Podstawy uczenia maszynowego

07.03.2024

Laboratorium 1

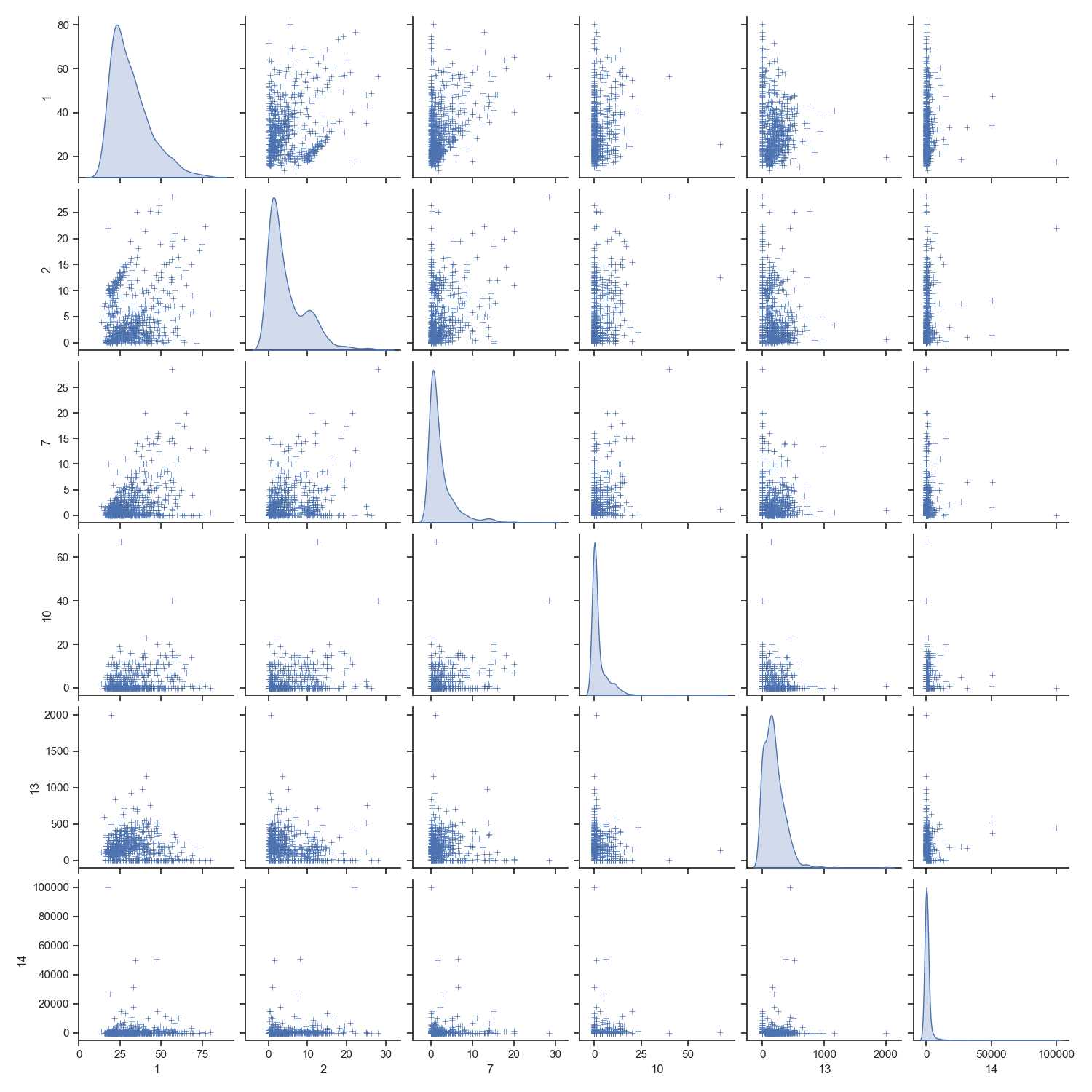
Preprocessing danych

Łukasz Stępień, Kacper Fus

1. Cel zadania.

Celem zadania jest przeanalizowanie technik wstępnego przetwarzania danych i ich wpływu na wyniki klasyfikacji. Pracować będziemy na zbiorze Credit approval dataset1. Jako klasyfikatorów użyj metod: regresji logistycznej, naiwnego Bayesa (NB), najbliższych sąsiadów (klasyfikatora k-NN), metody wektorów nośnych (Support Vector Machines, SVM) oraz lasów losowych (ang. random forests) z ich domyślnymi parametrami.

1. Implementacja.

* Brakujące dane zostały uzupełnione … . (Jak będzie wyglądało uzupełnianie brakujących wartości dla danych numerycznych, a jak dla danych nominalnych?)
* Wygenerowano macierz rozrzutu.
* Zakodowano wartości nominalne za pomocą label encoding oraz one hot encoding. (Które reprezentacje cech są dopuszczalne dla każdego z klasyfikatorów? Jeśli obie reprezentacje są dopuszczalne, która reprezentacja będzie bardziej efektywna?)
* Przeprowadzono skalowanie cech za pomocą normalizacji oraz standaryzacji (jak skalowanie cech wpływa na dokładność klasyfikatorów k-NN oraz lasów losowych?)
* Porównanie klasyfikatorów przy pomocy 5-krotnej walidacji krzyżowej
* Wygenerowano wykres precyzji w funkcji pełności (ang. precision-recall curve) dla klasyfikatora … oraz wykres charakterystyki roboczej odbiornika (ang. receiver operating characteristic, ROC).