Subset Dataset	precision test	precision training	precision validation	recall test	recall training	recall validation	f1-score test	f1-score training	f1-score validation	support test	support training	support validation
BPIC12-100	1.000	0.999	0.999	1.000	0.999	0.999	1.000	0.999	0.999	60.000	1000.000	841.000
BPIC12-25	0.808	0.770	0.765	0.750	0.742	0.733	0.738	0.733	0.723	60.000	1000.000	1000.000
BPIC12-50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	60.000	1000.000	819.000
BPIC12-75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	60.000	1000.000	841.000
DiCE4EL	0.780	0.806	0.821	0.700	0.755	0.749	0.677	0.744	0.739	60.000	1000.000	1000.000
Sepsis-100	0.259	0.246	0.250	0.509	0.496	0.500	0.343	0.329	0.333	55.000	123.000	42.000
Sepsis-25	0.478	0.511	0.528	0.483	0.508	0.519	0.449	0.482	0.495	60.000	1000.000	873.000
Sepsis-50	0.250	0.240	0.261	0.500	0.490	0.511	0.333	0.322	0.346	60.000	1000.000	1000.000
Sepsis-75	0.207	0.254	0.300	0.455	0.504	0.548	0.284	0.338	0.388	55.000	123.000	42.000
TrafficFines	1.000	0.987	0.984	1.000	0.987	0.983	1.000	0.987	0.983	60.000	1000.000	1000.000

Table 1: The evaluation metrics for the prediction component on all datasets. Includes precision, recall and f1 score for test, training and validation data.

We list the predictions of our prediction component in Table 1. The F1-Scores on the test sets are generally higher for the BPIC dataset. Furthermore, in the case of the BPIC datasets, the length of the dataset determines whether the prediction model always predicts correctly or not. It is fair to assume that the length of a loan application process determines the chance of getting rejected or not.