MSCS-634 – Advanced Big Data and Data Mining

MSCS-631

Nischal Pokharel

Swarna Anjani Devershetty

Vishnu Mallam

Tejeswar Reddy Chemikala

Shyam Nath

Introduction and Dataset Overview

This project explored the application of advanced data mining techniques using a real-world medical dataset focused on stroke prediction. Our main goal was to gain actionable insights and patterns that could help identify patients at high risk of stroke using a combination of predictive modeling, clustering, and pattern mining.

We selected the Stroke Prediction Dataset from Kaggle, which contains 5,110 records with demographic and medical features. This dataset is relevant because strokes are a serious global health issue, and being able to predict or understand risk factors could support early intervention.

The dataset includes features such as age, gender, work type, glucose levels, BMI, heart disease, hypertension, and smoking status. The target variable is a binary label indicating whether a person had a stroke.

Data Preparation and Preprocessing

We performed a range of preprocessing tasks:

* Missing Values: The bmi column had missing values, which we filled using median imputation.
* Duplicates: Duplicate rows were removed to ensure data integrity.
* Inconsistent Data: We filtered out categories like 'Other' in the gender column to avoid ambiguity.
* Categorical Encoding: Variables like work\_type, gender, and smoking\_status were encoded using one-hot encoding.
* Scaling: Numerical features were standardized using StandardScaler for model compatibility.

We also performed binning of continuous features (age, glucose, BMI) for association rule mining.

Modeling and Analysis

Regression Models:  
We used Linear Regression and Ridge Regression to predict stroke probability as a continuous outcome. While both models produced similar scores (R² around 0.09), they offered minimal predictive power. This is likely due to class imbalance and the binary nature of the target variable.

Classification Models:  
We used Decision Trees and k-Nearest Neighbors (k-NN). After hyperparameter tuning:

* Decision Trees showed better interpretability and slightly better F1-scores.
* ROC curves and confusion matrices revealed performance gaps caused by class imbalance.

Clustering:  
We applied K-Means clustering (with 3 clusters) and visualized the results using PCA. This allowed us to identify distinct patient groups based on attributes like age, glucose level, and lifestyle.

Association Rule Mining:  
Using the Apriori algorithm, we found meaningful patterns. For example:

* Patients with high glucose, high BMI, and hypertension often appeared in the 'had\_stroke' group.
* These patterns may help in designing public health strategies.

Insights and Recommendations

* Glucose levels, age, and hypertension were major indicators of stroke risk in our models.
* Work type and marital status also appeared frequently in stroke-related patterns.
* Model performance was limited by class imbalance, which suggests future work should include balancing techniques like SMOTE.
* Clustering helped visualize how at-risk patients differ from others.
* Association rules uncovered useful patterns for risk prediction and patient profiling.

Ethical Considerations

Throughout the project, we were aware of several ethical concerns:

* Bias and Fairness: Some attributes (like gender or work type) might introduce social biases. We ensured we didn't draw conclusions that generalized or discriminated based on these features.
* Data Privacy: The dataset was anonymized and used only for academic purposes. No identifiable information was accessed.
* Medical Responsibility: Our models are not intended for clinical use. They're exploratory and educational. Misuse in real-world settings could be harmful without proper validation.

We recommend any future use of this work be paired with strong ethical review and domain expertise.

References

1. Kaggle. “Stroke Prediction Dataset.” https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
2. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python.
3. Mlxtend Documentation – Association Rule Mining: https://rasbt.github.io/mlxtend/