Indukcyjne metody analizy danych
Ćwiczenie 5

Zespoły klasyfikatorów

Prowadzący: dr inż. Paweł Myszkowski
Student: Piotr Bielak, 218137
WT 17:05

Wrocław, 29 maja 2018r.

Spis treści

1 Wprowadzenie

1.1 Cel ćwiczenia

Celem ćwiczenia było poznanie algorytmów zespołów klasyfikatrów (bagging, boosting, random firest) oraz zbadanie i ocena ich działania na 3 określonych zbiorach danych. W trakcie badań należało uwzględnić różne parametry algorytmów. Należało również zaobserwować wpływ tych parametrów na wartości zadanej miary (F1-Score).

1.2 Zespoły klasyfikatorów

Zespół klasyfikatorów składa się, jak sama nazwa wskazuje, z kilku klasyfikatorów, określanych często jako klasyfikatory bazowe. Mogą nimi być różne klasyfikatory proste, typu: naiwny Bayes, drzewo decyzyjne, SVM (Support Vector Machine) itp. W zależności od konkretnego algorytmu, bazowe klasyfikatory są uczone na pełnym lub części zbioru treningowego, dodatkowo uczenie może odbywać się sekwencyjnie (boosting) lub równolegle (bagging). Ostateczna decyzja zespołu jest podejmowana na podstawie decyzji wszystkich klasyfikatorów bazowych. Głosowanie może być równouprawione lub ważone (np. błędem klasyfikacji danego klasyfikatora).

Wśród podstawowych algorytmów zespołów klasyfikatorów można wyróżnić m.in.:

- bagging – metodą boostrap wylosuj zbiór treningowy dla każdego klasyfikatora; równolegle ucz klasyfikatory bazowe; zlicz liczbę głosów w każdej klasie i podejmij ostateczną decyzję,
- boosting – każdej instancji zbioru uczącego przypisz równą co do wartości wagę, wylosuj zbiór uczący dla pierwszego klasyfikatora bazowego; na podstawie błędnie zklasyfikowanych instancji wyznacz nowe wagi instancji (błędnie zklasyfikowane dostaną wyższe wagi i przez to mają większą szansę na pojawienie się w nowym zbiorze uczącym kolejnego klasyfikatora bazowego); podejmij decyzję na podstawie decyzji wszystkich kasyfikatorów bazowych,
- random forest – idea jest podobna jak w przypadku baggingu, jednak tutaj istnieje również możliwość wyboru / losowania podzbioru atrybutów; dodatkowym ustalony jest również klasyfikator bazowy – jest nim drzewo decyzyjne (stąd m.in. nazwa algorytmu).

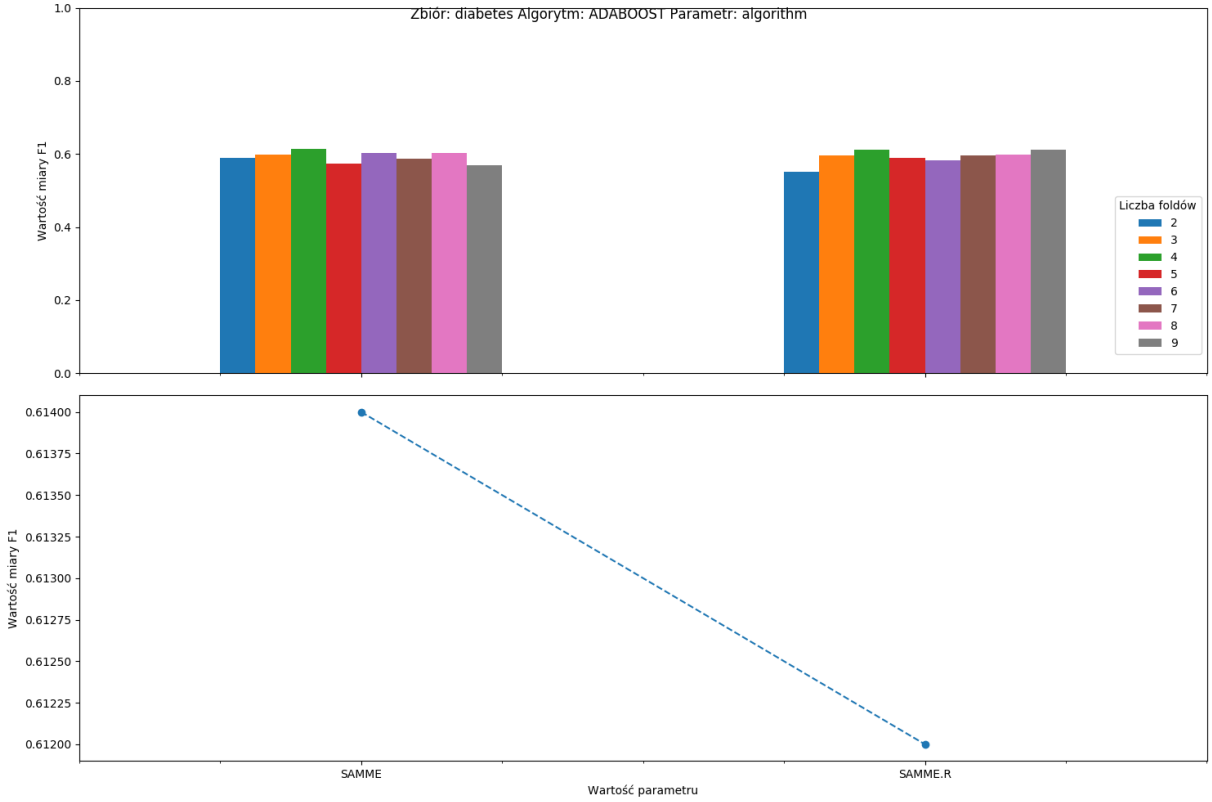
1.3 Badane parametry

- bagging:
  - bootstrap: [True, False],
  - max_samples: [0.25, 0.50, 0.75, 1.0],
  - n_estimators: [10, 25, 50, 75, 99],
- boosting:
  - algorithm: [SAMME, SAMME.R],
  - learning_rate: [0.1, 0.01, 0.001, 0.0001],
  - n_estimators: [10, 25, 50, 75, 99],
- random forest:
  - bootstrap: [True, False],
  - criterion: [gini, entropy],
  - n_estimators: [10, 25, 50, 75, 99],

2 Zbiór Diabetes

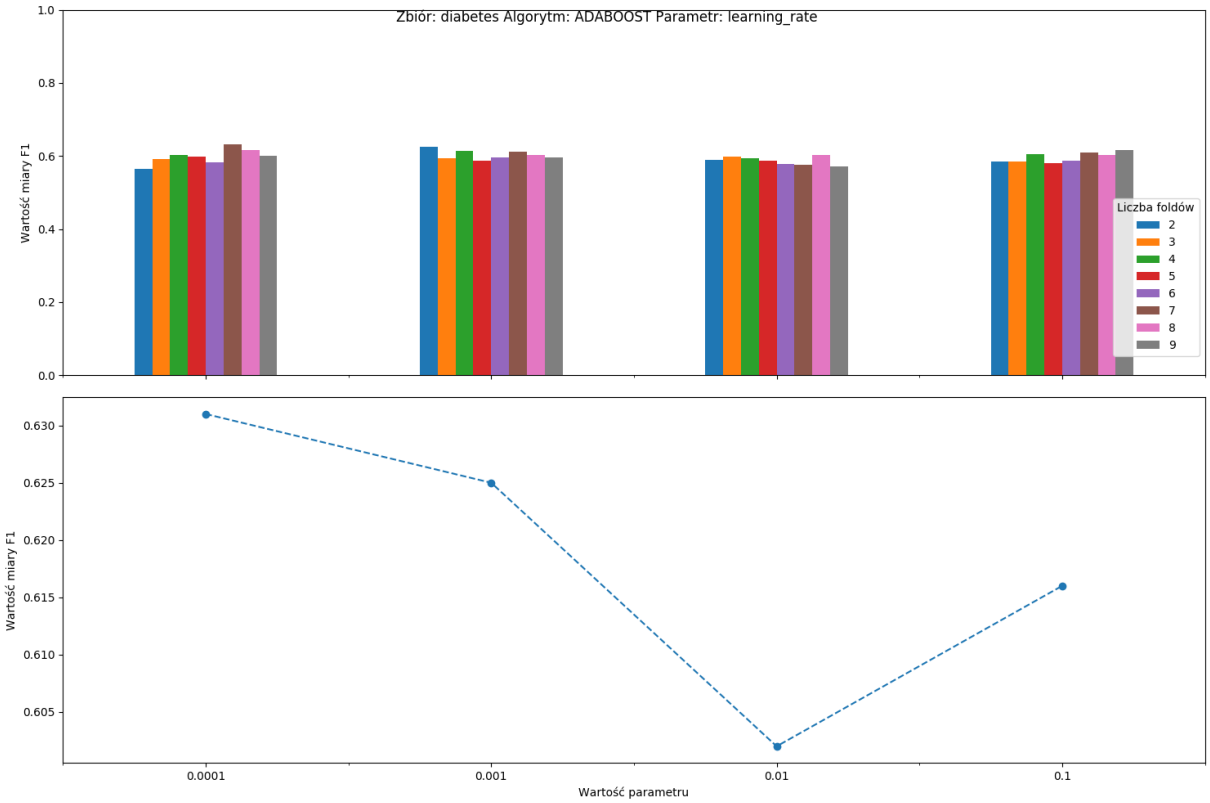
2.1 Algorytm Adaboost

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
algorithm	SAMME	0.589	0.598	0.614	0.574	0.603	0.588	0.602	0.570
	SAMME.R	0.551	0.597	0.612	0.589	0.583	0.596	0.599	0.611



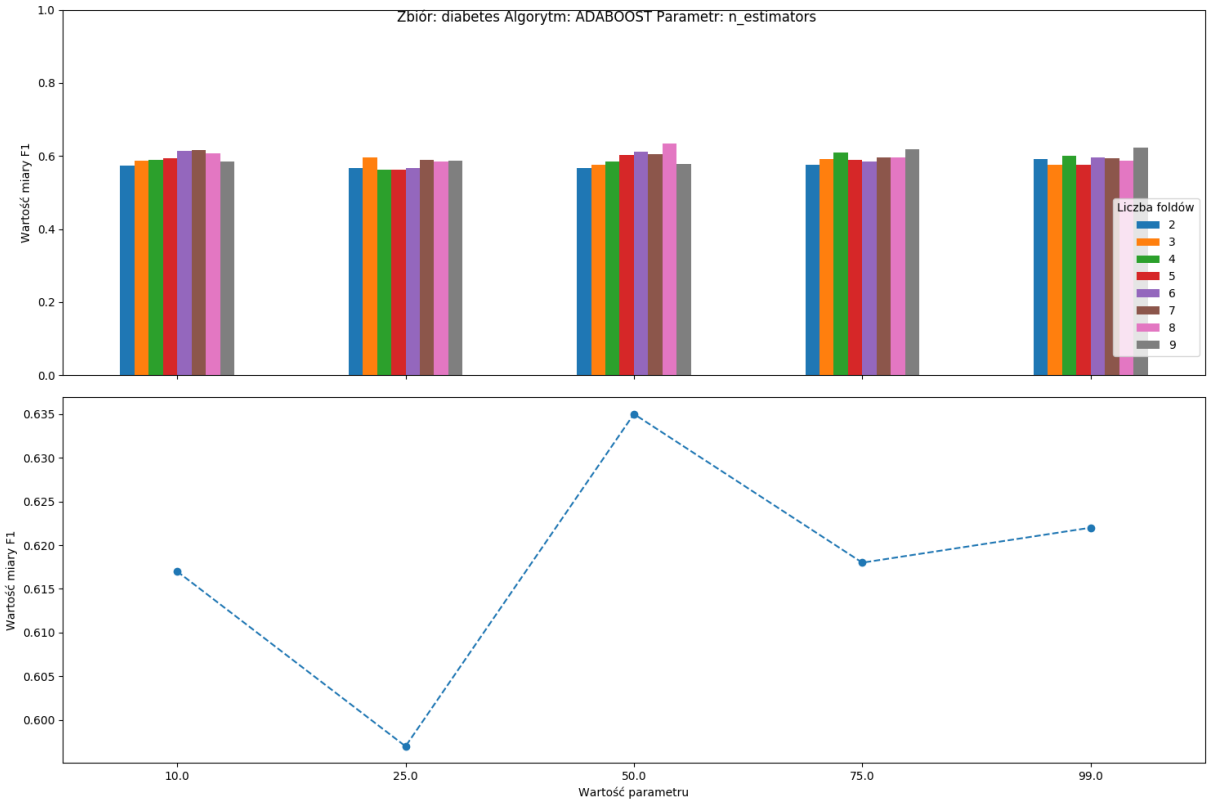
Rysunek 1: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Adaboost" przy ustalonym parametrze "algorithm".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
learning_rate	0.0001	0.565	0.591	0.603	0.598	0.583	0.631	0.617	0.600
	0.001	0.625	0.593	0.614	0.588	0.596	0.612	0.602	0.597
	0.01	0.589	0.598	0.593	0.588	0.579	0.577	0.602	0.572
	0.1	0.585	0.586	0.604	0.581	0.588	0.609	0.602	0.616



Rysunek 2: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Adaboost" przy ustalonym parametrze "learning_rate".

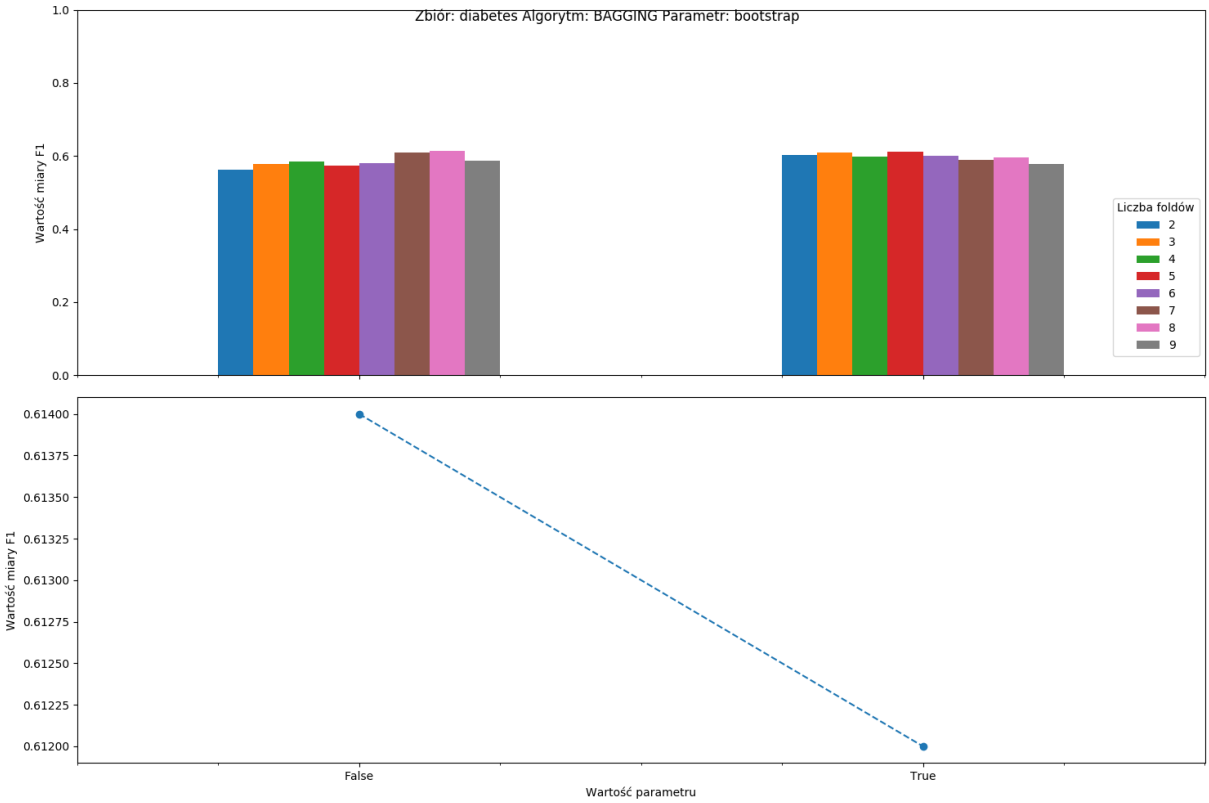
	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10.0	0.574	0.587	0.590	0.593	0.615	0.617	0.607	0.586
	25.0	0.567	0.597	0.563	0.563	0.566	0.590	0.584	0.587
	50.0	0.568	0.576	0.585	0.602	0.612	0.605	0.635	0.579
	75.0	0.576	0.591	0.609	0.590	0.585	0.595	0.597	0.618
	99.0	0.592	0.576	0.600	0.577	0.595	0.594	0.588	0.622



Rysunek 3: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Adaboost" przy ustalonym parametrze "n_estimators".

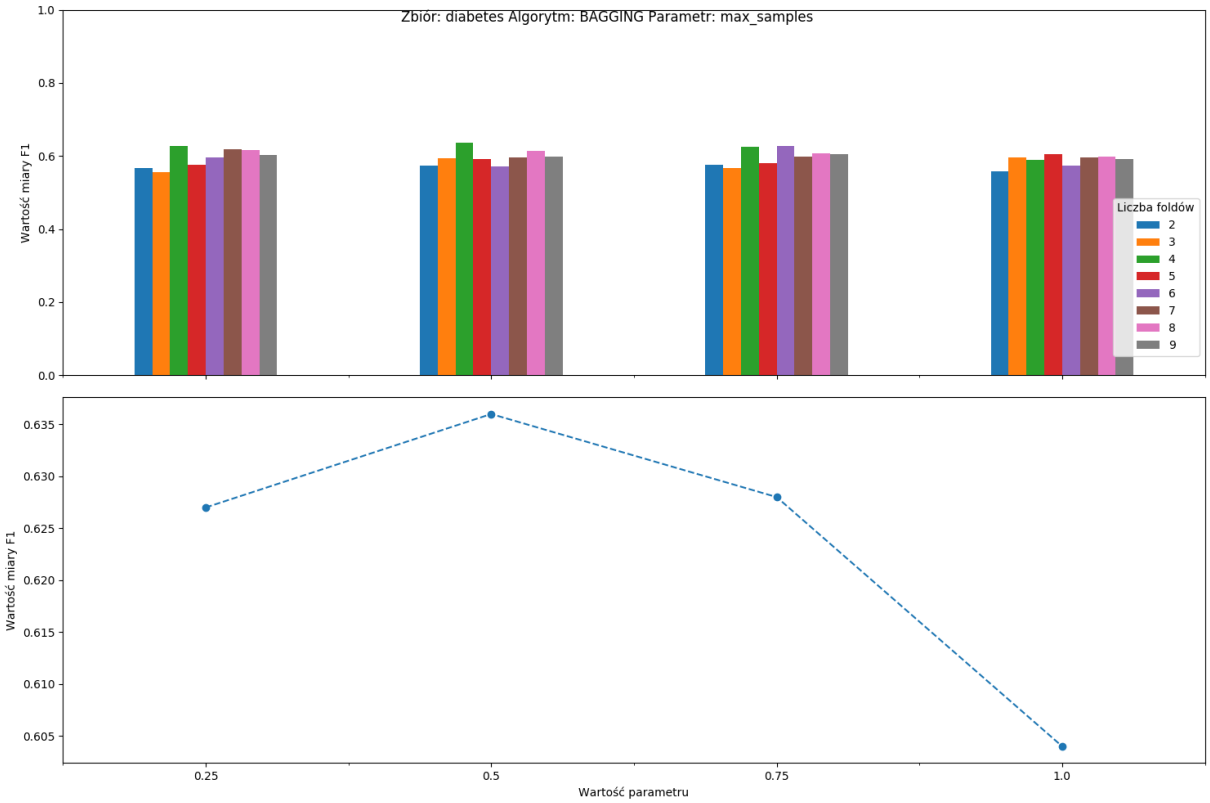
2.2 Algorytm Bagging

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
bootstrap	False	0.563	0.579	0.586	0.573	0.580	0.61	0.614	0.588
	True	0.603	0.610	0.598	0.612	0.601	0.59	0.597	0.578



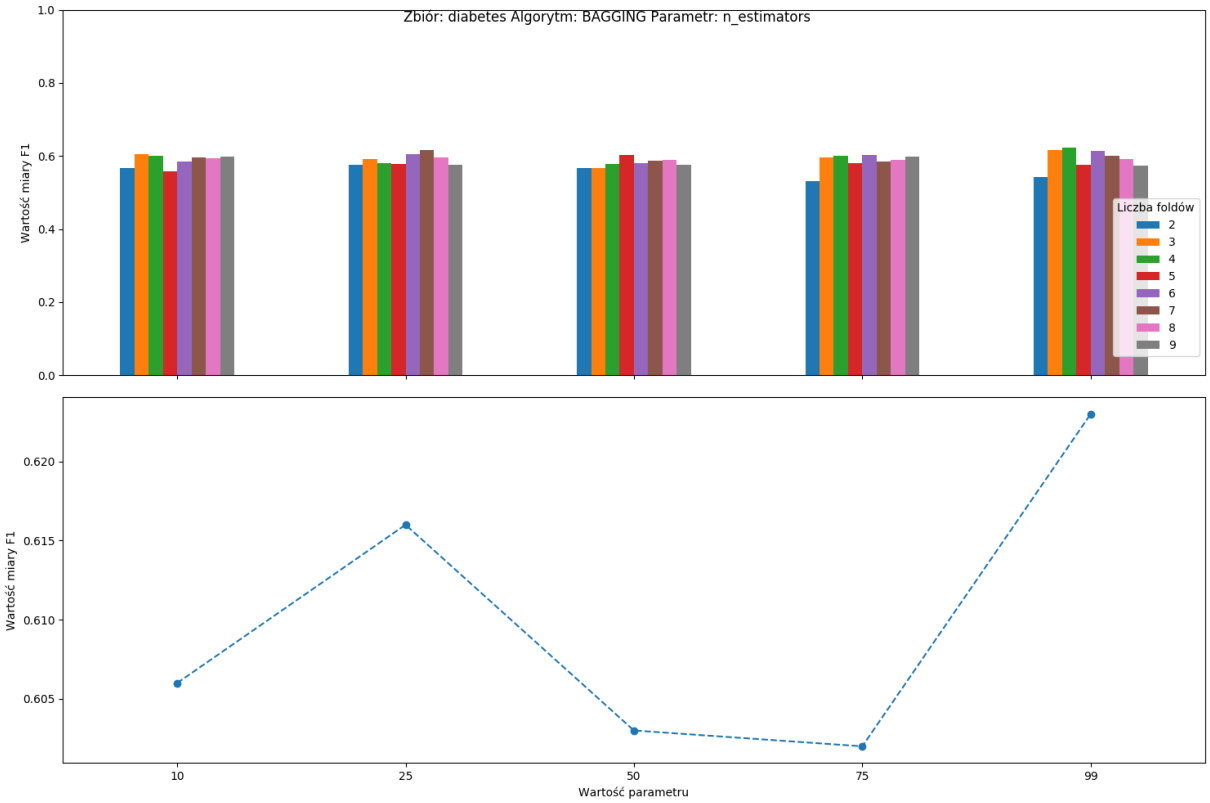
Rysunek 4: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Bagging" przy ustalonym parametrze "bootstrap".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
max_samples	0.25	0.567	0.556	0.627	0.575	0.596	0.619	0.617	0.603
	0.5	0.574	0.594	0.636	0.592	0.572	0.597	0.615	0.598
	0.75	0.576	0.566	0.626	0.580	0.628	0.598	0.608	0.604
	1.0	0.559	0.596	0.589	0.604	0.573	0.595	0.599	0.591



Rysunek 5: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Bagging" przy ustalonym parametrze "max_samples".

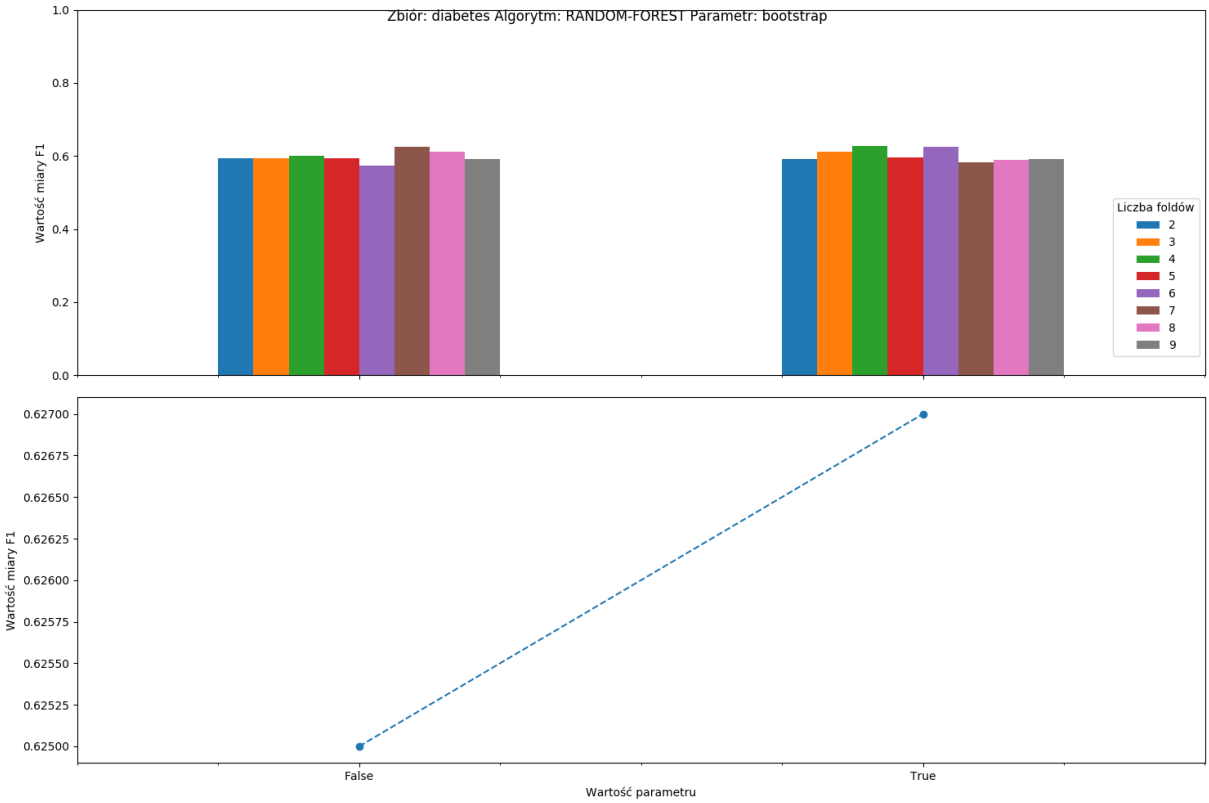
	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.566	0.606	0.601	0.557	0.585	0.595	0.594	0.598
	25	0.575	0.591	0.580	0.578	0.606	0.616	0.595	0.575
	50	0.568	0.567	0.578	0.603	0.581	0.587	0.590	0.577
	75	0.531	0.597	0.600	0.581	0.602	0.584	0.589	0.599
	99	0.542	0.616	0.623	0.575	0.615	0.600	0.592	0.573



Rysunek 6: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Bagging" przy ustalonym parametrze "n_estimators".

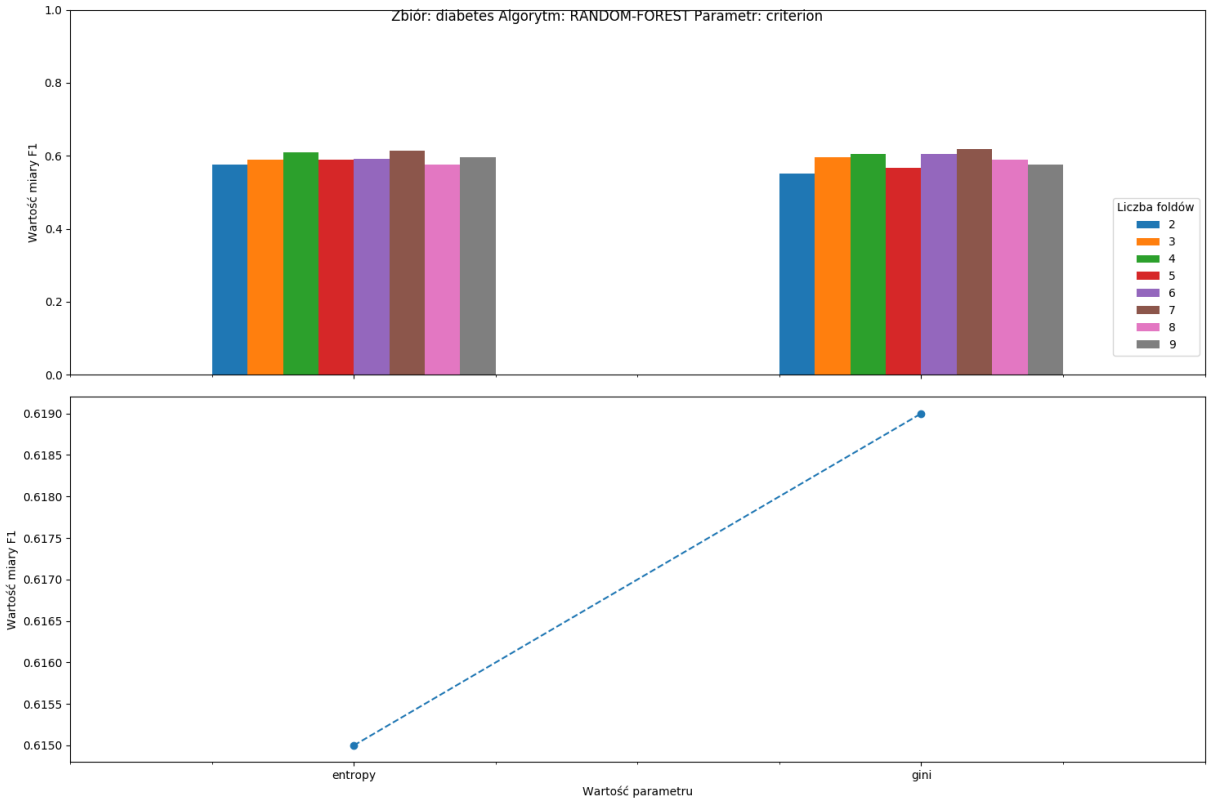
2.3 Algorytm Random-forest

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
bootstrap	False	0.593	0.593	0.600	0.593	0.573	0.625	0.612	0.591
	True	0.591	0.612	0.627	0.596	0.626	0.583	0.589	0.592



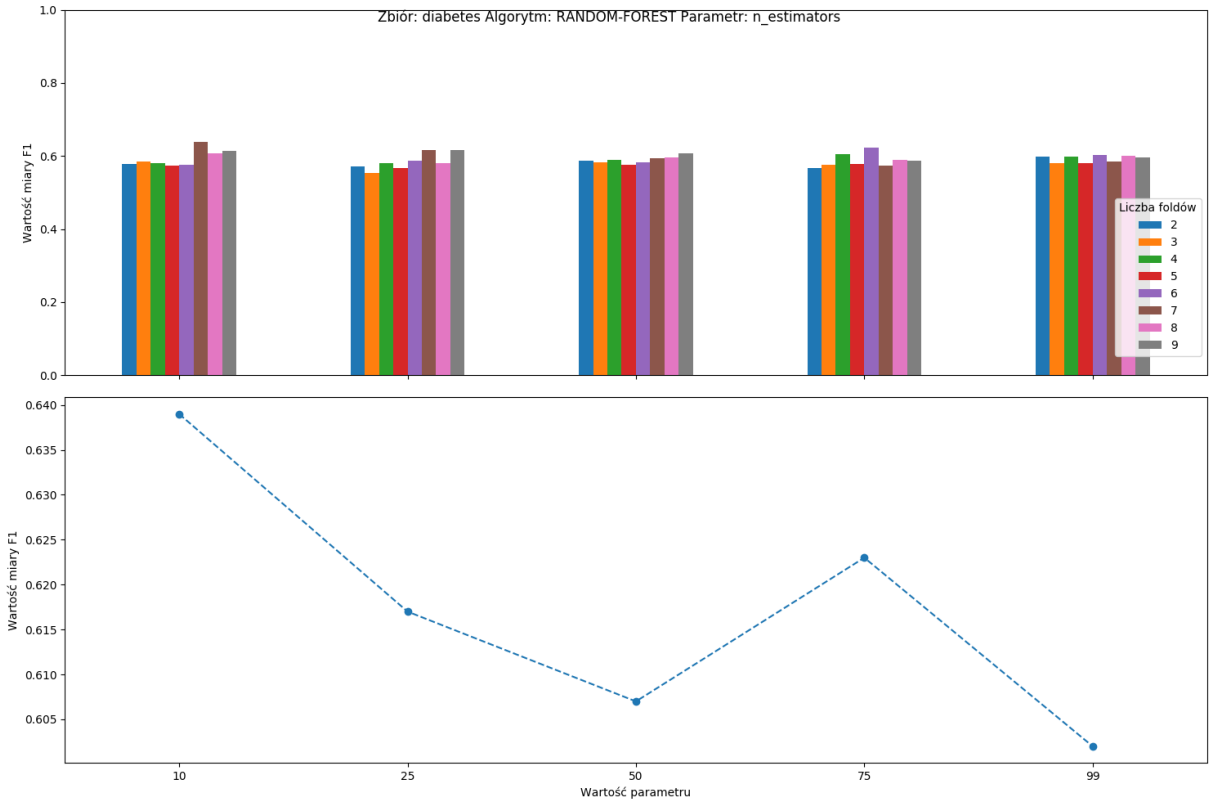
Rysunek 7: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Random-forest" przy ustalonym parametrze "bootstrap".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
criterion	entropy	0.577	0.590	0.610	0.589	0.592	0.615	0.577	0.595
	gini	0.551	0.597	0.606	0.567	0.604	0.619	0.589	0.575



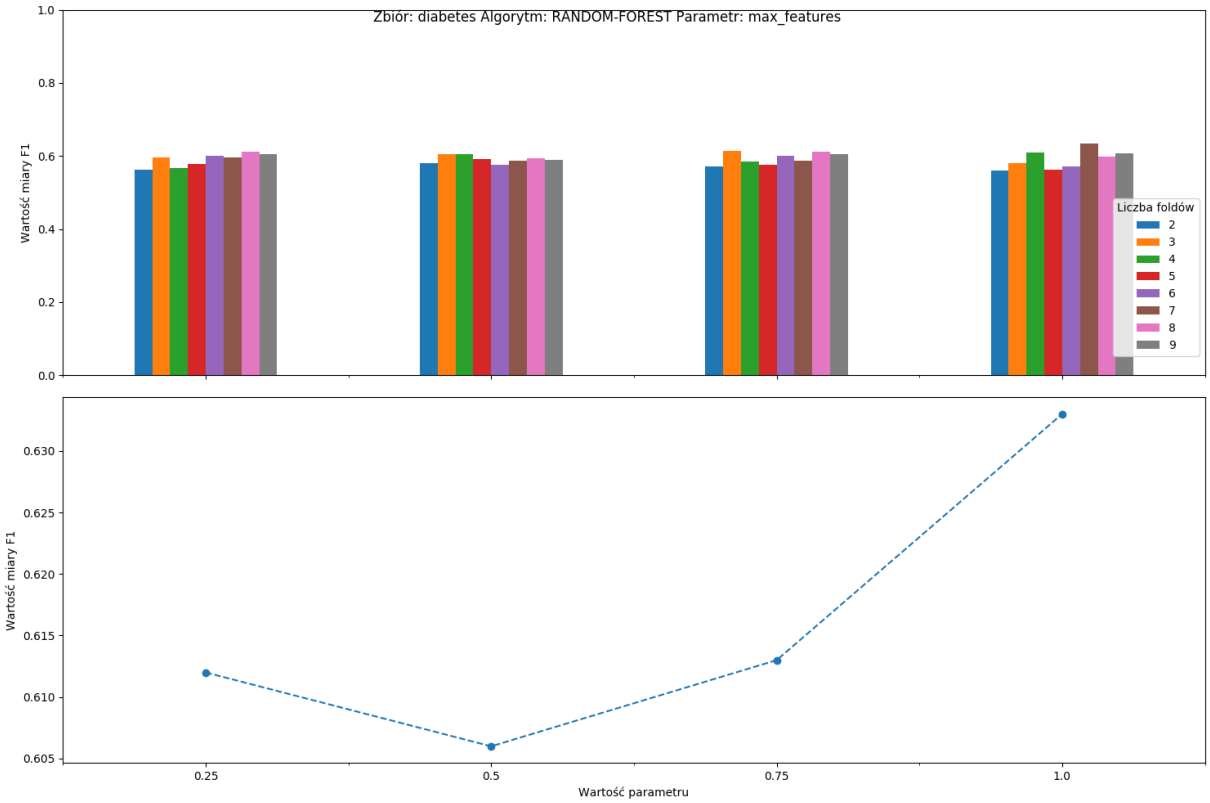
Rysunek 8: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Random-forest" przy ustalonym parametrze "criterion".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.579	0.585	0.581	0.574	0.577	0.639	0.608	0.615
	25	0.571	0.554	0.581	0.567	0.588	0.617	0.580	0.616
	50	0.587	0.583	0.589	0.575	0.583	0.593	0.597	0.607
	75	0.568	0.576	0.604	0.578	0.623	0.574	0.589	0.587
	99	0.598	0.581	0.598	0.580	0.602	0.586	0.600	0.597



Rysunek 9: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Random-forest" przy ustalonym parametrze "n_estimators".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
max_features	0.25	0.563	0.596	0.568	0.578	0.601	0.595	0.612	0.605
	0.5	0.580	0.604	0.606	0.591	0.577	0.588	0.594	0.590
	0.75	0.572	0.613	0.585	0.576	0.601	0.587	0.611	0.606
	1.0	0.561	0.580	0.610	0.563	0.572	0.633	0.598	0.607

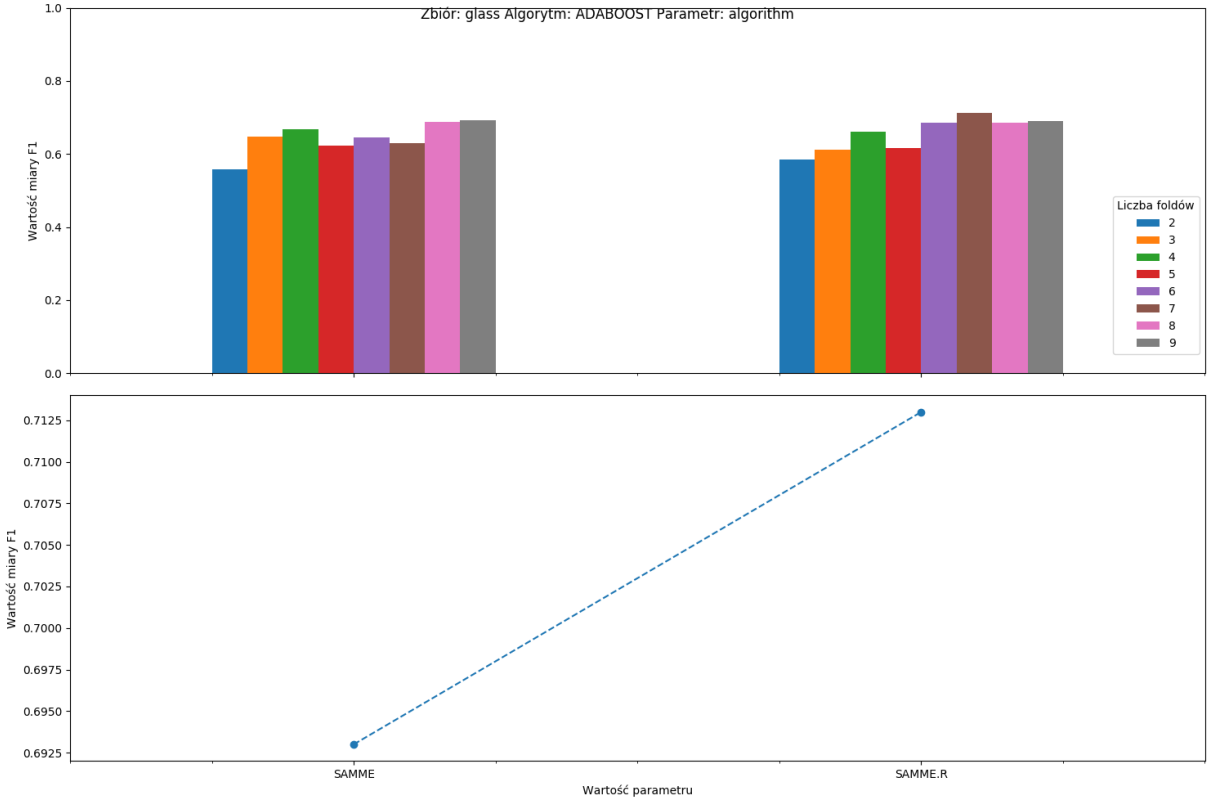


Rysunek 10: Wykres wartości miary F1 dla zbioru "Diabetes" algorytmu "Random-forest" przy ustalonym parametrze "max_features".

3 Zbiór Glass

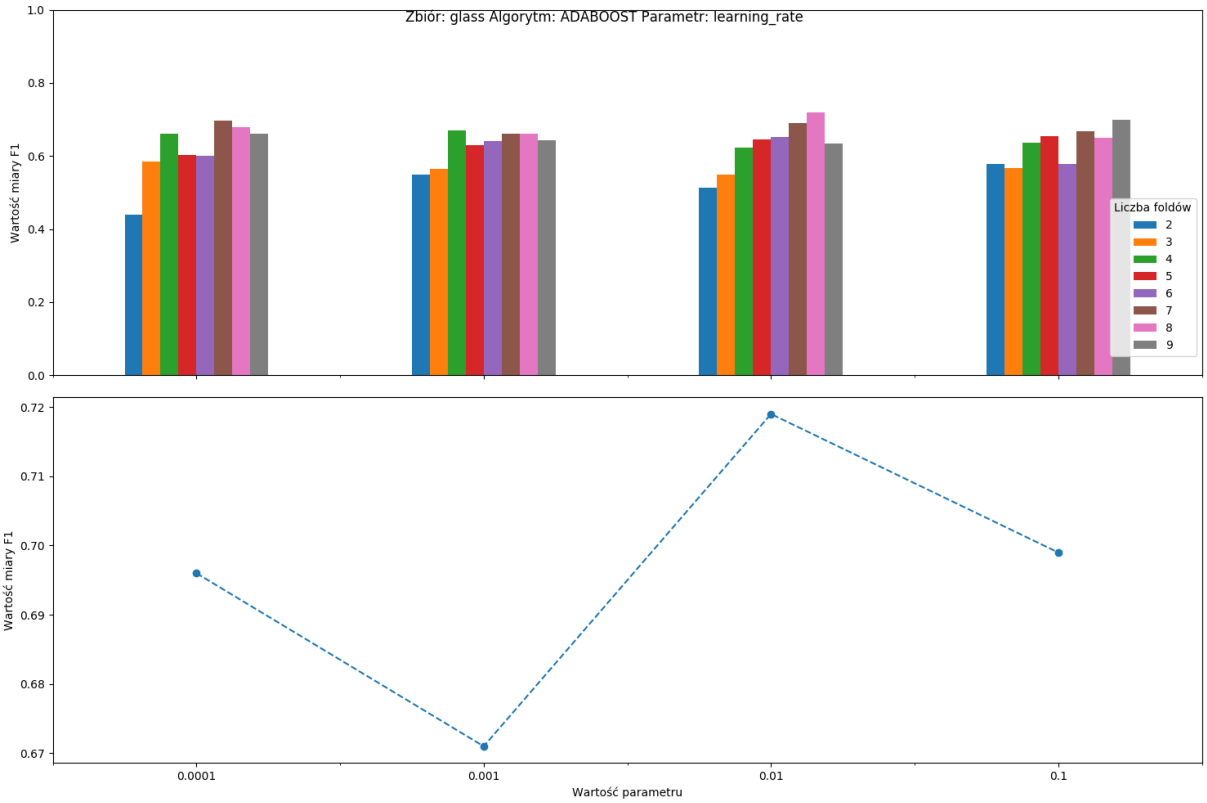
3.1 Algorytm Adaboost

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
algorithm	SAMME	0.557	0.647	0.667	0.624	0.645	0.630	0.687	0.693
	SAMME.R	0.586	0.611	0.662	0.616	0.685	0.713	0.686	0.689



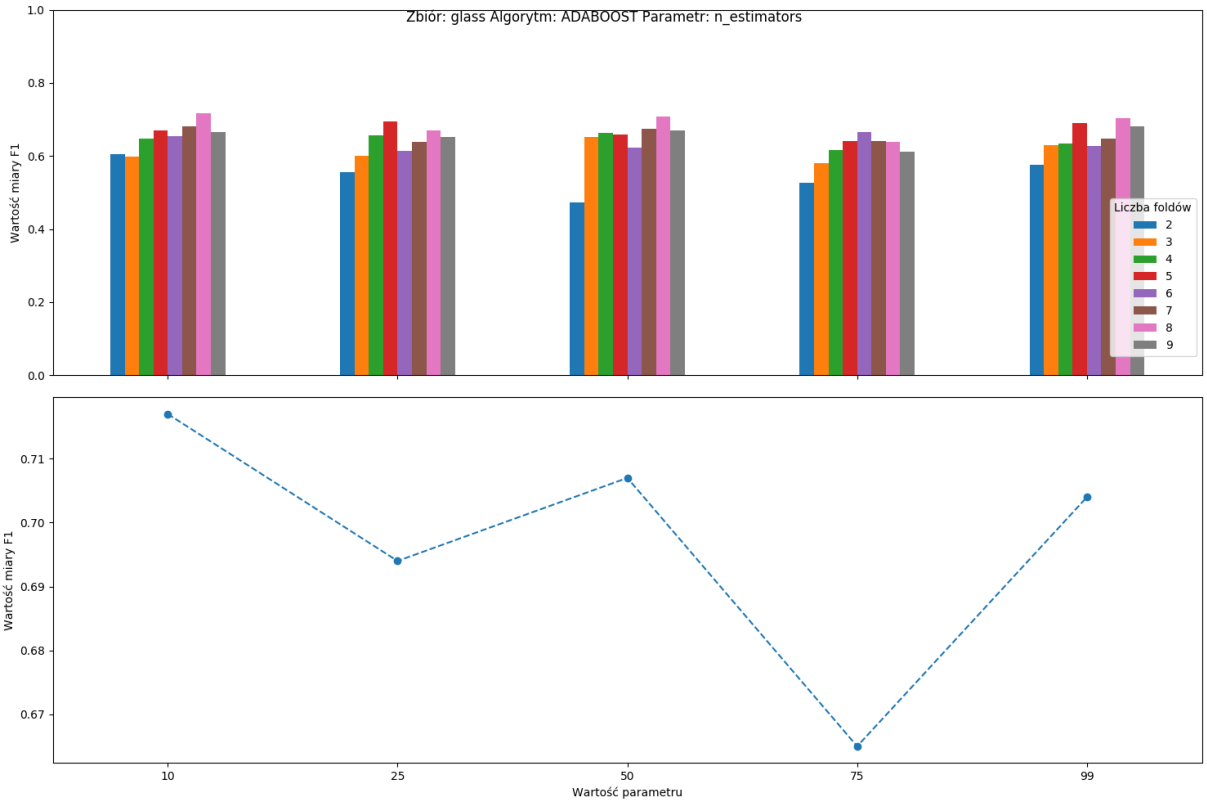
Rysunek 11: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Adaboost" przy ustalonym parametrze "algorithm".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
learning_rate	0.0001	0.440	0.585	0.661	0.602	0.600	0.696	0.679	0.661
	0.001	0.548	0.565	0.671	0.629	0.641	0.661	0.662	0.644
	0.01	0.513	0.550	0.623	0.645	0.653	0.689	0.719	0.635
	0.1	0.579	0.568	0.636	0.655	0.578	0.668	0.650	0.699



Rysunek 12: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Adaboost" przy ustalonym parametrze "learning_rate".

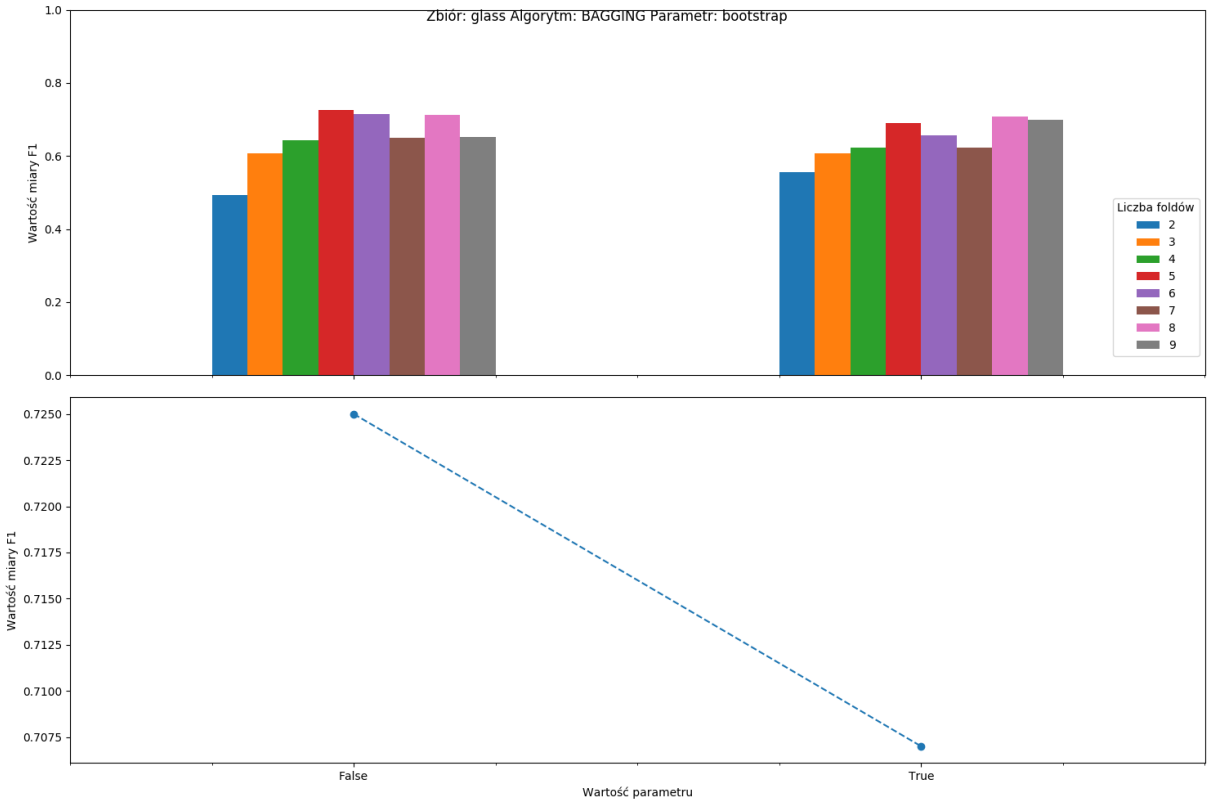
	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.606	0.598	0.647	0.669	0.655	0.682	0.717	0.666
	25	0.555	0.601	0.657	0.694	0.613	0.638	0.669	0.651
	50	0.473	0.651	0.663	0.659	0.622	0.675	0.707	0.669
	75	0.526	0.580	0.617	0.641	0.665	0.641	0.639	0.612
	99	0.576	0.629	0.634	0.690	0.627	0.647	0.704	0.682



Rysunek 13: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Adaboost" przy ustalonym parametrze "n_estimators".

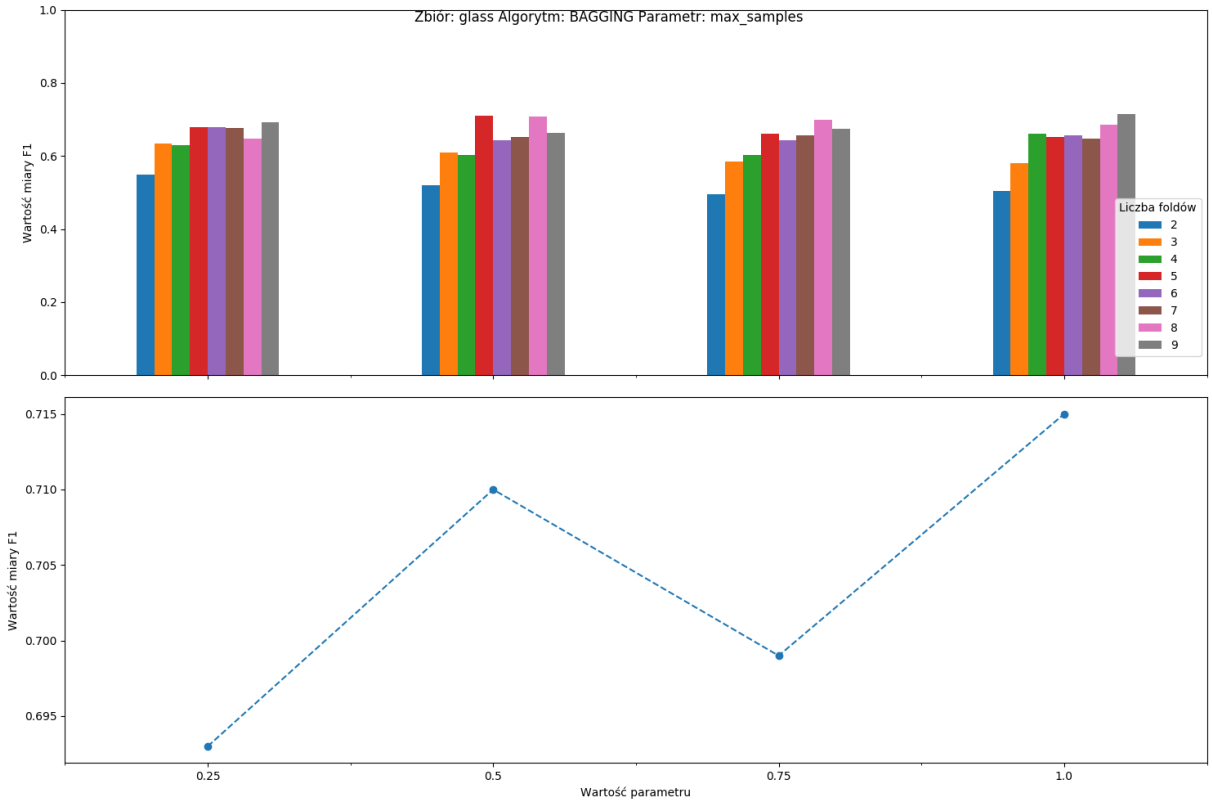
3.2 Algorytm Bagging

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
bootstrap	False	0.494	0.607	0.644	0.725	0.714	0.649	0.713	0.653
	True	0.555	0.607	0.622	0.689	0.656	0.624	0.707	0.698



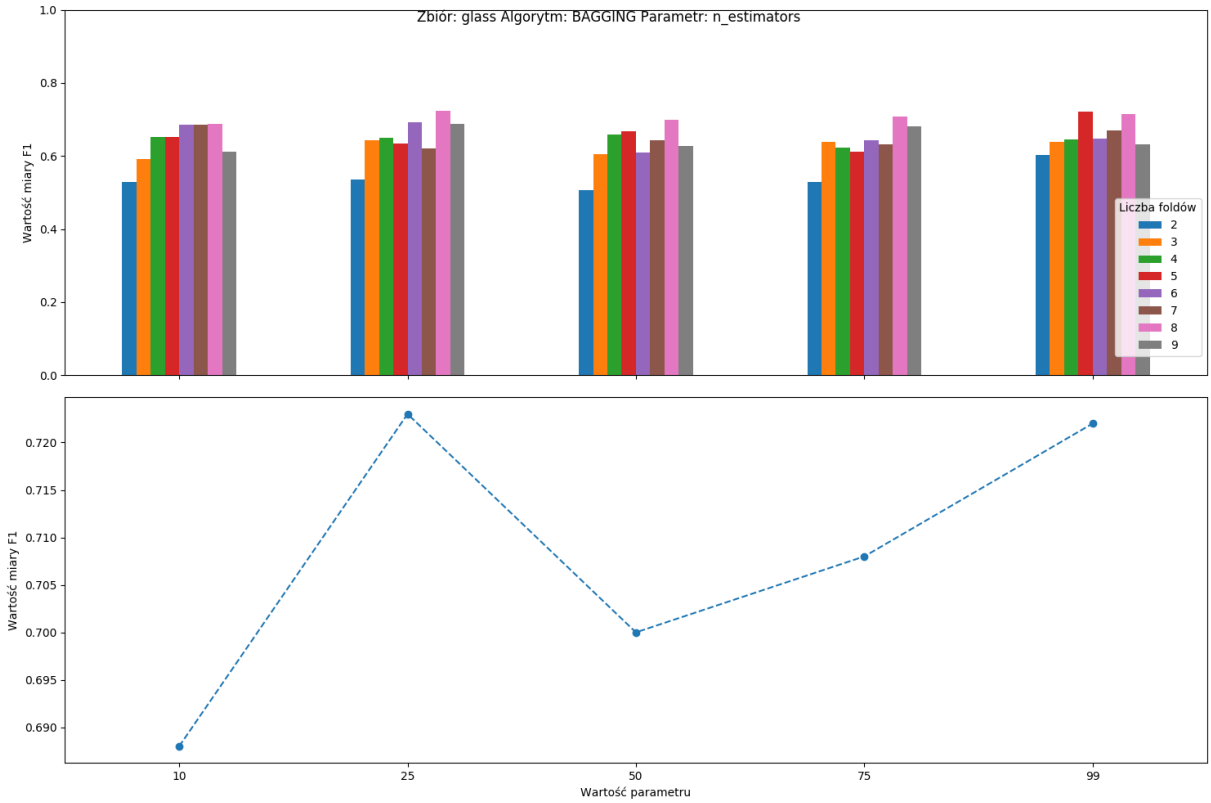
Rysunek 14: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Bagging" przy ustalonym parametrze "bootstrap".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
max_samples	0.25	0.550	0.633	0.630	0.679	0.678	0.677	0.648	0.693
	0.5	0.520	0.610	0.602	0.710	0.642	0.652	0.707	0.664
	0.75	0.496	0.584	0.603	0.662	0.643	0.656	0.699	0.675
	1.0	0.505	0.581	0.661	0.651	0.656	0.648	0.686	0.715



Rysunek 15: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Bagging" przy ustalonym parametrze "max_samples".

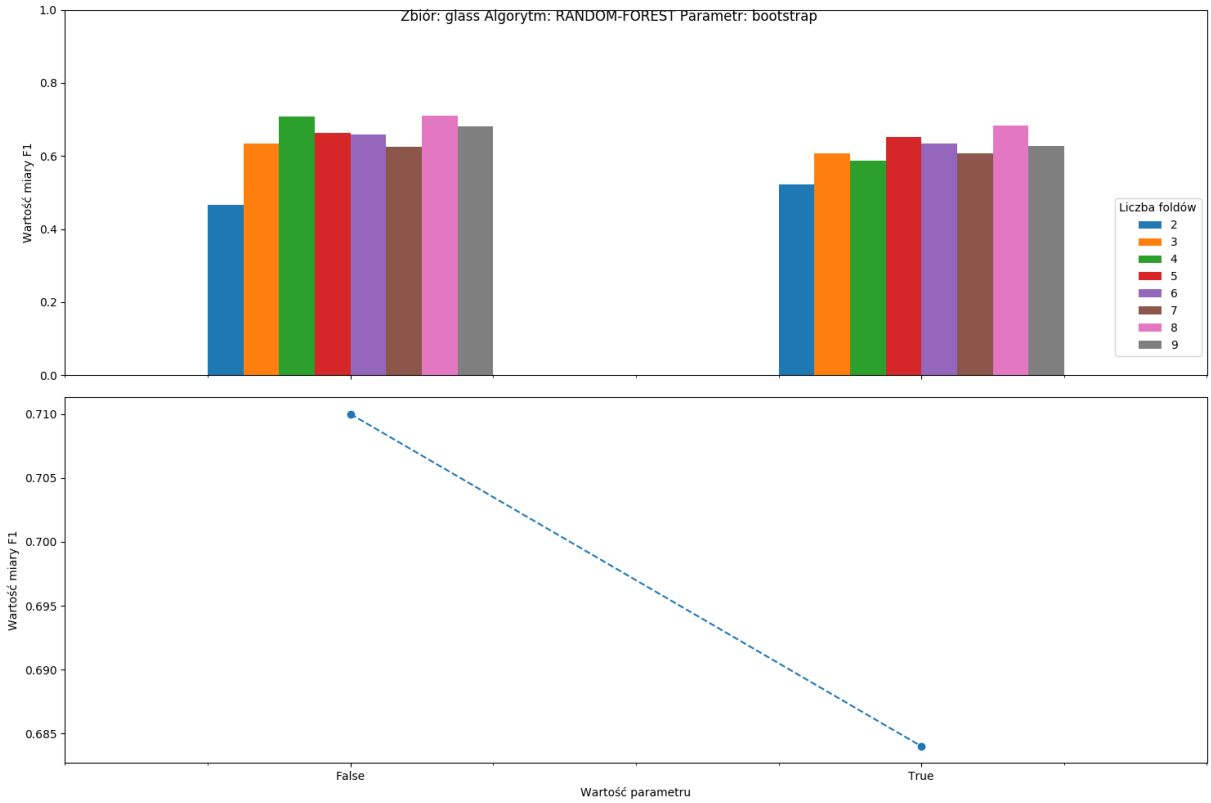
	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.528	0.591	0.652	0.651	0.685	0.686	0.688	0.611
	25	0.535	0.644	0.649	0.634	0.693	0.621	0.723	0.687
	50	0.507	0.606	0.659	0.667	0.610	0.643	0.700	0.627
	75	0.530	0.638	0.623	0.612	0.643	0.632	0.708	0.682
	99	0.603	0.639	0.645	0.722	0.648	0.671	0.714	0.632



Rysunek 16: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Bagging" przy ustalonym parametrze "n_estimators".

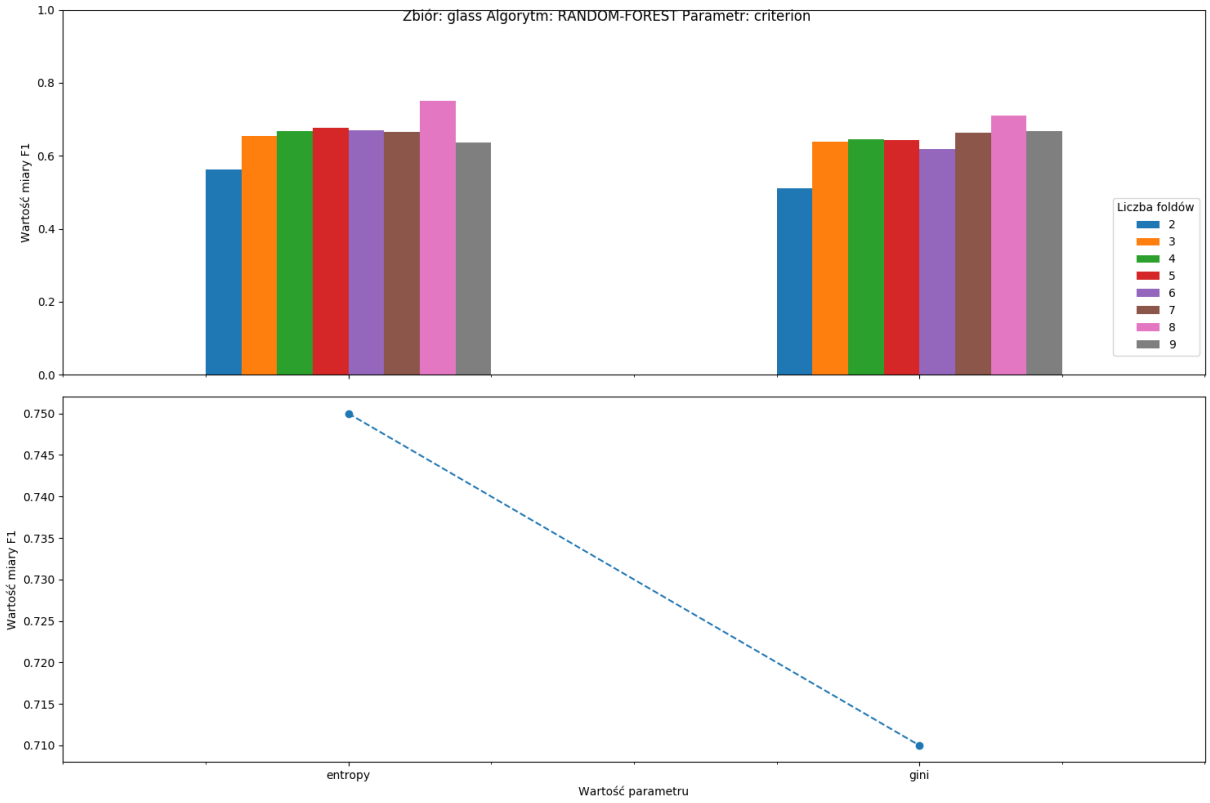
3.3 Algorytm Random-forest

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
bootstrap	False	0.467	0.635	0.708	0.663	0.658	0.625	0.710	0.682
	True	0.523	0.607	0.587	0.651	0.633	0.608	0.684	0.627



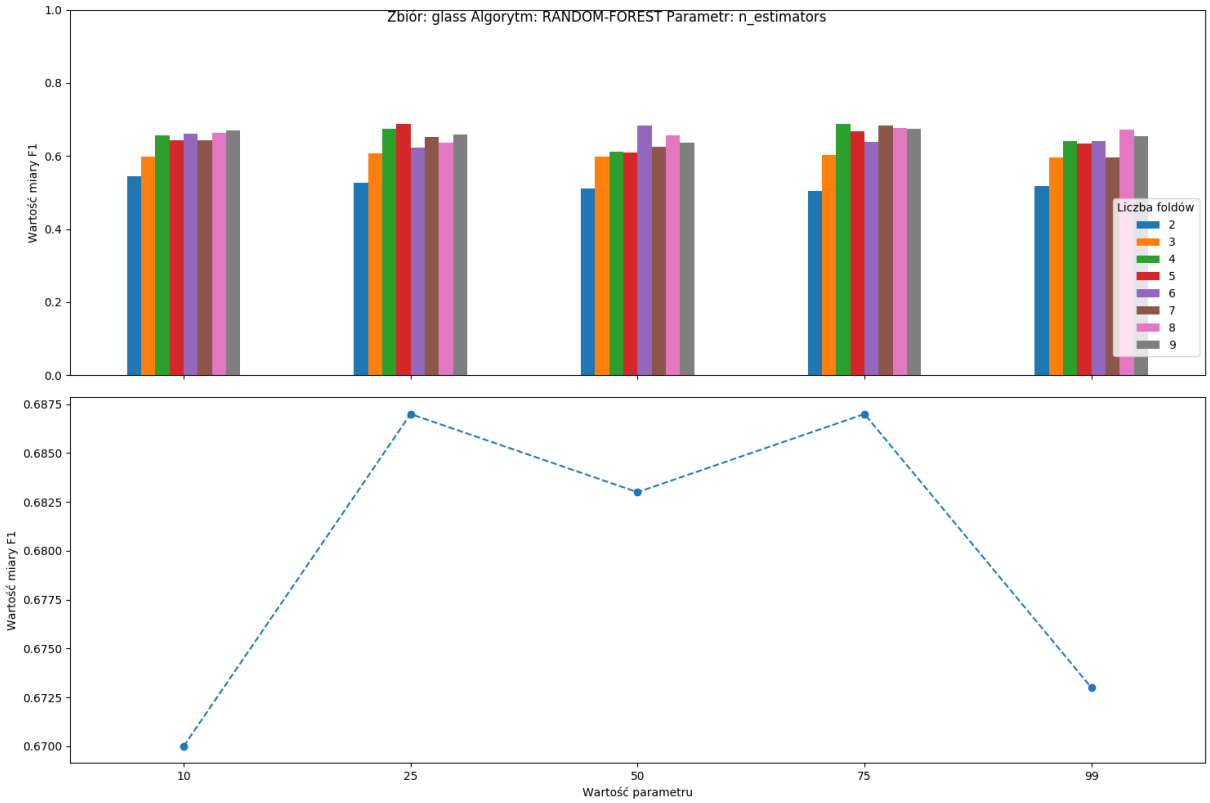
Rysunek 17: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Random-forest" przy ustalonym parametrze "bootstrap".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
criterion	entropy	0.563	0.654	0.668	0.676	0.669	0.666	0.75	0.637
	gini	0.510	0.639	0.645	0.643	0.619	0.663	0.71	0.668



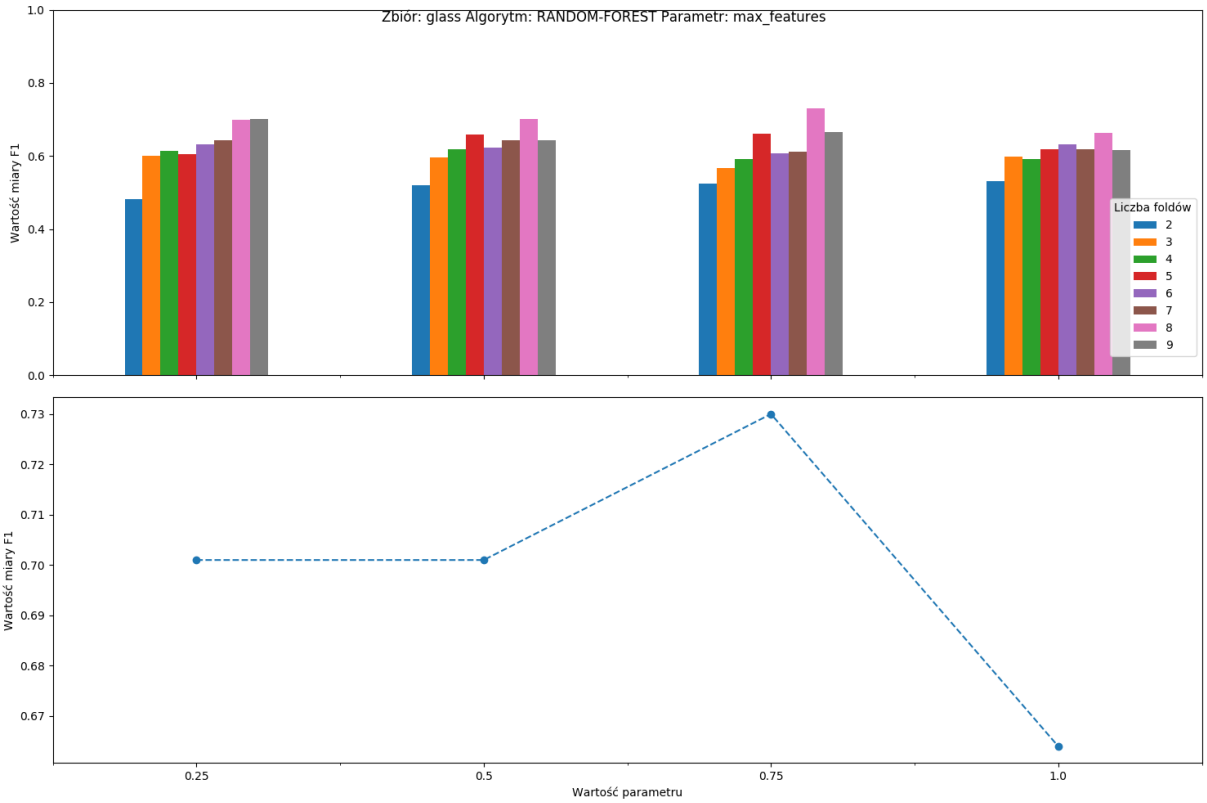
Rysunek 18: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Random-forest" przy ustalonym parametrze "criterion".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.545	0.598	0.656	0.643	0.660	0.644	0.664	0.670
	25	0.527	0.608	0.674	0.687	0.624	0.651	0.636	0.658
	50	0.511	0.599	0.612	0.610	0.683	0.625	0.657	0.637
	75	0.504	0.603	0.687	0.667	0.639	0.684	0.677	0.674
	99	0.517	0.597	0.641	0.635	0.641	0.596	0.673	0.655



Rysunek 19: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Random-forest" przy ustalonym parametrze "n_estimators".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
max_features	0.25	0.482	0.600	0.615	0.606	0.632	0.643	0.698	0.701
	0.5	0.520	0.597	0.619	0.658	0.622	0.644	0.701	0.644
	0.75	0.525	0.568	0.592	0.661	0.607	0.612	0.730	0.666
	1.0	0.532	0.598	0.592	0.618	0.631	0.619	0.664	0.616

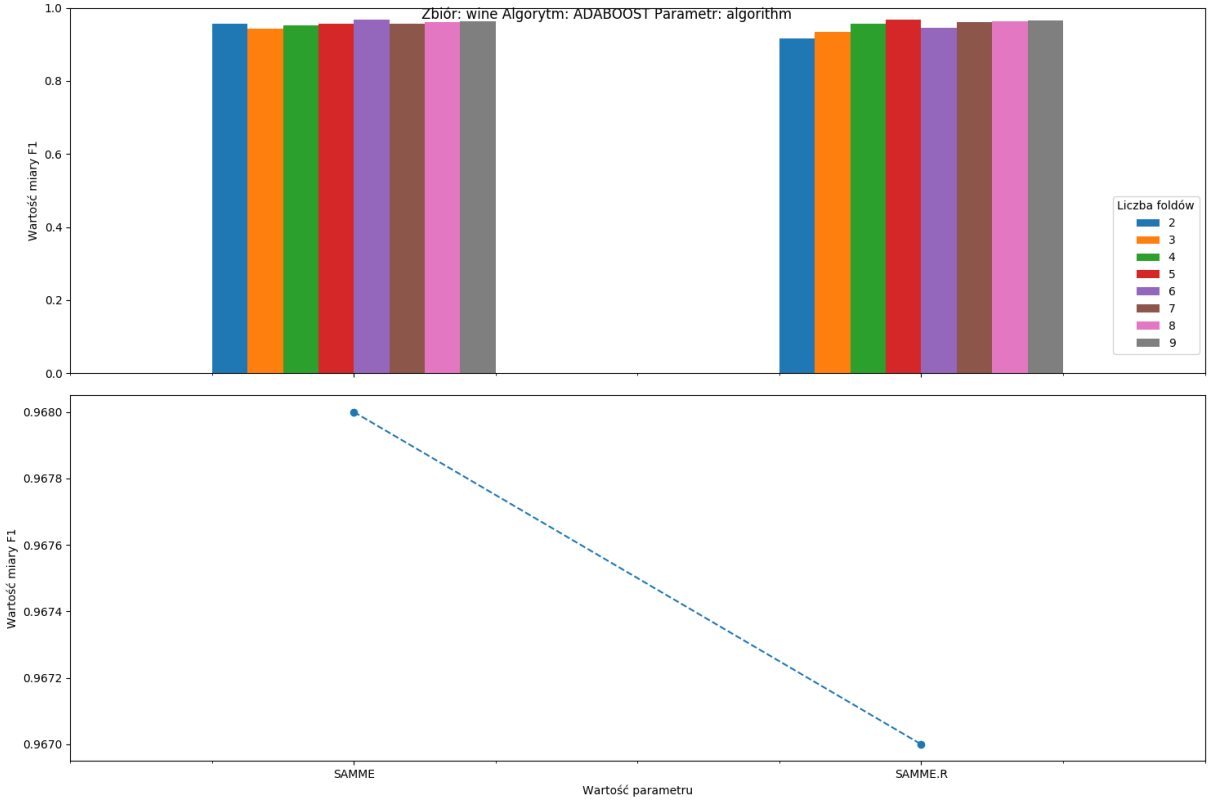


Rysunek 20: Wykres wartości miary F1 dla zbioru "Glass" algorytmu "Random-forest" przy ustalonym parametrze "max_features".

4 Zbiór Wine

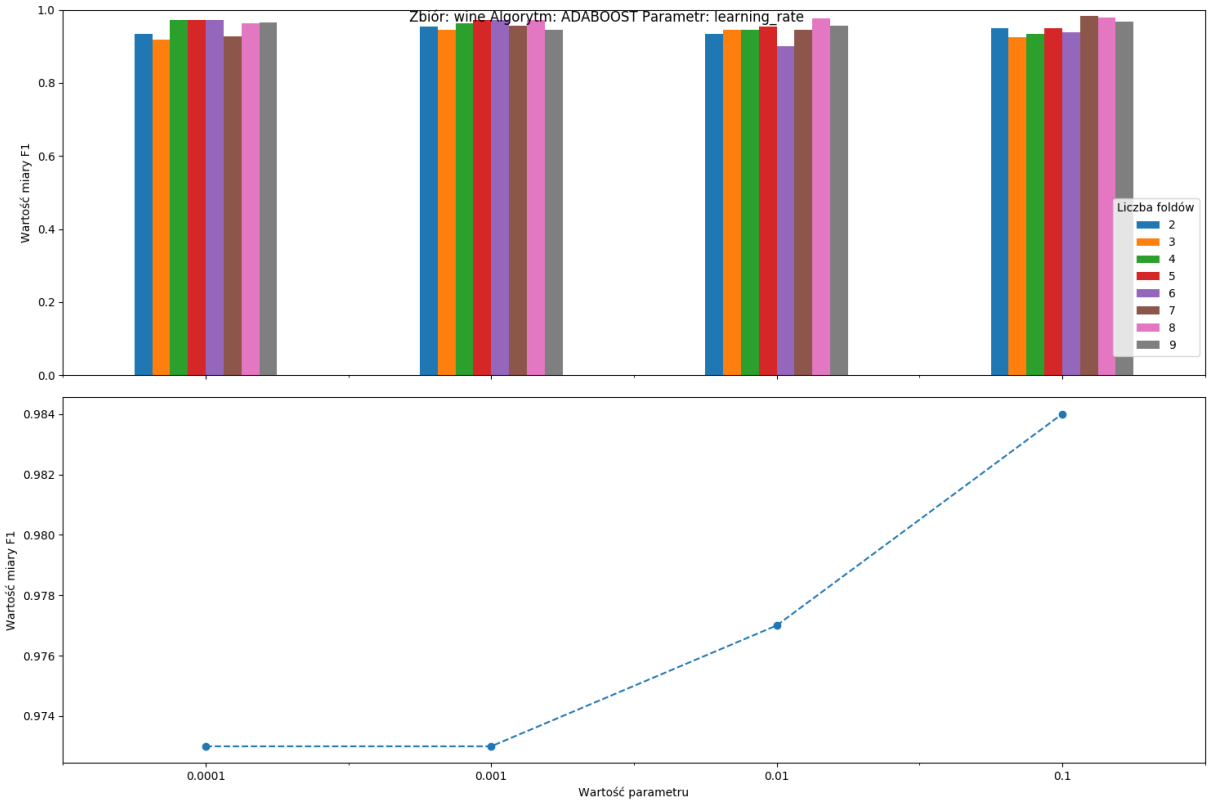
4.1 Algorytm Adaboost

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
algorithm	SAMME	0.956	0.944	0.951	0.956	0.968	0.957	0.961	0.962
	SAMME.R	0.915	0.934	0.957	0.967	0.945	0.961	0.962	0.966



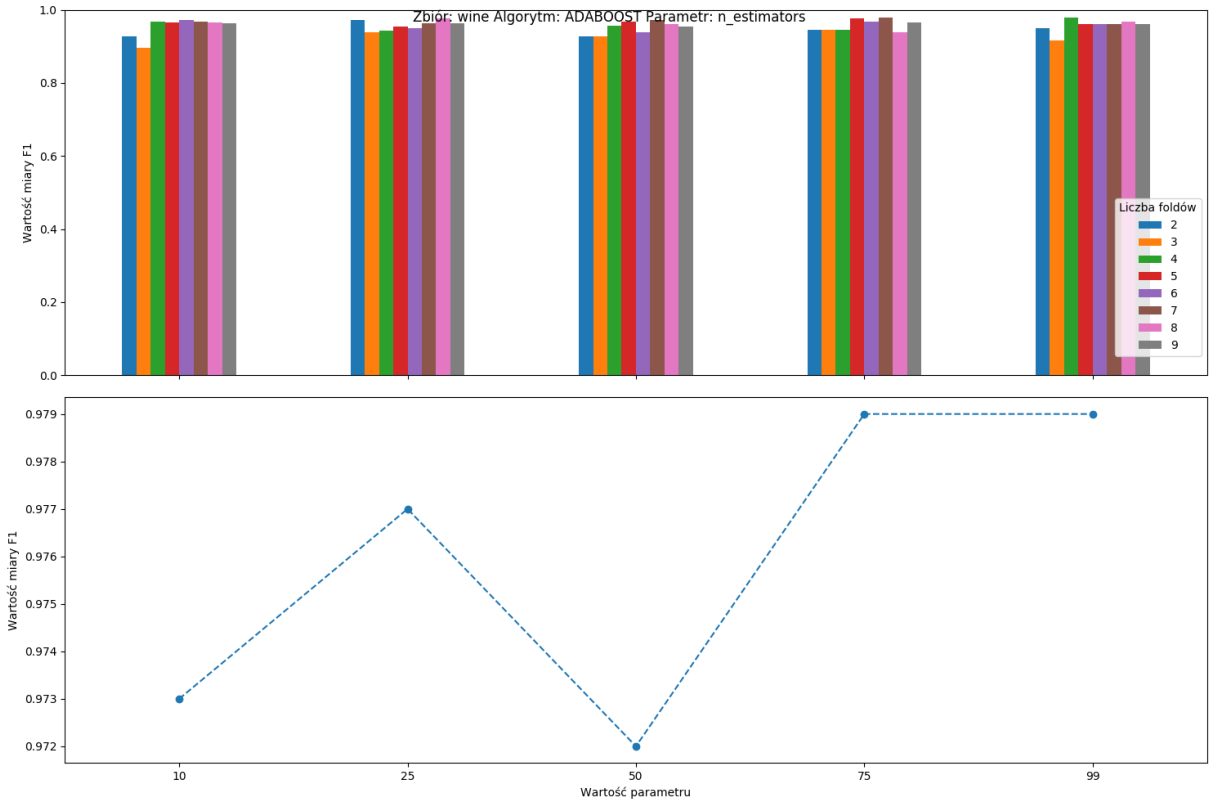
Rysunek 21: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Adaboost" przy ustalonym parametrze "algorithm".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
learning_rate	0.0001	0.934	0.918	0.972	0.972	0.973	0.928	0.962	0.966
	0.001	0.955	0.945	0.962	0.972	0.973	0.956	0.972	0.945
	0.01	0.933	0.946	0.945	0.955	0.900	0.946	0.977	0.956
	0.1	0.950	0.924	0.934	0.950	0.939	0.984	0.978	0.968



Rysunek 22: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Adaboost" przy ustalonym parametrze "learning_rate".

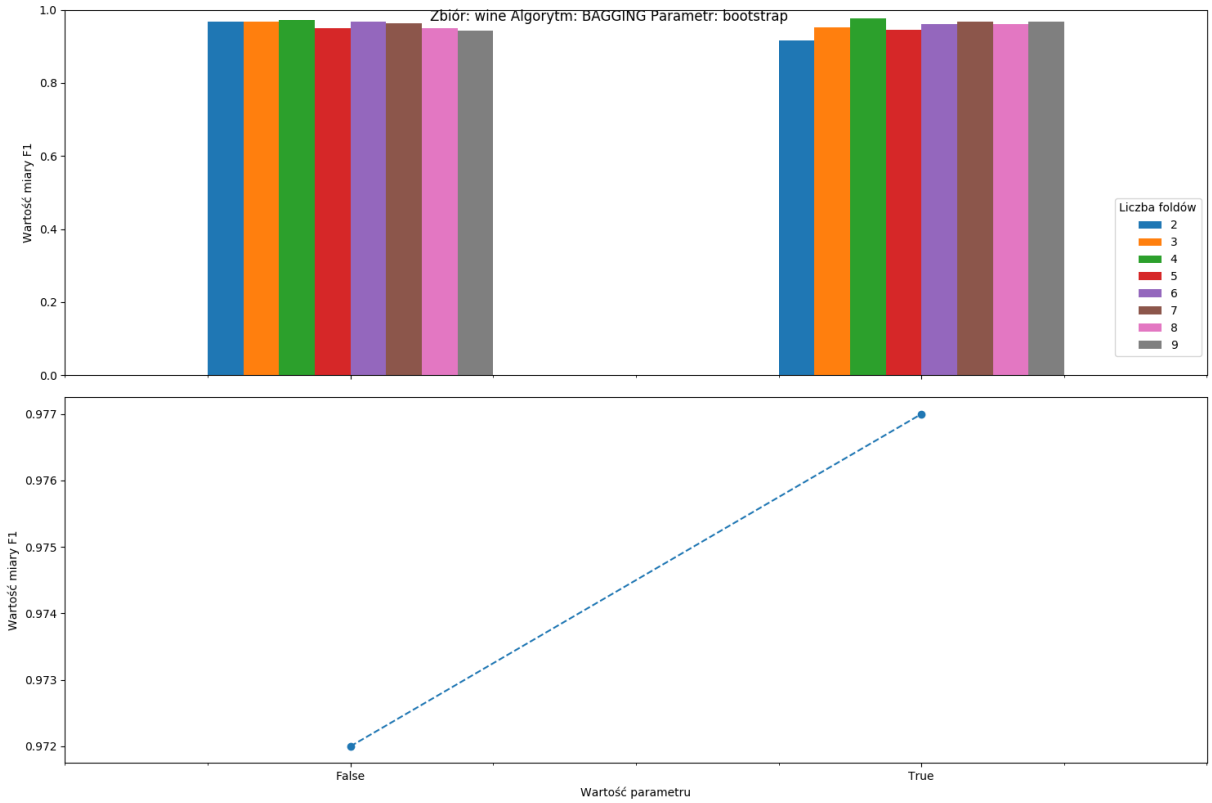
	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.928	0.896	0.968	0.966	0.973	0.967	0.966	0.962
	25	0.972	0.939	0.944	0.955	0.949	0.962	0.977	0.962
	50	0.928	0.927	0.957	0.967	0.939	0.972	0.960	0.955
	75	0.945	0.945	0.946	0.977	0.968	0.979	0.938	0.966
	99	0.950	0.917	0.979	0.961	0.961	0.961	0.968	0.960



Rysunek 23: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Adaboost" przy ustalonym parametrze "n_estimators".

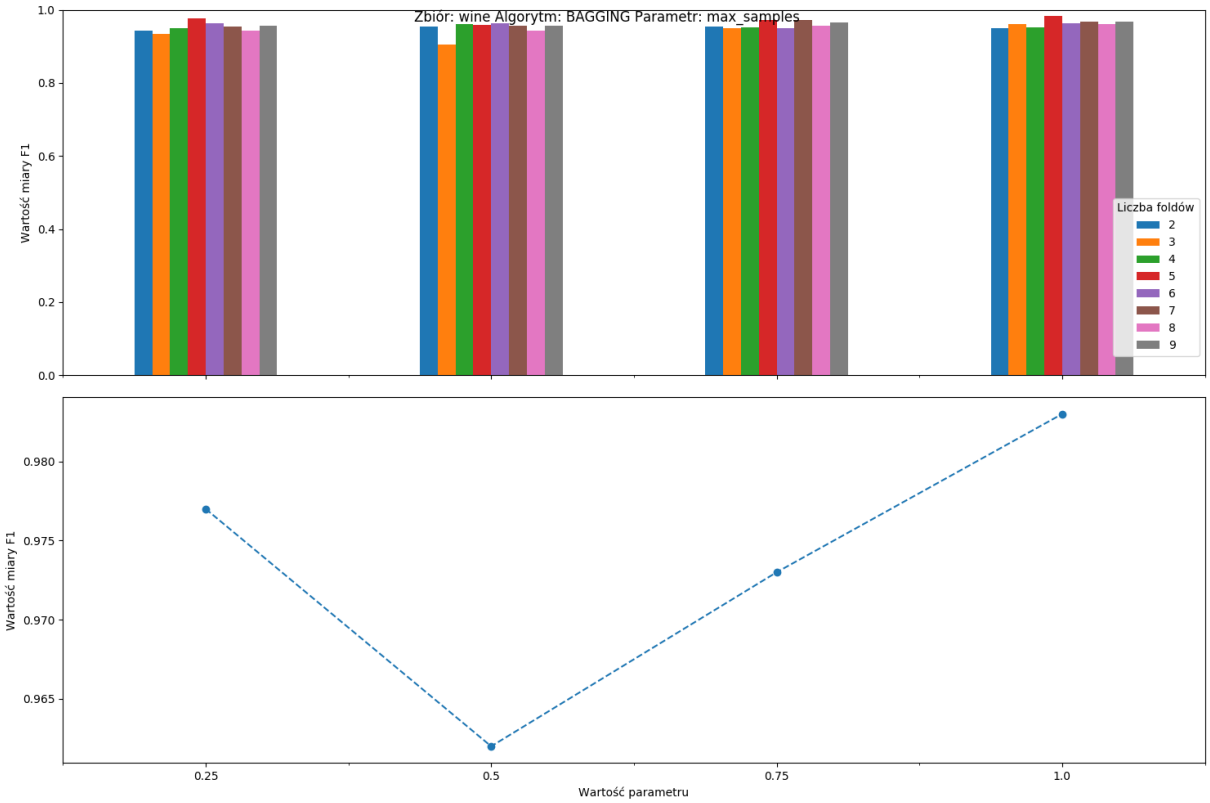
4.2 Algorytm Bagging

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
bootstrap	False	0.967	0.968	0.972	0.950	0.967	0.962	0.949	0.944
	True	0.915	0.951	0.977	0.945	0.961	0.967	0.961	0.967



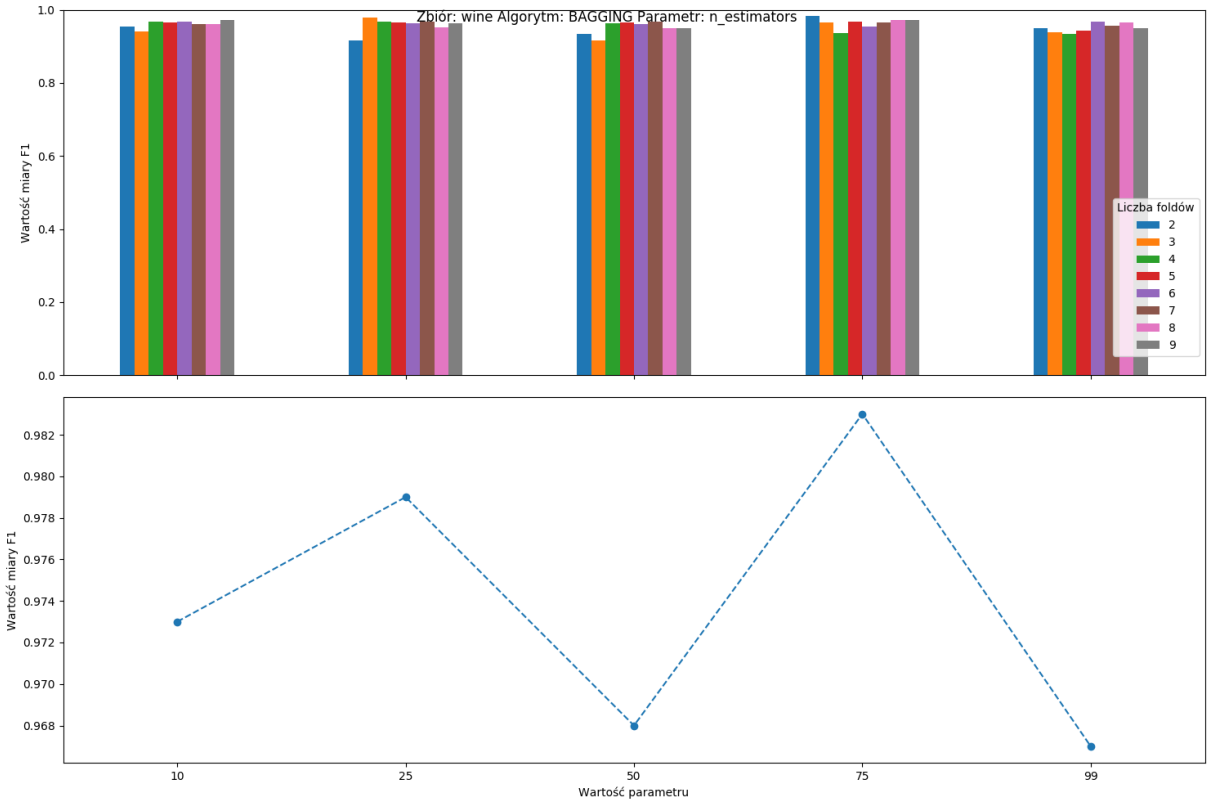
Rysunek 24: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Bagging" przy ustalonym parametrze "bootstrap".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
max_samples	0.25	0.943	0.935	0.950	0.977	0.962	0.955	0.944	0.957
	0.5	0.955	0.905	0.961	0.959	0.962	0.957	0.944	0.957
	0.75	0.954	0.949	0.952	0.972	0.950	0.973	0.956	0.966
	1.0	0.949	0.961	0.951	0.983	0.962	0.967	0.961	0.967



Rysunek 25: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Bagging" przy ustalonym parametrze "max_samples".

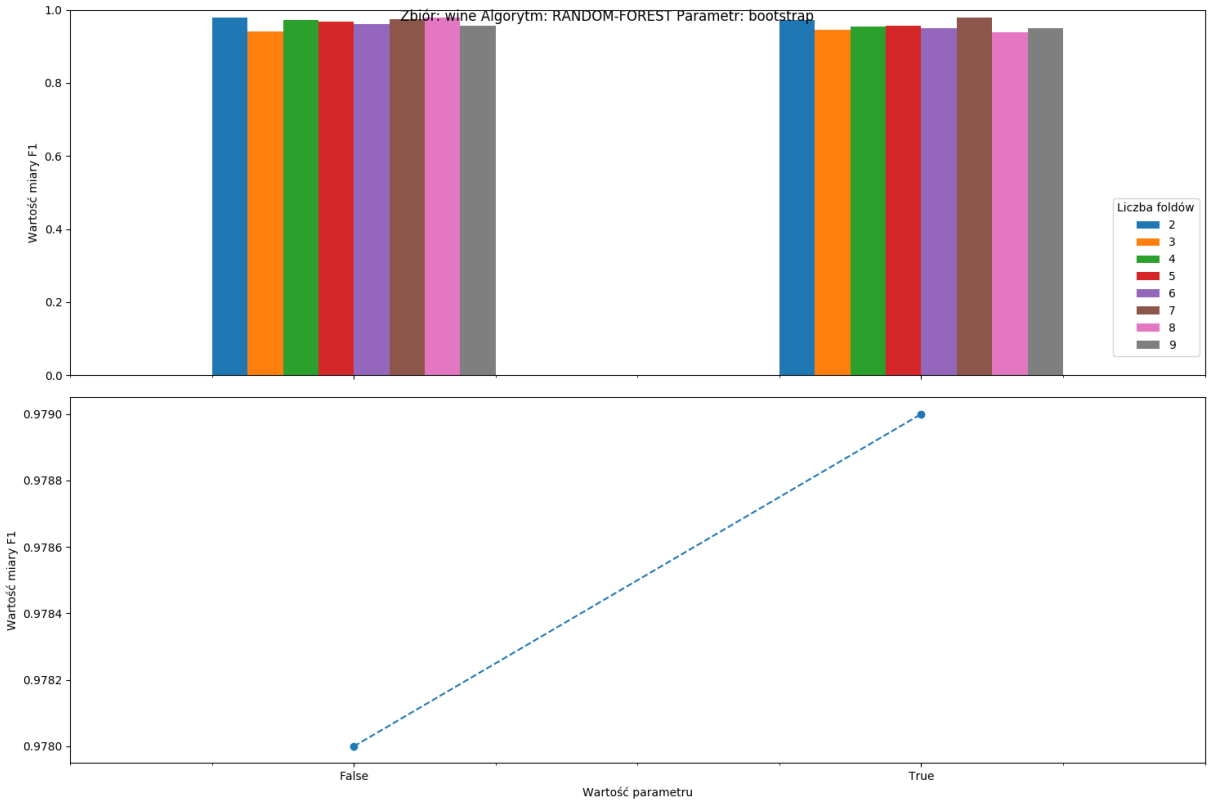
	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.955	0.940	0.967	0.966	0.967	0.961	0.961	0.973
	25	0.915	0.979	0.968	0.966	0.962	0.967	0.951	0.962
	50	0.934	0.915	0.962	0.966	0.961	0.968	0.949	0.950
	75	0.983	0.966	0.937	0.968	0.955	0.966	0.971	0.972
	99	0.949	0.939	0.935	0.944	0.967	0.957	0.966	0.950



Rysunek 26: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Bagging" przy ustalonym parametrze "n_estimators".

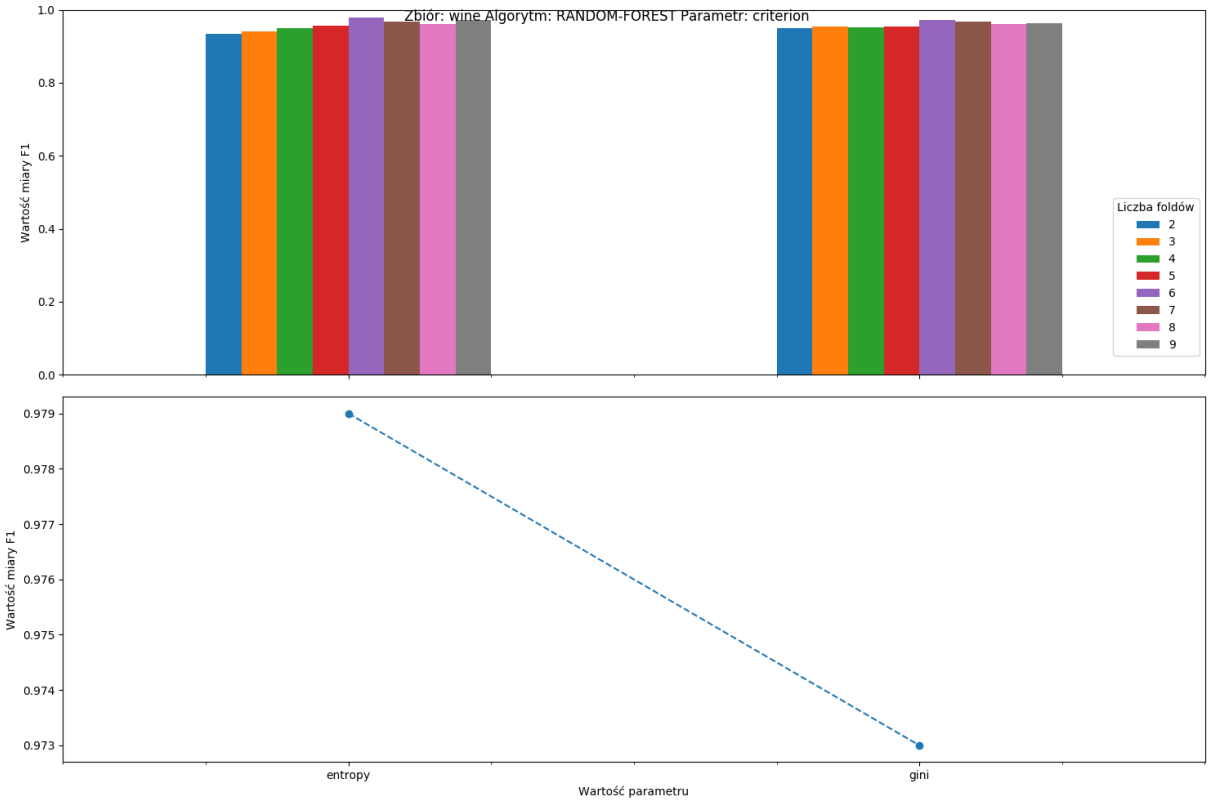
4.3 Algorytm Random-forest

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
bootstrap	False	0.978	0.941	0.972	0.967	0.961	0.974	0.978	0.957
	True	0.973	0.945	0.955	0.956	0.949	0.979	0.938	0.950



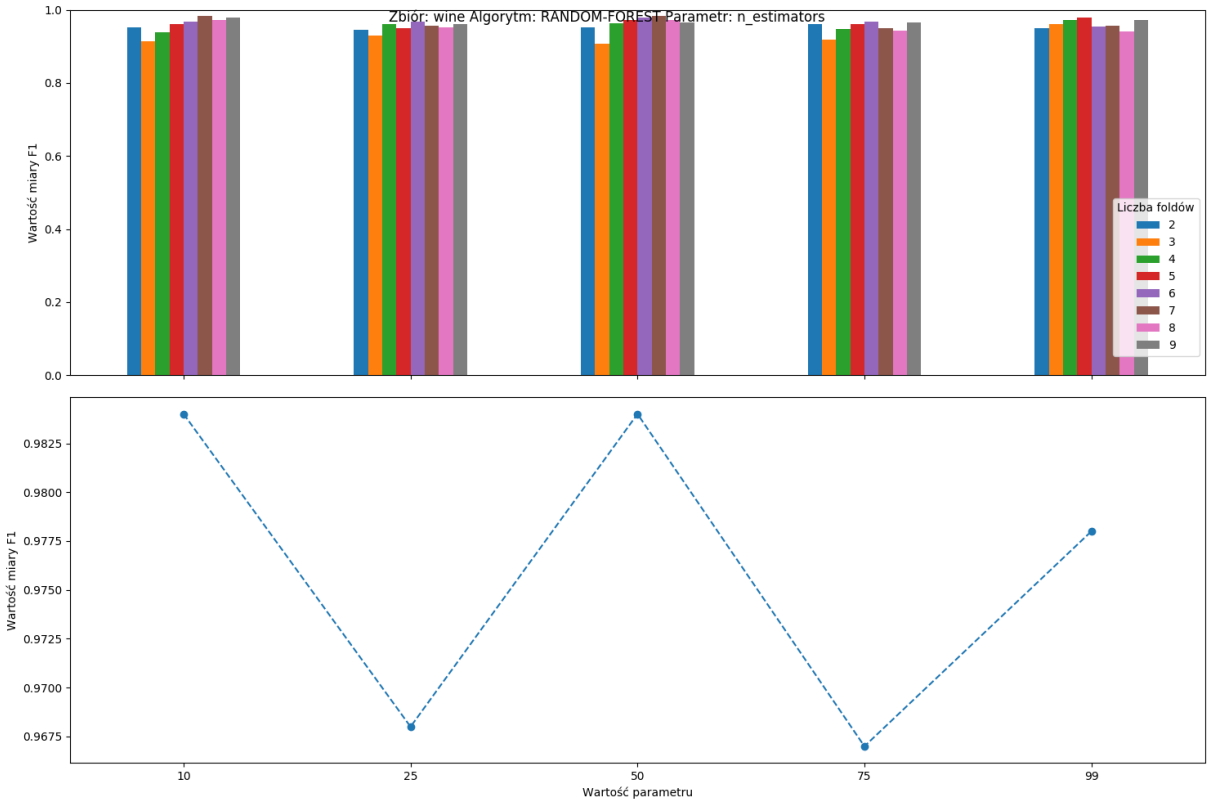
Rysunek 27: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Random-forest" przy ustalonym parametrze "bootstrap".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
criterion	entropy	0.933	0.941	0.949	0.956	0.979	0.968	0.961	0.973
	gini	0.950	0.954	0.951	0.955	0.973	0.967	0.961	0.962



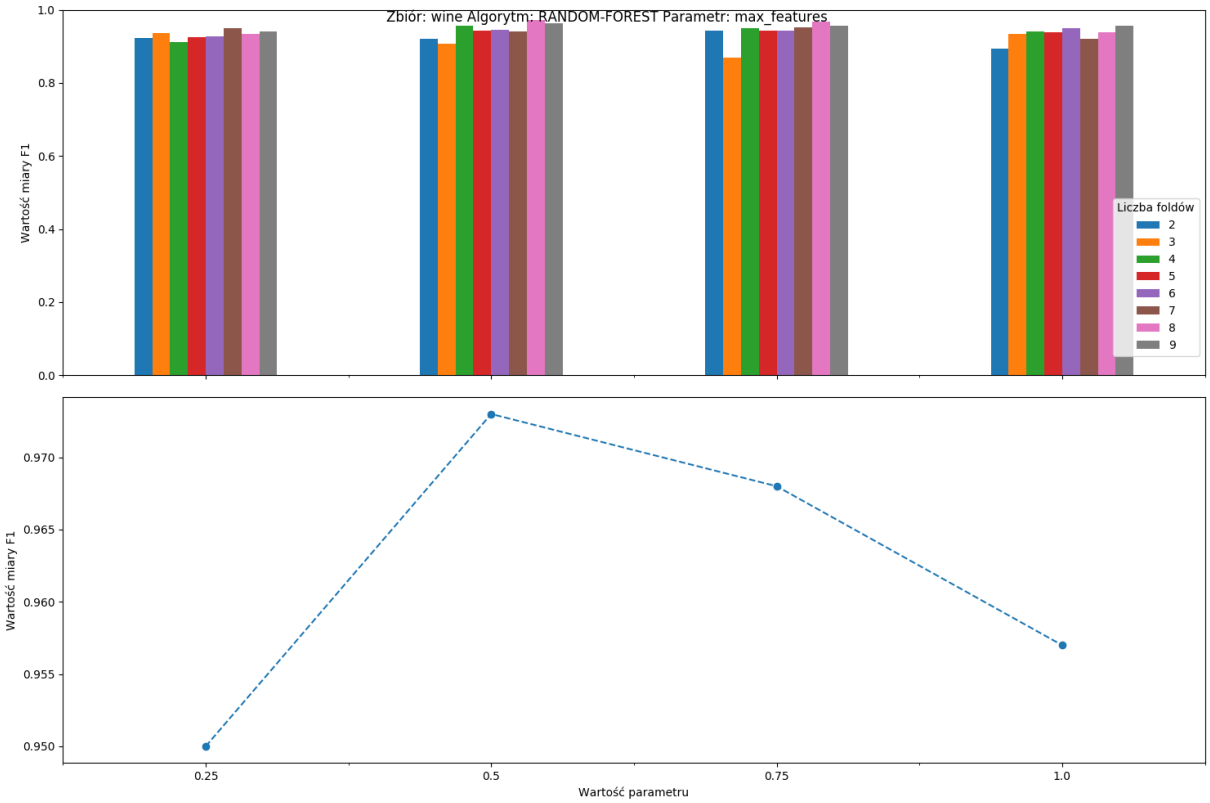
Rysunek 28: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Random-forest" przy ustalonym parametrze "criterion".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
n_estimators	10	0.951	0.913	0.939	0.961	0.968	0.984	0.972	0.979
	25	0.945	0.929	0.960	0.950	0.968	0.956	0.951	0.961
	50	0.951	0.907	0.963	0.972	0.979	0.984	0.972	0.966
	75	0.961	0.919	0.947	0.961	0.967	0.950	0.943	0.966
	99	0.950	0.961	0.972	0.978	0.955	0.956	0.941	0.973



Rysunek 29: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Random-forest" przy ustalonym parametrze "n_estimators".

	{}	Miara F1							
	Liczba foldów	2	3	4	5	6	7	8	9
Parametr	Wartość parametru								
max_features	0.25	0.922	0.937	0.911	0.925	0.928	0.950	0.933	0.940
	0.5	0.920	0.908	0.956	0.944	0.946	0.941	0.973	0.962
	0.75	0.944	0.868	0.950	0.944	0.943	0.952	0.968	0.957
	1.0	0.893	0.934	0.940	0.939	0.949	0.921	0.938	0.957



Rysunek 30: Wykres wartości miary F1 dla zbioru "Wine" algorytmu "Random-forest" przy ustalonym parametrze "max_features".

5 Porównanie klasyfikatorów

Klasyfikator	F1	Komentarz
C4.5	0.82	CV = 6, C3
Bagging	0.64	max_samples = 0.5, CV = 4 (strat.)
Random-Forest	0.64	n_estimators = 10, CV = 7 (strat.)
Adaboost	0.63	learning_rate = 0.0001, CV = 7 (strat.)
Naiwny Bayes	0.63	CV = 6, brak dyskr.
KNN	0.58	CV = 5 (strat.), k = 3, euklides, głos. równ.

Tabela 1: Najlepsze wyniki klasyfikatorów dla zbioru "Diabetes".

Klasyfikator	F1	Komentarz
C4.5	0.77	CV = 8, C1
Random-Forest	0.75	criterion = entropy, CV = 8 (strat.)
Bagging	0.73	bootstrap = False, CV = 5 (strat.)
Adaboost	0.72	learning_rate = 0.01, CV = 8 (strat.)
KNN	0.63	CV = 5 (strat.), k = 1, euklides, głos. równ.
Naiwny Bayes	0.61	CV = 5, CAIM

Tabela 2: Najlepsze wyniki klasyfikatorów dla zbioru "Glass".

Klasyfikator	F1	Komentarz
Adaboost	0.98	learning_rate = 0.1, CV = 7 (strat.)
Bagging	0.98	max_samples = 1.0, CV = 5 (strat.)
Random-Forest	0.98	n_estimators = 10, CV = 7 (strat.)
Naiwny Bayes	0.97	CV = 2, brak dyskr.
C4.5	0.95	CV = 7, C1
KNN	0.72	CV = 5 (strat.), k = 5, manhatt., głos. równ.

Tabela 3: Najlepsze wyniki klasyfikatorów dla zbioru "Wine".