Group Y:
Nicolò Alberto Pellizzari - 63585
Bastien Gobet - 52255
Tomás Leite Barbosa Oliveira - 56466
Drey Tengan - 59840

Enhancing Stroke Prevention with Machine Learning: A Business Case for Insurance RPM

Final Group Project For the Machine Learning Course (2767-2425_T3)
Professor Nuno André Inácio Rodrigues Da Silva

Nova School of Business and Economics

Carcavelos, Lisbon

March, 2025

NOVA SCHOOL OF BUSINESS & ECONOMICS

Table of Contents

1. Introduction

Stroke is a major cause of disability and mortality, imposing significant costs on healthcare and insurance systems. Early detection is essential to reducing its impact. This report explores a predictive analytics approach using a stroke risk dataset, incorporating key health indicators such as age, hypertension, and heart disease. By applying machine learning, we assess an individual's stroke risk, supporting proactive medical interventions and data-driven decision-making in healthcare and insurance.

For insurance companies, integrating stroke prediction into a Remote Patient Monitoring (RPM) system enhances risk assessment and preventive care. By leveraging wearable devices and digital health tools, insurers can monitor policyholders in real time, enabling early intervention and reducing long-term healthcare costs. Patients who experience chronic, on-going risk factor symptoms could be recommended to adopt RPM devices, while awareness and recognition for acute stroke symptoms can be encouraged for patients (with conversations during check-ups or additional informational materials) once they reach specific risk-levels.

2. Exploratory Data Analysis

The dataset was developed using a combination of medical literature, expert consultations, and statistical modeling to ensure clinical validity and balance. The dataset is comprised of 15 key symptoms which were chosen in line with guidelines from the American Stroke Association (ASA), Mayo Clinic, and Cleveland Clinic, along with established textbooks such as *Harrison's Principles of Internal Medicine* and *Stroke Prevention, Treatment, and Rehabilitation* (Oxford University Press). The data also includes patient age, (a key determinant of risk), and the target variables 'Stroke Risk (%)' and 'At Risk (Binary)'.

The symptoms have binary outcomes and balanced distributions, with roughly a 50/50 rate of occurrence, with no missing values. Thus, the Number of Symptoms per patient follows a binomial distribution. Across the board, the correlation heat map shows each of the symptoms have roughly a 0.18 correlation with Stroke Risk. Age is by far the most correlated variable with a 0.73 correlation. Across the dataset, roughly 65% is identified as 'At Risk', which is a binary transformation of the 'Stroke Risk (%)' greater than 50%. Stroke Risk is relatively normal in its distribution with mean 55.5% and standard deviation of 14.3%. Because the risk of stroke is so strongly correlated with patient age, we might consider age adjusted risk of stroke as a key variable. In this case, age-adjusted risk of stroke is positively skewed with mean 1.14 and median 1.03. See appendix for detailed graphs.

3. Unsupervised Learning - Clustering

This section presents the results of a clustering analysis aimed at identifying meaningful subgroups in the dataset based on health-related features. KMeans clustering was used, with validation via the Elbow Method, Silhouette Score, and Hierarchical Clustering.

Methodology & Key Steps

1. *Preprocessing & Feature Engineering:*
   'Age' was binned into categorical groups and one-hot encoded. A new feature, *Number of Symptoms*, was created by summing binary symptom indicators. Feature selection via VarianceThreshold retained all features.
2. *Dimensionality Reduction:*
   PCA showed that reducing dimensions to 2–3 components preserved too little variance. Since all 22 features contributed valuable information, PCA was not used for final clustering
3. *Optimal K Selection:*
   - Elbow Method: Sharp WCSS drop at K=3 suggests optimal value.
   - Silhouette Score: Moderate (~0.17) at K=3, indicating reasonably well-separated clusters.
   - Hierarchical Clustering: Dendrogram supports K=3 for more granular subgrouping; K=2 oversimplifies.

Cluster Interpretation

- **Cluster 0:** High symptom prevalence (~0.60), high stroke risk (63.4%), 85.1% at risk.
- **Cluster 1:** Low symptoms (~0.29), low stroke risk (40.4%), 25.0% at risk.
- **Cluster 2:** Moderate symptoms (~0.40), moderate stroke risk (50.9%), 53.0% at risk.

Clusters are well-defined, reflecting symptom severity and risk profiles. Notably, mean age was similar across clusters when *Number of Symptoms* was included—potentially masking age effects. Removing that feature revealed distinct age patterns (Cluster 0: 30, Cluster 1: 55, Cluster 2: 79) and similar mean values for symptoms-related features, indicating that age is a meaningful segmenting factor. This suggests that engineered features can obscure or alter patterns, and further investigation is needed to understand interactions between variables.

4. Model Comparison and Interpretation of Results

4.1. Regression

Building a **Linear Regression** model to predict stroke risk as a continuous percentage allows for a more detailed understanding on how various factors contribute to an individual's likelihood of experiencing a stroke. The results confirm that age is the dominant predictor, exerting an overwhelming influence on the model's output. This is expected, as age is a well-established risk factor for cardiovascular events. Notably, the model initially produced an R^2 score of 1, indicating perfect prediction accuracy.

While keeping the variable age in our models could result in high model performance, it also raised concerns about the overreliance on a single variable, reducing the interpretability and clinical utility of the model.

The model chosen in this case was a **Random Forest Regression** which, optimized without age, achieved a R^2 score of 0.42, indicating reduced predictive power but a more symptom-driven approach. Chest pain, fatigue & weakness, and high blood pressure emerged as key predictors, reflecting known clinical risk factors.
Unlike the age-driven model, where a single variable dominated predictions, this approach ensures that multiple symptoms collectively influence risk assessment, making the model more interpretable and clinically relevant for identifying high-risk patients based on their presenting symptoms.

From a business perspective, these results highlight the shift from a demographic-driven approach to a symptom-focused model for stroke risk assessment. While age is a known risk factor, relying on it too heavily can overshadow the importance of immediate clinical symptoms that might indicate a stroke risk regardless of age. This model prioritizes chest pain, fatigue, high blood pressure, and breathing issues, aligning more closely with how medical professionals assess patients in real time. This makes the model more practical in emergency settings, where symptoms take precedence over demographic details.

4.2. Classification

To develop a reliable stroke risk classification model, we focused on identifying individuals most at risk, ensuring that the model prioritizes sensitivity over specificity. In a medical context, especially for stroke prediction, missing a high-risk patient can have severe consequences, potentially leading to delayed intervention or lack of preventive measures. The goal was to minimize false negatives, reducing the chances of missing high-risk

patients, to do this we tested **Logistic Regression**, **Random Forest**, and **Gradient Boosting**, selecting hyperparameters that balanced accuracy, sensitivity, and efficiency.

Gradient Boosting achieves the highest recall (0.91), making it the best model for detecting stroke cases, though it has lower precision (0.733), leading to more false positives. Logistic Regression provides a balanced approach with strong recall (0.84) and F1-score (0.81), while Random Forest has the highest precision (0.80) but lower recall (0.72), missing more true stroke cases. Since recall is the priority in medical risk assessment, Gradient Boosting is the best choice, while Logistic Regression remains a viable alternative for a balance between recall and precision.

Key symptoms consistently identified across models include Cold Hands/Feet, Excessive Sweating, High Blood Pressure, Chest Pain, and Anxiety/Feeling of Doom. From a business standpoint, these results emphasize the importance of choosing a predictive model that aligns with real-world healthcare priorities. The selection of Gradient Boosting, due to its high recall, ensures that the model is geared toward minimizing missed stroke cases, which is crucial in clinical settings where early detection can reduce hospitalizations, improve patient outcomes, and lower long-term healthcare costs. While this comes at the expense of higher false positives, in a preventive care and insurance context, flagging more potential stroke cases can lead to proactive interventions, reducing the incidence of severe strokes and associated treatment expenses. Moreover, the identification of key risk factors (e.g., chest pain, high blood pressure, cold hands/feet) allows healthcare providers, insurers, and digital health platforms to refine risk stratification models, enhance preventive screening programs, and personalize health recommendations for at-risk individuals.

5. Actionable Insights and Concluding Recommendations

Integrating **symptom-based risk prediction** into clinical triage protocols can significantly enhance early stroke detection and intervention. Implementing a digital scoring system that flags high-risk symptoms—such as chest pain, fatigue, and high blood pressure—can help clinicians prioritize urgent cases in real time. Additionally, **wearable devices and remote monitoring programs** tracking fluctuations in blood pressure, heart rate irregularities, and persistent fatigue could enable proactive interventions before symptoms escalate. Our analysis confirms that these symptoms, along with cold hands/feet, anxiety, and dizziness, are among the most influential predictors of stroke risk. However, aside from age, **no single symptom overwhelmingly determines risk,** highlighting the importance of a multivariate approach rather than reliance on individual factors.

To enhance predictive accuracy, businesses should **avoid rigid classification thresholds** (e.g., 65% risk), **incorporate explainability tools** like SHAP, and implement continuous model validation to account for evolving risk patterns. Expanding risk assessments beyond traditional demographic metrics and collaborating with healthcare providers for preventive care initiatives can further strengthen both risk management and patient outcomes.

In addition, insurers should consider **partnering with telemedicine platforms** for continuous health monitoring and integrating risk scores into tiered insurance products to support customized wellness programs. These initiatives not only reduce the incidence of major stroke events but also foster long-term client retention and satisfaction.

6. Limitations and Future Work

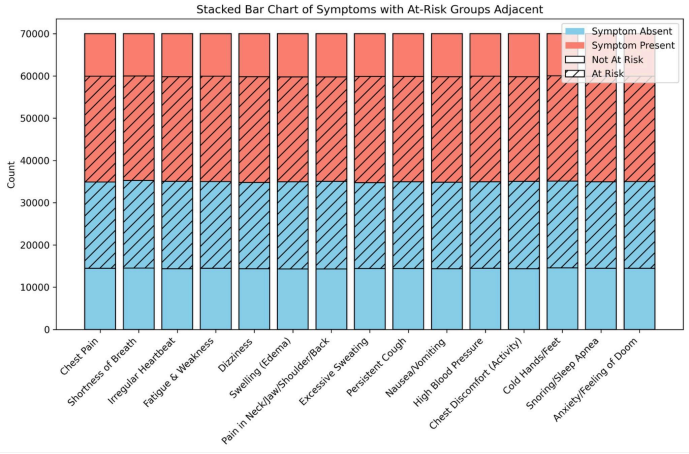
Our study, while providing valuable insights into stroke prediction, has a few notable limitations:

1. **Real-world Data Complexity:**
    - Clinical data often contains noise, missing values, and inconsistencies that our idealized dataset does not account for.
2. **Population-level Biases:**
    - Certain demographic groups may be underrepresented in stroke literature due to healthcare access disparities, delayed diagnosis patterns, or varying symptom presentation across different populations. These biases could affect the generalizability of our model to diverse patient populations and potentially perpetuate existing inequities in stroke detection and care.
3. **Time-series Granularity:**
    - The temporal relationship between symptoms—including their onset, duration, and progression—may be crucial predictive factors that our static dataset cannot capture. Additionally, the sequence in which symptoms appear could provide valuable diagnostic information that our current approach overlooks.

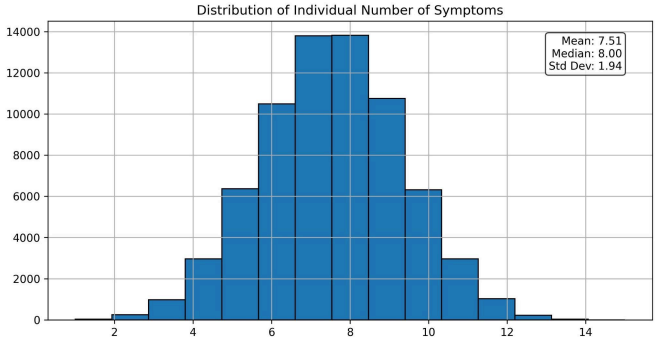
Despite these limitations, our findings demonstrate the potential value of machine learning in stroke risk assessment, particularly when integrated with RPM systems for insurance applications. Future work should focus on validating these models with real-world clinical data, incorporating temporal features, and ensuring equitable performance across diverse patient populations.

7. Appendix

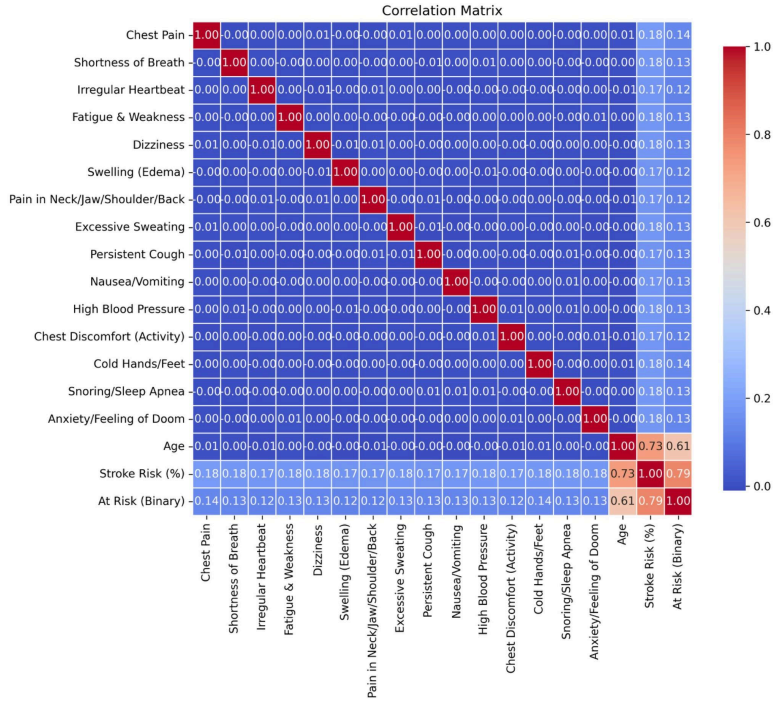
Appendix 2.1 - Symptom Ratio with At-Risk Measure



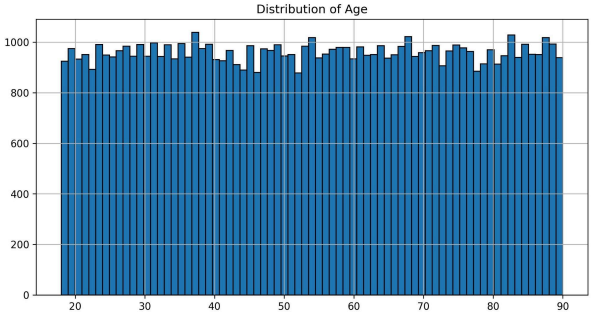
Appendix 2.2 - Distribution of Number of Symptoms

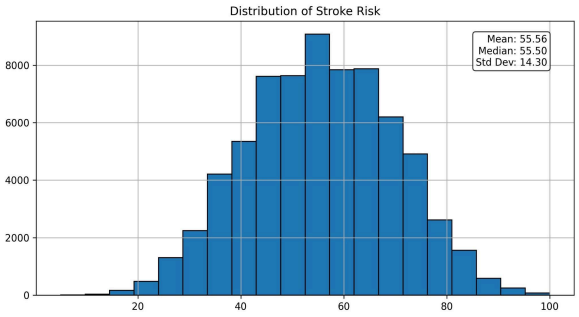


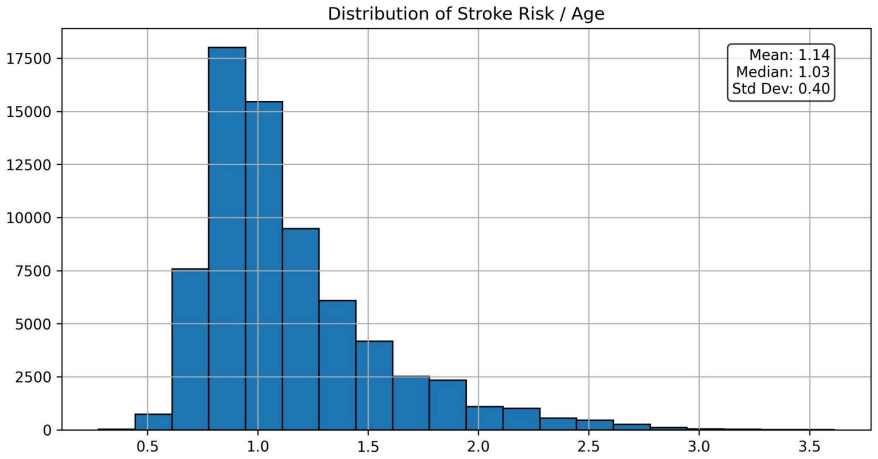
Appendix 2.3 - Variable Correlation



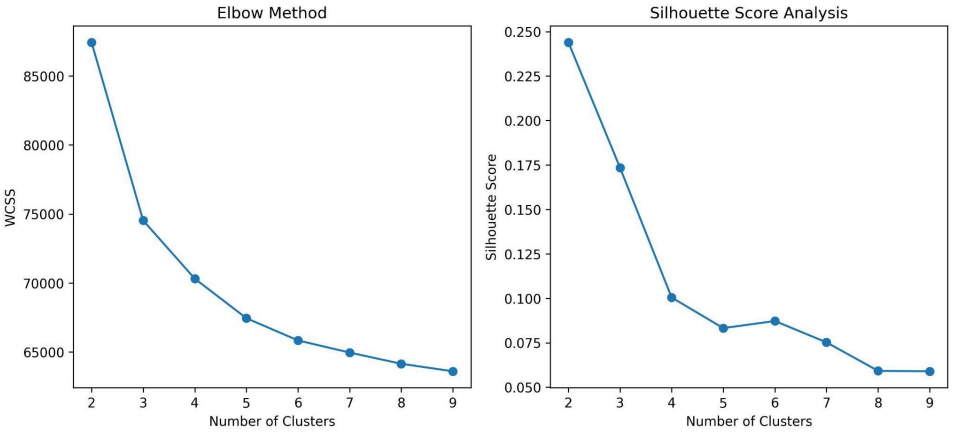
Appendix 2.4 - Distribution of Age, Stroke Risk and Age-Adjusted Stroke Risk



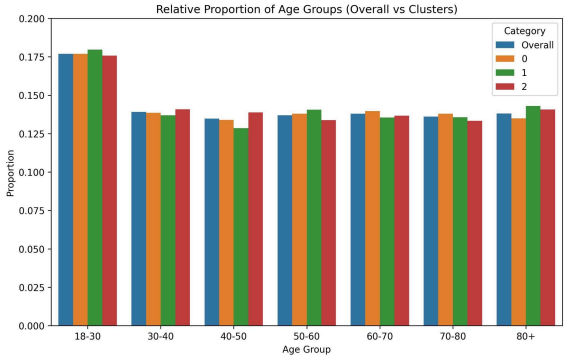


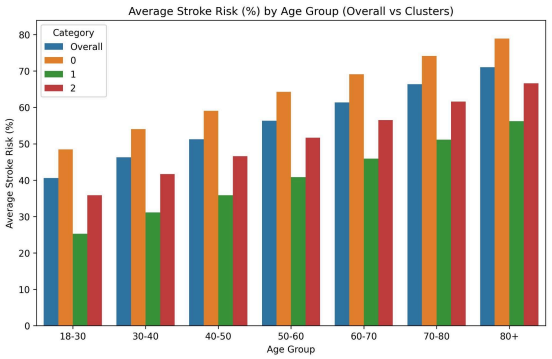


Appendix 3.1 - Clustering Elbow and Silhouette Scores

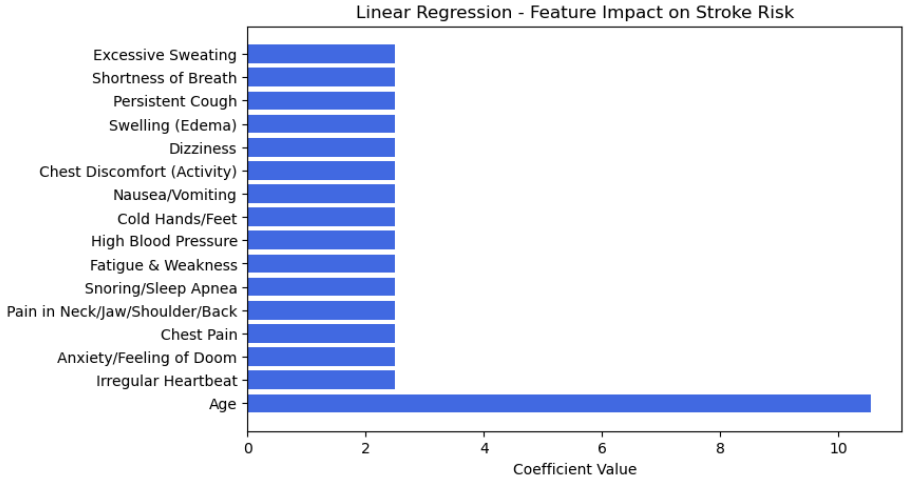


Appendix 3.2 - Clusters by Age Group and Stroke Risk

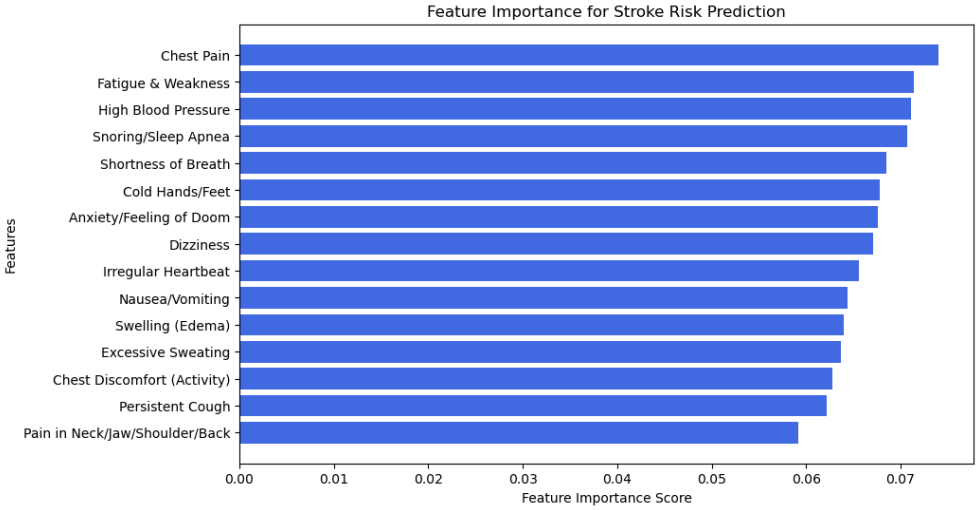




Appendix 4.1.1- Feature Importance of Linear Regression (with variable age)



Appendix 4.1.2- Feature Importance of Random Forest Regressor



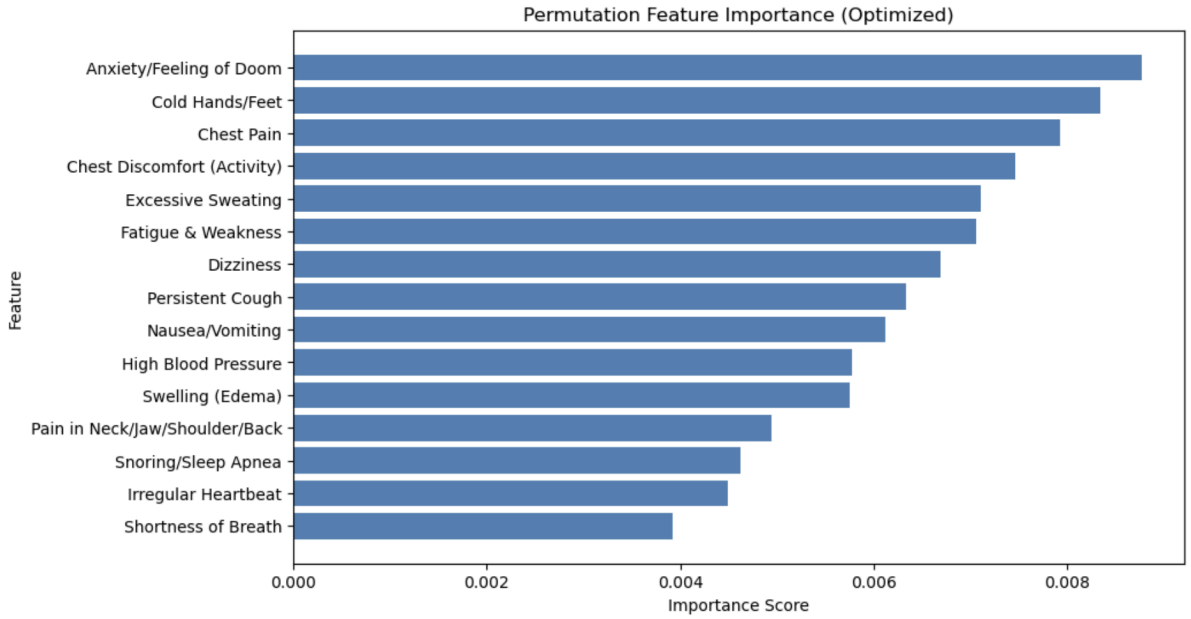
Appendix 4.2.1- Model Performance Per Age Group-Fairness Metrics[1]

Fairness Metrics Across Age Groups:

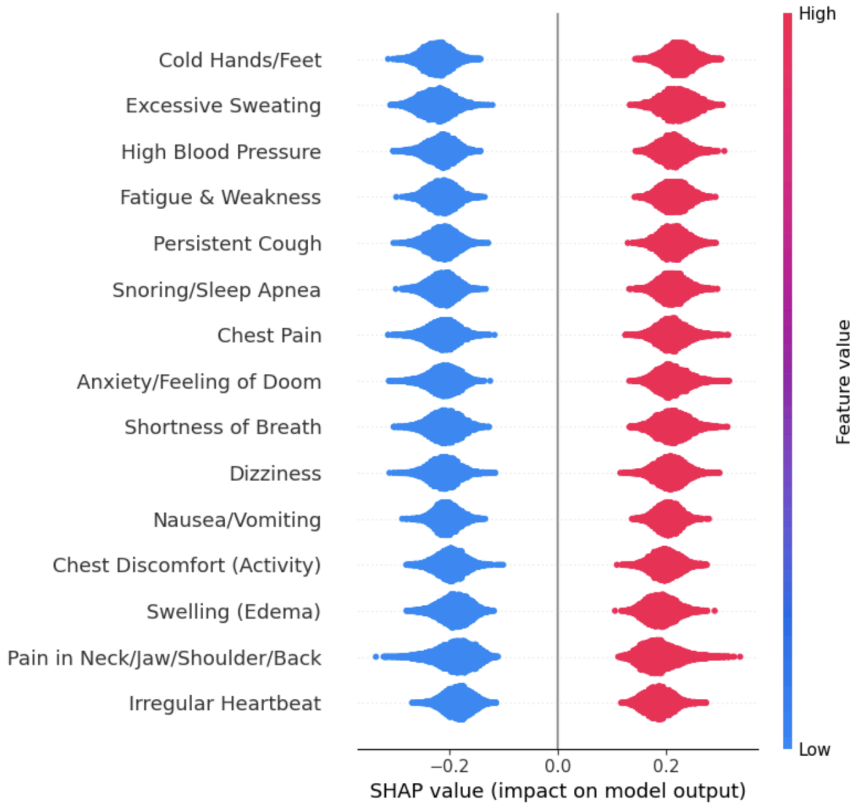
Age	accuracy	precision	recall
40–60	0.917713	0.910015	0.972245
60–80	0.779767	1.000000	0.761249
80+	0.717922	1.000000	0.714211
<40	0.558410	0.367157	1.000000

[1]The fairness metrics across age groups reveal that the model's recall is highest for younger individuals (<40) but lower for older groups (60–80 and 80+), indicating that the model correctly identifies stroke cases more frequently in younger populations. However, precision is much lower for younger individuals (0.367), meaning a higher false positive rate. Conversely, precision is perfect (1.000) for older groups, but their recall is lower (0.714–0.761), suggesting more missed stroke cases. This trade-off highlights a potential bias where the model is more cautious in predicting stroke for older individuals, possibly due to the higher prevalence of stroke-like symptoms unrelated to actual stroke risk.

Appendix 4.2.2- Feature Importance of Gradient Boosting Method[2]

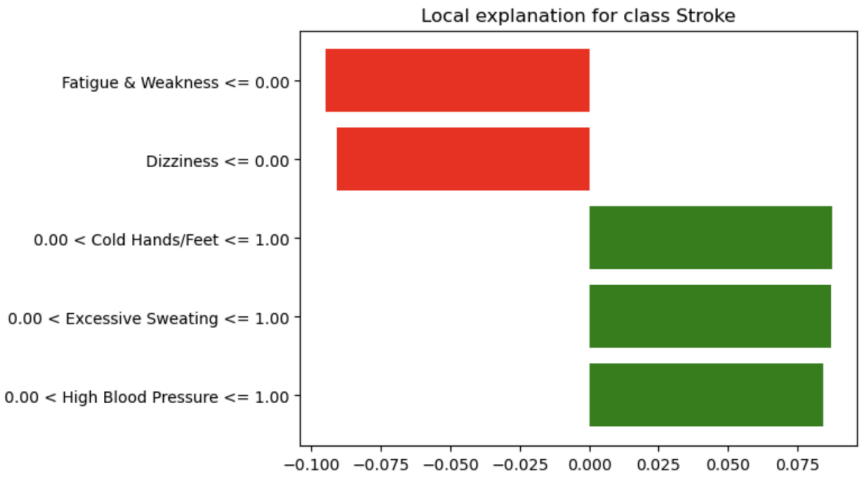


Appendix 4.2.3- Shap Values of Gradient Boosting Method[3]



[2]Optimized for recall, using a random subsample of 1,000 observations for computational efficiency and 15 feature swaps. The score is calculated as the baseline recall minus the average recall when the given feature is swapped out.
[3]Optimized for recall, SHAP values were computed using a random subsample of 1,000 observations for efficiency. The analysis measures the difference in prediction when a feature is present versus absent, considering all possible feature combinations

Appendix 4- Lime [4] (Local Interpretable Model-agnostic Explanations)



8. References

Mahatir Ahmed Tusher. (2025). Stroke Risk Prediction Dataset Based on Symptoms [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/10754870

[4]LIME provides a local interpretation by analyzing how small perturbations in feature values affect the model's decision. For the selected individual (Instance 19), Cold Hands/Feet, Excessive Sweating, and High Blood Pressure positively contributed to the stroke prediction (green bars), increasing the model's confidence. Conversely, the lack of Fatigue & Weakness and Dizziness negatively impacted the prediction (red bars), slightly lowering the stroke probability. However, the positive contributions outweighed the negative, leading to the final classification as stroke-positive