

Statistical analysis

1 Results of the normality test using the KS test

Tables containing the results of normality tests conducted for each training scenario are presented, using the balanced 2022 dataset. The normality test, employing the KS-test method, was carried out to assess the distribution of relevant variables. The presented tables are from the normality test conducted on the literature 16, experts 17, literature by health experts 18, and SFS 19 scenarios.

Table 16 – Results of the KS-test on the models tested in the balanced dataset from the year 2022 in the literature scenario.

EXPERIMENTS	MODELS	accuracy	precision	recall	specificity	f1-score	AUC ROC	MCC
LitDup+	Adaboost	9.17e-17	1.85e-17	1.75e-14	3.62e-19	1.01e-15	9.17e-17	3.04e-07
	Decision Tree	1.30e-16	5.65e-17	3.84e-15	3.92e-18	5.77e-16	1.30e-16	3.23e-07
	Gradient Boosting	8.83e-17	1.58e-17	2.15e-14	2.65e-19	1.10e-15	8.83e-17	3.10e-07
	Randon Forest	1.64e-16	6.95e-17	6.98e-15	3.75e-18	8.47e-16	1.64e-16	2.85e-07
	XGBoost	9.56e-17	2.17e-17	1.44e-14	4.93e-19	9.31e-16	9.56e-17	2.97e-07
LitDup-	Adaboost	8.89e-16	1.53e-15	8.43e-19	5.72e-13	6.52e-17	8.89e-16	3.89e-07
	Decision Tree	1.12e-15	1.84e-15	8.65e-19	8.53e-13	7.45e-17	1.12e-15	2.89e-07
	Gradient Boosting	1.46e-15	2.35e-15	3.52e-19	2.93e-12	6.48e-17	1.46e-15	3.10e-07
	Randon Forest	1.53e-15	2.38e-15	7.52e-19	2.01e-12	8.31e-17	1.53e-15	2.91e-07
	XGBoost	1.28e-15	2.10e-15	5.01e-19	1.74e-12	6.75e-17	1.28e-15	3.14e-07
LitDupMaj-	Adaboost	1.10e-15	1.66e-15	4.27e-18	2.04e-13	1.25e-16	1.10e-15	3.15e-07
	Decision Tree	1.18e-15	1.67e-15	1.11e-17	9.96e-14	1.81e-16	1.18e-15	2.75e-07
	Gradient Boosting	1.24e-15	1.82e-15	5.30e-18	2.12e-13	1.43e-16	1.24e-15	2.88e-07
	Randon Forest	1.30e-15	1.77e-15	1.73e-17	8.01e-14	2.23e-16	1.30e-15	3.21e-07
	XGBoost	1.31e-15	1.85e-15	8.75e-18	1.50e-13	1.76e-16	1.31e-15	2.89e-07

Table 17 – Results of the KS-test on the models tested in the balanced dataset of the year 2022 in the scenario of health experts.

EXPERIMENTS	MODELS	accuracy	precision	recall	specificity	f1-score	AUC ROC	MCC
ExpDup	Adaboost	1.21e-15	2.18e-15	7.11e-20	7.81e-12	3.65e-17	1.21e-15	3.31e-07
	Decision Tree	5.62e-16	1.00e-15	1.64e-18	1.41e-13	6.27e-17	5.62e-16	2.82e-07
	Gradient Boosting	1.35e-15	2.21e-15	4.02e-19	3.21e-12	6.49e-17	1.35e-15	3.10e-07
	Randon Forest	1.12e-15	2.03e-15	1.89e-19	5.08e-12	4.18e-17	1.12e-15	3.45e-07
	XGBoost	1.10e-15	1.59e-15	8.45e-18	1.11e-13	1.58e-16	1.10e-15	3.07e-07
ExpDup-	Adaboost	1.09e-15	1.47e-15	2.74e-17	3.78e-14	2.41e-16	1.09e-15	3.20e-07
	Decision Tree	3.44e-15	1.79e-15	7.51e-11	3.05e-16	1.36e-12	3.44e-15	3.47e-07
	Gradient Boosting	8.26e-16	9.37e-17	9.35e-12	2.98e-20	2.31e-13	8.26e-16	2.92e-07
	Randon Forest	5.72e-16	3.86e-17	2.17e-11	1.70e-20	3.77e-13	5.72e-16	2.94e-07
	XGBoost	5.03e-16	7.57e-17	1.32e-12	9.29e-20	3.95e-14	5.03e-16	2.69e-07
ExpDupMsj-	Adaboost	7.09e-16	5.44e-16	5.18e-15	9.29e-17	1.79e-15	7.09e-16	3.49e-07
	Decision Tree	4.65e-16	2.61e-18	8.16e-11	4.25e-22	1.25e-12	4.65e-16	3.08e-07
	Gradient Boosting	4.44e-16	2.63e-18	8.99e-11	5.71e-22	1.34e-12	4.44e-16	2.93e-07
	Randon Forest	4.57e-16	1.85e-21	1.27e-09	9.17e-24	2.89e-11	4.57e-16	3.01e-07
	XGBoost	5.79e-16	1.76e-19	6.01e-10	4.42e-23	1.25e-11	5.79e-16	3.05e-07

Table 18 – Results of the KS-test on the models tested in the balanced dataset from the year 2022 in the scenario literature by health experts.

EXPERIMENTS	MODELS	accuracy	precision	recall	specificity	f1-score	AUC ROC	MCC
ExpDup	Adaboost	9.36e-17	3.05e-17	5.35e-15	1.39e-18	5.55e-16	9.36e-17	3.39e-07
	Decision Tree	2.37e-16	1.02e-16	1.30e-13	1.37e-17	5.02e-15	2.37e-16	2.83e-07
	Gradient Boosting	9.98e-17	2.32e-17	1.46e-14	5.29e-19	9.64e-16	9.98e-17	3.14e-07
	Randon Forest	1.62e-16	7.05e-17	6.39e-15	5.12e-18	8.35e-16	1.62e-16	2.76e-07
	XGBoost	1.20e-16	4.33e-17	6.04e-15	2.04e-18	6.87e-16	1.20e-16	2.99e-07
ExpDup-	Adaboost	4.29e-15	5.33e-15	2.10e-20	1.79e-10	5.00e-17	4.29e-15	2.62e-07
	Decision Tree	5.08e-15	5.89e-15	6.22e-20	1.05e-10	7.41e-17	5.08e-15	2.87e-07
	Gradient Boosting	5.42e-15	6.27e-15	3.03e-21	1.08e-09	3.37e-17	5.42e-15	2.84e-07
	Randon Forest	7.10e-15	7.36e-15	8.72e-21	8.29e-10	5.06e-17	7.10e-15	3.36e-07
	XGBoost	5.78e-15	6.50e-15	5.29e-21	8.08e-10	4.02e-17	5.78e-15	3.02e-07
ExpDupMsj-	Adaboost	5.55e-15	6.21e-15	1.48e-19	6.23e-11	9.98e-17	5.55e-15	2.99e-07
	Decision Tree	5.97e-15	6.49e-15	3.65e-19	3.46e-11	1.36e-16	5.97e-15	3.46e-07
	Gradient Boosting	5.85e-15	6.44e-15	1.20e-19	7.99e-11	9.64e-17	5.85e-15	2.76e-07
	Randon Forest	5.14e-15	5.87e-15	3.11e-19	3.16e-11	1.20e-16	5.14e-15	3.10e-07
	XGBoost	5.22e-15	5.96e-15	1.76e-19	4.88e-11	1.02e-16	5.22e-15	2.64e-07

Table 19 – Results of the KS-test on the models tested in the balanced dataset of the year 2022 in the SFS scenario.

EXPERIMENTS	MODELS	accuracy	precision	recall	specificity	f1-score	AUC ROC	MCC
SFSDup	Adaboost	7.09e-17	1.46e-17	1.06e-14	3.68e-19	6.60e-16	7.09e-17	3.19e-07
	Decision Tree	1.44e-16	1.18e-17	1.41e-11	6.92e-20	1.29e-13	1.44e-16	3.17e-07
	Gradient Boosting	5.76e-17	1.28e-17	6.29e-15	4.24e-19	4.44e-16	5.76e-17	3.10e-07
	Randon Forest	9.93e-17	2.43e-17	3.20e-14	6.09e-19	1.42e-15	9.93e-17	3.23e-07
	XGBoost	5.76e-17	9.73e-18	1.15e-14	2.18e-19	6.04e-16	5.76e-17	3.22e-07
SFSDup-	Adaboost	2.84e-16	3.91e-16	3.22e-17	2.40e-15	1.20e-16	2.84e-16	3.10e-07
	Decision Tree	3.45e-16	4.90e-16	2.53e-17	4.37e-15	1.23e-16	3.45e-16	3.07e-07
	Gradient Boosting	4.88e-16	7.08e-16	2.03e-17	1.06e-14	1.38e-16	4.88e-16	3.11e-07
	Randon Forest	3.97e-16	4.75e-16	1.09e-16	1.56e-15	2.31e-16	3.97e-16	3.95e-07
	XGBoost	4.51e-16	6.64e-16	1.80e-17	1.02e-14	1.26e-16	4.51e-16	3.30e-07
SFSDupMsj-	Adaboost	3.94e-16	5.06e-16	5.98e-17	2.50e-15	1.83e-16	3.94e-16	3.27e-07
	Decision Tree	3.71e-16	3.55e-16	5.03e-16	2.74e-16	4.23e-16	3.71e-16	3.52e-07
	Gradient Boosting	4.04e-16	4.86e-16	1.05e-16	1.54e-15	2.31e-16	4.04e-16	3.32e-07
	Randon Forest	3.82e-16	3.60e-16	5.69e-16	2.58e-16	4.54e-16	3.82e-16	3.75e-07
	XGBoost	5.05e-16	5.72e-16	1.91e-16	1.32e-15	3.35e-16	5.05e-16	3.48e-07

2 Results of the Wilcoxon statistical test

The tables display the results of applying the Wilcoxon method to compare the F1-scores obtained from different models. We aim to identify whether there are statistically significant differences, thereby contributing to a robust and reliable interpretation of the effectiveness of the evaluated methods. The tables pertain to the results obtained in the scenarios of literature 20, experts 21, literature by health experts 22, and SFS 23.

Table 20 – Wilcoxon test results for the literature scenario in the F1-score metric.

Experiments	Model	Adaboost	Decision Tree	Gradient Boosting	Random Forest	XGBoost
LitDup	Adaboost	——	1.69e-14	1.69e-14	1.21e-12	1.69e-14
	Decision Tree	1.69e-14	——	1.69e-14	1.21e-12	1.69e-14
	Gradient Boosting	1.69e-14	1.69e-14	——	1.21e-12	1.69e-14
	Random Forest	1.21e-12	1.21e-12	1.21e-12	——	1.21e-12
	XGBoost	1.69e-14	1.69e-14	1.69e-14	1.21e-12	——
LitDup-	Adaboost	——	1.69e-14	1.69e-14	1.21e-12	1.69e-14
	Decision Tree	1.69e-14	——	1.69e-14	4.56e-03	1.69e-14
	Gradient Boosting	1.69e-14	1.69e-14	——	1.21e-12	1.69e-14
	Random Forest	1.21e-12	4.56e-03	1.21e-12	——	3.36e-11
	XGBoost	1.69e-14	1.69e-14	1.69e-14	3.36e-11	——
LitDupMaj-	Adaboost	——	4.63e-13	1.69e-14	1.16e-12	1.69e-14
	Decision Tree	4.63e-13	——	4.63e-13	1.40e-11	4.63e-13
	Gradient Boosting	1.69e-14	4.63e-13	——	1.16e-12	1.69e-14
	Random Forest	1.16e-12	1.40e-11	1.16e-12	——	1.16e-12
	XGBoost	1.69e-14	4.63e-13	1.69e-14	1.16e-12	——

Table 21 – Wilcoxon test results for the health experts’ scenario in the f1-score metric.

EXPERIMENTS	Model	Adaboost	Decision Tree	Gradient Boosting	Random Forest	XGBoost
ExpDup	Adaboost	——	4.63e-13	1.04e-12	3.47e-01	1.69e-14
	Decision Tree	4.63e-13	——	4.92e-05	1.45e-11	4.63e-13
	Gradient Boosting	1.04e-12	4.92e-05	——	2.69e-11	1.04e-12
	Random Forest	3.47e-01	1.45e-11	2.69e-11	——	1.21e-12
	XGBoost	1.69e-14	4.63e-13	1.04e-12	1.21e-12	——
ExpDup-	Adaboost	——	1.21e-12	1.15e-12	1.21e-12	1.69e-14
	Decision Tree	1.21e-12	——	1.04e-07	1.47e-07	5.89e-02
	Gradient Boosting	1.15e-12	1.04e-07	——	1.43e-03	1.15e-12
	Random Forest	1.21e-12	1.47e-07	1.43e-03	——	1.21e-12
	XGBoost	1.69e-14	5.89e-02	1.15e-12	1.21e-12	——
ExpDupMsj-	Adaboost	——	1.69e-14	1.21e-12	1.20e-12	1.69e-14
	Decision Tree	1.69e-14	——	1.33e-08	1.20e-12	1.69e-14
	Gradient Boosting	1.21e-12	1.33e-08	——	3.00e-11	1.21e-12
	Random Forest	1.20e-12	1.20e-12	3.00e-11	——	1.20e-12
	XGBoost	1.69e-14	1.69e-14	1.21e-12	1.20e-12	——

Table 22 – Wilcoxon test results for the scenario literature by health experts in the f1-score metric.

EXPERIMENTS	Model	Adaboost	Decision Tree	Gradient Boosting	Random Forest	XGBoost
LitExpDup	Adaboost	——	7.47e-10	1.69e-14	1.21e-12	1.69e-14
	Decision Tree	7.47e-10	——	2.05e-05	6.53e-08	7.47e-10
	Gradient Boosting	1.69e-14	2.05e-05	——	1.21e-12	1.69e-14
	Random Forest	1.21e-12	6.53e-08	1.21e-12	——	2.05e-05
	XGBoost	1.69e-14	7.47e-10	1.69e-14	2.05e-05	——
LitExpDup-	Adaboost	——	1.69e-14	1.69e-14	1.33e-08	1.69e-14
	Decision Tree	1.69e-14	——	1.69e-14	1.21e-12	1.69e-14
	Gradient Boosting	1.69e-14	1.69e-14	——	1.21e-12	1.69e-14
	Random Forest	1.33e-08	1.21e-12	1.21e-12	——	1.21e-12
	XGBoost	1.69e-14	1.69e-14	1.69e-14	1.21e-12	——
LitExpDupMsj-	Adaboost	——	1.69e-14	1.69e-14	1.21e-12	1.69e-14
	Decision Tree	1.69e-14	——	1.69e-14	1.21e-12	1.69e-14
	Gradient Boosting	1.69e-14	1.69e-14	——	1.21e-12	1.69e-14
	Random Forest	1.21e-12	1.21e-12	1.21e-12	——	1.21e-12
	XGBoost	1.69e-14	1.69e-14	1.69e-14	1.21e-12	——

Table 23 – Wilcoxon test results for the SFS scenario in the f1-score metric.

EXPERIMENTS	Model	Adaboost	Decision Tree	Gradient Boosting	Random Forest	XGBoost
SFSDup	Adaboost	——	1.21e-12	1.69e-14	1.21e-12	1.69e-14
	Decision Tree	1.21e-12	——	1.21e-12	3.02e-11	1.21e-12
	Gradient Boosting	1.69e-14	1.21e-12	——	1.21e-12	1.69e-14
	Random Forest	1.21e-12	3.02e-11	1.21e-12	——	1.21e-12
	XGBoost	1.69e-14	1.21e-12	1.69e-14	1.21e-12	——
SFSDup-	Adaboost	——	1.69e-14	4.16e-14	1.21e-12	1.69e-14
	Decision Tree	1.69e-14	——	4.16e-14	1.21e-12	1.69e-14
	Gradient Boosting	4.16e-14	4.16e-14	——	2.35e-12	4.16e-14
	Random Forest	1.21e-12	1.21e-12	2.35e-12	——	1.21e-12
	XGBoost	1.69e-14	1.69e-14	4.16e-14	1.21e-12	——
SFSDupMsj-	Adaboost	——	1.69e-14	4.63e-13	9.58e-13	1.69e-14
	Decision Tree	1.69e-14	——	4.63e-13	3.12e-07	1.69e-14
	Gradient Boosting	4.63e-13	4.63e-13	——	1.20e-11	4.63e-13
	Random Forest	9.58e-13	3.12e-07	1.20e-11	——	9.58e-13
	XGBoost	1.69e-14	1.69e-14	4.63e-13	9.58e-13	——