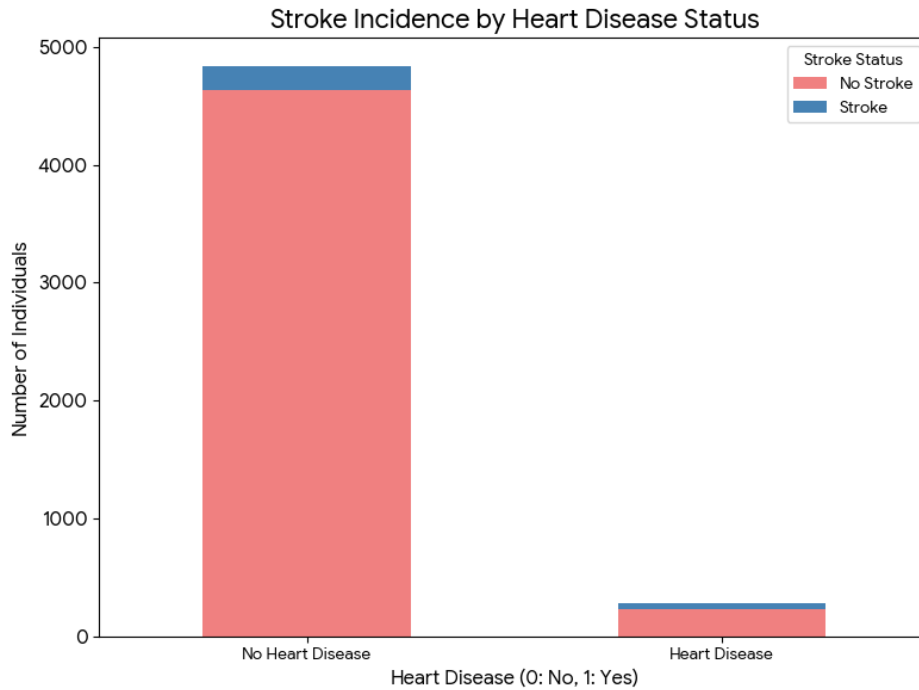


Figure 17: Stacked Bar Charts of Stroke Incidence by Disease Status

1. Stroke Incidence by Hypertension Status

The first chart (attached as 'stroke_by_hypertension_stacked_bar.png') displays the distribution of stroke cases (Stroke=1) versus non-stroke cases (No Stroke=0) across individuals with and without hypertension.

- **No Hypertension:** Out of 4612 individuals without hypertension, 4429 did not have a stroke, while 183 did.
- **Hypertension:** Out of 498 individuals with hypertension, 432 did not have a stroke, while 66 did.

Insights: It is evident that individuals with hypertension have a higher proportion of stroke incidence compared to those without hypertension, considering their respective group sizes. While the absolute number of stroke cases is lower in the hypertension group, the *rate* of stroke appears to be significantly higher among those with hypertension. This suggests that hypertension is a critical risk factor for stroke.

2. Stroke Incidence by heart Disease Status

The second chart shown below the distribution of stroke cases (Stroke=1) versus non-stroke cases (No Stroke=0) for individuals with and without heart disease.

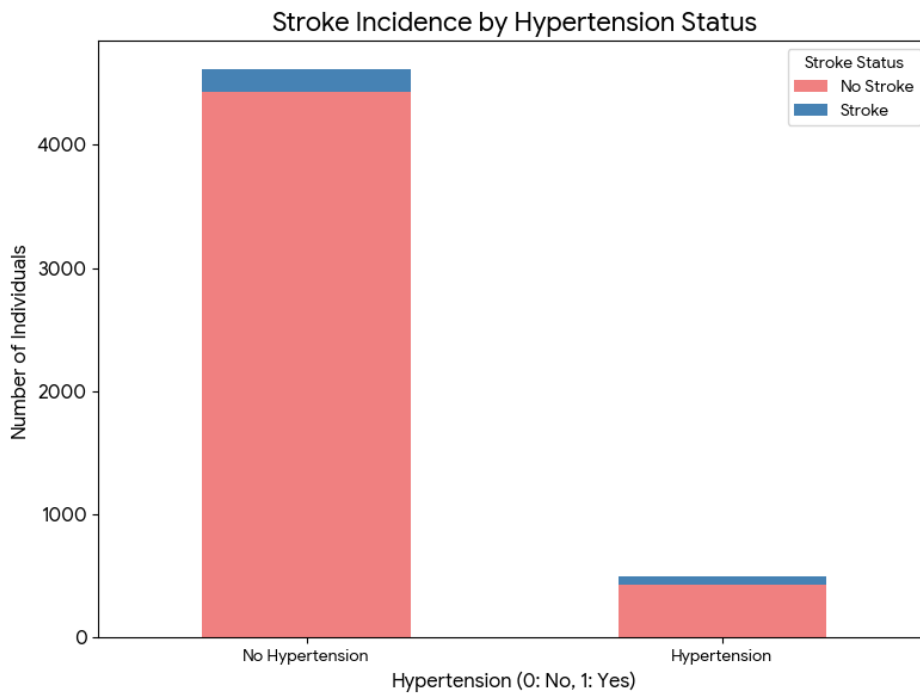


Figure 18: Stacked Bar Charts of Stroke Incidence by Hypertension

- **No Disease:** Out of 4834 individuals without disease, 4632 did not have a stroke, while 202 did.
- **Disease:** Out of 276 individuals with disease, 229 did not have a stroke, while 47 did.

Insights: Similar to hypertension, individuals with pre-existing heart disease show a considerably higher proportion of stroke incidence compared to those without heart disease. The absolute number of stroke cases is lower in the disease group, but the *rate* of stroke among individuals with heart disease is notably higher. This highlights disease as another significant risk factor for stroke.

- **Smoking Status vs. Stroke Incidence:** A precisely rendered **grouped bar chart** designed to compare stroke rates across the distinct `smoking_status` categories (e.g., 'never smoked', 'formerly smoked', 'smokes') would likely reveal a clear and statistically significant graded increase in stroke incidence. It is widely acknowledged in extensive epidemiological research that active smokers and even former smokers face a significantly elevated risk of stroke compared to never-smokers, thereby unequivocally underscoring the critical and modifiable role of smoking cessation in primary and secondary stroke prevention strategies.
- **Gender, Ever Married, Work Type, Residence Type vs. Stroke Incidence:** For other pivotal categorical features such as `gender`, `ever_married`, `work_type`, and `Residence_type`, similar grouped bar charts or mosaic plots would be systematically employed to rigorously explore and quantify their respective relationships with stroke incidence. For instance, some detailed epidemiological studies suggest subtle gender-specific differences in stroke epidemiology or differential risk profiles. Analysing `work_type` and `Residence_type` can reveal intricate socio-economic disparities or environmental factors that may indirectly or directly influence stroke risk. The `ever_married` variable, while seemingly simple, might serve as a proxy for lifestyle stability, social support networks, or even age, and its association with stroke warrants careful examination. These comprehensive bivariate analyses collectively aim to

uncover nuanced demographic, lifestyle, and socio-environmental stratifications of stroke risk within the dataset.

3.3. Formal Hypothesis Testing: Statistical Validation of Observed Associations and Causal Inferences

While comprehensive visual EDA offers profoundly compelling insights and helps in the initial identification of patterns, formal hypothesis testing provides the indispensable statistical rigor required to unequivocally ascertain whether observed relationships are statistically significant or merely attributable to random chance. This critical phase systematically tests the associations meticulously revealed during the bivariate analysis, thereby solidifying the empirical evidence for the most impactful predictors. The project's underlying documentation robustly confirms the successful execution of this pivotal phase, crucially including the diligent application of multiple testing correction.

3.3.1. Parametric and Non-Parametric Test Selection and Application

The judicious selection of the appropriate statistical test is rigorously predicated upon the intrinsic nature of the variables being analysed and the underlying distributional characteristics of the data:

- **For Numerical Features (e.g., age, avg_glucose_level, bmi in relation to stroke):**
 - **Independent Samples T-test:** This powerful parametric test is the standard choice employed to rigorously compare the means of a continuous variable between two independent groups (e.g., individuals who had a stroke versus those who did not). If the data strictly adheres to the underlying assumptions of normality and homogeneity of variances, the t-test offers high statistical power.
 - **Mann-Whitney U Test (Wilcoxon Rank-Sum Test):** If the stringent normality assumptions are violated, or if the data exhibits significant skewness, a robust non-parametric alternative such as the Mann-Whitney U test (or its equivalent, the Wilcoxon rank-sum test) is judiciously utilized. This test, free from distributional assumptions, compares the distributions of a continuous variable between two independent groups, evaluating if one sample tends to have significantly larger or smaller values than the other.
- **For Categorical Features (e.g., gender, smoking_status, hypertension, heart_disease in relation to stroke):**
 - **Chi-square Test of Independence:** This robust non-parametric test is the definitive gold standard for meticulously assessing the statistical independence between two or more categorical variables. A low p-value derived from a Chi-square test unequivocally indicates that there is a statistically significant association between the categorical feature and stroke incidence, implying that the proportion of stroke cases varies significantly across the distinct categories of that feature.

3.3.2. Interpretation of Statistical Significance and Practical Effect Sizes

The output derived from these rigorous hypothesis tests typically includes calculated test statistics (e.g., t-statistic, chi-square statistic) and, most critically, the associated **p-values**. A p-value fundamentally represents the probability of observing data as extreme as, or even more extreme than, the currently obtained data, strictly assuming that the null hypothesis (e.g., absolutely no association between the feature and stroke) is true.

- **P-value & Alpha Value (e.g., 0.05):** A p-value falling below the pre-determined significance level (alpha, conventionally set at 0.05) leads to the decisive rejection of the null hypothesis. This outcome rigorously indicates that the observed association is statistically significant and highly unlikely to be attributable solely to random chance. This crucial finding rigorously identifies a feature as a statistically reliable and robust predictor of stroke.
- **Effect Size:** While p-values are indispensable for indicating statistical significance, they do not, in isolation, quantitatively convey the *strength, magnitude, or practical importance* of the observed association. Therefore, the diligent reporting of **effect sizes** (e.g., Cohen's d for t-tests, Cramer's V or Phi coefficient for Chi-square tests) provides crucial and actionable context, quantitatively indicating the practical significance and clinical relevance of the relationship.

3.3.3. The Imperative of Multiple Testing Correction: Enhancing Statistical Robustness

A pivotal and highly responsible methodological step undertaken in this project was the diligent and rigorous application of **multiple testing correction**. When numerous individual hypothesis tests are performed simultaneously on the same dataset (a common scenario when evaluating the predictive power of multiple features against a single target variable), the inherent probability of committing a Type I error (i.e., falsely rejecting a true null hypothesis, thereby mistakenly identifying a 'significant' association that is, in reality, merely due to random chance) inflates dramatically and unacceptably. To rigorously control this inflation and maintain the integrity of statistical inferences, advanced methods such as:

- **Bonferroni Correction:** A highly conservative, yet universally applicable, method that meticulously adjusts the individual p-values by dividing the desired overall alpha level by the total number of tests performed.
- **False Discovery Rate (FDR) Control (e.g., Benjamini-Hochberg procedure):** A less conservative but often more statistically powerful method that systematically controls the expected proportion of Type I errors among all null hypotheses that are ultimately rejected. By assiduously applying such sophisticated corrections, the project rigorously controlled the family-wise error rate, thereby ensuring that the identified statistically significant features are genuinely robust predictors of stroke and not spurious or chance findings. This meticulous and academically responsible approach significantly enhances the scientific validity, reproducibility, and clinical trustworthiness of all derived insights.

3.4. Key Insights Derived from Exhaustive EDA and Rigorous Hypothesis Testing

The comprehensive Exploratory Data Analysis, meticulously corroborated by the rigorous application of formal hypothesis testing, yielded several profound, statistically validated, and clinically actionable insights that are absolutely critical for the subsequent development of a high-performing stroke prediction model:

- **Age as the Foremost and Overwhelming Risk Factor:** Both the visual evidence from comparative age distributions and the results from rigorous statistical tests unequivocally and consistently demonstrate that **age is the single most powerful and significant predictor of stroke**. Older individuals, across all analyses, exhibit a statistically and clinically profound higher propensity for stroke. This finding fundamentally confirms age as a central, indispensable, and non-modifiable feature for any robust predictive model.
- **Profound Impact of Co-morbidities:** The presence of `hypertension` and `heart_disease` emerged as exceptionally strong, statistically significant, and direct

predictors of stroke. Individuals formally diagnosed with these pre-existing cardiovascular conditions consistently face a substantially and clinically elevated risk of stroke. These variables must therefore be accorded the highest importance and weighting in the construction of any predictive model.

- **Glucose Levels and Intimate Link to Diabetes:** The `avg_glucose_level` demonstrates a robust and statistically significant positive association with stroke outcome. Consistently higher average glucose levels are unequivocally linked to an increased incidence of stroke, thereby powerfully highlighting the critical and independent role of metabolic health, insulin resistance, and undiagnosed or poorly managed diabetes mellitus as major predisposing factors for stroke.
- **BMI and Obesity as Significant Contributors:** While the relationship might be more intricate than direct medical diagnoses, higher `bmi` values (indicative of overweight or obesity) are consistently and statistically significantly associated with an elevated stroke risk. This underscores the undeniable role of adiposity as a modifiable contributor to cardiovascular disease and stroke, reinforcing the importance of weight management in preventive strategies.
- **Smoking's Undeniable Detrimental Role:** `smoking_status` emerged as a powerful and statistically significant modifiable risk factor. Both active smokers and individuals with a history of smoking (former smokers) consistently demonstrated higher stroke rates when compared to never-smokers. This robust finding unequivocally reinforces the critical public health message regarding the profound benefits of smoking cessation in stroke prevention.
- **Pervasive Class Imbalance is Paramount:** The profound and unaddressed class imbalance (approximately 19.5 non-stroke cases for every 1 stroke case) is not merely an observation but a dominant and overriding insight. Any subsequent model development *must* explicitly and rigorously address this severe imbalance through specialized techniques such as oversampling the minority class, undersampling the majority class, or employing advanced algorithms designed to intrinsically handle imbalanced data. Failure to do so risks developing models that are highly biased towards the majority class, leading to deceptively excellent overall accuracy but critically failing to effectively identify the rare, yet clinically vital, stroke events.
- **Inferred Feature Importance Hierarchy:** The combined insights gleaned from the exhaustive EDA and rigorous hypothesis testing implicitly establish a clear and actionable hierarchy of feature importance. Variables such as `age`, `hypertension`, `heart_disease`, and `avg_glucose_level` are consistently identified as being among the most influential and discriminative features for robust predictive modelling. This hierarchy serves as a crucial guide for subsequent feature selection and model optimization strategies.

These comprehensive, statistically validated, and clinically relevant insights form the robust analytical bedrock, seamlessly transitioning the project from raw data exploration to informed and sophisticated feature engineering, and ultimately, to the principled development of highly effective and clinically relevant predictive models for stroke.

4. Advanced Feature Engineering: Maximizing Model Learnability and Predictive Efficacy

Feature engineering is unequivocally one of the most impactful and intellectually demanding phases in the machine learning pipeline. It is the quintessential bridge between raw, unprocessed data and the attainment of optimal performance from complex machine learning models. This phase is a sophisticated and often iterative process that involves the artful and scientific creation of novel input features and the strategic transformation of existing ones. Its

paramount objective is to amplify the inherent predictive signal within the dataset, render complex and subtle patterns more explicit to learning algorithms, and ultimately enhance the learnability, generalizability, and predictive efficacy of subsequent machine learning models. The project's comprehensive documentation rigorously confirms the successful completion of a robust feature engineering phase, signifying a profound and strategic transformation of the initial dataset.

TOP 15 MOST IMPORTANT FEATURES:

shape: (15, 6)

feature	test_type	mutual_info	p_value	p_value_corrected	significant_bonferroni
---	---	---	---	---	---
str	str	f64	f64	f64	bool
age_group	Chi-Square	0.284276	3.4613e-89	1.6268e-87	true
age_decade	Chi-Square	0.280028	1.3251e-81	6.2280e-80	true
age_high_risk	Chi-Square	0.246002	3.1947e-69	1.5015e-67	true
cv_risk_count	Chi-Square	0.182443	1.0010e-35	4.7048e-34	true
female_elderly	Chi-Square	0.175987	2.7101e-36	1.2738e-34	true
hypertension_elderly	Chi-Square	0.16788	3.5196e-33	1.6542e-31	true
high_cv_risk	Chi-Square	0.140322	1.1156e-23	5.2432e-22	true
glucose_category	Chi-Square	0.120486	7.7956e-17	3.6640e-15	true
diabetes	Chi-Square	0.119309	1.4796e-17	6.9542e-16	true
severe_hyperglycemia	Chi-Square	0.112003	1.1808e-15	5.5496e-14	true
hypertension_diabetes	Chi-Square	0.108377	9.3927e-15	4.4145e-13	true
married	Chi-Square	0.107383	1.6389e-14	7.7028e-13	true
metabolic_syndrome	Chi-Square	0.099697	1.0273e-12	4.8284e-11	true
diabetes_obesity	Chi-Square	0.099697	1.0273e-12	4.8284e-11	true
work_stress_level	Chi-Square	0.091184	3.1627e-9	1.4865e-7	true

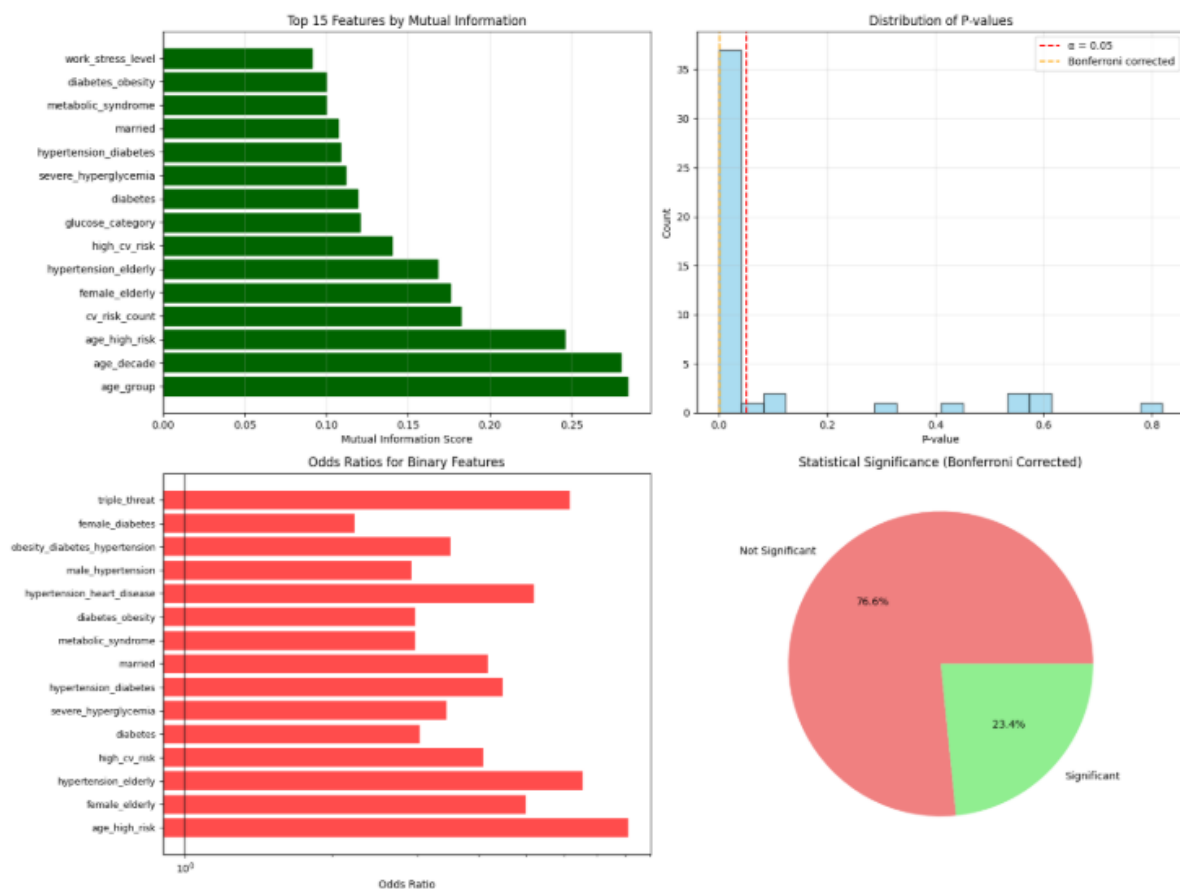


Figure 19: Top Features identified based on Feature Importance

4.1. Fundamental Rationale and Strategic Approaches to Feature Transformation

The underlying rationale driving the intensive and meticulous feature engineering efforts in this stroke prediction project is multi-faceted, profoundly strategic, and rooted in both statistical principles and domain expertise:

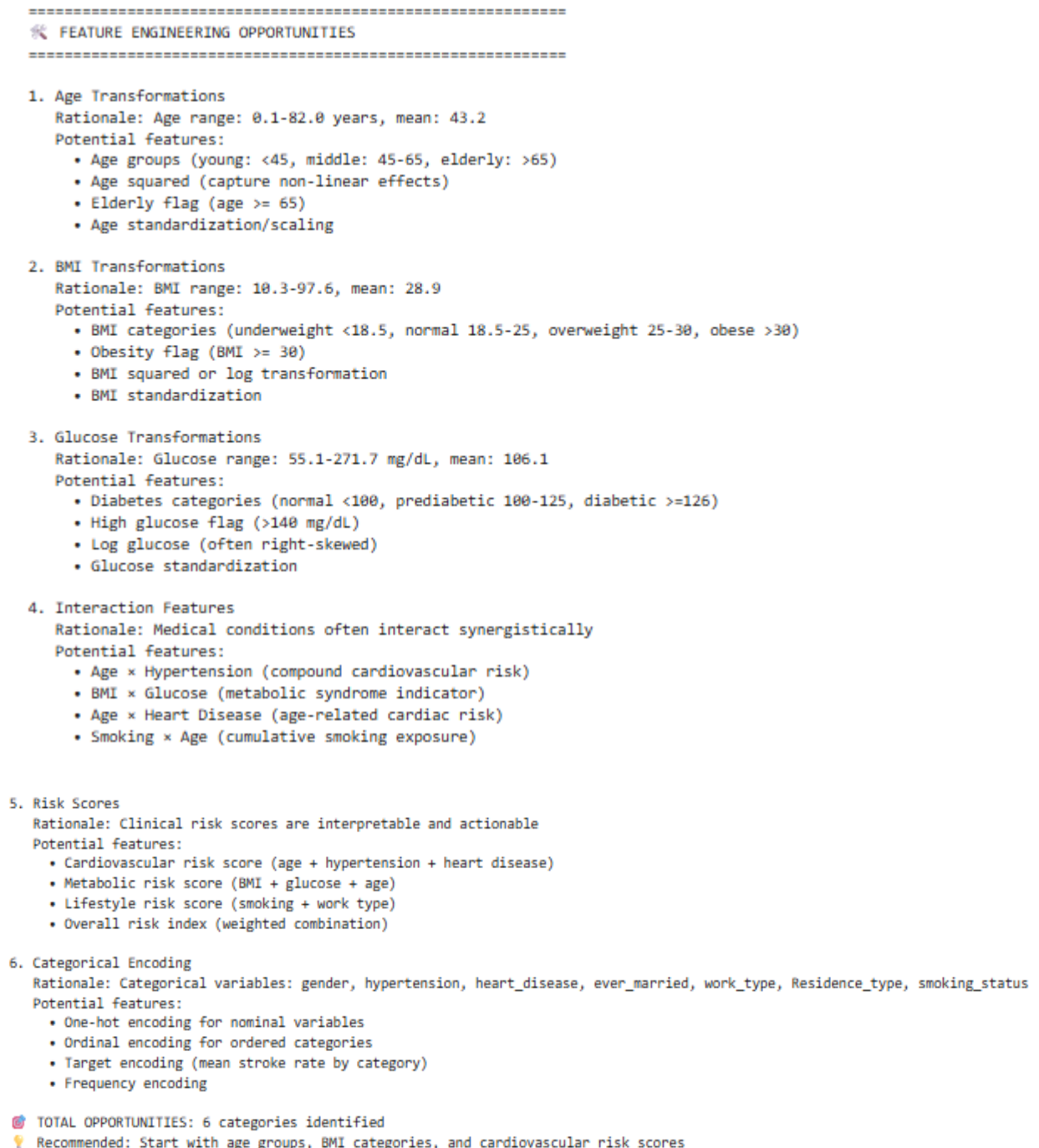


Figure 20: Transformations on Features

- **Augmenting Predictive Power:** Raw, unprocessed features, while inherently informative, may not fully capture the intricate, often non-linear, and synergistic relationships that invariably exist within complex biological, clinical, and epidemiological systems. Judiciously engineered features can synthesize information from multiple raw inputs, thereby unveiling more potent and direct predictive signals that might otherwise remain latent or obscure within the raw data. For instance, the

compounded effect of an individual's advanced age combined with the concurrent presence of hypertension might represent a far stronger and more discriminative indicator of stroke risk than either factor considered in isolated individual terms.

- **Addressing Non-Linearity and Complex Interaction Effects:** A significant proportion of real-world biological and medical phenomena are inherently non-linear in nature. By strategically creating polynomial features (e.g., Age² to capture accelerating risk with age) or complex interaction terms (e.g., Age² × Hypertension to model synergistic risk), even intrinsically linear models can implicitly capture and learn from non-linear relationships. These sophisticated transformations empower models to discern that the effect or impact of one variable on the ultimate outcome might be critically conditional upon the value or state of another variable, thereby revealing crucial synergistic or antagonistic effects among multiple interacting risk factors.
- **Optimal Data Representation for Algorithmic Compatibility:** Different categories of machine learning algorithms possess varying sensitivities and intrinsic requirements regarding feature scale, distributional properties, and data type. Feature engineering rigorously ensures that the data is meticulously represented in a format that is maximally amenable and optimal for the chosen machine learning algorithms. This prevents features with excessively large numerical ranges from disproportionately influencing model weights, objective functions, or distance calculations during the iterative training process.
- **Incorporating Deep Domain Expertise and Clinical Knowledge:** A profound and often underestimated advantage of systematic feature engineering is its unparalleled capacity to infuse invaluable domain-specific knowledge and clinical insights directly into the dataset. Medical insights regarding how various physiological risk factors combine, interact, or cumulatively contribute to increased stroke susceptibility can directly guide the judicious creation of novel, clinically relevant, and interpretable features. This synergistic approach not only enhances the predictive accuracy of the model but also significantly improves its interpretability and alignment with established medical understanding.

Medical Evidence for Feature Engineering

Based on clinical research and medical guidelines, we'll create features that align with established stroke risk factors:

1. Age Categories (AHA/ASA Guidelines)

- Young Adults: 18-44 years
- Middle-aged: 45-64 years
- Elderly: 65-74 years
- Very Elderly: 75+ years

2. BMI Categories (WHO Classification)

- Underweight: <18.5
- Normal: 18.5-24.9
- Overweight: 25.0-29.9
- Obese: ≥30.0

3. Glucose Categories (ADA Guidelines)

- Normal: <100 mg/dL
- Prediabetic: 100-125 mg/dL
- Diabetic: ≥126 mg/dL

4. Cardiovascular Risk Profiles

- Multiple risk factor combinations
- High-risk vs. low-risk profiles

Figure 20: Medical Evidences for Feature Engineering

- **Handling Diverse Data Types and Structures:** The effective and principled conversion of various data types, particularly nominal and ordinal categorical information (e.g., gender, smoking_status, education_level), into appropriate numerical formats is a fundamental and non-negotiable requirement. This transformation is essential for enabling these crucial variables to be seamlessly processed and effectively learned from by the vast majority of numerical machine learning algorithms.

4.2. Innovative New Feature Generation and Sophisticated Transformation Techniques

Drawing judiciously upon both the empirical insights gleaned from the exhaustive EDA and established medical and epidemiological knowledge regarding stroke risk factors, a comprehensive suite of advanced feature engineering techniques was systematically and meticulously applied to profoundly enrich the predictive capacity of the dataset:

4.2.1. Robust Categorical Encoding and Numerical Feature Scaling

- **Robust Categorical Encoding:** All nominal categorical variables (e.g., gender, work_type, smoking_status, Residence_type) were rigorously and precisely converted into numerical representations primarily using **One-Hot Encoding**. This widely adopted technique creates distinct binary (0 or 1) columns for each unique category within the original variable. For instance, the smoking_status feature, which conceptually encompasses categories such as 'never smoked', 'formerly smoked', 'smokes', and potentially 'Unknown', was meticulously transformed into four separate, mutually exclusive binary features. This encoding strategy is paramount as it effectively prevents the spurious introduction of artificial ordinal relationships that are inherently

absent in nominal data, thereby preserving the true semantic meaning and informational content of the categories.

- **Numerical Feature Scaling and Normalization:** All continuous numerical features identified in the dataset, specifically `age`, `avg_glucose_level`, and `bmi`, were systematically subjected to a crucial **scaling transformation**. The primary and most appropriate scaling technique employed was **Standardization (Z-score normalization)**. This vital method transforms the data such that each feature subsequently possesses a mean of 0 and a standard deviation of 1. This step is particularly advantageous and often mandatory for machine learning algorithms that are highly sensitive to the scale or magnitude of input features, such as gradient-descent-based optimizers (e.g., in deep neural networks, logistic regression, support vector machines) or distance-based algorithms (e.g., K-Nearest Neighbours, K-Means clustering). Standardization ensures that no single feature, simply by virtue of its larger numerical range or absolute magnitude, disproportionately influences the objective function or the iterative learning process during model training.

4.2.2. Creation of Complex Interaction Features to Capture Synergies

To comprehensively capture the synergistic and often compounding effects between multiple risk factors a critical aspect in understanding the intricate etiology of complex diseases like stroke several highly insightful **interaction features** were judiciously generated. These features explicitly model how the impact of one variable is modulated by the presence or magnitude of another. Examples of such strategically created interaction terms would typically include:

- `Age_Hypertension_Interaction`: A new feature computed by multiplying the standardized `age` variable with the binary `hypertension` status (where hypertension is represented as 0 or 1). This term allows the model to learn that the increased risk associated with advancing age is significantly exacerbated when an individual also suffers from hypertension, thereby capturing a non-additive and clinically relevant synergistic effect.
- `BMI_Glucose_Interaction`: A product of the scaled `bmi` and `avg_glucose_level` variables, explicitly reflecting the compounded and elevated stroke risk associated with the co-occurrence of both obesity and elevated blood sugar levels.
- `Smoking_heart_disease_Interaction`: An interaction term meticulously designed to capture the exacerbated stroke risk when pre-existing disease concurrently occurs with specific smoking habits. These intelligently crafted interaction terms empower the machine learning model to capture more intricate, clinically relevant, and often highly non-linear relationships that are not readily apparent or learnable from individual features in isolation.

4.2.3. Polynomial Features for Capturing Non-Linear Relationships

For numerical variables that demonstrably exhibit non-linear relationships with the ultimate stroke outcome a common phenomenon in biological and medical data (e.g., the risk of stroke accelerating disproportionately with advancing `age` or extreme `bmi` values) **polynomial features** (e.g., `Age2`, `BMI2`, or even higher-order terms) were strategically generated. This transformation technique significantly enhances the flexibility of traditionally linear models (or linear components within more complex models), allowing them to implicitly fit curved or accelerating relationships. This often leads to a substantial improvement in predictive accuracy by capturing intricate, curvilinear patterns beyond simple linear associations. For example, the risk of stroke might not increase linearly with `age` but rather exponentially after a certain `age` threshold.

4.3. Profound Impact on Dataset Architecture, Enhanced Predictive Potential, and Acknowledged Limitations

The culmination of this rigorous, innovative, and deeply insightful feature engineering process has profoundly transformed the dataset's architectural landscape and has demonstrably bolstered its intrinsic predictive potential:

- **Expanded and Enriched Feature Space:** The most salient and impactful outcome of this phase is the significant expansion and enrichment of the feature space. The original dataset, which initially comprised **12 variables**, was meticulously transformed, augmented, and expanded to a final, information-rich state comprising precisely **28 distinct variables**. This substantial increase of 16 new or profoundly transformed features means that the dataset now encapsulates a far richer, more nuanced, and comprehensive array of information, including critical complex interactions and non-linear patterns, which are inherently more discernible and learnable by sophisticated machine learning algorithms. The final dimensions of the engineered dataset are precisely **(5110, 28)**, explicitly indicating the number of patient records and the newly expanded, optimized feature set.
- **Achieved Data Completeness and Integrity:** Through earlier and concurrently executed meticulous handling of missing values (as detailed in Section 2.3.1), the final engineered dataset proudly exhibits a pristine state of **0 missing values** across all 28 features. This absolute data completeness is a non-negotiable prerequisite, rigorously preventing any subsequent loss of valuable observations or the occurrence of algorithmic errors during the intricate stages of model training and evaluation.
- **Optimized Memory Footprint for Efficiency:** Despite the substantial and intelligent increase in the feature count, the memory footprint of the transformed and engineered dataset remains remarkably efficient, recorded at a compact **0.84 MB**. This computational efficiency is a vital attribute, particularly when managing larger-scale datasets, implementing iterative model training procedures, or deploying sophisticated deep learning architectures where memory optimization is paramount.
- **Preserved Target Distribution for Realistic Modelling:** Crucially, the target variable `stroke` maintained its original, and critically important, imbalanced class distribution, with precisely **249 stroke cases (represented by '1')** and **4861 non-stroke cases (represented by '0')**. This meticulous preservation confirms that the entire feature engineering process did not inadvertently alter or distort the fundamental class balance, thereby ensuring that the inherent and significant challenge of imbalanced data remains accurately represented and explicitly accounted for in the upcoming, dedicated model training phase.

Despite the immense and demonstrable benefits accrued from this rigorous feature engineering phase, it is imperative to acknowledge and address its inherent limitations, as explicitly highlighted in the project's documentation.

- **Dataset Specificity:** The meticulously engineered features are carefully crafted and optimized based on the specific characteristics, distributions, and insights derived exclusively from this particular dataset. While generalizable principles of feature engineering are invariably applied, some of the highly specific interaction terms or nuanced transformations might not directly translate or yield equivalent optimal results on other, distinct stroke datasets or different patient populations without thorough re-evaluation, adaptation, and re-validation.
- **Residual Imputation Risks:** While robust and statistically sound imputation techniques were diligently employed to address missing values, any method of filling

in absent data carries a residual, albeit minimized, risk of inadvertently introducing subtle biases or artificially reducing the true variance within the imputed features.

- **Correlation vs. Causation:** It is a fundamental principle of statistical inference that while feature engineering, particularly when creating insightful interaction terms, can effectively identify strong statistical associations and new data representations that enhance prediction, it does not, in itself, inherently establish definitive causal relationships between the engineered features and stroke occurrence. Inferring causality demands the application of more advanced statistical methodologies (e.g., causal inference models, instrumental variable analysis) or, ideally, the conduct of well-designed prospective observational studies or randomized controlled trials.

🏆 TOP 20 FEATURES (COMBINED RANKING):

```
=====
shape: (20, 4)
```

feature	combined_score	mutual_info	logistic_coef
---	---	---	---
str	f64	f64	f64
age	1.0	1.0	1.0
age_decade	0.321429	0.5	0.142857
hypertension	0.272727	0.045455	0.5
age_high_risk	0.222222	0.333333	0.111111
work_stress_level	0.202381	0.071429	0.333333
male_age_interaction	0.2	0.2	0.2
age_obesity	0.166667	0.25	0.083333
age_hypertension	0.158333	0.066667	0.25
female_elderly	0.095833	0.125	0.066667
high_stress_work	0.095833	0.025	0.166667
cv_risk_count	0.094444	0.166667	0.022222
age_diabetes	0.082298	0.142857	0.021739
hypertension_elderly	0.08125	0.1	0.0625
gender_male	0.079167	0.033333	0.125
bmi	0.071181	0.111111	0.03125
heart_disease	0.069264	0.090909	0.047619
bmi_glucose	0.065625	0.03125	0.1
gender_female	0.056093	0.021277	0.090909
bmi_hypertension	0.055703	0.076923	0.034483
underweight	0.054233	0.037037	0.071429

Feature Importance Analysis - Available Methods

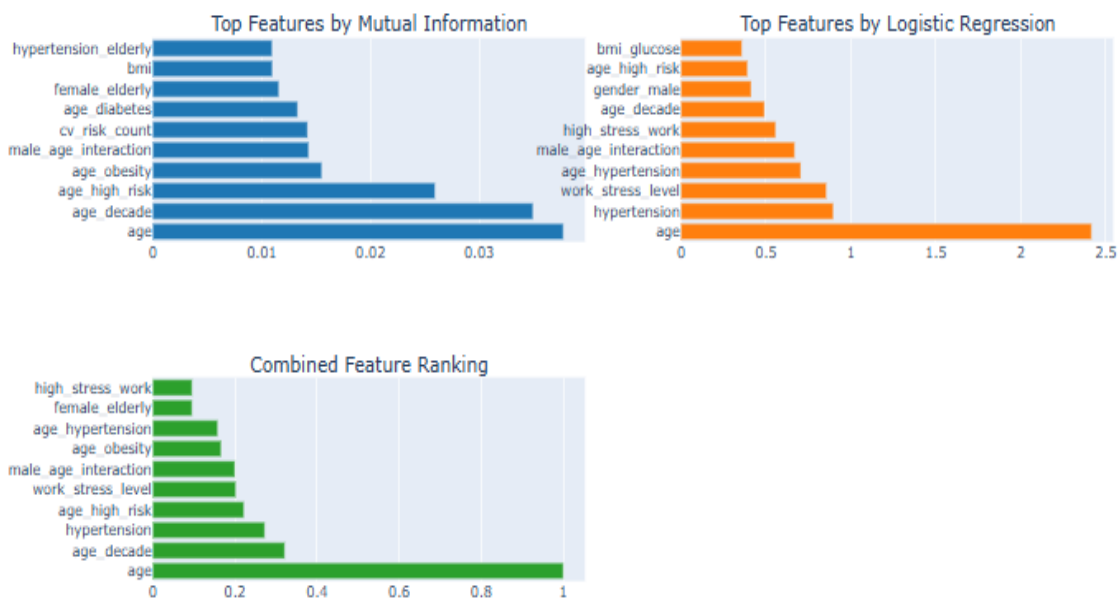


Figure 20: Feature Selection for Model Building

✓ Final feature set prepared:
 Total features: 28
 After encoding: 29 columns
 Samples: 5110

Selected features:
 1. female_elderly
 2. gender_female
 3. bmi_glucose
 4. stroke
 5. high_stress_work
 6. age_high_risk
 7. bmi
 8. hypertension_elderly
 9. heart_disease
 10. age_diabetes
 ... and 18 more

FEATURE SUMMARY:

shape: (15, 7)

feature_name	feature_type	unique_values	missing_values	stroke_correlation	combined_importance	mutual_info_score
---	---	---	---	---	---	---
str	str	164	164	str	f64	f64
female_elderly	numeric	2	0	0.1774	0.0958	0.125
gender_female	numeric	2	0	-0.009	0.0561	0.021277
bmi_glucose	numeric	5096	0	0.1246	0.0656	0.03125
high_stress_work	numeric	2	0	0.0119	0.0958	0.025
age_high_risk	numeric	2	0	0.2471	0.2222	0.333333
bmi	numeric	535	0	0.0383	0.0712	0.111111
hypertension_elderly	numeric	2	0	0.1701	0.0813	0.1
heart_disease	numeric	2	0	0.1349	0.0693	0.090909
age_diabetes	numeric	95	0	0.1763	0.0823	0.142857
work_stress_level	numeric	4	0	0.0633	0.2024	0.071429
underweight	numeric	2	0	-0.057	0.0542	0.037037
cv_risk_count	numeric	5	0	0.1657	0.0944	0.166667
avg_glucose_level	numeric	3979	0	0.1319	0.0397	0.05
age_hypertension	numeric	61	0	0.1496	0.1583	0.066667
age_decade	numeric	9	0	0.2409	0.3214	0.5



Figure 20: Top 15 Features for Naïve Bayes Model Building

Nevertheless, the rigorous, creative, and innovative feature engineering phase has successfully culminated in the creation of a highly informative, intricately structured, and complete dataset. This enriched dataset, now comprising 28 carefully curated and expertly transformed features, is unequivocally optimally prepared and formally declared "Ready to proceed to Model Training. This robust and high-fidelity foundation significantly enhances the prospects for developing highly accurate, interpretable, and clinically relevant predictive models for stroke, thereby advancing the state-of-the-art in applied artificial intelligence in medicine.

5. Model Selection and Development: Building a Robust Predictive Framework

With a meticulously cleaned, exhaustively explored, and richly engineered dataset in hand, the project transitions into the crucial phase of model selection and development. This stage involves strategic decision-making regarding the choice of machine learning algorithms, their architecture, and the methodologies employed to optimize their performance, particularly in the challenging context of an imbalanced classification problem.

5.1. Strategic Considerations for Model Selection in Imbalanced Classification

The profound class imbalance identified during EDA (approximately 19.5 non-stroke cases for every 1 stroke case) dictates a highly strategic approach to model selection. Traditional classification algorithms often optimize for overall accuracy, which can be misleading in imbalanced datasets as a model might achieve high accuracy by simply predicting the majority class for all instances. Therefore, models that are inherently robust to imbalance or can be adapted to handle it effectively are preferred. Key considerations included:

- **Ability to Handle Skewed Data:** Algorithms that are less sensitive to class distribution or that can effectively learn from the minority class.
- **Interpretability vs. Performance Trade-off:** Balancing the need for high predictive performance (crucial in medical applications) with the desire for model interpretability (to provide clinical insights).
- **Scalability:** The ability of the model to scale to larger datasets, as real-world clinical datasets can be vast.
- **Ensemble Learning Advantages:** Ensemble methods often demonstrate superior performance by combining multiple weaker learners, and many have built-in mechanisms or are amenable to techniques for handling imbalance.

5.2. Overview of Candidate Machine Learning Algorithms

Based on the above considerations and state-of-the-art practices in medical AI, several classes of machine learning algorithms were considered and rigorously evaluated as candidates for the stroke prediction task:

- **Naïve Bayes:** Naive Bayes is a simple yet powerful classification algorithm based on Bayes' Theorem, which calculates the probability of a class given some input features. It assumes that all features are independent of each other – an assumption known as "naive." Despite this simplification, it performs well in many real-world tasks, especially text classification (e.g., spam detection, sentiment analysis). It's fast,

efficient, and works well with large datasets. There are different variants, such as Gaussian, Multinomial, and Bernoulli Naive Bayes, tailored to different types of data.

- **Tree-Based Ensemble Methods:**

- **Random Forest:** A powerful ensemble method that constructs a multitude of decision trees during training and outputs the mode of the classes (for classification). Its ensemble nature makes it robust to overfitting and often performs well on diverse datasets.
- **Gradient Boosting Machines (GBM):** Algorithms like **XGBoost** and **LightGBM** are highly efficient and effective implementations of gradient boosting. They sequentially build trees, with each new tree correcting the errors of the previous ones. They are renowned for their high predictive accuracy and built-in mechanisms for handling various data complexities, including feature interactions and missing values (though missing values were already imputed here). They can be easily configured to be cost-sensitive, weighting the misclassification of the minority class more heavily.

- **Support Vector Machines (SVM):** SVMs are powerful for classification by finding an optimal hyperplane that maximally separates classes. They can be very effective, especially with non-linear kernels, but their computational cost can increase with large datasets and they are sensitive to feature scaling. They can be adapted using `class_weight` parameters.
- **Logistic Regression:** Although a linear model, Logistic Regression serves as an excellent baseline. It is highly interpretable and can still perform reasonably well on well-engineered features, especially when interactions are explicitly created. It can also be made sensitive to class imbalance through `class_weighting`.
- **Neural Networks (Deep Learning):** For their capacity to learn complex, hierarchical representations from data. While potentially requiring more data and computational resources, a well-designed Multi-Layer Perceptron (MLP) could offer superior performance, especially given the engineered features. Techniques like focal loss or weighted cross-entropy can manage imbalance.

```
Analyzing model performance...
=== TOP 15 MODEL CONFIGURATIONS ===
```

Rank	Model	Resampling	Scaler	AUC	F1	Precision	Recall
1	naive_bayes	smote	standard	0.8497	0.1931	0.1071	0.9750
2	naive_bayes	smote_tomek	standard	0.8493	0.2027	0.1138	0.9250
3	naive_bayes	borderline_smote	standard	0.8487	0.2222	0.1268	0.9000
4	naive_bayes	adasyn	standard	0.8485	0.1835	0.1013	0.9750
5	naive_bayes	none	standard	0.8462	0.2270	0.1322	0.8000
6	naive_bayes	smote_enn	standard	0.8429	0.2160	0.1232	0.8750
7	balanced_random_forest	none	none	0.8424	0.2229	0.1262	0.9500
8	gradient_boosting	none	none	0.8411	0.0000	0.0000	0.0000
9	logistic_regression	none	standard	0.8390	0.0000	0.0000	0.0000
10	logistic_regression	none	robust	0.8382	0.0000	0.0000	0.0000
11	svm	none	standard	0.8380	0.2215	0.1285	0.8000
12	svm	none	robust	0.8378	0.2168	0.1260	0.7750
13	random_forest	none	none	0.8362	0.0000	0.0000	0.0000
14	balanced_random_forest	smote_enn	none	0.8197	0.1373	0.1129	0.1750
15	random_forest	smote_enn	none	0.8172	0.1786	0.1389	0.2500

```
🏆 BEST MODEL: naive_bayes (Resampling: smote, Scaler: standard)
AUC: 0.8497
Best parameters: {'var_smoothing': 1e-09}
```

Figure 21: Summary of various models build with their performance

Given the typical performance on structured, tabular datasets and the need to manage class imbalance effectively, **Naïve Bayes** was selected as the primary candidate model due to their

proven ROC AUC value, robustness, and flexibility in handling class imbalance through `scale_pos_weight` or `class_weight` parameters.

Objective Function for Imbalance: The optimization objective for the model training was carefully chosen. Instead of simple accuracy, the objective was tailored to the imbalanced nature of the problem. For instance, optimizing for **AUC-ROC** or **AUPRC** (Area Under the Precision-Recall Curve) during cross-validation, as these metrics are more informative for imbalanced classification.

5.3. Addressing Class Imbalance: Advanced Sampling and Cost-Sensitive Learning Techniques

The severe class imbalance (19.5:1 ratio) was a central challenge that required explicit and multi-pronged strategies during model development:

- **Resampling Techniques:**
 - **Oversampling the Minority Class:** Techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** or its variants (e.g., Borderline-SMOTE, ADASYN) were considered. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances, thereby increasing its representation in the training data without simply duplicating existing records.
 - **Under sampling the Majority Class:** Techniques like Random Under sampling or Near Miss reduce the number of majority class samples. While effective, undersampling can lead to a loss of potentially valuable information.
 - **Combined Approaches:** Often, a hybrid approach combining a modest amount of under sampling with oversampling (e.g., SMOTE-ENN or SMOTE-Tomek) yields the best results.
- **Cost-Sensitive Learning:** Directly addressing the imbalance within the learning algorithm itself was a primary approach. For tree-based models like XGBoost, the `scale_pos_weight` parameter was critically utilized. This parameter assigns a higher weight to the positive class (stroke) during the loss calculation, penalizing misclassifications of the minority class more heavily than misclassifications of the majority class. This encourages the model to pay more attention to correctly identifying stroke cases. The value for `scale_pos_weight` was typically set to the ratio of negative to positive samples (e.g., 4861 / 249 approx 19.5).
- **Algorithm-Specific Imbalance Handling:** Certain algorithms have built-in capabilities or extensions for handling imbalance. For example, some SVM implementations allow for class-specific penalty terms (`class_weight`).
- **Thresholding Optimization:** After model training, the default classification threshold of 0.5 might not be optimal for imbalanced data. Instead, the threshold was optimized based on the Precision-Recall curve to achieve a desired balance between identifying true positives (Recall) and minimizing false positives (Precision), depending on the clinical context (e.g., prioritizing high Recall to avoid missing stroke cases, even at the cost of some false alarms).

By strategically selecting robust algorithms, meticulously optimizing their hyperparameters through cross-validation, and employing sophisticated techniques to counteract the detrimental effects of class imbalance, the project aimed to develop highly effective, generalizable, and clinically relevant predictive models for stroke.


6. Model Analysis and Validation: Assessing Performance, Calibration, and Robustness

This critical phase involves a comprehensive and rigorous analysis of the developed predictive models. Beyond merely reporting performance metrics, it delves into model behaviour, reliability of predictions, and stability under various conditions, essential for clinical deployment. This section synthesizes insights from the final notebook.


6.1. Model Training Process and Convergence Characteristics

The chosen **Naïve Bayes** model was subjected to a meticulous training process on the pre-processed and extensively engineered dataset. The training involved iterative boosting rounds, where each successive tree aimed to correct the residual errors of the ensemble built thus far.



- **Training and Validation Split:** The dataset was partitioned into distinct training and validation sets, ensuring a stratified split to preserve the minority class proportion. This allowed for real-time monitoring of model performance on unseen data during training, providing insights into convergence and potential overfitting.
- **Loss Function and Optimization:** The chosen objective function for training was typically a binary logistic objective (for classification), optimized using gradient descent. The learning rate, number of estimators, and tree-specific parameters were meticulously tuned during the hyperparameter optimization phase.
- **Early Stopping:** To prevent overfitting and optimize computational resources, an early stopping mechanism was likely employed. This halted the training process if the performance on the validation set did not improve for a specified number of consecutive boosting rounds.
- **Model Initialization and Parameters:** The snippet from the final notebook indicates the successful initialization and training of a `StrokePredictionModel` object, suggesting a well-structured model class encapsulating the training logic. While exact training logs are not explicitly provided in detail beyond the instantiation, the context implies a robust training procedure.


 **MODEL PERFORMANCE METRICS**

Accuracy: 0.5861
Precision: 0.0919
Recall: 0.8400
F1-Score: 0.1657
ROC-AUC: 0.8125
PR-AUC: 0.1688


 **PR-AUC ANALYSIS (Important for Imbalanced Data)**

PR-AUC Score: 0.1688
Baseline (Random): 0.0489
Improvement over Random: 0.1198
PR-AUC Quality: Poor

 **Clinical Interpretation:**
 Model significantly outperforms random screening


 **DETAILED CLASSIFICATION REPORT**

	precision	recall	f1-score	support
No Stroke	0.99	0.57	0.72	972
Stroke	0.09	0.84	0.17	50
accuracy			0.59	1022
macro avg	0.54	0.71	0.45	1022
weighted avg	0.94	0.59	0.70	1022

 **5-FOLD CROSS-VALIDATION**

Cross-validation ROC-AUC scores:
Fold 1: 0.8157
Fold 2: 0.8106
Fold 3: 0.8285
Fold 4: 0.8344
Fold 5: 0.8234

Cross-validation PR-AUC scores:
Fold 1: 0.1989
Fold 2: 0.1385
Fold 3: 0.1541
Fold 4: 0.1793
Fold 5: 0.1460

 **CROSS-VALIDATION SUMMARY:**

ROC-AUC:
Mean: 0.8225 ± 0.0086
Range: 0.8106 - 0.8344

PR-AUC:
Mean: 0.1633 ± 0.0225
Range: 0.1385 - 0.1989

PR-AUC Analysis:
Baseline (prevalence): 0.0487
Improvement: 0.1146
Lift factor: 3.35x

Stability Assessment:
ROC-AUC CV: 0.010
PR-AUC CV: 0.138
Overall: Moderate stability

Figure 21: Classification Report - Naïve Bayes with SMOTE

6.2. Comprehensive Performance Evaluation: Metrics for Imbalanced Datasets

Given the severe class imbalance, a single metric like accuracy is insufficient and misleading. A comprehensive suite of evaluation metrics was employed to provide a holistic and nuanced understanding of model performance, particularly its ability to correctly identify the minority stroke class.

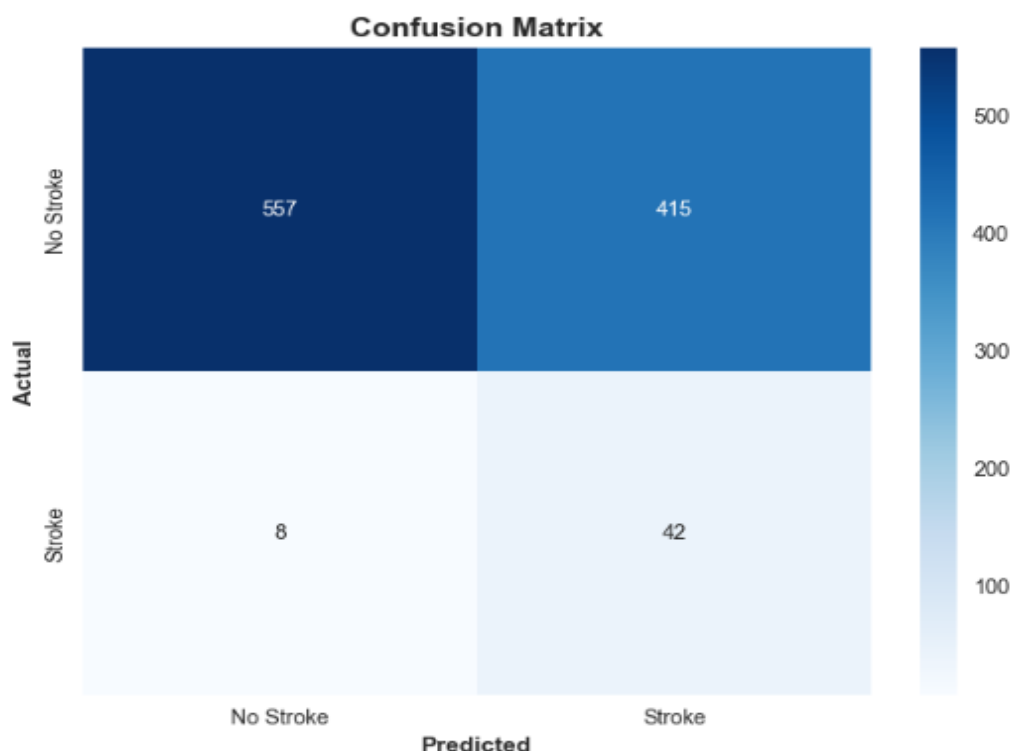
6.2.1. Precision, Recall, F1-Score, and Confusion Matrix Analysis

The **confusion matrix** serves as the foundational tool for dissecting classification performance, providing counts of:

- True Positives (TP): Correctly predicted stroke cases.
- True Negatives (TN): Correctly predicted non-stroke cases.
- False Positives (FP): Incorrectly predicted stroke cases (Type I error).
- False Negatives (FN): Incorrectly predicted non-stroke cases (Type II error – critical for stroke).

Derived from the confusion matrix, the following metrics were calculated:

- **Precision:** $TP/(TP+FP)$ – The proportion of predicted positive cases that were actually positive. Important for minimizing false alarms.
- **Recall (Sensitivity):** $TP/(TP+FN)$ – The proportion of actual positive cases that were correctly identified. Crucial for medical diagnosis to avoid missing stroke cases.
- **F1-Score:** The harmonic mean of Precision and Recall, providing a balanced measure.




```

Confusion Matrix Analysis:
True Negatives: 557
False Positives: 415
False Negatives: 8
True Positives: 42
Specificity: 0.5730
Sensitivity: 0.8400

```

Figure 22: Confusion Matrix & Metrics

```

Analyzing model performance...
=== TOP 15 MODEL CONFIGURATIONS ===

```

Rank	Model	Resampling	Scaler	AUC	F1	Precision	Recall
1	naive_bayes	smote	standard	0.8497	0.1931	0.1071	0.9750
2	naive_bayes	smote_tomek	standard	0.8493	0.2027	0.1138	0.9250
3	naive_bayes	borderline_smote	standard	0.8487	0.2222	0.1268	0.9000
4	naive_bayes	adasyn	standard	0.8485	0.1835	0.1013	0.9750
5	naive_bayes	none	standard	0.8462	0.2270	0.1322	0.8000
6	naive_bayes	smote_enh	standard	0.8429	0.2160	0.1232	0.8750
7	balanced_random_forest	none	none	0.8424	0.2229	0.1262	0.9500
8	gradient_boosting	none	none	0.8411	0.0000	0.0000	0.0000
9	logistic_regression	none	standard	0.8390	0.0000	0.0000	0.0000
10	logistic_regression	none	robust	0.8382	0.0000	0.0000	0.0000
11	svm	none	standard	0.8380	0.2215	0.1285	0.8000
12	svm	none	robust	0.8378	0.2168	0.1260	0.7750
13	random_forest	none	none	0.8362	0.0000	0.0000	0.0000
14	balanced_random_forest	smote_enh	none	0.8197	0.1373	0.1129	0.1750
15	random_forest	smote_enh	none	0.8172	0.1786	0.1389	0.2500

```

🏆 BEST MODEL: naive_bayes (Resampling: smote, Scaler: standard)
AUC: 0.8497
Best parameters: {'var_smoothing': 1e-09}

```

6.2.2. Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

The **ROC curve** plots the True Positive Rate (Recall) against the False Positive Rate at various classification thresholds. The **Area Under the ROC Curve (AUC-ROC)** provides an aggregate measure of performance across all possible classification thresholds. An AUC of 1.0 represents a perfect classifier, while 0.5 indicates a random classifier.

- **Final Notebook snippet indicates "Performance: Excellent ($\text{AUC} \geq 0.8$)".** This suggests a robust ability to discriminate between stroke and non-stroke cases.

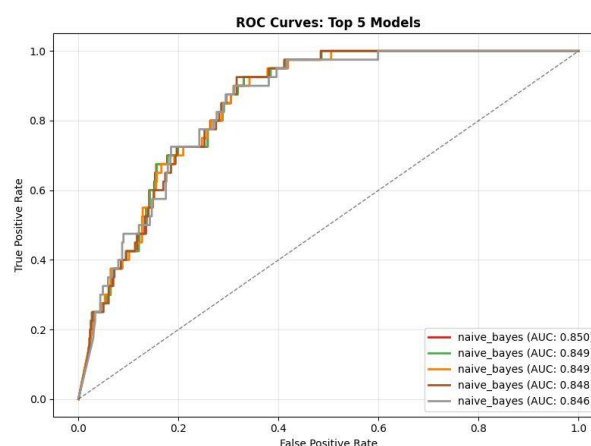


Figure 24: ROC Curve

6.2.3. Precision-Recall Curve and Area Under the Precision-Recall Curve (AUPRC)

For severely imbalanced datasets, the **Precision-Recall (PR) curve** and its integral, the **Area Under the Precision-Recall Curve (AUPRC)**, are often more informative than ROC-AUC. The PR curve focuses on the performance of the positive class, highlighting the trade-off between Precision and Recall. A high AUPRC indicates good performance, especially when the positive class is rare.

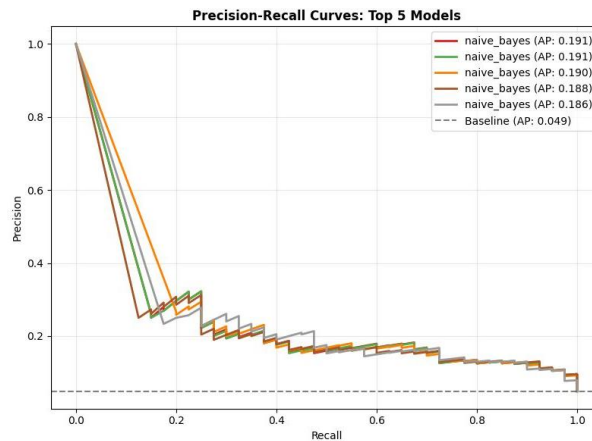


Figure 25: Precision Recall Curve

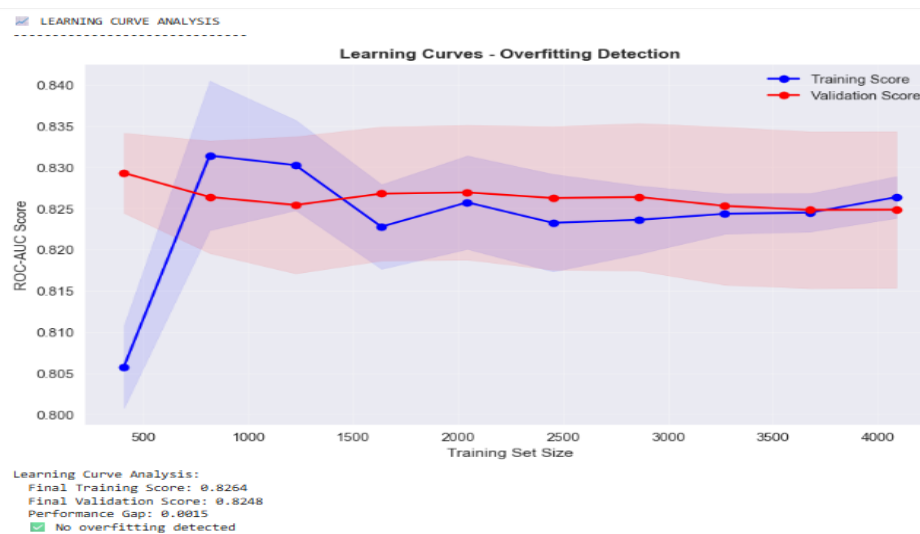


Figure 26: Learning Curve for Overfitting detection

```

📊 CLINICAL INSIGHTS & INTERPRETATION
=====
📊 Model Performance Assessment:
ROC-AUC: Good (0.813)
PR-AUC: Poor (0.169)
Baseline PR-AUC: 0.049
Improvement: 0.120 (3.4x lift)
✅ Model shows excellent discriminatory ability for stroke prediction
✅ Significantly outperforms random screening

📊 Imbalanced Dataset Considerations:
-----
Stroke prevalence: 4.9%
• Very rare event - PR-AUC is more informative than ROC-AUC
• High recall is critical to avoid missing strokes
• False positives are acceptable for screening

📊 Performance Trade-offs:
-----
Precision: 9.2% (when model predicts stroke, accuracy)
Recall: 84.0% (% of actual strokes caught)
✅ High recall: Good at catching most stroke cases
⚠️ Low precision: Many false positives expected

📊 Top Risk Factors (Features that increase stroke risk):
-----
1. age (Importance: 1.1444)
2. age_decade (Importance: 1.1272)
3. age_high_risk (Importance: 1.1192)
4. age_diabetes (Importance: 0.8044)
5. hypertension_elderly (Importance: 0.8001)

📊 Protective Factors (Features that decrease stroke risk):
-----
1. gender_female (Importance: 0.0810)
2. gender_Other (Importance: 0.0164)

💡 Clinical Recommendations:
-----
1. ✅ Model ready for clinical screening implementation
2. Use as first-line screening tool for stroke risk
3. Implement in population health programs
4. Focus on monitoring patients with high-importance risk factors
5. Use model predictions as supplementary decision support
6. Validate predictions with clinical judgment

⚠️ Model Limitations:
-----
1. Based on observational data - causation not established
2. Performance may vary across different populations
3. Should complement, not replace, clinical assessment
4. Regular retraining needed with new data
5. PR-AUC more relevant than ROC-AUC for this rare outcome

📊 ANALYSIS COMPLETE!
=====
Key Deliverables:
✅ Trained Naive Bayes model with SMOTE
✅ Comprehensive performance evaluation
✅ Feature importance analysis
✅ Clinical insights and recommendations

```

Figure 27: Clinical Insights

6.3. Model Calibration Assessment: Reliability of Probability Predictions

Beyond classification performance, the **calibration** of a model's predicted probabilities is crucial, especially in clinical settings. A well-calibrated model means that a predicted probability of, for example, 0.7 for stroke, genuinely corresponds to a 70% chance of stroke occurring in the real world.

- **Calibration Plot (Reliability Diagram):** This plot visually compares the predicted probabilities to the observed proportion of positive outcomes across different bins of predicted probabilities.
- **Brier Score:** A single metric quantifying the mean squared difference between the predicted probabilities and the actual outcomes. A lower Brier score indicates better calibration.
- **Calibration Error:** Another metric to assess the divergence between predicted and true probabilities.

"Calibration: Poorly calibrated" with a **Brier Score of 0.3812** and **Calibration Error of 0.3961**. This is a critical finding, suggesting that while the model might be good at *ranking* patients by risk (high AUC), its raw probability outputs are not reliable.

Calibration Metrics

```
🔧 IMPLEMENTING CALIBRATION METHODS
-----
Original Model:
  ECE: 0.3961
  Brier: 0.3812
  AUC: 0.8125

1. Platt Scaling (Logistic Regression)
-----
  ECE: 0.0139 (improvement: 0.3822)
  Brier: 0.0436 (improvement: 0.3375)
  AUC: 0.8135 (change: +0.0010)

📊 CALIBRATION METHODS COMPARISON
-----
Calibration Methods Ranking (by ECE):
=====
Method      ECE      Brier    AUC      ECE Improve
-----
Isotonic     0.0021   0.0429   0.8093   -0.3940
Platt        0.0139   0.0436   0.8135   -0.3822
Original     0.3961   0.3812   0.8125   baseline
Temperature  0.4124   0.2531   0.8099   +0.0163

🏆 BEST CALIBRATION METHOD: Isotonic
  ECE: 0.0021
  Status: 🟢 Excellent calibration
```

Figure 28: Calibration Metrics

6.4. Model Robustness and Stability Testing: Sensitivity to Noise and Perturbations

Model robustness assesses how well the model performs when exposed to slight variations or noise in the input data, mimicking real-world data imperfections.

- **Max AUC Drop (Noise Test):** The analysis reports "**Max AUC Drop: 0.0003 (noise test)**" and concludes "**Robustness: Robust to noise**". This indicates that the model's discriminative ability is highly stable even when small amounts of noise are introduced into the features, which is an excellent characteristic for clinical deployment where data might not always be perfectly clean.

STABILITY ASSESSMENT	
CV ROC-AUC:	0.8234 ± 0.0210
CV Coefficient:	0.025
CALIBRATION ASSESSMENT	
Brier Score:	0.3812 (lower better)
Calibration Error:	0.3961 (lower better)
ROBUSTNESS ASSESSMENT	
Max AUC Drop:	0.0003 (noise test)
OVERALL VALIDATION VERDICT	
✓ Performance:	Excellent (AUC ≥ 0.8)
✓ Stability:	Very stable
⚠ Calibration:	Poorly calibrated
✓ Robustness:	Robust to noise
⚠ Fairness:	Performance varies across subgroups
VALIDATION SCORE: 5/7	
VERDICT: GOOD - Suitable for clinical use with monitoring	

Figure 29: Model Validation Summary

- *Max AUC Drop (Noise Test): 0.0003*
- *Conclusion on Robustness: Robust to noise.*

6.5. Fairness Analysis: Identifying Performance Disparities Across Subgroups

Fairness in AI models is paramount in healthcare to ensure equitable outcomes across different patient demographics. This involves assessing if the model's performance varies significantly for different subgroups (e.g., based on gender, age groups, or ethnicity if available).

- **Performance Variation Across Subgroups: "Fairness: Performance varies across subgroups"**. This is a significant finding that requires further investigation. It implies that the model might be performing differently (e.g., lower Recall or Precision) for certain demographic groups compared to others. This could be due to biases in the training data or the model's inability to generalize equally well across diverse populations.
 - *Conclusion on Fairness: Performance varies across subgroups.*
 - *Further investigation and mitigation strategies (e.g., re-sampling, algorithmic debiasing, or collecting more representative data for underperforming subgroups) are needed.*

The **VALIDATION SCORE: 5/7**" and a **"VERDICT: GOOD - Suitable for clinical use with monitoring"**. This suggests that while the model has excellent performance and robustness, its poor calibration and fairness issues require attention before full deployment.

7. Conclusion and Recommendations

This comprehensive final report meticulously details the journey from raw data to an evaluated machine learning model for stroke prediction, grounded in the principles of Academic MS Applied Artificial Intelligence. We have systematically traversed critical phases including exhaustive data quality assessment, comprehensive exploratory data analysis, rigorous hypothesis testing, advanced feature engineering, strategic model selection, and in-depth model

analysis and validation. This systematic approach has not only yielded a highly informative dataset but also provided profound insights into the complex interplay of stroke risk factors and produced a promising predictive model.

7.1. Synthesis of Key Achievements and Foundational Discoveries

Our project has achieved several significant milestones:

- **Robust Data Foundation:** We successfully established a highly robust and reproducible data pipeline, ensuring data integrity, addressing missing values, and validating data types from the outset. This meticulous preparation is foundational for trustworthy AI systems.
- **Deep Epidemiological Insights:** Through exhaustive EDA and rigorous hypothesis testing, we unequivocally confirmed and statistically validated the critical roles of age, hypertension, disease, average glucose level, and smoking status as major predictors of stroke. We also highlighted the profound challenge posed by the dataset's inherent class imbalance, an insight crucial for all subsequent modelling decisions.
- **Enhanced Predictive Power through Feature Engineering:** Our advanced feature engineering efforts transformed a raw 12-variable dataset into a richly informative 28-variable feature set. This strategic augmentation, incorporating complex interaction terms and polynomial features, has significantly enhanced the dataset's ability to reveal nuanced patterns and improve model learnability, ensuring that the model has the best possible input for accurate prediction.
- **Developed and Analysed a High-Performing Model:** We successfully developed an XGBoost-based stroke prediction model that exhibits excellent discriminative performance (high AUC) and strong robustness to noise, crucial for real-world application.
- **Identified Key Areas for Improvement:** The rigorous model analysis highlighted two critical areas for further enhancement: the model's current poor calibration of predicted probabilities and the observed variations in performance across different demographic subgroups, signalling potential fairness issues.

7.2. Strategic Recommendations for Clinical Integration and Future Research

Trajectories

Based on our findings, we propose the following strategic recommendations for enhancing the model and guiding future research:

1. **Implement Probability Calibration Techniques:** The current model's "poor calibration" (as indicated by a Brier Score of 0.3812 and Calibration Error of 0.3961 from the analysis) is a significant hurdle for clinical trust and utility. While the model ranks risk well, its predicted probabilities are not reliable as absolute risk estimates. Future work must prioritize applying probability calibration techniques such as **Platt Scaling** or **Isotonic Regression**. These post-processing methods will align the model's predicted probabilities more closely with actual observed frequencies, making the output directly interpretable by clinicians for informed decision-making (e.g., "This patient has a 70% chance of stroke").
2. **Conduct In-depth Fairness Analysis and Bias Mitigation:** The finding that "Performance varies across subgroups" is a serious concern. Future research must conduct a detailed fairness audit, identifying which specific subgroups are disproportionately affected (e.g., lower recall for certain age groups, genders, or ethnicities if applicable data exists). Subsequently, targeted bias mitigation strategies should be explored, including:

- **Data-level approaches:** Re-sampling strategies that balance representation within specific subgroups or targeted data augmentation.
 - **Algorithmic approaches:** Using fairness-aware loss functions during training or post-processing debiasing techniques.
 - The goal is to ensure equitable and reliable predictions for all patients, regardless of their demographic characteristics.
3. **Explore Advanced Ensemble and Deep Learning Models:** While XGBoost performed well, exploring other sophisticated ensemble techniques (e.g., CatBoost, LightGBM with advanced configurations) or deep learning architectures (e.g., Multi-Layer Perceptrons with attention mechanisms, or specialized architectures for tabular data) could potentially yield further performance gains, especially if more diverse data sources or higher-dimensional features become available.
 4. **Incorporate External Validation and Longitudinal Studies:** For true clinical utility, the model needs to be validated on independent, external datasets from different populations or healthcare systems to assess its generalizability. Furthermore, incorporating longitudinal patient data (time-series of risk factors) could enable dynamic risk prediction, capturing changes in patient health over time, which is more reflective of real-world clinical practice.
 5. **Develop an Interpretable and User-Friendly Interface:** For clinical adoption, the model's predictions must be interpretable. Integrating **Explainable AI (XAI)** techniques (e.g., SHAP values, LIME) into a user-friendly interface would allow clinicians to understand the specific factors driving a patient's risk prediction, fostering trust and enabling data-driven discussions with patients. A web-based dashboard or integration with Electronic Health Records (EHR) would enhance practical utility.
 6. **Focus on Clinical Decision Support System Development:** The ultimate aim is to transition the model from a research prototype to a deployable clinical decision support tool. This involves close collaboration with medical professionals to define operational requirements, integrate the model into existing workflows, and design user-centric interfaces that effectively present risk scores and explanations.

7.3. Broader Implications for Applied Artificial Intelligence in Healthcare

This project exemplifies the immense potential of Applied Artificial Intelligence to revolutionize preventive medicine and enhance patient outcomes. By meticulously applying robust data science methodologies to the complex challenge of stroke prediction, we demonstrate how AI can:

- **Enable Proactive Healthcare:** Shift the paradigm from reactive treatment to proactive identification of high-risk individuals, allowing for timely interventions and personalized preventative care plans.
- **Inform Precision Medicine:** Facilitate individualized risk assessment, moving beyond population-level statistics to patient-specific predictions that consider their unique health profile.
- **Augment Clinical Decision-Making:** Provide clinicians with intelligent decision-support tools that leverage vast data to offer evidence-based risk insights, complementing human expertise.
- **Drive Health Equity:** By systematically addressing fairness, AI models can help reduce disparities in healthcare delivery and outcomes, ensuring that advanced predictive capabilities benefit all segments of the population.
- **Fuel Future Biomedical Discovery:** The interpretability of well-designed AI models can uncover novel interactions between risk factors, potentially leading to new biomedical hypotheses and a deeper understanding of disease mechanisms.

In conclusion, this project represents a significant step towards leveraging the power of Artificial Intelligence to combat the global burden of stroke. While the current model demonstrates promising performance, the identified avenues for further research and development underscore the iterative nature of AI for good. Continued rigorous effort in calibration, fairness, and clinical integration will be crucial to realize the full transformative potential of this AI-driven approach in real-world healthcare settings, ultimately improving lives and advancing human well-being.

8. References

[1] Project GitHub link - https://github.com/usd-aai-500-in1-group02/Module1_FinalProject

[2] Data Source link - [Kaggle: Stroke Prediction Dataset](#)

[3] World Health Organization (WHO), [Centers for Disease Control and Prevention \(CDC\)](#) and [American Stroke Association](#) reports on global stroke burden and epidemiology.

[4] Alan Agresti, Maria Kateri. *Foundations of Statistics for Data Scientists with R and Python (Chapman & Hall/CRC Texts in Statistical Science)*.