10-fold CV on 9 Classification Models:

Models were fitted after a Correlation Based Feature Subset Selection (CFS) was done. Note that the logistic regression can perform better with different features (accuracy up to 85%). It is reasonable that parametric approaches (Logistic Regression, LDA, QDA and Naive Bayes) do not work well.

Own identified flaky tests only:

model	acc	prec	reca	f1
ctree	0.912461204220981	0.915487518187015	0.908986311095405	0.912023380134223
randForest	0.963593265052762	0.950659470694769	0.977701739688671	0.963954803696308
gbm	0.802811918063315	0.757502765628565	0.889950734981117	0.818310242149595
adaBoost	0.942744413407821	0.931755669327411	0.955822195316473	0.943422775332107
xgBoost	0.965400372439479	0.951596691940133	0.980591692920368	0.965850563734405
naiveBayes	0.697622594661701	0.840121211688658	0.48802817028396	0.616953243545695
logReg	0.636453289882061	0.733738545892744	0.428613934150542	0.540805910889612
LDA	0.634387026691496	0.758422366586332	0.394196159837502	0.518479952112681
QDA	0.532365766604593	0.643820724372047	0.146137029265856	0.237685330605914

Recorded flaky tests from iDFlakies only:

model	acc	prec	reca	f1
ctree	0.914675667287399	0.925020049605696	0.903202629086355	0.913784214481047
randForest	0.962482929857231	0.948435831962635	0.978027985861057	0.962975657503108
gbm	0.802392147734327	0.7563664246296	0.892523527786812	0.818442510212297
adaBoost	0.943443513345748	0.927808488322791	0.961640436714159	0.944315955298992
xgBoost	0.961232153941651	0.947970395197155	0.975658544534968	0.961558917807216
naiveBayes	0.732907355679702	0.824285532482184	0.591999931124767	0.688690225344676
logReg	0.624943358162632	0.714451912305406	0.416814932897162	0.526011447158615
LDA	0.623828367473619	0.736619393886403	0.386103368092534	0.506154786536639
QDA	0.561409062693979	0.764990429941438	0.177170240932081	0.287319981228549

Own identified flaky tests or recorded by iDFlakies data set (union):

model	acc	prec	reca	f1
ctree	0.906194909993793	0.896287449669467	0.919482247232907	0.907274617433777
randForest	0.959280726256983	0.944092964395372	0.976625425644403	0.96002541756802
gbm	0.805032588454376	0.759278905776473	0.893093162497655	0.8206379113205
adaBoost	0.943163407821229	0.932703739514942	0.955601390764212	0.943879771182444
xgBoost	0.961777622594662	0.947771939112234	0.977565256902487	0.962416231764135
naiveBayes	0.700535381750466	0.834178870772968	0.500926607949033	0.625558631412343
logReg	0.631319832402235	0.727536270031593	0.420417199385706	0.532368457954317
LDA	0.62869025450031	0.744026240323778	0.392539676805183	0.513630319715814
QDA	0.535313469894475	0.650065228248132	0.152214790960374	0.246243523658026

The intersection of the data sets was not done since there are only two test cases. This means that the replication of the recorded flakiness by the iDFlakies data set is not possible (at least in my setting with 10 iterations of the whole test suites and a single core server with 1GB of RAM).