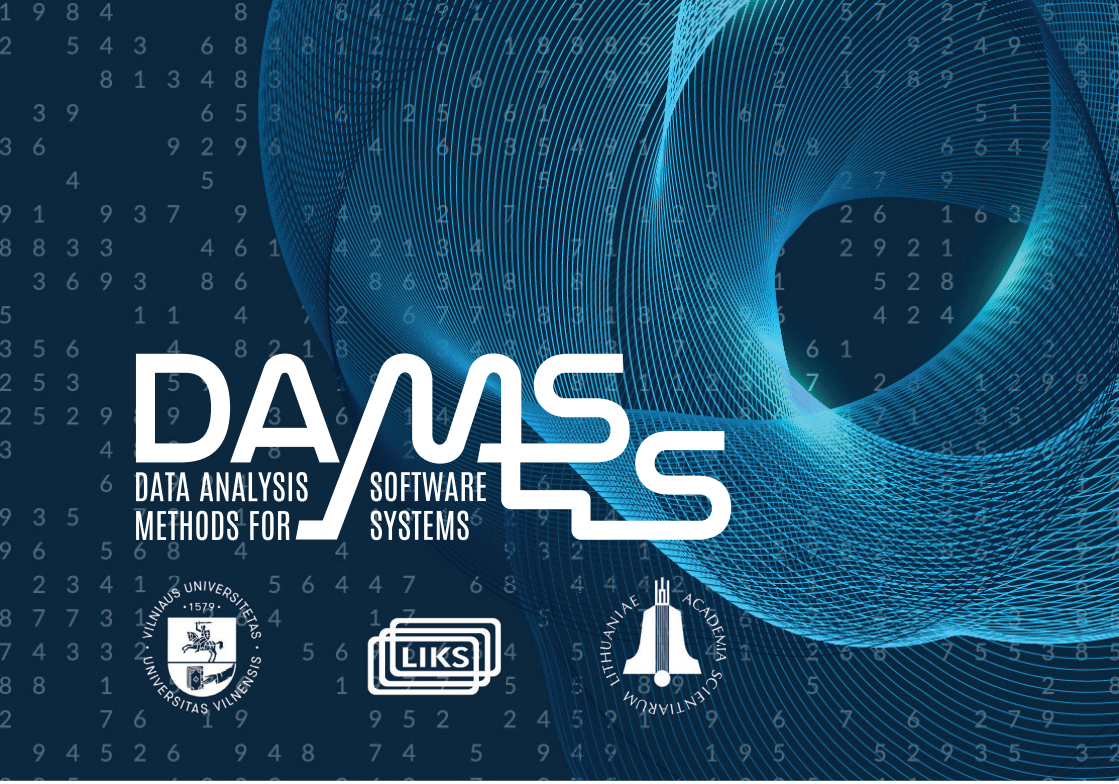

15th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 28–30, 2024

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Tel. (+370 5) 2109 315
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt

https://doi.org/10.15388/DAMSS.15.2024
ISBN 978-609-07-1112-5 (digital PDF)

Evaluating ML Binary Classification for Predicting Stroke-Related Mortality

Dalia Kriksciuniene, Virgilijus Sakalauskas

Vilnius University

dalia.kriksciuniene@knf.vu.lt

In this study, we evaluated the effectiveness of five binary classification machine learning models – Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Networks (MLPClassifier) – for predicting stroke-related mortality using a dataset of clinical and demographic features from the neurology department of Clinical Centre of Montenegro. By comparing these models across key performance metrics such as accuracy, precision, recall, and F1-score, we gained insights into their predictive power and reliability. Additionally, we examined the importance of features to identify the most critical factors driving mortality predictions. Our results indicate that ensemble-based models like Random Forest and XGBoost outperformed other methods, delivering higher accuracy and interpretability. These models consistently identified Health Status, Age and Stroke Symptoms as the most influential predictors, underscoring the importance of these variables in stroke outcome prediction. While Neural Networks showed competitive performance, particularly in terms of precision and recall, the model's lack of interpretability remains a limitation in clinical applications where understanding the driving factors is crucial. Interestingly, simpler models like Logistic Regression, while offering less accuracy, provided clearer insights into feature importance, making them potentially valuable in settings where transparency and ease of interpretation are critical. Conversely, SVM, while delivering good results for certain metrics, struggled with generalizability across different test set sizes. Ultimately, our findings highlight the potential of machine learning models in predicting stroke-related mortality, with Random Forest and XGBoost standing out as the most robust options. These models offer both strong performance and the ability to interpret feature importance, making them suitable for real-world clinical applications. This research underscores the promise of data-driven approaches in improving stroke care by facilitating early identification of high-risk patients and guiding more personalised treatment strategies.