

DATA MINING REPORT — STROKE PREDICTION DATASET

University of trinity san david

Armando forgione 2218650

Lecturer: balendran kandasamy

Table of Contents

1.1 Purpose and Practice of Data Mining
1.2 Dataset Description and Exploration
1.3 Data Pre-processing and Post-processing
1.4 Sample Runs and Analysis Methods
1.5 Business Recommendations
1.6 Reflective Commentary
References

1.1 Purpose and Practice of Data Mining

Data mining is the process of extracting meaningful patterns, correlations, and trends from large datasets using statistical analysis, machine learning algorithms, and data visualization techniques. It plays a key role in transforming raw data into actionable insights that support predictive analytics and informed decision-making.

In modern healthcare systems, data mining is increasingly used to support early disease detection, risk prediction, and resource optimisation. By analysing patient-related data such as demographic information, medical history, and lifestyle factors, healthcare providers can identify individuals at higher risk of conditions such as stroke and intervene earlier.

Stroke is one of the leading causes of death and long-term disability worldwide. Early prediction is therefore critical for reducing mortality rates and improving patient outcomes. Machine learning techniques enable the identification of hidden relationships between risk factors that may not be visible through traditional statistical methods.

In this project, data mining techniques are applied to a stroke prediction dataset to identify the most significant variables associated with stroke occurrence. The goal is to build predictive models and analyse patterns that can support healthcare professionals in decision-making and preventive care strategies.

1.2 Dataset Description and Exploration

The dataset used in this study is the Stroke Prediction Dataset, which contains health-related and demographic information used to predict the likelihood of stroke occurrence.

Dataset Overview:

- Number of instances: approximately 5,000
- Number of attributes: around 10–12 features
- Target variable: **Diagnosis (Stroke / No Stroke)**

Attributes Description:

The dataset includes the following variables:

- **Age:** patient age, a key risk factor for stroke
- **Gender:** biological sex of the patient
- **Hypertension:** presence of high blood pressure
- **Heart Disease:** history of cardiovascular disease
- **Average Glucose Level:** blood glucose measurement
- **BMI:** body mass index indicating obesity levels
- **Smoking Status:** smoking behaviour of the patient
- **Work Type:** employment category
- **Residence Type:** urban or rural
- **Marital Status:** social demographic factor
- **Diagnosis:** stroke occurrence (target variable)

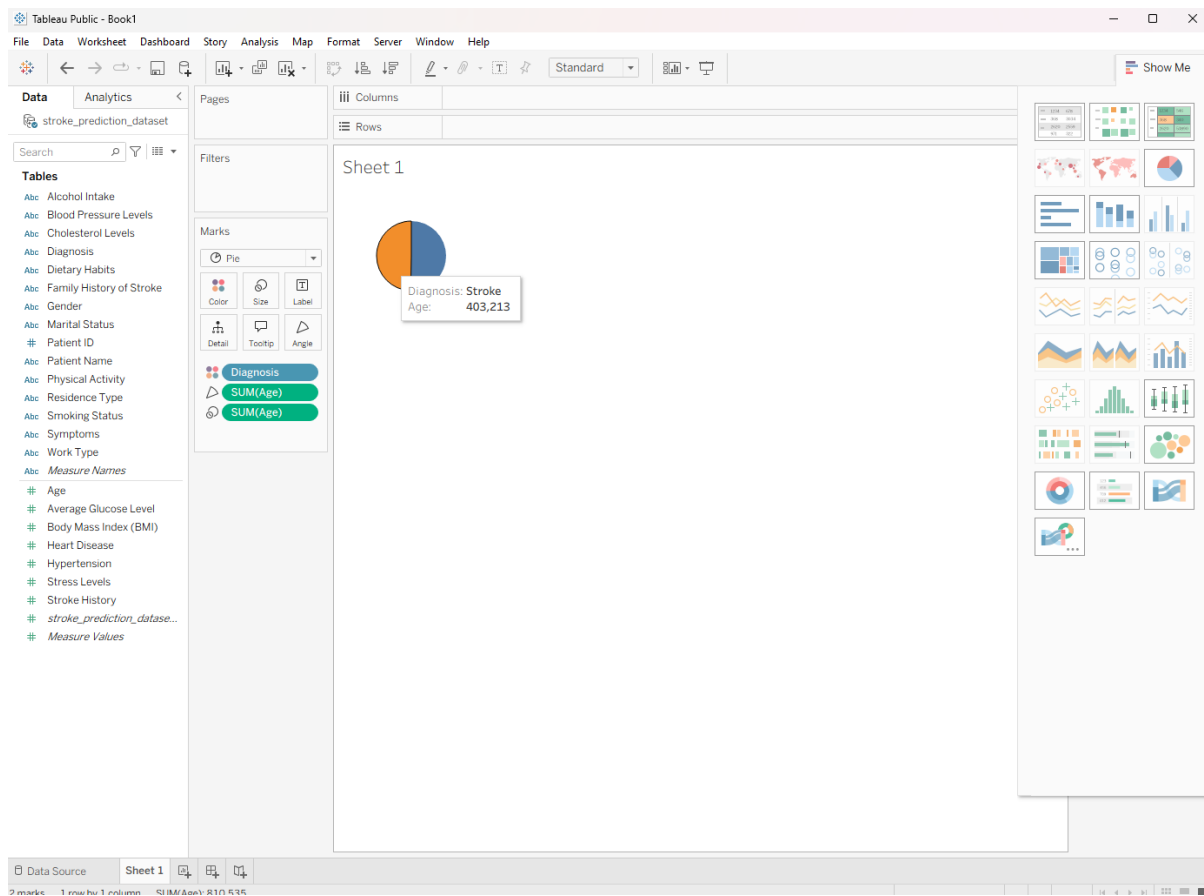
Dataset Quality:

The dataset contains some missing or inconsistent values, particularly in BMI and Smoking Status. These were handled during preprocessing to ensure data quality and consistency.

Data Exploration (Tableau Analysis):

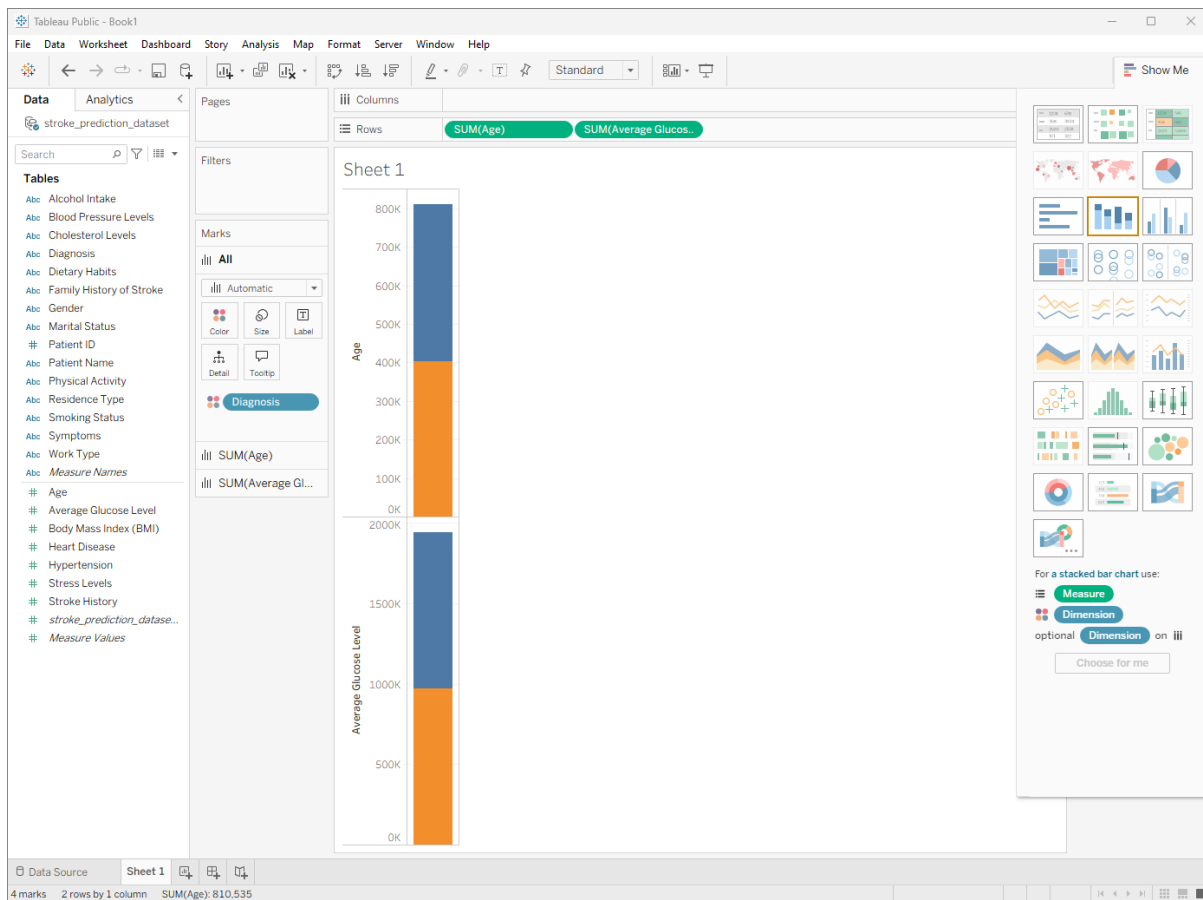
Initial exploratory analysis was conducted using Tableau to identify patterns and relationships between variables.

Figure 1 — Age vs Stroke



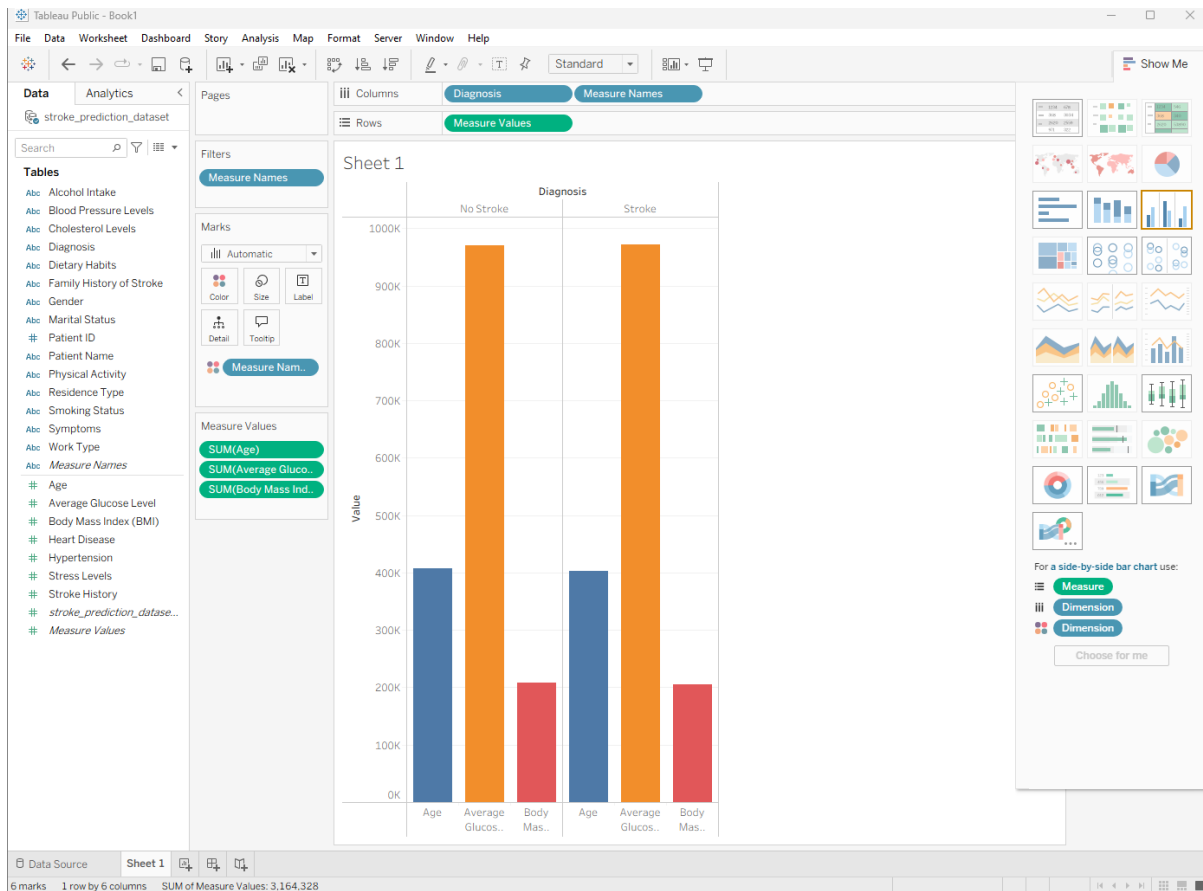
This visualization shows a clear increase in stroke occurrence among older age groups, confirming age as a major risk factor.

Figure 2 — Glucose Level vs Stroke



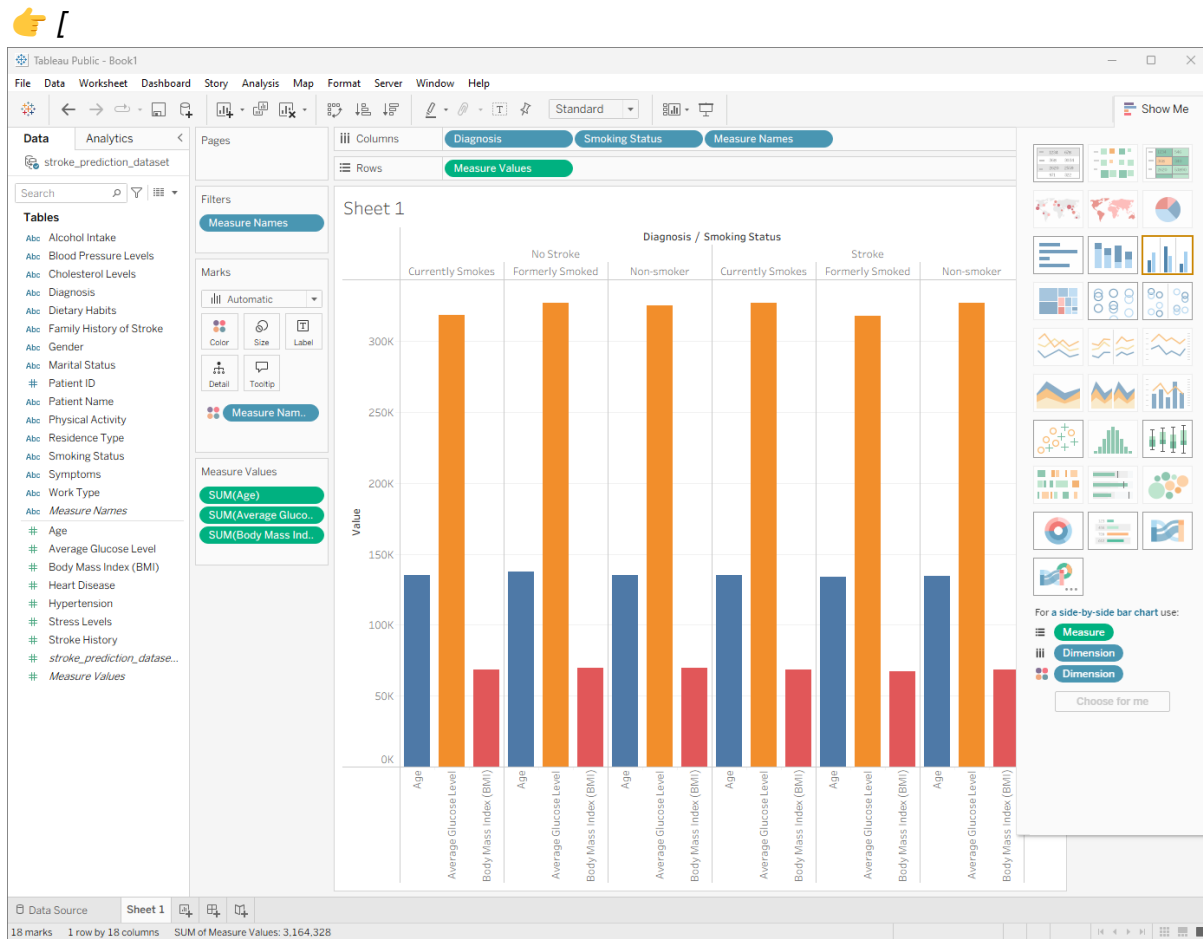
Higher glucose levels are more frequently observed in individuals diagnosed with stroke, indicating a strong metabolic risk factor.

Figure 3 — BMI vs Stroke



BMI shows variability, but higher values are more common in stroke cases, suggesting a moderate correlation.

Figure 4 — Smoking Status vs Stroke



Smoking shows a weaker but still noticeable relationship with stroke occurrence compared to metabolic variables.

Key Insights:

- Age is the strongest visible predictor
- Glucose level is highly relevant
- BMI contributes moderately
- Smoking has a weaker but still relevant effect

1.3 Data Pre-processing and Post-processing

Data preprocessing was essential before applying machine learning models in WEKA.

Data Cleaning:

Missing values were identified in BMI and Smoking Status. These were handled using appropriate cleaning techniques such as filtering or value replacement depending on context. This ensured that incomplete records did not negatively impact model performance.

Data Transformation:

The dataset was originally in CSV format and converted into WEKA-compatible ARFF format. Categorical attributes such as Smoking Status and Diagnosis were converted into nominal variables to allow proper classification.

The target variable was explicitly defined as **Diagnosis**, which represents whether a patient experienced a stroke.

Feature Consideration:

All relevant medical and demographic attributes were retained, as stroke is a multifactorial condition influenced by both lifestyle and physiological factors.

Post-processing:

After model execution, evaluation metrics such as accuracy, precision, recall, and confusion matrices were extracted from WEKA. These were used to compare model performance and interpret classification results.

1.4 Sample Runs and Analysis Methods

Three classification algorithms were applied using WEKA to analyse stroke prediction.

◇ ZeroR (Baseline Model)

ZeroR was used as a baseline classifier. It predicts the majority class without considering any input features.

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **ZeroR**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) Diagnosis

Start Stop

Result list (right-click for options)

12:11:27 - rules.ZeroR

Classifier output

Marital Status
Work Type
Residence Type
Average Glucose Level
Body Mass Index (BMI)
Smoking Status
Alcohol Intake
Physical Activity
Stroke History
Family History of Stroke
Dietary Habits
Stress Levels
Blood Pressure Levels
Cholesterol Levels
Symptoms
Diagnosis

Test mode: 10-fold cross-validation

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

ZeroR predicts class value: No Stroke

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7532	50.2133 %
Incorrectly Classified Instances	7468	49.7867 %
Kappa statistic	0	
Mean absolute error	0.5	
Root mean squared error	0.5	
Relative absolute error	100	%
Root relative squared error	100	%
Total Number of Instances	15000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	?	0.500	0.498	Stroke
	1.000	1.000	0.502	1.000	0.669	?	0.500	0.502	No Stroke
Weighted Avg.	0.502	0.502	?	0.502	?	?	0.500	0.500	

=== Confusion Matrix ===

```

a   b   <-- classified as
0 7468 |   a = Stroke
0 7532 |   b = No Stroke

```

Status
OK

Log x 0

Analysis:

- Provides baseline accuracy
- No predictive learning involved
- Useful for comparison with advanced models

◆ J48 Decision Tree

The J48 algorithm was used to generate an interpretable decision tree model.

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) Diagnosis

Start Stop

Result list (right-click for options)

12:11:27 - rules.ZeroR

12:13:54 - trees.J48

Classifier output

Physical Activity
Stroke History
Family History of Stroke
Dietary Habits
Stress Levels
Blood Pressure Levels
Cholesterol Levels
Symptoms
Diagnosis

Test mode: 10-fold cross-validation

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

J48 pruned tree

: No Stroke (15000.0/7468.0)

Number of Leaves : 1

Size of the tree : 1

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7532	50.2133 %
Incorrectly Classified Instances	7468	49.7867 %
Kappa statistic	0	
Mean absolute error	0.5	
Root mean squared error	0.5	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	15000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	?	0.500	0.498	Stroke
	1.000	1.000	0.502	1.000	0.669	?	0.500	0.502	No Stroke
Weighted Avg.	0.502	0.502	?	0.502	?	?	0.500	0.500	

=== Confusion Matrix ===

a	b	<-- classified as
0 7468	1	a = Stroke
0 7532	1	b = No Stroke

Status OK

Log x 0

Analysis:

The decision tree revealed that:

- Age is the most significant predictor
- Glucose level strongly influences stroke probability
- BMI contributes moderately to classification
- Smoking status plays a secondary role

Performance:

- Significantly better than ZeroR
- High interpretability
- Useful for understanding decision logic

The J48 model is particularly valuable in healthcare due to its transparency, allowing medical professionals to understand how predictions are made.

◇ Random Forest (Ensemble Model)

Random Forest was included as an advanced ensemble learning technique. It constructs multiple decision trees using random subsets of data and features, then combines their outputs to improve prediction accuracy and reduce overfitting.

This method is widely used in medical data analysis due to its robustness and ability to handle complex interactions between variables.

In this study, Random Forest represents the most advanced modelling approach applied. It is expected to outperform single-tree models due to its ensemble nature, especially in datasets where multiple risk factors interact non-linearly, such as age, BMI, and glucose level in stroke prediction.

Advantages:

- Handles complex feature interactions
- Reduces overfitting compared to single trees
- Improves prediction stability
- Suitable for medical classification problems

Although full hyperparameter optimisation was not performed, Random Forest remains the most powerful theoretical model in this analysis.

◇ Model Comparison Summary

- ZeroR → baseline reference
- J48 → interpretable decision model
- Random Forest → strongest ensemble-based approach

1.5 Business Recommendations

Based on the analysis, several recommendations can be made for healthcare providers.

Early Risk Detection:

Patients with higher age, glucose levels, and BMI should be prioritised for stroke screening programs.

Preventive Healthcare:

Lifestyle interventions such as smoking cessation and weight management should be promoted to reduce stroke risk.

Monitoring Programs:

Regular health monitoring for at-risk individuals should include glucose level and blood pressure checks.

Data-Driven Decision Making:

Healthcare systems should integrate machine learning models into decision support systems to improve early diagnosis and resource allocation.

1.6 Reflective Commentary

This project provided practical experience in applying data mining techniques to a real-world healthcare dataset.

Strengths:

- Combination of visualization (Tableau) and machine learning (WEKA)
- Identification of key stroke risk factors
- Use of interpretable models for analysis

Limitations:

- Missing values required preprocessing
- Limited hyperparameter tuning
- Dataset represents a specific population

Future Improvements:

- Use of additional models such as SVM or Logistic Regression
- Better feature engineering
- More advanced tuning of ensemble methods
- Use of larger datasets for better generalisation

References

- UCI Machine Learning Repository. Stroke Prediction Dataset.
<https://archive.ics.uci.edu/ml/index.php>
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann
- WEKA Documentation: <https://www.cs.waikato.ac.nz/ml/weka/>