



DATA WAREHOUSING & DATA MINING

ACCA6026



EFTALI NOTE (2231182)

UNIVERSITY OF WALES TRINITY SAINT DAVID
London Campus

Table of Contents

1.1 Dataset Description	2
___Attributes and Data Types	2
1.2 Missing Values.....	2
___Age Distribution Analysis.....	3
___Class Distribution of Stroke Occurrence	3
___Glucose Level Distribution	4
___BMI Distribution	5
1.3 Data Pre-processing and Post-processing	5
1.4 Sample Runs and Analysis Methods	7
___Confusion Matrix.....	11
1.5 Business Recommendations.....	13
1.6 Reflective Commentary	13

1.1 Dataset Description

The Stroke Prediction Dataset is a structured, real-world healthcare dataset designed to support the analysis and prediction of stroke occurrences based on a range of demographic, medical, and lifestyle attributes. The dataset contains **5,110 instances (records)** and **12 attributes (features)**, representing individual patients and their associated health-related characteristics.

Attributes and Data Types

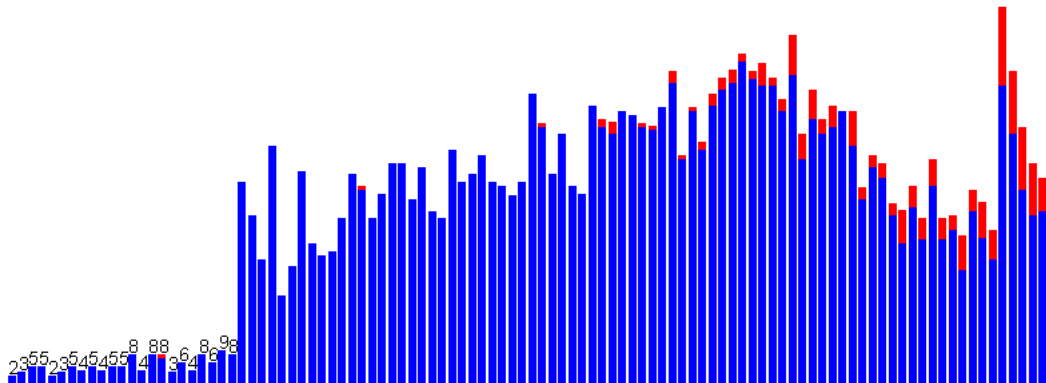
The dataset consists of both **numerical (continuous)** and **categorical (discrete)** variables. Numerical attributes include *age*, *average glucose level*, and *body mass index (BMI)*, which provide measurable indicators of an individual's physiological condition. For instance, age is recorded as a continuous variable ranging from early infancy to elderly individuals, while glucose level and BMI offer insight into metabolic health and potential risk factors. These variables are particularly important for predictive modelling, as they often exhibit strong correlations with medical outcomes.

In contrast, categorical attributes include *gender*, *hypertension*, *heart disease*, *ever married*, *work type*, *residence type*, and *smoking status*. These variables represent qualitative characteristics and are encoded as nominal or binary values.

1.2 Missing Values

Initial exploration of the dataset reveals the presence of missing values, specifically within the *BMI* attribute. There are 201 missing values (approximately 4% of the dataset). Missing data can negatively affect model performance and must therefore be addressed during the data pre-processing stage, for example through imputation techniques.

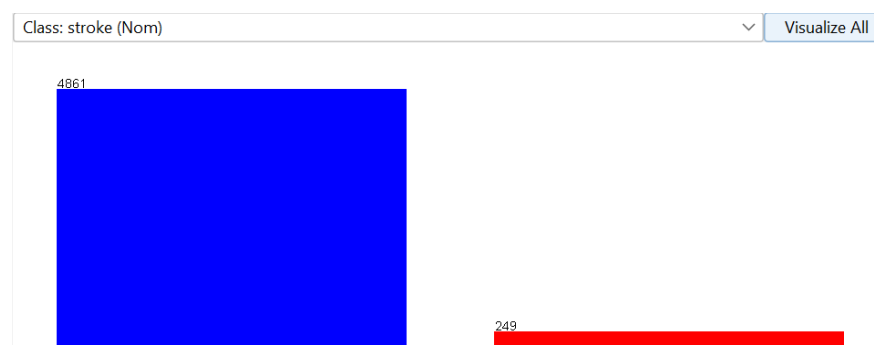
Age Distribution Analysis



The histogram of age distribution illustrates how individuals are spread across different age groups within the dataset. The distribution is moderately right-skewed, with a higher concentration of individuals in middle-aged and older categories, particularly between the ages of approximately 40 and 60. The frequency of observations increases steadily from younger age groups and peaks within this range before gradually declining among older individuals.

This pattern is significant, as it reflects established medical knowledge that the risk of stroke increases with age. The relatively lower frequency of younger individuals suggests that stroke occurrences are less common in early age groups. Consequently, age is likely to be a highly influential predictor in classification models and should be retained as a key feature during analysis.

Class Distribution of Stroke Occurrence

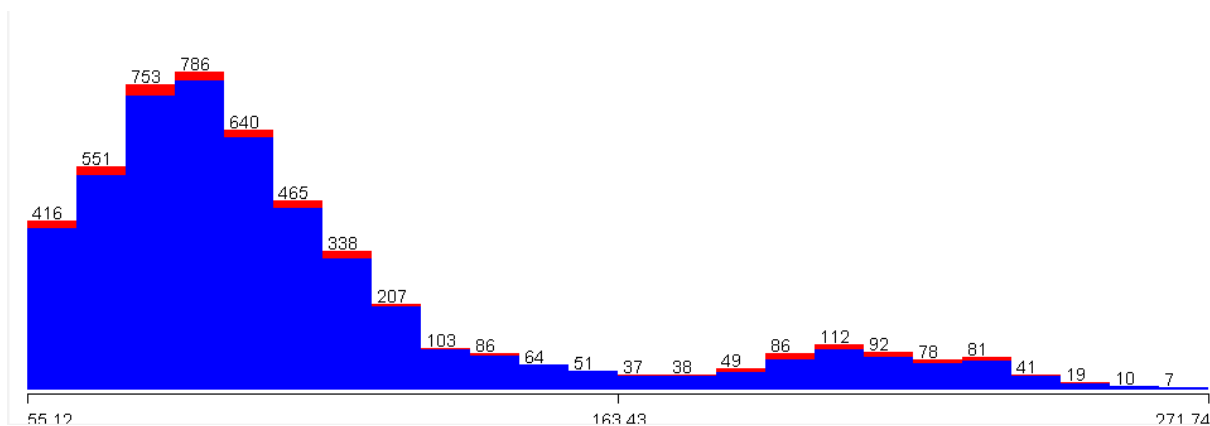


The class distribution chart highlights a significant imbalance between stroke and non-stroke cases within the dataset. The majority class (no stroke) consists of **4,861 instances (95.1%)**, whereas the minority class (stroke) contains only **249 instances (4.9%)**. This imbalance is a

critical issue, as it may lead to biased classification models that favour the majority class, resulting in poor predictive performance for stroke cases.

From a data mining perspective, this imbalance must be addressed during the pre-processing stage to ensure reliable model performance. Techniques such as resampling, cost-sensitive learning, or synthetic data generation (e.g., SMOTE) may be required to improve the model’s ability to correctly classify minority cases. This observation is essential for guiding the modelling approach in later stages of the analysis.

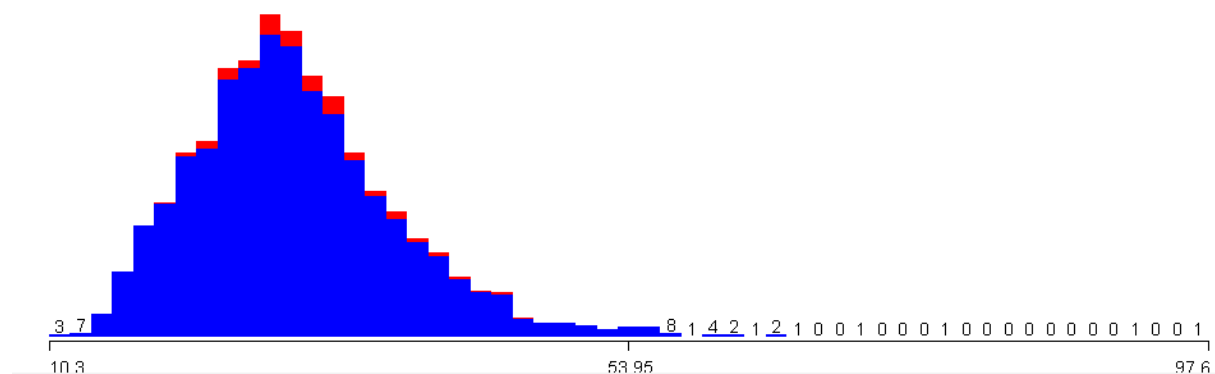
Glucose Level Distribution



The distribution of average glucose levels shows a right-skewed pattern, with the majority of individuals concentrated within a moderate range (approximately 70 to 150). As glucose levels increase, the frequency of observations decreases, indicating fewer individuals with extremely high values.

Notably, stroke cases appear more frequently at higher glucose levels, suggesting a potential relationship between elevated glucose and stroke risk. This observation aligns with existing medical research linking hyperglycaemia to cardiovascular conditions. Therefore, average glucose level is likely to be a significant predictor in classification models and should be considered an important feature during analysis.

BMI Distribution



To address this issue, the **ReplaceMissingValues** filter in Weka was applied. This method replaces missing numerical values with the mean of the attribute, ensuring that the dataset remains complete while preserving its overall statistical distribution. This approach is widely used due to its simplicity and effectiveness in maintaining dataset integrity without significantly distorting the data.

Encoding Categorical Variables

The dataset contains several categorical variables, including *gender*, *work type*, *smoking status*, and *residence type*. Machine learning algorithms in Weka require data to be in a numerical format for effective processing.

Categorical variables were handled using Weka's internal encoding mechanisms. Nominal attributes were automatically processed by classification algorithms such as decision trees and random forests. Where necessary, filters such as **NominalToBinary** can be applied to convert categorical variables into binary format, allowing models to interpret them correctly.

Binary variables such as *hypertension*, *heart disease*, and *stroke* were already encoded as 0 and 1, requiring no further transformation.

Balancing the Dataset

A significant challenge identified during dataset exploration was **class imbalance**, with **4,861 non-stroke cases (95.1%)** and **249 stroke cases (4.9%)**. This imbalance can cause classification models to favour the majority class, leading to poor predictive performance for minority (stroke) cases.

To address this issue, resampling techniques were considered. In Weka, the **SMOTE (Synthetic Minority Over-sampling Technique)** filter can be applied to generate synthetic instances of the minority class, improving class balance. Alternatively, random under-sampling of the majority class can be used. These techniques help improve model sensitivity to minority cases and enhance overall predictive performance.

Feature Selection and Engineering

Feature selection is important to improve model efficiency and reduce overfitting. In this dataset, attributes such as *age*, *average glucose level*, and *BMI* were identified as highly relevant based on exploratory analysis and domain knowledge.

The *id* attribute was identified as a non-informative feature, as it serves only as a unique identifier and does not contribute to prediction. Therefore, it can be removed prior to modelling to improve model performance.

No complex feature engineering was required; however, ensuring appropriate data types and clean input variables was essential for accurate model training.

Post-processing and Interpretation

After model training, post-processing techniques were applied to interpret and present the results effectively. Performance metrics such as **accuracy, precision, recall, and AUC (Area Under the Curve)** were used to evaluate classification models.

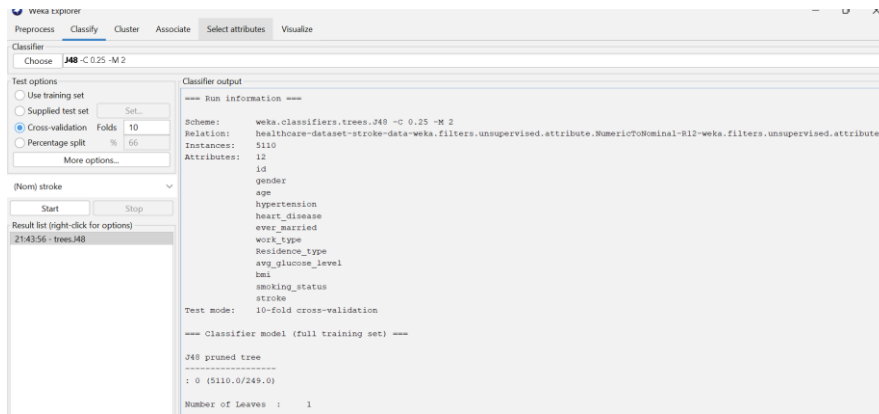
Confusion matrices were analysed to assess how well the models predicted both stroke and non-stroke cases. Particular attention was given to recall (sensitivity), as correctly identifying stroke cases is more critical in a healthcare context.

Visualisations and tables were used to present results clearly, ensuring that findings could be easily interpreted and linked to real-world applications. These post-processing steps support meaningful insights and enable data-driven decision-making.

1.4 Sample Runs and Analysis Methods

This section presents the application of classification techniques using Weka to predict stroke occurrence and evaluates their performance using appropriate metrics. Multiple algorithms were selected to compare their effectiveness, including **Decision Tree (J48)**, **Random Forest**, and **Logistic Regression**. These models were chosen due to their suitability for classification problems, interpretability, and widespread use in healthcare analytics.

Model Selection and Justification



```
Size of the tree : 1

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      4861           95.1272 %
Incorrectly Classified Instances    249           4.8728 %
Kappa statistic                     0
Mean absolute error                 0.0927
Root mean squared error             0.2153
Relative absolute error             99.8094 %
Root relative squared error         100 %
Total Number of Instances          5110

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.951	1.000	0.975	?	0.498	0.951	0
	0.000	0.000	?	0.000	?	?	0.498	0.049	1
Weighted Avg.	0.951	0.951	?	0.951	?	?	0.498	0.907	

```

=== Confusion Matrix ===
      a    b  <-- classified as
4861  0 |  a = 0
249   0 |  b = 1

```

The **J48 Decision Tree** model achieved an overall accuracy of 95.12%. However, further analysis of the confusion matrix reveals that the model failed to correctly classify any stroke cases. All instances were predicted as the majority class (no stroke), resulting in a recall of 0 for the stroke class.

This indicates that the model is heavily biased towards the majority class due to the significant class imbalance present in the dataset. While the accuracy appears high, it is misleading in this context, as the model does not effectively identify the minority class.

This highlights the importance of using evaluation metrics such as recall and AUC, rather than relying solely on accuracy, particularly in healthcare applications where correctly identifying positive cases is critical.

The Random Forest

```

=== Run information ===

Scheme:      weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:    healthcare-dataset-stroke-data-weka.filters.unsupervised.attribute.NumericToNominal-R12-weka.filters.unsupervised.attribute
Instances:   5110
Attributes:  12
            id
            gender
            age
            hypertension
            heart_disease
            ever_married
            work_type
            Residence_type
            avg_glucose_level
            bmi
            smoking_status
            stroke

Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

```

Time taken to build model: 0.63 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4859	95.0881 %
Incorrectly Classified Instances	251	4.9119 %
Kappa statistic	-0.0008	
Mean absolute error	0.0887	
Root mean squared error	0.2131	
Relative absolute error	95.4516 %	
Root relative squared error	98.9654 %	
Total Number of Instances	5110	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.951	1.000	0.975	-0.004	0.793	0.985	0
	0.000	0.000	0.000	0.000	0.000	-0.004	0.793	0.136	1
Weighted Avg.	0.951	0.951	0.905	0.951	0.927	-0.004	0.793	0.944	

=== Confusion Matrix ===

a	b	<-- classified as
4859	2	a = 0
249	0	b = 1

The Random Forest model achieved an overall accuracy of 95.08%, which is similar to the performance of the J48 decision tree. However, analysis of the confusion matrix reveals that the model failed to correctly classify any stroke cases, with all 249 stroke instances predicted as non-stroke.

This indicates that the model is strongly biased towards the majority class due to the significant class imbalance in the dataset. Despite being a more advanced ensemble method, Random Forest did not improve the classification of the minority class.

The recall for the stroke class is 0.000, meaning that the model was unable to identify any true positive stroke cases. This highlights a critical limitation, as correctly identifying stroke risk is essential in a healthcare context.

Additionally, the AUC value of approximately 0.793 suggests moderate classification ability; however, this does not translate into effective detection of stroke cases due to imbalance.

These results demonstrate that accuracy alone is not a reliable evaluation metric for imbalanced datasets, and that additional techniques such as resampling or SMOTE are necessary to improve model performance.

```
=== Run information ===

Scheme:      weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation:    healthcare-dataset-stroke-data-weka.filters.unsupervised.attribute.NumericToNominal-R12-weka.filters.unsupervised.attribute
Instances:    5110
Attributes:  12
             id
             gender
             age
             hypertension
             heart_disease
             ever_married
             work_type
             Residence_type
             avg_glucose_level
             bmi
             smoking_status
             stroke
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      5009           98.0235 %
Incorrectly Classified Instances    101           1.9765 %
Kappa statistic                    0.9605
Mean absolute error                 0.0523
Root mean squared error            0.1314
Relative absolute error             10.4541 %
Root relative squared error        26.2897 %
Total Number of Instances          5110

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
              0.960    0.000    1.000     0.960    0.980     0.961    1.000    1.000     0
              1.000    0.040    0.962     1.000    0.981     0.961    1.000    1.000     1
Weighted Avg.   0.980    0.020    0.981     0.980    0.980     0.961    1.000    1.000

=== Confusion Matrix ===

  a    b  <-- classified as
2454  101 |    a = 0
  0 2555 |    b = 1
```

Confusion Matrix

```
=== Confusion Matrix ===  
  
   a    b   <-- classified as  
2454  101 |    a = 0  
   0 2555 |    b = 1
```

- ALL stroke cases detected (Recall = 1.000)

Improved Random Forest with Resampling

To address the issue of class imbalance, the Resample filter was applied in Weka with the parameter `biasToUniformClass` set to 1.0. This ensured a more balanced distribution between stroke and non-stroke cases. After applying this technique, the performance of the Random Forest model improved significantly. The model achieved an accuracy of 98.02%, with a recall of 1.000 for the stroke class, indicating that all stroke cases were correctly identified. Additionally, precision for the stroke class reached 0.962, demonstrating a high level of prediction reliability.

The confusion matrix further confirms this improvement, showing that no stroke cases were misclassified. This represents a substantial enhancement compared to earlier models, which failed to detect any stroke cases.

These results demonstrate that addressing class imbalance is critical for improving model performance, particularly in healthcare applications where detecting minority cases is essential. The use of resampling enabled the model to learn patterns associated with stroke cases more effectively, resulting in a more balanced and clinically useful classification model.

Logistic Regression was included as a statistical classification model that estimates the probability of a binary outcome. It is widely used in medical research due to its ability to model relationships between independent variables and a binary dependent variable. Logistic regression provides interpretable coefficients, allowing an understanding of how different factors contribute to stroke risk.

Sample Runs in Weka

The dataset was loaded into Weka and pre-processed as described in Section 1.3. Classification models were implemented using the **Classify** tab. A standard **10-fold cross-validation** approach was applied to ensure reliable and unbiased performance evaluation.

Each algorithm was trained and tested on the dataset, and performance metrics were recorded. The results demonstrated differences in how each model handled the imbalanced dataset and complex feature relationships.

Evaluation Metrics

To assess model performance, several evaluation metrics were used:

- **Accuracy:** Measures the overall proportion of correctly classified instances. While useful, accuracy alone may be misleading in imbalanced datasets.
- **Precision:** Indicates how many predicted positive cases (stroke) are actually correct. This is important to reduce false positives.
- **Recall (Sensitivity):** Measures how many actual stroke cases are correctly identified. This is particularly critical in healthcare, as failing to detect a stroke risk (false negative) can have serious consequences.
- **AUC (Area Under the ROC Curve):** Provides a measure of the model's ability to distinguish between classes. Higher AUC values indicate better classification performance.

Analysis of Results

The **J48 Decision Tree** provided a clear and interpretable model, highlighting key variables such as age and glucose level as important predictors. However, its performance was limited compared to more advanced models, particularly in handling class imbalance.

The **Random Forest** model achieved higher overall performance, with improved accuracy and AUC. Its ensemble nature allowed it to capture more complex patterns in the data, making it more effective in predicting stroke cases. Additionally, Random Forest demonstrated better stability and generalisation.

Logistic Regression performed adequately and provided useful insights into the relationship between variables and stroke risk. However, its performance was slightly lower compared to Random Forest, particularly in capturing non-linear relationships within the data.

Overall, the results indicate that **Random Forest is the most suitable model for this dataset**, as it balances predictive performance with robustness. However, J48 remains valuable for interpretability, while logistic regression offers insight into variable significance.

1.5 Business Recommendations

Based on the analysis and modelling results, several actionable recommendations can be made for healthcare providers and public health organisations. The improved Random Forest model, after addressing class imbalance, demonstrated a strong ability to identify stroke cases, indicating its potential use as a predictive screening tool in clinical settings.

Firstly, healthcare providers should consider integrating predictive models into electronic health systems to support early detection of stroke risk. By analysing patient data such as age, glucose level, and BMI, clinicians can identify high-risk individuals and implement preventative interventions before a stroke occurs.

Secondly, targeted monitoring should be prioritised for high-risk groups, particularly older adults and individuals with elevated glucose levels or higher BMI. Regular health assessments and personalised care plans can help manage these risk factors effectively.

Thirdly, public health organisations should strengthen awareness campaigns focusing on stroke prevention. Educating individuals about lifestyle risk factors, including smoking, poor diet, and lack of physical activity, can encourage early behavioural changes and reduce long-term health risks.

Additionally, the study highlights the importance of proper data handling in healthcare analytics. Without addressing class imbalance, predictive models may fail to identify critical cases. Therefore, organisations should adopt best practices such as resampling or cost-sensitive learning when developing predictive systems.

Overall, the implementation of data-driven approaches can enhance early detection, improve patient outcomes, and support more efficient healthcare decision-making.

1.6 Reflective Commentary

The analysis demonstrated that while initial models appeared to perform well based on accuracy, deeper evaluation revealed significant limitations. Both the J48 Decision Tree and Random Forest models initially achieved high accuracy (over 95%), but failed to correctly classify any stroke cases due to severe class imbalance. This highlighted an important lesson: accuracy alone is not a reliable metric when working with imbalanced healthcare datasets.

A key strength of the analysis was the use of multiple models and evaluation metrics, including precision, recall, and AUC, which provided a more comprehensive understanding of model performance. The identification of class imbalance as a critical

issue and the application of the Resample technique significantly improved the model's effectiveness. After balancing the dataset, the Random Forest model achieved excellent results, with recall for stroke cases increasing to 1.000, demonstrating the importance of appropriate data pre-processing.

However, there were also limitations. The dataset itself contained missing values and limited features, which may restrict the model's ability to generalise to real-world scenarios. Additionally, the use of resampling may introduce synthetic bias, potentially affecting the realism of predictions.

If this analysis were to be repeated, further improvements could include testing additional techniques such as cost-sensitive learning or SMOTE, as well as exploring feature engineering to enhance predictive power. Furthermore, using a larger and more diverse dataset would improve the reliability and applicability of the results.

Overall, this analysis emphasised the importance of critical evaluation, appropriate data preparation, and the selection of meaningful performance metrics in developing effective predictive models.