Stroke Risk Analysis Report

Comprehensive Analysis of Healthcare Dataset for Stroke Prediction and Prevention

Dataset: healthcare-dataset-stroke-data.csv | Records: 5,110 | Features: 12

Executive Summary

This analysis examines healthcare data to identify key risk factors for stroke and develop predictive models for early detection. The dataset reveals that approximately **4.9% of patients** experienced a stroke, with age, cardiovascular conditions, and metabolic health being the primary drivers.

Total Records	Stroke Prevalence	Average Age	Hypertension
5,110	4.9%	43 yrs	9.7%

Key Insight: Stroke risk is strongly associated with age, hypertension, heart disease, elevated glucose levels, and smoking. Predictive models achieved up to **86% ROC-AUC** in identifying high-risk patients.

Analytical Approach

Descriptive Analytics

- Average patient age: ~43 years
- Gender distribution: 58.6% Female, 41.4% Male
- Hypertension prevalence: 9.7%
- Heart disease prevalence: 5.4%
- Average glucose level: ~106 mg/dL
- Average BMI: ~28.9 (overweight range)

Value: Establishes baseline understanding of population health characteristics.

Diagnostic Analytics

- Stroke patients disproportionately older (>60 years)
- Hypertension and heart disease significantly increase risk
- Higher BMI and glucose levels correlate with stroke occurrence
- Former and current smokers at elevated risk

Value: Identifies why strokes occur in specific population segments.

Predictive Analytics

- Multiple ML models tested (Logistic Regression, Decision Tree, Random Forest)
- Random Forest achieved best performance (ROC-AUC: 0.86)
- Key predictors: Age, glucose levels, BMI, hypertension, heart disease

Value: Enables forecasting of stroke risk for preventive healthcare.

Prescriptive Analytics

- Lifestyle interventions for high-risk patients
- Targeted screening for patients over 50
- Public health campaigns for at-risk demographics
- Integration of predictive models into healthcare systems

Value: Provides actionable steps to reduce risk and healthcare costs.

Key Risk Factors

Age	Average Glucose Level	BMI	Hypertension	Heart Disease	Smoking Status
High Impact	High Impact	High Impact	Medium Impact	Medium Impact	Low Impact

Interpretation: Traditional cardiovascular risk factors (age, hypertension, heart disease) combined with metabolic health indicators (glucose, BMI) are the primary drivers of stroke risk in this population.

Predictive Model Performance

Logistic Regression	Decision Tree	Random Forest
ROC-AUC: 0.81	ROC-AUC: 0.78	ROC-AUC: 0.86
Good baseline model with high interpretability	Easy to explain but prone to overfitting	Best performance with balanced metrics

Implementation Note: The Random Forest model provides the most accurate stroke risk predictions while maintaining interpretability through feature importance rankings.

Strategic Recommendations

For Healthcare Providers

- Implement targeted screening for patients over 50 with hypertension or heart disease
- Develop lifestyle intervention programs focusing on weight management and glucose control
- Integrate predictive models into Electronic Medical Records for real-time risk alerts
- Establish smoking cessation support programs

For Policymakers

- Fund community health programs in rural and underserved areas
- Launch public awareness campaigns about stroke warning signs and risk factors
- Incentivize workplace wellness programs through tax benefits
- Allocate resources for preventive care infrastructure

For Insurance Companies

- Develop risk-based premium models using predictive analytics
- Offer discounts for policyholders who complete preventive health screenings
- Cover preventive care and lifestyle intervention programs
- Partner with healthcare providers for early intervention initiatives

Limitations & Future Research

- **Imbalanced Dataset:** Only 4.9% stroke cases may lead to under-detection; future work should apply techniques like SMOTE
- **Self-reported Data:** Lifestyle factors (smoking, work type) may suffer from reporting bias
- **Cross-sectional Nature:** Lack of longitudinal data prevents observation of risk evolution over time
- **External Validation:** Models should be tested on independent datasets to ensure generalizability

Conclusion

This analysis demonstrates that **stroke risk is primarily driven by age, cardiovascular health (hypertension and heart disease), metabolic health (BMI and glucose levels), and lifestyle factors (smoking)**. The Random Forest model effectively identifies high-risk patients with an ROC-AUC of 0.86.

By integrating these insights into **preventive care strategies, public health campaigns, and insurance policies**, stakeholders can significantly reduce the burden of strokes on both individuals and healthcare systems. Early detection through predictive analytics combined with targeted interventions represents a cost-effective approach to improving population health outcomes.

Final Recommendation: Implement a multi-stakeholder approach combining predictive screening, lifestyle interventions, and public awareness to reduce stroke incidence and healthcare costs.