

Supervised Outlier Detection via Binary Classification: A Simulation Based Analysis

Arvand Jourabian*

February 2026

Abstract

Detecting outliers is important in many applications and often requires problem specific solutions. A common framing is as binary classification. In this study, we compare Support Vector Machines (SVM), Random Forest (RF) and k-Nearest Neighbors (kNN) across four simulated scenarios, with known class labels. The scenarios are designed to vary class separability through cluster overlap, heavy-tailed outliers, outlier spread and boundary overlap. Models are trained in a supervised setting and evaluated using the Area Under the ROC Curve (AUC), F1-Score (F1) and Balanced Accuracy (BA). Overall, Random Forest is the most robust, kNN performs weakest in the most difficult settings and SVM is often competitive but shows high variability. As overlap increases or outliers become less extreme, the performance declines, suggesting that method choice and tuning are task dependent.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: arvand.j@hotmail.com. Supervisor: Ola Hössjer, Johannes Heiny, Daniel Ahlberg.