ABSTRACT

Selecting a major in Informatics is a crucial process that determines a student's academic and career path. However, decision-making is often not based on objective academic data. This study aims to develop a machine learning-based major recommendation system utilizing the Principal Component Analysis (PCA) algorithm and the Decision Tree Classifier. The data used are the academic grades of Informatics students that have been converted from letters to numbers. The initial process includes data cleaning (removing irrelevant columns such as name and semester), imputing empty grades with column averages, and handling data imbalance using undersampling techniques. Dimensionality reduction was performed using PCA, resulting in 9 principal components representing 87% of the data variation. These components were used as input features to the Decision Tree Classifier model. The model was trained with a data split of 80% training and 20% testing, and evaluated using a confusion matrix, accuracy score, and classification report. Initial results showed good classification performance, which was then improved through hyperparameter tuning using GridSearchCV with cross-validation (cv=5). The best parameters were obtained at n_components=9, max_depth=3, min_samples_split=2, and min_samples_leaf=5, resulting in a test accuracy of 83.33%, with an explained variance of 89.50% and an average cross-validation accuracy of 46% ± 26.53%. This best model was then implemented into a GUI-based recommendation system using Python Tkinter, allowing students to receive direct interest recommendations based on their academic grades.

Keywords: accuracy, Decision Tree, academic data, Machine Learning, Principal Component Analysis.

.