Reyta Akdeniz

Advanced Data Analysis

Assignment 6 Report

ASSIGNMENT 6 REPORT

Three machine learning models used within this assignment are; knn, logistic regression and support vector machine models. Target variable that is aimed to be predicted is the income variable that has binary categories of <=50 and >50, and the features are the continuous variables of ones' total years of education attained and the total working hours in a week. Having a categorical target variable, classification based ML models are selected for the task. The data is split into train and test data with 20 percentage of test size and 80 percentage of training, aligning with the common practice.

The knn model predicted the income with an f1-macro score of 0.60. The features in this model seems to be predicting the first category of <=50 better than the second category, interpreting from the precision scores of 0.80 and 0.60, respectively. The cross-validation of the model across five different pieces of test portions shows a balanced ranging of the scores between around 0.73 and 0.74.

The second model, logistic regression, also displays the same f1-macro score with the knn model, 0.60. The precision scores, however, in this model, 0.79 and 0.63, for the <=50 and >50 categories of income. It can be observed, in these metrics that this model has a more balanced predicting performance for the two categories and can be preferred over the knn model in cases where the both categories are equally important for the final objective of the task. Cross validation scores found to be consistent for different testing and training matches in this model as well.

The last tried SVM model displayed the lowest quality of performance among the three with 0.43 f1-macro score at first try. However, the tuning the kernel from 'linear' to 'rbf', it fact, provided slightly higher results than the other two models, with 0.63 f1-macro score and 0.81 and 0.64 precision scores, while again cross validated with around 0.75 accuracy.