**A black and red logo

Description automatically generated**

**Marshall School of Business**

Final Project Report On

**“Stroke Prediction Modeling”**

Submitted in partial fulfillment for the subject

**DSO-568 : Healthcare Analytics**

**Academic Term:** Fall 2024

**Submission Date:** November 25, 2024

**Submitted by: Team 5**

Campbell Duncan, Jnana Kundur Prakash, Lindsay Yan,

Lucian Deng, Marliese Sanabria, Zoha Akhtar

**Course Instructor:**

Cosimo Arnesano, Assistant Professor

Data Sciences and Operations Department

USC Marshall School of Business

**Executive Summary:**

Stroke is a leading cause of death and long-term disability worldwide, affecting millions annually. Its unpredictability and rapid onset make it a critical area for preventive healthcare interventions. With timely prediction, healthcare systems can shift from reactive treatment to proactive care, potentially saving lives, reducing disability, and lowering healthcare costs.

In this project, we utilized the Kaggle Stroke Prediction Dataset, which contains 5,110 records and 12 features, to develop machine learning models aimed at predicting stroke risk. The dataset includes demographic, clinical, and lifestyle attributes, such as age, BMI, smoking status, and average glucose level, alongside stroke labels. The dataset reflects a significant class imbalance, with stroke cases constituting only 4.87% (249 cases) of the total records. Key challenges addressed include handling missing data, dealing with this class imbalance, and identifying significant predictive factors.

We began with data preprocessing, exploratory analysis, and feature engineering to prepare the dataset for machine learning. A series of models were evaluated, starting with Logistic Regression as a baseline. Advanced models like Decision Trees, Random Forests, and XGBoost were implemented to enhance prediction performance. To address the significant class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, creating a balanced training dataset with stroke and non-stroke cases.

The ensemble model combining Random Forest and XGBoost achieved the highest performance, with an accuracy of 95.47% and a ROC-AUC score of 99.04%. This result underscores the ensemble's ability to balance precision and recall while minimizing false positives and negatives. Age & average glucose level emerged as the most significant predictors of stroke risk.

This project highlights the potential of machine learning in transforming stroke prediction and preventive healthcare. Ethical considerations, such as data privacy and bias mitigation, were prioritized throughout the analysis. The outcomes not only enable healthcare providers to focus preventive measures on high-risk individuals but also align with public health goals of reducing the global burden of stroke. Future extensions could include real-time data integration, increasing the dataset size with additional records, and incorporating personalized treatment recommendations to enhance prediction accuracy further.

**1. Problem Definition:**

**Healthcare Process:**

The healthcare process targeted in this project is the early identification of patients who are at high risk for stroke. Stroke is one of the leading causes of death and long-term disability globally, creating a significant burden on individuals, families, and healthcare systems. Strokes occur when the blood supply to the brain is interrupted, often due to blocked or ruptured blood vessels, and they can lead to severe neurological impairments or death. Many stroke cases are preventable if high-risk patients are identified early and provided with appropriate interventions. This process is critical because timely identification allows healthcare providers to implement preventive strategies, such as lifestyle changes, medication, and regular monitoring, which can significantly reduce the risk of stroke and improve patient outcomes.

This process directly impacts multiple stakeholders:

* **Patients**: Early identification empowers patients to take preventive measures, improving their quality of life and reducing the risk of severe disability or mortality.
* **Healthcare Providers**: It enables clinicians to allocate resources more efficiently, focusing on high-risk individuals who require immediate attention.
* **Healthcare Systems**: Proactive stroke prevention can reduce the financial burden associated with acute stroke treatment and long-term care for stroke survivors.

**Significance:**  
Predictive modeling in this context is transformative because it shifts the focus of healthcare from reactive treatment to proactive prevention. Many healthcare systems are overwhelmed by the cost and resource demands of treating preventable conditions. Stroke prevention through predictive modeling offers several significant benefits:

1. **Enhancing Patient Risk Stratification**:

Predictive models allow for the identification of individuals who are at the greatest risk for stroke based on demographic, clinical, and lifestyle data. By stratifying patients according to their risk, healthcare providers can prioritize resources and interventions for those who need them the most.

1. **Facilitating Preventive Interventions**:

High-risk patients identified by the model can benefit from targeted interventions, such as blood pressure management, glucose monitoring, and smoking cessation programs. These measures can lower the incidence of strokes and reduce long-term complications.

1. **Optimizing Resource Allocation**:

Stroke prevention is far more cost-effective than treatment. Predictive models enable healthcare providers to focus their efforts on high-risk patients, reducing unnecessary interventions for low-risk individuals and ensuring that resources such as diagnostics, medications, and follow-ups are used efficiently. This not only improves outcomes but also reduces the financial burden on healthcare systems.

**Objective:**  
The primary goal of this project is to develop a robust and accurate machine learning model that predicts stroke risk based on an individual’s demographic, clinical, and lifestyle factors. The model aims to:

1. **Support Early Intervention**:

By predicting stroke risk early, healthcare providers can intervene before a stroke occurs. This can include medical interventions, lifestyle modifications, or regular monitoring to mitigate risk factors.

1. **Enhance Healthcare Outcomes**:

Reducing the number of strokes translates to improved health outcomes, decreased mortality rates, and fewer cases of long-term disability. This aligns with the broader goal of improving public health and patient well-being.

1. **Provide Actionable Insights for Resource Planning**:

Insights from the predictive model can guide healthcare providers and administrators in making data-driven decisions regarding resource allocation, such as focusing preventive efforts on high-risk populations, planning community health programs, and ensuring adequate staffing and infrastructure to handle potential stroke cases.

**4. Insights and Interpretation:**

**Key Findings:**

1. **Age and Glucose Levels as Strong Predictors:**
   * **Significance of Age:**
     + The analysis and model results confirm that age is a critical factor in predicting stroke risk.
     + Older individuals, especially those above 60 years of age, exhibited a significantly higher likelihood of experiencing strokes.
   * **Impact of Glucose Levels:**
     + Elevated glucose levels, particularly above 150 mg/dL, were strongly associated with increased stroke risk.
     + This finding underscores the importance of regular monitoring and management of blood sugar levels as part of preventive healthcare strategies.
2. **SMOTE for Addressing Class Imbalance:**
   * **Class Imbalance Problem:**
     + The dataset displayed a significant class imbalance, with far fewer stroke cases (minority class) compared to non-stroke cases (majority class).
   * **Effectiveness of SMOTE:**
     + By applying SMOTE (Synthetic Minority Oversampling Technique), the minority class was oversampled, leading to better model performance in identifying stroke cases.
     + This technique improved the model's recall for the stroke class, enabling it to accurately predict stroke occurrences without being biased toward the majority class.
   * **Outcome:**
     + The application of SMOTE ensured a fair representation of both classes, which is crucial for healthcare analytics where the minority class often represents critical outcomes.
3. **Superior Performance of the Ensemble Model:**
   * **Ensemble Model Advantages:**
     + The ensemble model, combining Random Forest and XGBoost, achieved the highest performance among all tested models.
     + It delivered the highest accuracy (95.47%) and ROC-AUC score (99.04%), making it the most reliable and robust model for stroke prediction.
   * **Key Strengths:**
     + By leveraging the complementary strengths of Random Forest and XGBoost, the ensemble model captured complex relationships in the data.
     + It demonstrated a balance between precision, recall, and overall predictive performance, outperforming standalone models.
   * **Recommendation:**
     + The ensemble approach is ideal for tasks requiring high accuracy and reliability, particularly in healthcare where predictive insights can significantly impact patient outcomes.

**Actionable Insights:**

1. **Focus on High-Risk Factors:**
   * **Age and Glucose Levels:**
     + These should be prioritized as critical features in stroke risk assessments and healthcare interventions.
     + Preventive measures, such as routine screenings for older individuals and glucose management programs, should be implemented.
2. . **Adopt Data Balancing Techniques:**
   * **Balanced Datasets:**
     + Techniques like SMOTE are essential for building fair and accurate predictive models, especially in healthcare datasets with imbalanced classes.
   * **Improved Model Performance:**
     + Balancing datasets enhances the model's ability to predict critical outcomes (e.g., stroke), reducing the likelihood of false negatives
3. **Implement Ensemble Models:**
   * **Reliability in Healthcare Analytics:**
     + Ensemble methods like the Random Forest-XGBoost combination offer significant advantages in predictive healthcare analytics.
     + These models ensure high accuracy, reliability, and robustness, making them suitable for healthcare tasks where precision is critical.
4. **Continuous Monitoring and Updates:**
   * Models should be regularly updated with new data to maintain accuracy and adapt to changing patient demographics and healthcare trends.
   * Incorporating additional features (e.g., cholesterol levels, physical activity) can further enhance predictive performance.

These insights can guide healthcare providers in designing targeted interventions, optimizing resource allocation, and improving patient outcomes, ultimately contributing to effective stroke prevention and care strategies.

**5. Ethical and Privacy Considerations:**

In healthcare analytics, ethical and privacy considerations are crucial to ensure that predictive models are effective while aligning with legal, moral, and social standards. Below are detailed considerations for the stroke prediction model:

**1) Patient Privacy and Data Confidentiality:**

* **Data Anonymization:** All personal identifiers in the dataset (e.g., name, address, social security number) must be removed or anonymized to protect patient identities. Compliance with regulations like HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) is essential for handling sensitive health data.
* **Secure Data Storage:** Data should be securely stored using encryption techniques and secure servers to prevent unauthorized access. Access must be restricted to authorized personnel with legitimate reasons for usage.
* **De-Identification Techniques:** Data attributes that could inadvertently identify individuals, such as rare demographic combinations or unique conditions, should be de-identified to maintain confidentiality.

**2) Fairness and Bias in Predictions:**

* **Addressing Model Bias:** Predictive models may inherit biases from historical data, such as underrepresentation of specific age groups, genders, or socioeconomic classes. Regular audits should be conducted to identify and mitigate such biases to ensure equitable predictions for all patient groups.
* **Avoiding Discrimination:** The model must not disproportionately benefit or harm specific demographic groups. Subgroup analyses should ensure consistent performance across various populations, including age, gender, and socioeconomic groups.
* **Transparency in Decision-Making:** The model should provide interpretable outputs so healthcare providers can understand the rationale behind predictions. Avoid reliance on "black-box" models without explainability, especially in critical healthcare applications.

**3) Informed Consent and Ethical Use of Data:**

* **Informed Consent from Patients:** Data used for training and predictions should be collected with explicit patient consent. Patients should be informed about how their data will be used, who will access it, and the potential benefits and risks involved.
* **Ethical Use of Predictive Models:** The stroke prediction model should support clinical decision-making rather than replace medical expertise. Predictions must be contextualized by healthcare professionals alongside other clinical factors.

**4) Transparency and Accountability:**

* **Transparency in Limitations:** The model’s limitations, such as its reliance on specific features like age and glucose levels, must be clearly communicated to stakeholders. Predictions should be supplemented with clinical judgment and not blindly trusted.
* **Accountability for Decisions:** Protocols should establish accountability for the model’s use in clinical settings. Responsibilities of healthcare providers and data scientists should be clearly defined to ensure ethical and effective application.

**5) Ethical Handling of False Positives and False Negatives:**

* **Balancing Errors:** In stroke prediction, false negatives (missing a high-risk patient) can have severe consequences, while false positives (misclassifying a low-risk patient) can lead to unnecessary anxiety or interventions. The model should prioritize reducing false negatives while minimizing false positives to avoid overburdening healthcare resources.
* **Impact on Patient Outcomes:** Ethical considerations should ensure that false predictions do not result in harm, such as unnecessary medical procedures or neglect of high-risk patients.

**6) Continuous Monitoring and Updates:**

* **Performance Monitoring:** Regularly monitor the model’s performance to address any drift in accuracy or fairness. Re-training with updated data will ensure the model reflects current population trends and medical advancements.
* **Stakeholder Involvement:** Healthcare professionals, patients, and ethicists should be included in discussions about the model’s use, limitations, and updates to ensure its continued ethical application.

By addressing these ethical and privacy considerations, the stroke prediction model can be responsibly implemented to enhance patient outcomes, build trust among stakeholders, and ensure compliance with healthcare standards. These measures are essential to balancing the benefits of predictive analytics with the ethical responsibility to protect patients’ rights and well-being.

**6. Recommendations:**

Based on the results and insights derived from the predictive models, the following recommendations are provided to optimize stroke prevention and healthcare decision-making:

**1) Focus on High-Risk Groups:**

* **Age-Based Interventions:**
  + Older individuals, particularly those above 60 years, are at a significantly higher risk of stroke. Healthcare providers should prioritize routine screenings and targeted preventive measures for this demographic.
* **High Glucose Monitoring:**
  + Patients with elevated glucose levels (>150 mg/dL) should be closely monitored and provided with lifestyle recommendations and medical interventions to manage blood sugar levels effectively.

**2) Enhance Preventive Care Strategies:**

* Use the model's predictions to identify patients who are at a higher risk of stroke and provide tailored preventive care. For example:
  + **Nutritional counseling** to manage BMI and glucose levels.
  + **Regular check-ups** and diagnostics for individuals with hypertension or heart disease.
  + **Smoking cessation programs** for patients identified as current or former smokers.

**3) Utilize the Ensemble Model in Clinical Settings:**

* The ensemble model, combining Random Forest and XGBoost, achieved the best performance with an accuracy of 95.47% and ROC-AUC of 99.04%. This model can be integrated into clinical workflows to predict stroke risks with high confidence.
* **Decision Support Tool:**
  + Deploy the model as a decision support tool for healthcare providers to complement their clinical judgment, focusing on high-risk cases flagged by the model.

**4) Address Class Imbalance in Future Applications:**

* The use of SMOTE in the project demonstrated the importance of addressing class imbalance in the dataset. This technique improved the model's ability to predict stroke cases without compromising overall accuracy.
* Future implementations should ensure balanced datasets to maintain fairness and predictive accuracy, particularly in datasets with rare but critical outcomes like stroke.

**5) Leverage Insights for Resource Allocation:**

* **Optimize Healthcare Resources:**
  + Insights from the model can guide resource allocation, such as assigning specialized care teams to high-risk patients or planning hospital admissions for potential stroke cases.
* **Preventive Outreach Programs:**
  + Community-based outreach programs can be designed to educate at-risk populations, emphasizing age, glucose level, and lifestyle management as critical factors.

**6) Incorporate Additional Features for Improved Accuracy:**

* The results indicate that variables such as age, glucose level, and hypertension are strong predictors of stroke. Future iterations of the model can incorporate:
  + **Additional clinical features,** such as cholesterol levels and blood pressure trends.
  + **Behavioral data,** such as physical activity levels, sleep patterns, and stress management.

**7) Ethical Usage and Regular Audits:**

* Regular audits should be conducted to ensure the model’s predictions are unbiased and equitable across all patient demographics.
* The model should be transparent in its predictions, with clear explanations provided to clinicians about the factors influencing the results.
* Patient privacy must be protected, and sensitive data should be anonymized to comply with regulatory standards such as HIPAA.

**8) Continuous Model Improvement:**

* Regularly update the model with new data to maintain accuracy and relevance in dynamic healthcare environments.
* Test the model on additional datasets from different populations to validate its generalizability and robustness.

**Conclusion:**

This project successfully developed a robust predictive model for identifying stroke risk among patients, addressing one of the most pressing healthcare challenges. By utilizing advanced machine learning techniques, the project overcame challenges such as class imbalance and feature selection, achieving exceptional performance metrics. The ensemble model, combining the strengths of Random Forest and XGBoost, emerged as the most effective predictive approach, with an accuracy of 95.47% and an ROC-AUC score of 99.04%. This model demonstrated superior reliability, interpretability, and predictive power, making it a valuable tool for stroke risk assessment.

The insights derived from this project emphasize the critical role of age and glucose levels in predicting stroke risk. Additionally, the use of SMOTE to address class imbalance highlighted the importance of balanced datasets in ensuring fair and accurate predictions, particularly for healthcare scenarios where the minority class often represents critical outcomes. These findings have direct implications for healthcare practices, supporting proactive measures to improve patient outcomes and optimize healthcare resource allocation.

This project not only contributes to the field of healthcare analytics but also underscores the potential of predictive modeling in addressing real-world medical challenges. By integrating these findings into clinical workflows, healthcare providers can enhance preventive care strategies, reduce stroke incidence, and improve overall patient well-being.