A single instance of the problem under consideration is a typical classification experiment employing stratified 5-fold cross-validation in the data division, whose aim is to compare the quality of two classification algorithms according to the basic accuracy metric. To assess the statistical significance of differences between the achieved accuracies, the student's T-test is used, taking the α significance threshold of 0.05.

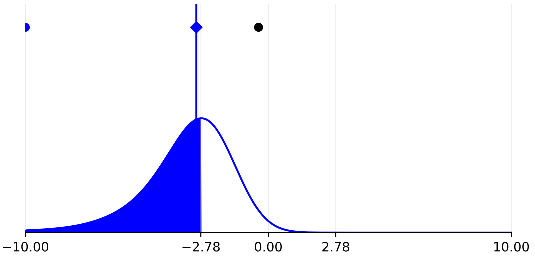
Eighteen popular datasets from the UCI ML repository were adopted for the set of problems being the basis for the experiments, and the research itself was carried out comparing, on a peer-to-peer basis, three simple classification algorithms, according to their implementations included in the scikit-learn library and assuming their default hyperparameters:

- GNB — Gaussian Naive Bayes,
- KNN — k-Nearest Neighbors,
- DTC — CART decision tree.

For each data set, ten thousand repetitions of division into five folds was performed and each one calculated (a) the average accuracy of each classifier, (b) T-statistics for each comparison between classifiers and (c) p-value of each T-statistic. The visualization of an example summary of results for a combination of data together with two compared classifiers is presented in Table 1.

Table 1: Example

SONAR



	GNB	KNN	T	p	I	D
≈	0.683	**0.805**	-2.96	0.04	—	—
=	**0.695**	**0.719**	-0.40	0.71	3354	0.4383
-	0.666	**0.811**	-106.50	0.00	4661	0.5617
+	—	—	—	—	—	—

The table header contains the name of a data set for which the experiment was carried out (in this case SONAR), and directly below it the T-statistics distribution in the population of ten thousand experiments, in the form of gaussian estimation, in the range from -10 to 10. Additional vertical lines represents the point 0 and the significance limits for the adopted α (0.05) and the division method used, with four points of freedom (approximately 2.78).

The chart is supplemented with four additional sets of values in the following rows of the table:

- $\approx$: the average value of the sample from ten thousand courses.
- $=$: the value of the sample with the T-statistic nearest to zero, contained in the statistical non-significant interval, symbolizing the lack of significant differences between the compared algorithms

- − : the value of the observation with the lowest T-statistic outside the statistical non-significant interval, symbolizing a statistically significant advantage of the algorithm from the right column,

- + : the value of the observation with the highest T-statistic outside the statistical non-significant interval, symbolizing a statistically significant advantage of the algorithm from the left column.

The next columns in the table are mean accuracies obtained by compared algorithms, T-statistics (column T), p-value (p), seed value which allowed to obtain a given division-observation (I) and the percentage of observations in which the situation described in the line occurred (D).

Therefore, the interpretation of the example table will be as follows. The average value of the observation indicates the predominance of the KNN algorithm over GNB, for the former the accuracy of approximately 81%, and for the second 68%. The average T-statistics value is -2.96, which gives p-value at 0.04, which makes it a statistically significant difference.

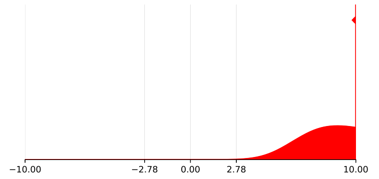
However, we may also observe situations in which a significant difference between algorithms does not occur (line =). An example is the partition instance obtained for random seed 3354, where T-statistic equal -0.4, which gives p-value at 0.71. Similar situations (no significant difference) occurs in 44% of the considered cases. Divisions which show a significant advantage of the KNN algorithm occur in 56% of cases, and the extreme value of T-statistics is -106.5, which is an outlier in the context of the problem under consideration.

In connection with the above observations, we can properly validate two contradictory research hypotheses by pulling the dataset into folds adequately many times. Moreover, with the average value of T-statistics close to the significance threshold, approximately half of the experiments will give us information about the KNN advantage, when the other half of the divisions will deny any statistical difference.

The simplest of situations encountered in performed experiments is that in which each of ten thousand experiments leads to the same conclusion, as is shown in the examples in the Table 2. The distribution of T-statistics in such cases is either narrow enough to fit within the statistical irrelevance interval (SOYBEAN) or far enough from it, so that even outlier observations leading to other conclusions do not happen. It is worth noting, however, that even in such cases, outliers reveal a major deviation from the mean value (for example, an average of 10.56 in the WINE dataset and an outlier observation of 114.79). It is particularly important to note at this point that only six of the fifty-four comparisons are characterized by such unambiguous conclusions.

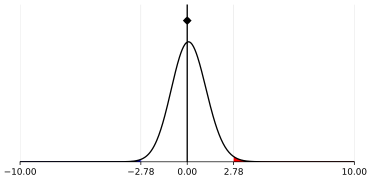
Table 2: Example

WINE



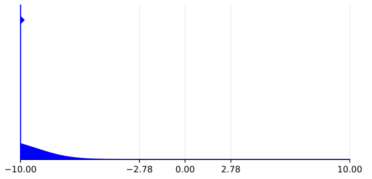
	GNB	KNN	T	p	I	D
≈	**0.972**	0.703	10.56	0.00	—	—
=	—	—	—	—	—	—
-	—	—	—	—	—	—
+	**0.976**	0.677	114.79	0.00	744	1.0000

SOYBEAN



	GNB	DTC	T	p	I	D
≈	**1.000**	**1.000**	0.00	1.00	—	—
=	**1.000**	**1.000**	0.00	1.00	0	1.0000
-	—	—	—	—	—	—
+	—	—	—	—	—	—

MONKONE

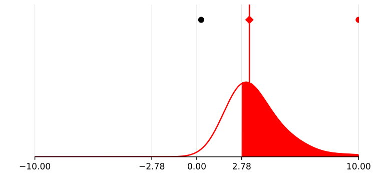


	GNB	KNN	T	p	I	D
≈	0.664	**0.946**	-15.84	0.00	—	—
=	—	—	—	—	—	—
-	0.667	**0.946**	-248.20	0.00	3069	1.0000
+	—	—	—	—	—	—

The dominant majority, as many as 39 out of 54 comparisons correspond to the situation presented in the example in Table 1, where we can draw two contradictory conclusions from the appropriate (let's emphasize - random) combination of patterns division into folds. Examples here are presented in Table 3. In the case of the WINE dataset, we may observe a situation in which the GNB algorithm in averaging achieves an advantage over the DTC, but in as many as three out of ten cases, random division of the data set will lead to the conclusion that there are no significant differences. A much more interesting case is the IRIS dataset, where in 99% of divisions there is no significant difference between the compared classifiers, but 1% of the problem instances leads to the conclusion that the KNN has a significant advantage over GNB. An even stronger example of this type is the AUSTRALIAN dataset, where, despite the predominance of the DTC over the KNN and the average difference in quality at 13%, we can still find two parts-per-thousand of situations in which the difference between them disappears.

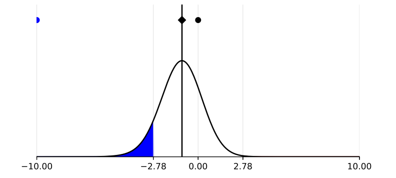
Table 3: Example

WINE



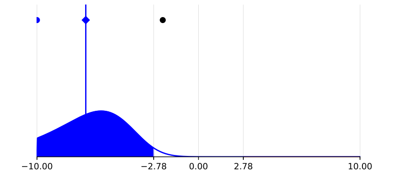
	GNB	DTC	T	p	I	D
≈	**0.972**	0.917	3.25	0.03	—	—
=	**0.972**	**0.966**	0.27	0.80	650	0.3246
-	—	—	—	—	—	—
+	**0.971**	0.886	58.74	0.00	6516	0.6754

IRIS



	GNB	KNN	T	p	I	D
≈	**0.967**	**0.967**	-1.00	0.37	—	—
=	**0.960**	**0.960**	0.00	1.00	0	0.9872
-	0.946	**0.980**	-146.00	0.00	213	0.0128
+	—	—	—	—	—	—

AUSTRALIAN



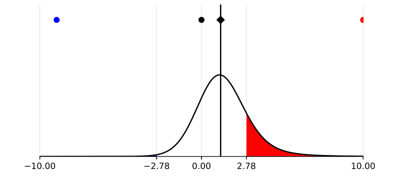
	KNN	DTC	T	p	I	D
≈	0.693	**0.820**	-6.98	0.00	—	—
=	**0.679**	**0.774**	-2.21	0.09	6575	0.0021
-	0.694	**0.839**	-72.59	0.00	3782	0.9979
+	—	—	—	—	—	—

The most interesting, however, is the third, last group of observations, which consists of 9 out of 54 examples (Table 4). There is a set of cases in which, depending on which of the random dataset divisions we select for the experiment, we can get the validation of each possible conclusion. The liver dataset is particularly interesting here, where only eight out of ten thousand cases show the DTC algorithm superiority over KNN, with approximately 90% of their equal quality and 10% of KNN advantage cases.

This may lead to the hypothesis that if we have sufficiently large computational power to repeat random dataset divisions, we will lead to a situation in which, using the standard approach to experiments, we will be able to reasonably substantiate any hypothesis about the statistical relationship between the compared classifiers. Moreover, on the basis of all separate groups of cases, we can certainly conclude that the existence of such combinations of data sets and classification algorithms is common, against which we can substantiate contradictory hypotheses, depending on the applied random division of the data set into folds.

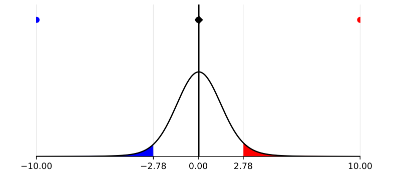
Table 4: Example

LIVER



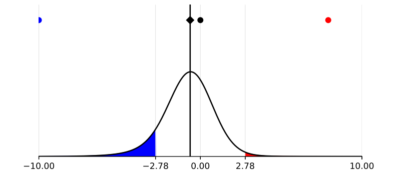
	KNN	DTC	T	p	I	D
≈	**0.667**	**0.638**	1.18	0.29	—	—
=	**0.637**	**0.637**	0.00	1.00	722	0.8828
-	0.663	**0.689**	-8.96	0.00	8873	0.0008
+	**0.669**	0.611	341.00	0.00	4553	0.1164

GERMAN



	KNN	DTC	T	p	I	D
≈	**0.690**	**0.690**	0.04	0.55	—	—
=	**0.678**	**0.678**	0.00	1.00	1082	0.9686
-	0.681	**0.710**	-9.98	0.00	9310	0.0136
+	**0.695**	0.662	12.92	0.00	4886	0.0178

DIABETES



	KNN	DTC	T	p	I	D
≈	**0.688**	**0.699**	-0.63	0.47	—	—
=	**0.687**	**0.687**	0.00	1.00	3549	0.9441
-	0.681	**0.733**	-14.88	0.00	8407	0.0529
+	**0.708**	0.674	7.92	0.00	9658	0.0030