Machine Learning 1 - project proposal

Ewelina Osowska, Ewa Sobolewska

Regression task (using small dataset)

Subject:
Comparison of performance of KNN algorithm and linear regression - case of violent crimes committed in different communities within the United States.

Source of dataset:
https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

Description of dataset:
The dataset combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

Description of analysed problem:
Predicting total number of violent crimes (variable ViolentCrimesPerPop) using a set of variables chosen using feature selection methods.

Classification task (using large dataset)

Subject:
Predicting the type of the forest - comparison of performance of different machine learning algorithms

Source of data:
https://archive.ics.uci.edu/ml/datasets/Covertype

Description of dataset:
The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled). Study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. We will limit the target variable to only two levels indicating whether the forest is pine or not.

Description of analysed problem:
Predicting forest cover type using different machine learning algorithms, that is SVM, KNN, discriminant analysis and logistic regression.