Appendix: Study 1

Figure 1 show the precision and recall performance of the Logistic Regression and Threshold-based classifiers on the datasets considered in the paper. The numbers in the legend correspond to indices of the used metrics and radial lines indicate *F1* measure values. Hexagons represent classifiers using metrics computed per cell, triangles represent formula cell-based metrics, inverted triangles are formula metrics, and bars indicate worksheet-based metrics.

Threshold models, even though optimized in terms of the used threshold, perform poorly when compared with the results that are obtained with decision trees. All threshold classifiers achieved recall scores below 50 %, and only two could reach precision values greater than 10 %. The results indicate that the faults cannot accurately be described by fixed thresholds of individual metrics, but require more flexibility in the used models. Logistic regression models provide such a flexibility, and thus perform noticeably better. Their results show performance levels that are similar to the decision tree models in terms of the recall, but they show worse results in terms of the precision. In general, we conclude that simple, linear models are not adequate to precisely classify faulty cells.

Table 1 provides a summary of the results obtained by the application of all three classifier types on the different datasets. Threshold models perform poorly in direct comparison with the results of decision trees, even though the used thresholds were optimized by the grid search. For all observations of the performance measures threshold models have noticeably lower mean, median and maximum values on all three datasets.

Appendix: Study 2

Figure 2 shows the performance of classifiers trained by Random Forests (RF), Adaptive Boosting (AB), Deep Neural Networks (DNN), and Support Vector Machines with stochastic gradient descent (SVM SDG). The plots indicate that the case of Enron Errors and INFO1 the *F1* values have very low variance. This means that the performance of classifiers does not depend on the cross validation split of the learning dataset into the training and test sets. In case of the EUSES dataset the performance is less stable and clearly depends on the split. As stated in the paper, this behavior can be counteracted by extending our catalog with additional metrics aiming at detection of mutation operations, which is a part of our future work.

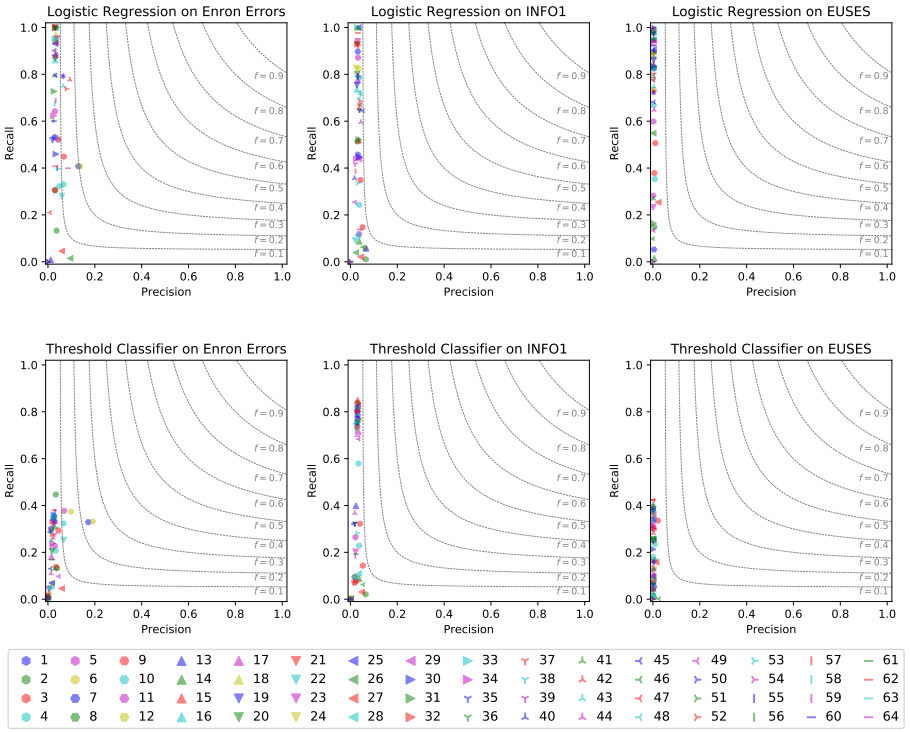
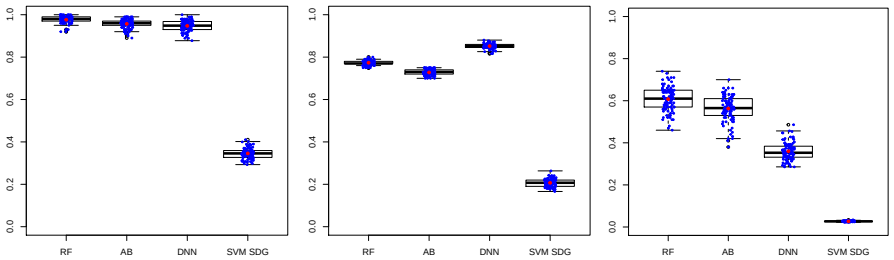


Figure 1: Performance of Logistic Regression and Threshold-based classifiers using individual metrics on the Enron Errors, INFO1 and EUSES corpora.



(a) Enron Errors dataset (b) INFO1 dataset (c) EUSES dataset

Figure 2: *F1* values recorded after running 10-fold cross validation 10 times.

Table 1: Evaluation results of the decision tree (DT), threshold-based (TR) and logistic regression (LR) classifiers.

		Dataset								
		Enron Errors			*INFO1*			*EUSES*		
		DT	TR	LR	DT	TR	LR	DT	TR	LR
Precision	Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1st Quantile	0.03	0.01	0.03	0.03	0.02	0.03	0.00	0.00	0.00
	Mean	0.10	0.03	0.04	0.05	0.02	0.03	0.00	0.00	0.00
	Median	0.08	0.02	0.03	0.05	0.03	0.03	0.00	0.00	0.00
	3rd Quantile	0.14	0.03	0.04	0.07	0.03	0.04	0.01	0.00	0.00
	Max	0.48	0.21	0.14	0.28	0.06	0.07	0.08	0.03	0.01
Recall	Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1st Quantile	0.50	0.06	0.41	0.67	0.11	0.35	0.62	0.08	0.64
	Mean	0.73	0.15	0.65	0.71	0.42	0.57	0.69	0.22	0.73
	Median	0.83	0.14	0.73	0.77	0.56	0.67	0.76	0.24	0.84
	3rd Quantile	0.96	0.22	0.93	0.88	0.66	0.79	0.92	0.35	0.94
	Max	1.00	0.37	1.00	1.00	0.74	1.00	1.00	0.54	1.00
F1-measure	Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1st Quantile	0.06	0.02	0.06	0.06	0.04	0.06	0.00	0.00	0.00
	Mean	0.15	0.04	0.07	0.10	0.04	0.06	0.01	0.00	0.00
	Median	0.13	0.04	0.06	0.08	0.06	0.06	0.00	0.00	0.00
	3rd Quantile	0.23	0.05	0.08	0.13	0.06	0.07	0.02	0.00	0.00
	Max	0.49	0.25	0.21	0.40	0.08	0.09	0.12	0.04	0.02