



PROCEEDINGS OF THE

3RD INTERNATIONAL CONFERENCE IN DATA SCIENCE

Autonomous AI Agents: Redefining Decision-Making in a Data-Driven World

25th and 26th of November 2025 at the University of Colombo, Sri Lanka

ICDS 2025



Organised By:



Department of Statistics
University of Colombo

Universiti Malaya
Malaysia

Silver Partner

Knowledge Partner





**Proceedings
of the
International Conference
in Data Science 2025
(ICDS 2025)**

25th - 26th November 2025
at the
**Department of Statistics, University of Colombo,
Colombo, Sri Lanka**

**ICDS 2025 is organised by the Center for Data Science jointly with
the
Department of Statistics, University of Colombo
and
Universiti Malaya, Malaysia**

Publisher: Center for Data Science, University of Colombo
ISSN: 3030-7163

A Comparative Study of Machine Learning Algorithms for Predicting Graduate Unemployment Duration in Sri Lanka

A M K S Aththanayake¹, R R L U I Rajapaksha^{2*}

¹ Faculty of Computing, NSBM Green University, Homagama, Sri Lanka

² Faculty of Computing and Technology, University of Kelaniya, Kelaniya, Sri Lanka

*rasikar@kln.ac.lk

Graduate unemployment continues to pose a significant socio economic challenge in Sri Lanka, with many graduates experiencing delays in securing suitable employment. This study aims to predict and compare the duration of post-graduation unemployment using a comprehensive dataset of approximately 49,000 records obtained from the 2019 Unemployed Graduates Training Program facilitated by the Presidential Secretariat. Unemployment duration was categorised into short, medium, and long term classes, with the dataset exhibiting substantial class imbalance. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was employed. Nine machine learning algorithms, Naive Bayes, Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest, XGBoost, LightGBM, CatBoost, and a Neural Network were evaluated using multiple performance metrics (accuracy, precision, recall, and F1-score). Although KNN yielded the highest initial test accuracy (94.12%), learning curve diagnostics indicated overfitting, leading to its exclusion from the final comparative analysis. Among the remaining models, Random Forest demonstrated the most favourable balance between predictive accuracy and generalisation, achieving a test accuracy of 90.61% and a cross validation accuracy of 92.86%. Class wise evaluation revealed strong performance for the majority class but reduced precision and recall for minority classes, consistent with the underlying imbalance. Macro averaged metrics (precision = 0.50, recall = 0.66, F1 = 0.54) and weighted averages (precision = 0.95, recall = 0.91, F1 = 0.92) provided a more informative representation of model behaviour. Feature importance analysis identified Age and Internal/External degree type as the most associated predictors, followed by District. The findings offer actionable insights to policymakers and higher education stakeholders for designing targeted employability interventions to reduce prolonged unemployment among graduates in Sri Lanka.

Keywords: Predictive modelling, Random Forest, SMOTE, Class imbalance, Machine learning.