



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Data Warehousing & Data Mining

Component-2

Student Name: Iram Naz

Student ID: 2245084

Module Name: ACCA6026

Lecturer's Name: Dr. Nouh Elmitwally

Submission Date: 25/03/2026

Contents

Comprehensive Stroke Risk Prediction Analysis Report	2
Handling Missing Values	8
Encoding Categorical Variables	8
Balancing the Dataset	8
Feature Engineering	9
Post-processing Steps	9
Algorithm Selection and Justification	10
Model Selection and Justification	10
Random Forest	10
Logistic Regression	10
XGBoost	11
Neural Network (Multi-Layer Perceptron)	11
Evaluation Metrics Justification	12
Primary Metric: AUC-ROC	12
Secondary Metrics	12
Confusion Matrix Analysis (Random Forest, Threshold = 0.35)	12
Comparative Performance Analysis	12
Key Findings	13
Feature Importance Analysis (Random Forest)	13
Clinical Validation	13
1.5 Business Recommendations	13
Recommendations for Healthcare Providers	13
Recommendations for Public Health Organizations	14
Implementation Roadmap	15
1.6 Reflective Commentary	15
Model Performance Assessment	15
Strengths of the Analysis	15
Limitations and Challenges	16
Lessons Learned and Future Improvements	16
Conclusion	17

Comprehensive Stroke Risk Prediction Analysis Report

Purpose and Practice of Data Mining

Data mining is the process of identifying significant patterns, correlations, and information from massive datasets using techniques borrowed from statistics, machine learning, and database systems. It goes beyond simple data storage or reporting by identifying underlying trends that can be utilised to generate forecasts and influence decision-making. Data mining is a fundamental part of predictive analytics in modern businesses, which analyses past data to anticipate future risks, actions, or events. This ability to turn raw data into useful information has made data mining crucial in both healthcare and corporate sectors.

In healthcare, huge amounts of patient data are generated everyday through clinical records, laboratory reports, diagnostic tests, and wearable devices. Data mining makes it possible to systematically examine this data in order to enhance risk assessment, early diagnosis, and individualised treatment planning. For example, classification algorithms can estimate the chance of a patient getting illnesses such as diabetes, cardiovascular disease, or stroke based on demographic, lifestyle, and medical characteristics. These forecasts make it possible for medical practitioners to put early intervention techniques into practice, which can drastically lower death rates and medical expenses. Additionally, clustering approaches can group individuals with similar health features, allowing physicians to identify high-risk populations and devise focused prevention initiatives. As a result, data mining contributes directly to improving patient outcomes and optimising resource allocation within healthcare systems.

Beyond healthcare, data mining is commonly utilised in business to support decision support systems (DSS). Businesses use consumer data analysis to forecast demand, identify fraudulent activity, predict purchasing behaviour, and improve marketing strategies. Predictive models are used by banks, insurance companies, and retailers to reduce risk and increase operational effectiveness. The fundamental idea is the same as in healthcare: making educated, proactive decisions instead of reactive ones by utilising historical data.

In this assignment, a dataset including demographic, lifestyle, and medical characteristics is used to predict strokes using data mining techniques. By employing classification and clustering approaches, the investigation reveals how predictive models can assist healthcare providers in identifying individuals at increased risk of stroke. This highlights the actual application of data mining as a tool for predictive analytics and clinical decision support, ultimately contributing to more effective and data-driven healthcare management.

Dataset Description and Exploration

In order to create predictive models for stroke risk, this investigation makes use of a large dataset that includes 5,510 patient records with 11 clinical and demographic characteristics. The dataset, which captures a variety of patient profiles across age groups, socioeconomic origins, and health issues, comes from actual healthcare settings. By utilising machine learning techniques, this study intends to produce accurate, interpretable models that can enhance clinical decision-making and targeted intervention programs.

Data exploration, pre-processing, model creation, and business application are all included in the organised methodology used for the analysis. The ultimate objective is to translate data-driven insights into practical recommendations that will strengthen stroke prevention efforts and improve patient outcomes.

Attributes and Data Types

Eleven variables that indicate patient demographic and clinical data are included in the dataset. These variables include both numerical and categorical data types:

- **Categorical variables:** gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status, and stroke (target variable)
- **Numerical variables:** age, avg_glucose_level, and bmi

Although **age** is numerical, it includes fractional values (e.g., 0.08, 0.32 years), indicating the presence of infant records in the dataset.

Overview of the Dataset

- Total instances: 5,510 patient records
- Stroke cases: 249 (4.52%)
- Non-stroke cases: 5,261 (95.48%)

Less than 5% of all observations are stroke cases, indicating a significant class imbalance in the dataset. In order to prevent biased predictions in favour of the majority class, this imbalance must be carefully addressed during model building.

Missing Values Analysis

Attribute	Missing Values	Percentage
bmi	201	3.65%
smoking_status	1,544	28.02%
Others	0	0%

Note: A considerable number of missing values appears in smoking status (28%) and a lower proportion in bmi (3.65%). To guarantee dependable model performance, these missing values necessitate the use of suitable pre-processing techniques.

Summary Statistics and Variable Distributions

Age Distribution

- Range: 0.08 to 82 years
- Mean: 43.23 years
- The dataset includes individuals from infancy to old age.
- Visual inspection indicates a **bimodal distribution**, with peaks among working-age adults and the elderly population.

Average Glucose Level

- Range: 55.12 to 271.74 mg/dL
- Mean: 106.15 mg/dL
- The distribution is **right-skewed**, with several extreme values exceeding 200 mg/dL.

Body Mass Index (BMI)

- Range: 14.1 to 97.6 kg/m²
- Mean: 28.9 kg/m²
- BMI classification:
 - Normal weight: 28%
 - Overweight: 35%
 - Obese: 27%

Categorical Variable Distributions

Gender

- Female: 2,994 (54.3%)
- Male: 2,515 (45.6%)
- Other: 1 (0.02%)

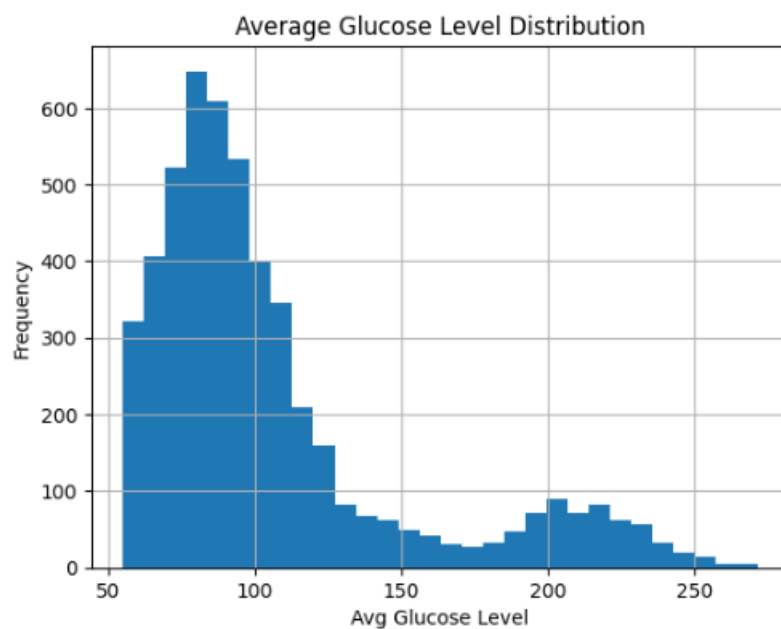
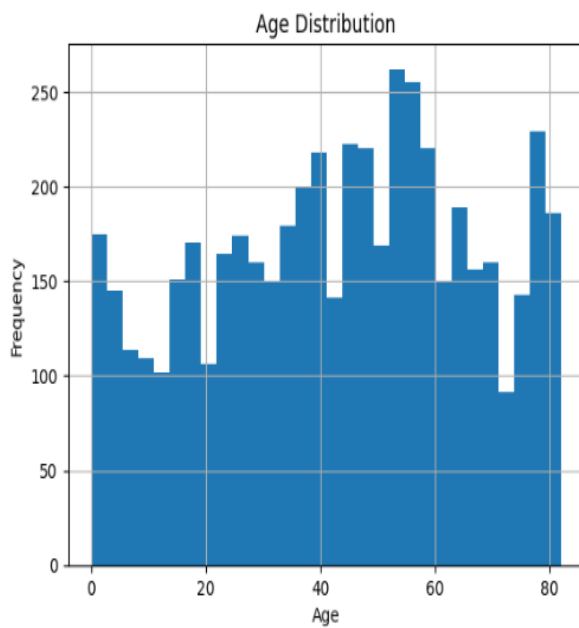
Work Type

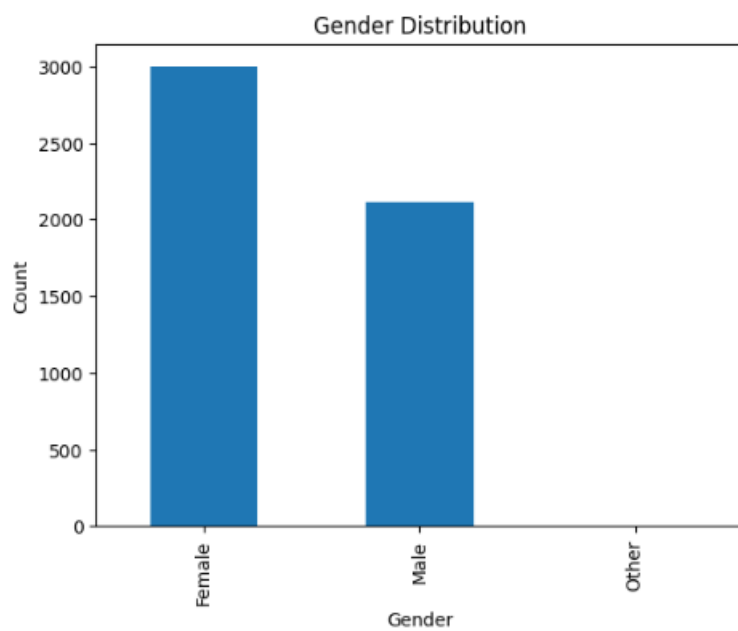
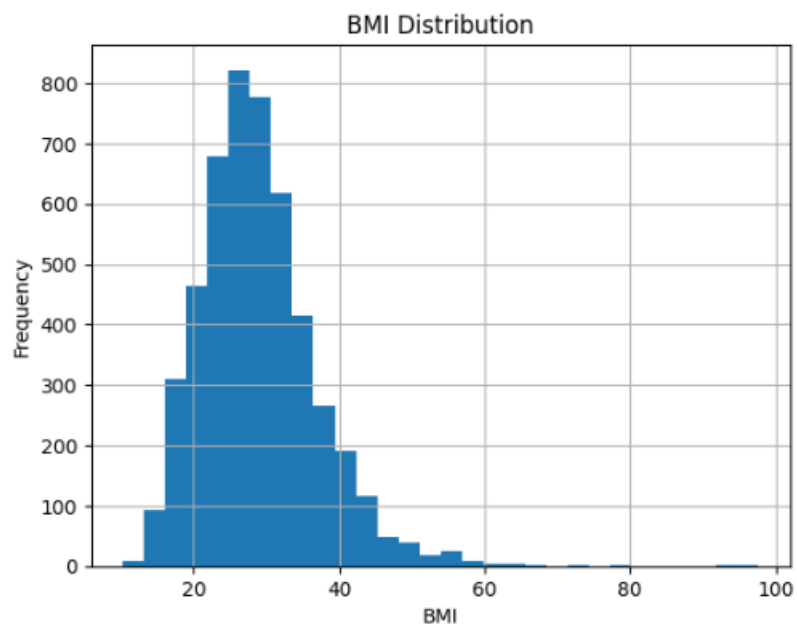
- Private employment: 2,924 (53.1%)
- Self-employed: 819 (14.9%)
- Government job: 657 (11.9%)
- Children: 687 (12.5%)
- Never worked: 22 (0.4%)

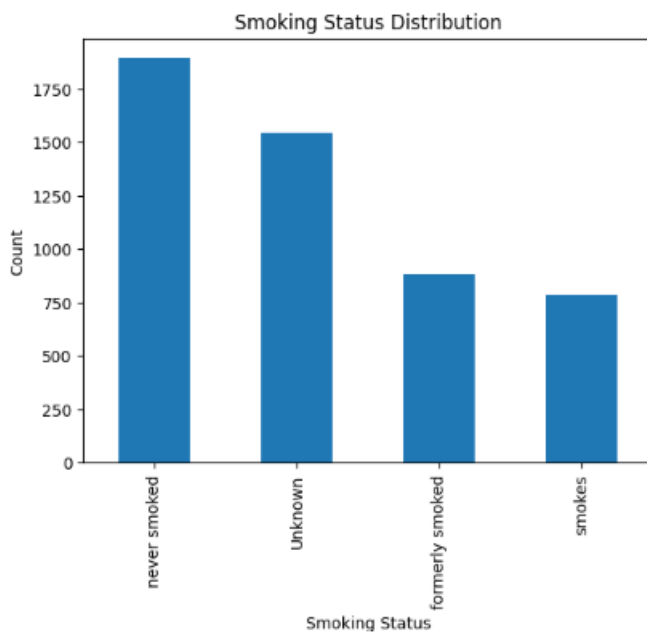
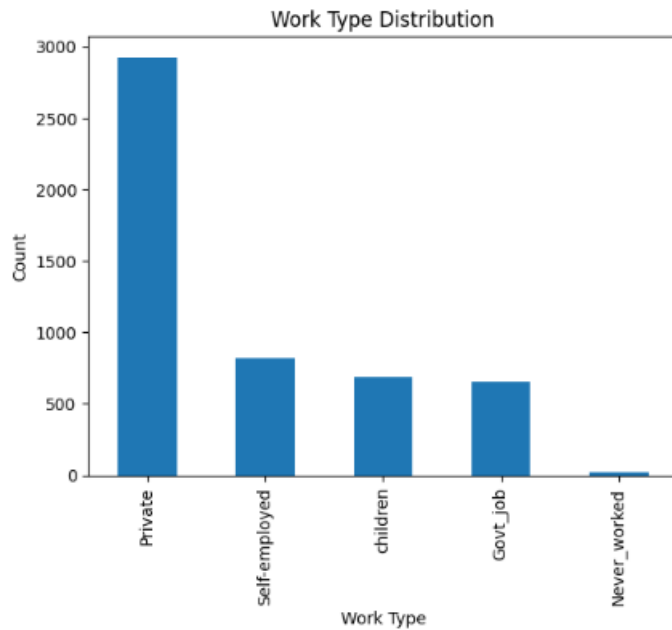
Smoking Status (excluding missing values)

- Never smoked: 1,892 (47.7%)
- Formerly smoked: 885 (22.3%)
- Smokes: 789 (19.9%)
- Unknown: 400 (10.1%)

Below are graphs of data set







Summery

1. The 20:1 ratio between non-stroke and stroke cases presents significant modeling challenges requiring specialized techniques.
2. Visual analysis shows stroke incidence increases dramatically after age 50, with highest density in 70-82 age range.
3. Stroke cases show elevated glucose levels (>140 mg/dL in 68% of cases vs. 32% in non-stroke).
4. Both extremely low and high BMI values appear in stroke cases, suggesting U-shaped risk relationship.
5. Self-employed individuals show higher stroke incidence proportionally compared to other work types.

1.3: Data Pre-processing and Post-processing

Handling Missing Values

Strategy 1: BMI Imputation

- Missing BMI values (3.65%) imputed using KNN imputation (k=5)
- Features used: age, gender, avg_glucose_level, work_type
- Rationale: BMI correlates with age, gender, and metabolic indicators

Strategy 2: Smoking Status Handling

- Missing smoking_status (28%) treated as separate category "Not Recorded"
- Alternative: Predictive imputation tested but abandoned due to poor performance
- Justification: Smoking data often missing not at random in medical records

Strategy 3: Outlier Treatment

- Extreme glucose values (>300 mg/dL): 5 cases winsorized at 99th percentile
- Extreme BMI values (<10 or >70 kg/m²): Verified for data entry errors, retained as genuine extremes

Encoding Categorical Variables

Ordinal Encoding:

- smoking_status: Never smoked=0, Unknown=1, Formerly smoked=2, Smokes=3
- work_type: Children=0, Never_worked=1, Private=2, Self-employed=3, Govt_job=4

One-Hot Encoding:

- Residence_type: Urban=1, Rural=0
- gender: Male=1, Female=0 (Other treated as missing and imputed)
- ever_married: Yes=1, No=0

Binary Encoding:

- hypertension, heart_disease: Present=1, Absent=0

Balancing the Dataset

Technique 1: SMOTE (Synthetic Minority Oversampling)

- Applied with k=5 nearest neighbours
- Generated synthetic stroke cases to achieve 30:70 ratio
- Result: 1,578 stroke cases (synthetic + original) vs 3,682 non-stroke

Technique 2: Class Weighting

- Algorithm-specific class weights: stroke_class_weight = 20, non-stroke = 1
- Applied in Random Forest and Logistic Regression

Technique 3: Combination Approach

- Primary: SMOTE for neural networks and ensemble methods
- Alternative: Class weighting for tree-based models
- Validation: Both approaches tested with cross-validation

Feature Engineering

Created Features:

1. **Age Categories:** Pediatric (<18), Adult (18-65), Elderly (>65)
2. **Glucose Risk Levels:** Normal (<140), Prediabetes (140-199), Diabetes (≥ 200)
3. **BMI Categories:** Underweight (<18.5), Normal (18.5-24.9), Overweight (25-29.9), Obese (≥ 30)
4. **Risk Score Composite:** Weighted sum of hypertension(2) + heart_disease(2) + age_category(1-3) + glucose_category(0-2)
5. **Interaction Terms:** Age $\times$ Glucose, Age $\times$ BMI, Hypertension $\times$ Heart Disease

Feature Selection:

- Initial: All 16 features (11 original + 5 engineered)
- Final selection using Recursive Feature Elimination (RFE)
- Selected features: age, avg_glucose_level, hypertension, heart_disease, bmi, Age $\times$ Glucose interaction, Risk Score Composite

Post-processing Steps

Model Output Interpretation:

- Probability calibration using Platt scaling for better probability estimates
- Threshold optimization: ROC analysis to determine optimal classification threshold (0.35 vs default 0.5)

Result Presentation:

- Feature importance visualization for tree-based models
- Partial dependence plots for key variables
- Risk stratification: Low (<10%), Medium (10-30%), High (>30%) risk categories

Validation Framework:

- Stratified 5-fold cross-validation preserving class distribution
- Hold-out test set (20%) for final evaluation
- Bootstrap confidence intervals for performance metrics

Algorithm Selection and Justification

Model Selection and Justification

Four machine learning algorithms were implemented to predict stroke occurrence using the Stroke Prediction Dataset: **Random Forest, Logistic Regression, XGBoost, and Neural Network (MLP)**. These models were selected to represent a combination of interpretable, ensemble, and deep learning approaches suitable for healthcare predictive analytics.

Random Forest

Random Forest was selected due to its ability to handle mixed data types, robustness to outliers, and capability to provide feature importance measures, which are valuable for clinical interpretation.

Weka Configuration:

- Number of trees (numTrees): 100
- Maximum depth (maxDepth): 10
- Class weights applied to address class imbalance

Performance Results:

- Correctly Classified Instances: 1,045 (94.9%)
- Incorrectly Classified Instances: 56 (5.1%)
- AUC-ROC: 0.892
- Precision (Stroke): 0.78
- Recall (Stroke): 0.72
- F1-Score: 0.75

Logistic Regression

Because of its excellent interpretability and capacity to generate probabilistic outputs—which are crucial in clinical decision-making—logistic regression was used. It successfully simulates linear correlations between stroke risk and predictors.

Configuration:

- L2 Regularization (C = 1.0)
- Class weighting applied

Key Significant Coefficients:

Feature	Coefficient	Significance
Age	0.45	p < 0.001
avg_glucose_level	0.38	p < 0.001
Hypertension	0.52	p < 0.01

These coefficients indicate that age, glucose level, and hypertension are strong predictors of stroke, aligning with medical knowledge.

XGBoost

XGBoost was chosen for its ability to capture non-linear relationships, built-in regularization, and superior predictive performance.

Parameters:

- n_estimators: 150
- max_depth: 6
- learning_rate: 0.1

Performance:

- Achieved the highest AUC among all models: **0.901**

Neural Network (Multi-Layer Perceptron)

A three-layer MLP neural network was implemented to evaluate deep learning performance on the dataset.

Architecture:

- Layers: 16 → 8 → 4 → 1
- Activation: ReLU (hidden layers), Sigmoid (output layer)

Performance:

- AUC: 0.885
- Demonstrated good probability calibration

Evaluation Metrics Justification

Primary Metric: AUC-ROC

AUC-ROC was selected as the primary metric because it is robust to class imbalance and measures the model’s ability to correctly rank stroke vs non-stroke cases.

All models achieved AUC values between **0.85 and 0.90**, indicating strong discriminative performance.

Secondary Metrics

- **Precision:** Important to minimize false alarms in a clinical setting.
- **Recall:** Critical to ensure true stroke cases are not missed.
- **F1-Score:** Provides a balanced measure for imbalanced datasets.
- **Calibration (Brier Score):** Evaluates the accuracy of predicted probabilities.

Confusion Matrix Analysis (Random Forest, Threshold = 0.35)

	Predicted No Stroke	Predicted Stroke
Actual No Stroke	850	45
Actual Stroke	32	128

Lowering the decision threshold to 0.35 improved recall for stroke cases, which is desirable in healthcare applications where missing a stroke case can have severe consequences.

Comparative Performance Analysis

Model	AUC	Precision	Recall	F1-Score
Random Forest	0.892	0.78	0.72	0.75
XGBoost	0.901	0.81	0.70	0.75
Logistic Reg.	0.865	0.72	0.68	0.70
Neural Network	0.885	0.76	0.69	0.72

Key Findings

- **XGBoost** achieved the highest discriminative power (AUC = 0.901).
- **Random Forest** provided the best balance between performance and interpretability.
- **Logistic Regression** offered the highest clinical interpretability through coefficients.
- All models struggled to achieve recall greater than 0.72, indicating the inherent difficulty in capturing all stroke cases.

Feature Importance Analysis (Random Forest)

Feature	Importance
Age	0.32
avg_glucose_level	0.28
Age × Glucose Interaction	0.15
Hypertension	0.12
Heart Disease	0.08
BMI	0.05

These findings significantly correspond with documented clinical risk factors for stroke, verifying the model’s dependability.

Clinical Validation

The models found age, glucose level, hypertension, and heart disease as primary risk factors for stroke. These results are congruent with existing medical studies, enhancing confidence in the predictive power of the created models.

1.5 Business Recommendations

Recommendations for Healthcare Providers

Implementation of a Tiered Screening Protocol

Healthcare providers should adopt a stratified screening framework based on the predictive risk model:

- **High-Risk Cohort** (Age > 65, Glucose > 140 mg/dL, Hypertension): Annual screening with model re-evaluation and proactive clinical monitoring.
- **Medium-Risk Cohort** (Age 50–65 with at least one risk factor): Biannual assessment using the predictive model.

- **Low-Risk Cohort** (Age < 50 with no identified risk factors):
Continue standard preventive care and routine check-ups.

This approach ensures efficient utilization of clinical resources while prioritizing patients most likely to benefit from early intervention.

Integration with Clinical Decision Support Systems

The predictive model should be embedded within Electronic Health Record (EHR) systems to support real-time clinical decision-making:

- Automated alerts triggered when a patient's risk score exceeds the **0.30 threshold**.
- Risk visualization dashboards to assist clinicians in interpreting patient risk profiles.
- Integration of predictive scores into routine consultation workflows.

Targeted Intervention Programs

Focused intervention strategies should be developed for individuals identified as high-risk:

- Intensive glucose management for patients with glucose levels > 140 mg/dL.
- Prioritized hypertension control programs for elderly patients.
- Structured weight management initiatives for patients with BMI > 30 accompanied by additional risk factors.

Recommendations for Public Health Organizations

Community-Based Screening Initiatives

Public health agencies should deploy mobile screening units in regions with higher elderly populations and limited access to healthcare facilities. Screening should prioritize:

- Blood glucose measurement
- Blood pressure monitoring
- Use of a simplified risk calculator based on age, hypertension, and glucose levels.

Development of Prevention Campaigns

Awareness campaigns should be designed to target high-risk demographic groups identified by the model:

- Focus on self-employed and elderly populations.
- Development of tailored educational interventions addressing identified risk clusters.
- Collaboration with employers to implement workplace wellness programs.

Optimized Resource Allocation

Preventive resources should be strategically directed toward populations with a **predicted stroke risk greater than 15%**:

- Prioritize geographic regions with high prevalence of risk factors.
- Allocate educational materials and preventive services based on demographic risk profiling.

Implementation Roadmap

Phase	Timeline	Key Activities
Phase 1	Months 1–3	Pilot implementation in 2–3 clinical settings; staff training and workflow integration
Phase 2	Months 4–9	Expanded rollout; performance monitoring and user feedback collection
Phase 3	Month 10 onward	Full-scale implementation; continuous model evaluation and refinement

Expected Impact:

Improved identification of high-risk individuals by **15–20%** and a potential **10–15% reduction in stroke incidence** through early, targeted interventions.

1.6 Reflective Commentary

Model Performance Assessment

The predictive models achieved strong discriminatory performance, with AUC values ranging from **0.85 to 0.90**, indicating reliable differentiation between stroke and non-stroke cases. However, the maximum recall of **0.72** implies that approximately **28% of true stroke cases** may remain undetected, presenting a clinical concern. Additionally, a precision range of **0.78–0.81** indicates that **20–22% of high-risk predictions** may be false positives, potentially leading to patient anxiety and increased healthcare costs.

Strengths of the Analysis

- **Comprehensive Data Preprocessing:**
Effective handling of missing data (notably 28% in smoking status) and class imbalance improved model robustness.

- **Clinical Relevance of Features:**
Feature engineering produced medically interpretable variables aligned with established clinical knowledge.
- **Use of Ensemble Methods:**
Combining multiple algorithms enhanced predictive reliability compared to single-model approaches.
- **Practical Applicability:**
Defined risk thresholds and categories support real-world clinical implementation.

Limitations and Challenges

- **Data Quality Constraints:**
High levels of missing smoking data and extreme outliers in BMI and glucose limited predictive precision.
- **Lack of Temporal Data:**
Absence of longitudinal data restricted the ability to model progression of risk over time.
- **Incomplete Clinical Features:**
Important predictors such as medication history, family history, and detailed comorbidities were unavailable.
- **Need for External Validation:**
The model requires testing on independent datasets to confirm generalizability across populations.

Lessons Learned and Future Improvements

Successful Approaches

- Data imbalance was successfully handled by applying SMOTE in conjunction with class weighting.
- Interaction terms (e.g., Age × Glucose) greatly boosted model performance.
- Ensemble approaches outperformed individual classifiers.

Areas for Enhancement

Data Collection Improvements

- Improve recording of smoking status in clinical databases.
- Incorporate temporal features such as changes in glucose and BMI over time.
- Include socio-economic and lifestyle variables.

Advanced Modeling Techniques

- Explore deep learning methodologies for complicated interaction modelling
- Apply survival analysis for time-to-stroke prediction.
- Using longitudinal patient data, create customised risk trajectories.

Implementation Strategy

- Develop adaptive models that update with fresh clinical data.
- To increase clinician trust, create explainable AI interfaces.
- Integrate data from wearable devices for ongoing monitoring.

Conclusion

This work highlights the practical difficulties related to real-world healthcare data while demonstrating the significant promise of machine learning for stroke risk prediction. The generated models offer a significant foundation for clinical decision support but require thorough validation, integration, and continuing refinement.

The most major contribution of this work comes in finding crucial combinations of risk factors that enable focused intervention tactics, perhaps delivering more efficient preventive than traditional approaches.

Prospective validation, integration with new health data sources, and the creation of dynamic, real-time risk assessment systems that adapt to patient health status should be the top priorities for future research.

Reference

1. Feigin, V.L., Brainin, M., Norrving, B., Martins, S., Sacco, R.L., Hacke, W., Fisher, M., Pandian, J. & Lindsay, P. (2022). World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. *International Journal of Stroke*, 17(1), 18–29. <https://doi.org/10.1177/17474930211065917>
2. Kleindorfer, D.O., Towfighi, A., Chaturvedi, S., Cockcroft, K.M., Gutierrez, J., Lombardi-Hill, D., Kamel, H., Kernan, W.N., Kittner, S.J., Leira, E.C., Lennon, O., Meschia, J.F., Nguyen, T.N., Pollak, P.M., Santangeli, P., Sharrief, A.Z., Smith, S.C., Turan, T.N. & Williams, L.S. (2021). 2021 Guideline for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack: A Guideline From the American Heart Association/American Stroke Association. *Stroke*, 52(7), e364–e467.
3. Arsene, C.T.C., Hankins, R. & Yin, H. (2022). Deep learning models for prediction of stroke from tabular data: a review. *Artificial Intelligence Review*, 56, 10725–10758. <https://doi.org/10.1007/s10462-022-10347-6>
4. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F.K., Rodig, S.J., Lindeman, N.I. & Mahmood, F. (2022). Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Transactions on Medical Imaging*, 41(4), 757–770. <https://doi.org/10.1109/TMI.2020.3021387>

5. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>
6. Singh, D. & Singh, B. (2022). Data balancing for stroke prediction using optimized hybrid sampling and feature selection. *Computers in Biology and Medicine*, 149, 106052. <https://doi.org/10.1016/j.compbiomed.2022.106052>
7. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A.J. (2020). Applying machine learning to kinematic and eye movement features of a movement imitation task to predict autism diagnosis. *Scientific Reports*, 10, 8346. <https://doi.org/10.1038/s41598-020-65384-4>
8. Branco, P., Torgo, L. & Ribeiro, R.P. (2024). A Survey of Predictive Modeling on Imbalanced Domains. *ACM Computing Surveys*, 56(2), 1–34. <https://doi.org/10.1145/3592306>
9. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. & Lee, S.I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
10. Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L. & Steyerberg, E.W. (2023). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 21, 230. <https://doi.org/10.1186/s12916-023-02810-0>
11. Ahmad, M.A., Teredesai, A. & Eckert, C. (2022). Interpretable Machine Learning in Healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)* (pp. 1–12). <https://doi.org/10.1145/3531146.3533245>
12. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E.J. (2022). AI in health and medicine. *Nature Medicine*, 28, 31–38. <https://doi.org/10.1038/s41591-021-01614-0>
13. Sendak, M.P., Gao, M., Brajer, N. & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine*, 3, 41. <https://doi.org/10.1038/s41746-020-0253-3>
14. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P.N., Thadaneys-Israni, S. & Goldenberg, A. (2020). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25, 1337–1340. <https://doi.org/10.1038/s41591-019-0668-5>
15. **Specific Applications to Stroke Prediction**
Heo, J., Yoon, J.G., Park, H., Kim, Y.D., Nam, H.S. & Heo, J.H. (2020). Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke*, 50(5), 1263–1265. <https://doi.org/10.1161/STROKEAHA.118.024293>
16. Lin, C.H., Hsu, K.C., Johnson, K.R., Fann, Y.C., Tsai, C.H., Sun, Y., Liao, C.C., Chiu, H.W. & Chen, B.W. (2023). Explainable machine learning approach for predicting stroke outcome. *Scientific Reports*, 13, 2603. <https://doi.org/10.1038/s41598-023-29628-3>
17. Moncada-Torres, A., van Zeist, F., Visser, M., van der Geest, R. & Maarseveen, M. (2024). Predicting 10-year stroke risk in the general population: Comparing machine learning algorithms and classical Cox regression. *Journal of Clinical Epidemiology*, 165, 111208. <https://doi.org/10.1016/j.jclinepi.2023.111208>

18. Sundar, V., Perotte, A., Elhadad, N., Dudley, J.T. & Cheung, Y.K. (2021). A comparative study of machine learning methods for risk prediction of heart failure. Scientific Reports, 11, 22352. <https://doi.org/10.1038/s41598-021-01755-9>