Prediction of Depression for Undergraduate Students Based on Imbalanced Data Using Data Mining Techniques

Introduction

This study focuses on the prevalence and prediction of depression among undergraduate students using data mining techniques. Depression is a widespread mental health issue with significant implications for young populations. The research utilized various machine learning models to predict depression risk based on socio-demographic data, internet addiction, alcohol use disorder, and stress levels. The study addresses the challenge of imbalanced data by applying advanced sampling and feature selection techniques to enhance prediction accuracy.

Key Findings

1. Prevalence of Depression:
   - Among the 380 participants surveyed, 77.9% were identified as being at risk for depression.
   - Key contributing factors included internet addiction, alcohol use disorder, and stress.
2. Data Sampling Techniques:
   - Synthetic Minority Over-Sampling Technique (SMOTE) and bootstrapping were applied to balance the dataset.
   - Bootstrapping was found to outperform SMOTE in improving model accuracy for predicting depression.
3. Feature Selection Methods:
   - Three feature selection algorithms (Correlation, Gain Ratio, and ReliefF) were employed to identify the most relevant predictors of depression.
   - ReliefF, when combined with bootstrapping, provided the highest accuracy.
4. Performance of Models:
   - Five classifiers were used: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree.

   - The Random Forest classifier, paired with the ReliefF feature selection method and bootstrapping, achieved the highest accuracy (93.16%).

5. Significant Predictors:
   - Socio-demographic factors such as income adequacy and family status.
   - Internet behaviors, including time spent online and forming online relationships.
   - Alcohol consumption patterns, particularly binge drinking and related harm.
   - Stress levels, as assessed by a validated stress evaluation form.

Implications

The findings demonstrate that machine learning techniques can effectively identify students at risk for depression. By integrating socio-demographic, behavioral, and psychological data, this approach enables early intervention and targeted mental health support. The use of advanced sampling and feature selection techniques is critical for handling imbalanced datasets in mental health research.

Recommendations

- Policy: Implement data-driven screening tools in university mental health programs to identify at-risk students.
- Future Research: Extend the study to assess the severity of depression and incorporate additional predictors such as academic performance and physical health metrics.
- Interventions: Develop targeted mental health initiatives focusing on key risk factors like internet addiction and stress management.

Reference: Narkbunnum, W., & Wisaeng, K. (2022). Prediction of Depression for Undergraduate Students Based on Imbalanced Data by Using Data Mining Techniques. *Applied System Innovation, 5*(120). https://doi.org/10.3390/asi5060120