Capstone Project

An In-Depth Analysis of Stroke Risk Factors Using Predictive Analytics Final Report

1. **Define the Problem Statement:**

Stroke is one of the leading causes of death and long-term disability worldwide, with high medical costs and lasting social and economic impacts. Early detection and prevention remain crucial. However, healthcare organizations often lack data-driven tools to proactively identify high-risk individuals before an event occurs.

This project seeks to analyze a publicly available stroke dataset to uncover key demographic, lifestyle, and health factors associated with stroke occurrence. By developing predictive and interpretable models, the study aims to support data-informed decision-making that could help healthcare teams identify patients at elevated risk and tailor preventive care strategies.

1. **Model Outcomes or Predictions:**

The identification of the top risk factors for stroke includes critical elements such as age, glucose levels, body mass index (BMI), and hypertension. By comparing the performance of various predictive models, we can determine which algorithm proves to be the most effective in forecasting stroke risk. Additionally, creating an interactive dashboard can effectively visualize the likelihood of stroke based on demographic and health-related attributes.

Ultimately, these predictive insights can be leveraged to enhance health monitoring practices and inform community wellness programs, leading to more targeted interventions and improved health outcomes. The stroke prediction project uses supervised learning, because I am training the models on known labels (stroke = 0 or 1) to predict outcomes for new data.

1. **Data Acquisition:**

A dataset sufficient to answer descriptive, diagnostic, and predictive questions should include both medical and behavioral variables commonly associated with stroke. An appropriate stroke dataset should include the following information:

* Demographics: gender, age, residence type, marital status
* Medical history: hypertension, heart disease, BMI, average glucose level
* Lifestyle: work type, smoking status
* Target variable: stroke (yes or no)

Source: kaggle.com

1. **Data Preprocessing/Preparation:**

Before moving into modeling, exploratory data analysis (EDA) was conducted to gain a clear understanding of the stroke dataset. The goal was to understand the data, check for missing values and patterns, and examine how factors such as age, glucose levels, and BMI relate to stroke occurrence. This process helped reveal initial trends and potential key variables, providing a solid foundation for building and interpreting predictive models.

The EDA provided a clear understanding of the stroke dataset and revealed several important patterns. The data showed that stroke cases accounted for a small proportion of the sample, confirming a significant class imbalance. Key variables such as age, average glucose level, and BMI displayed visible differences between individuals who experienced a stroke and those who did not. In particular, higher glucose levels and older age appeared more common among stroke cases.

Some missing values were identified, especially in BMI, but these were handled through imputation. Overall, the EDA established a solid foundation for the modeling phase by highlighting which factors may be most influential in predicting stroke occurrence.

Then, the modeling and analysis phase is as follows.

* 1. Split the data into training and testing sets (80/20) using stratification to keep stroke cases balanced.
  2. Prepare a baseline Logistic Regression model, then proceed to Random Forest, SVM, and Gradient Boosting for comparison.
  3. Evaluate performance using Accuracy, ROC-AUC, Precision, Recall, and F1-score.
  4. Identify the best-performing model and key predictive features.

1. **Modeling:**

The Logistic Regression model will serve as the baseline, providing an interpretable way to understand how each variable affects the likelihood of a stroke. It will set a benchmark for both accuracy and explainability. Following this, the Random Forest model will be employed, which utilizes an ensemble of decision trees to capture nonlinear relationships and determine the most critical factors influencing stroke outcomes.

The third model we will implement is the Support Vector Machine (SVM), which uses a margin-based approach to classify cases as either stroke or non-stroke. By employing a kernel function such as the Radial Basis Function (RBF), SVM can effectively handle complex, nonlinear data patterns while addressing class imbalance through class weighting.

Lastly, we will develop a Gradient Boosting model, such as XGBoost or LightGBM, to achieve the highest predictive accuracy. This model will build trees sequentially to minimize errors and will provide feature-level explanations using SHAP (Shapley Additive Explanations) values.

All models will be evaluated using consistent performance metrics, accuracy, ROC-AUC, precision, recall, and F1-score, with cross-validation to ensure reliability. The best-performing model will be selected based on a balance of accuracy, interpretability, and ethical fairness.

1. **Model Evaluation:**

The following four supervised learning models have been trained and tested: Logistic Regression, Random Forest, Support Vector Machine (SVM with RBF kernel), and Gradient Boosting. All models followed the same steps to prepare the data, which included filling in missing values, scaling, and one-hot encoding. Their performance were evaluated on the test dataset using two key metrics: ROC-AUC and PR-AUC (Average Precision). These metrics help us understand how well each model performs, especially with imbalanced data, where simply looking at accuracy can be misleading.

1. **Logistic Regression (Baseline)**

Logistic Regression provides a good starting point for understanding the factors that predict strokes. It shows a moderate ability to spot stroke cases, but it struggles with identifying true positives due to a large imbalance in the data, with very few strokes compared to the overall number of cases. While it can generally tell the difference between cases, its low performance in finding positive cases limits its usefulness. Despite these challenges, Logistic Regression helps us understand how different factors contribute, but it cannot handle complex, nonlinear relationships.

1. **Random Forest**

The Random Forest model performed better than Logistic Regression because it can capture complex relationships between variables and reduce overfitting by averaging results from multiple trees. It received a higher ROC-AUC and improved PR-AUC scores, which means it is better at identifying stroke cases while minimizing false positives and false negatives. The Random Forest model also ranks feature importance, showing which factors most impact stroke prediction, such as age, glucose level, and hypertension.

1. **Support Vector Machine (SVM – RBF Kernel)**

The SVM model produced moderate results, performing slightly worse than Random Forest and Gradient Boosting in both metrics. SVMs work well with complex decision boundaries, but they are sensitive to how data is scaled and can require a lot of computation, especially with one-hot encoded categorical data. Additionally, SVMs are not very interpretable, making them less suitable for healthcare applications, where explainability and transparency are essential.

1. **Gradient Boosting**

Gradient Boosting performed the best among all the models. It had the highest scores for ROC-AUC and PR-AUC, which measure how well the model can classify cases. This model learns in steps, allowing it to fix mistakes made by earlier trees. As a result, it is better at telling apart stroke cases from non-stroke cases. It was particularly good at identifying stroke cases, which are less common. Additionally, it shows which features are most important using SHAP values. This combination of strong prediction and clear explanations makes Gradient Boosting a great choice for predicting health risks.

**Conclusion**

Overall, Gradient Boosting proved to be the most effective model for predicting stroke occurrence, striking an optimal balance between accuracy and sensitivity to minority cases. Random Forest also performed well, providing valuable interpretive insights. In contrast, Support Vector Machine (SVM) and Logistic Regression offered useful benchmarks but exhibited lower predictive power. It is worthwhile to note that accuracy and F1-score can be misleading for imbalanced datasets, like the stroke dataset, so we should focused instead on ROC-AUC and PR-AUC, which provide a more realistic picture of performance. Because stroke cases are rare, ROC-AUC and PR-AUC tell us far more about how well the model identifies the people most at risk.