Klasyfikacja ziaren fasoli

MAKSIMOWICZ MARTYNA

WYDZIAŁ INFORMATYKI, POLITECHNIKA BIAŁOSTOCKA

Charakterystyka danych – problem

Zbiór danych „Dry Bean Dataset" pochodzący z repozytorium UCI Machine Learning Repository. (Źródło: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset).

Fasola to jedno z najczęściej produkowanych roślin strączkowych w Turcji. Odgrywa zatem kluczową rolę w rolnictwie. Jednocześnie jest wrażliwa na wpływ zmian klimatycznych.

Odporność i/lub tolerancję roślin na czynniki zewnętrzne można zwiększyć poprzez hodowlę nowych odmian i określenie cech ich nasion.

Jakość nasion ma zdecydowanie wpływ na produkcję roślinną. Dlatego też klasyfikacja nasion ma zasadnicze znaczenie dla powodzenia uprawy (oraz marketingu).

Charakterystyka danych

W Turcji uprawiane nasiona fasoli dzieli się na odmiany, biorąc pod uwagę cechy formy czy kształtu.

Wykonano zdjęcia 13 611 ziaren 7 różnych odmian fasoli o podobnych cechach za pomocą kamery o wysokiej rozdzielczości. Obrazy poddano etapom segmentacji i ekstrakcji cech.

Automatyczna identyfikacja ziaren fasoli miała pomoc rolnikom w procesie technologicznym, ponieważ ręczna klasyfikacja i sortowanie nasion to proces trudny, bardzo czasochłonny i o niskiej efektywności, zwłaszcza przy dużych nakładach produkcyjnych.

Charakterystyka danych – klasy

Badane odmiany fasoli:

- Seker – 2027 obiektów,
- Barbunya – 1322 obiektów,
- Bombay – 522 obiektów,
- Cali – 1630 obiektów,
- Dermason – 3546 obiektów,
- Horoz – 1928 obiektów,
- Sira – 2636 obiektów.

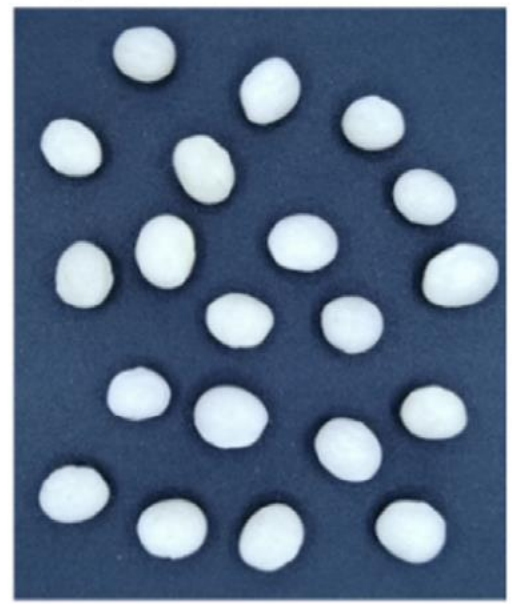
Barbunya

Bombay

Cali

Dermason

Horoz



Seker

Sira

Źródło: Multiclass classification of dry beans using computer vision and machine learning techniques, Murat Koklu, Ilker Ali Ozkan, Department of Computer Engineering, Selcuk University, Turkey, Konya, Turkey

Charakterystyka danych – cechy

Cechy wyodrębnione ze zdjęć:

- Area A(Pole) – obszar strefy fasoli i liczba pikseli w jej granicach
- Perimeter P (Obwód) – obwód ziarna fasoli (długość granicy)
- Major Axis Length L (Długość osi głównej) – odległość między końcami najdłuższej linii, którą można wyciągnąć z fasoli
- Minor Axis Length l (Długość osi mniejszej) – najdłuższy odcinek, prostopadły do osi głównej
- Aspect Ratio K (Współczynnik proporcji) – proporcja długości osi głównej do osi mniejszej $K = \frac{L}{l}$
- Eccentricity Ec (Mimośród) – mimośród elipsy (stosunek długości ogniskowej do długości półosi wielkiej)
- Convex Area C (Obszar wypukły) – liczba pikseli w najmniejszym wielokącie wypukłym, w którym zawarte jest ziarno
- Equivalent Diameter Ed (Średnica ekwiwalentna) – średnica koła o tej samej powierzchni, co obszar ziarna: $d = \sqrt{\frac{4A}{\pi}}$
- Extent Ex (Zakres) – stosunek pikseli w obwiedni (minimalny prostokąt ograniczający) do powierzchni ziarna

Charakterystyka danych – cechy

- Solidity S (Solidność) – wypukłość, stosunek pikseli w wypukłej łupinie do pikseli znajdujących się w całym ziarnie: $S = \frac{A}{C}$
- Roundness R (Zaokrąglenie) – zaokrąglenie obiektu obliczane według następującego wzoru: $R = \frac{4 \cdot \pi \cdot A}{P^2}$
- Compactness CO (Kompaktowość) – kompaktowość mierzy okrągłość obiektu według wzoru: $CO = \frac{Ed}{L}$
- Shape Factors (Współczynniki kształtu) – bezwymiarowe wielkości używane w analizie obrazu, które liczbowo opisują kształt obiektu, niezależnie od jego wymiarów. Współczynniki kształtu są obliczane na podstawie zmierzonych wymiarów, takich jak średnica, długość cięciwy, powierzchnia, obwód, środek ciężkości, momenty itp. Znormalizowane wielkości reprezentują stopień odchylenia od idealnego kształtu, takiego jak okrąg, kula lub wielościan równoboczny
  - Shape Factor 1: $SF1 = \frac{L}{A}$
  - Shape Factor 2: $SF2 = \frac{l}{A}$
  - Shape Factor 3: $SF3 = \frac{A}{\left(\frac{L}{2}\right)^2 \cdot \pi}$
  - Shape Factor 4: $SF4 = \frac{A}{\frac{L}{2} \cdot \frac{l}{2} \cdot \pi}$

Proces analizy zbioru danych

- Wstępna analiza danych
  - Przegląd próbki danych
  - Wstępna wizualizacja zmiennych z uwzględnieniem klas
- Podział danych na zbiory: treningowy i testowy w proporcji 75% i 25%
- Standaryzacja cech
- Wybór klasyfikatorów i ustalenie optymalnych parametrów
- Wytrenowanie modeli
- Predykcja wartości
- Zbadanie jakości klasyfikacji

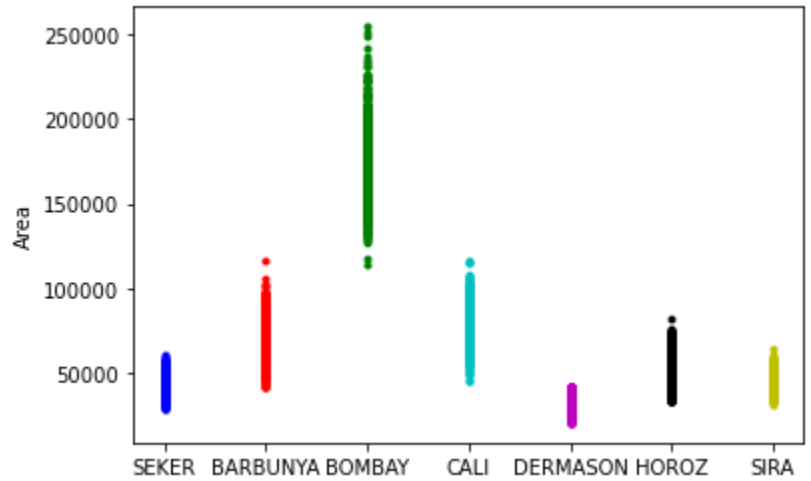
Wstępna analiza danych

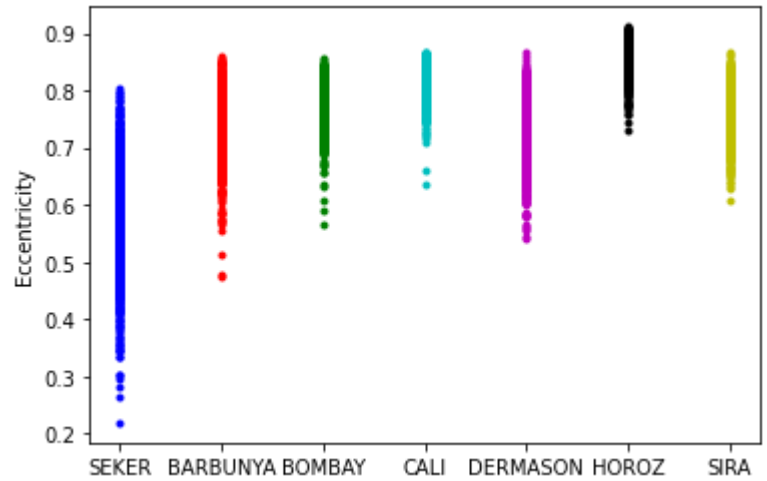
Próbka danych

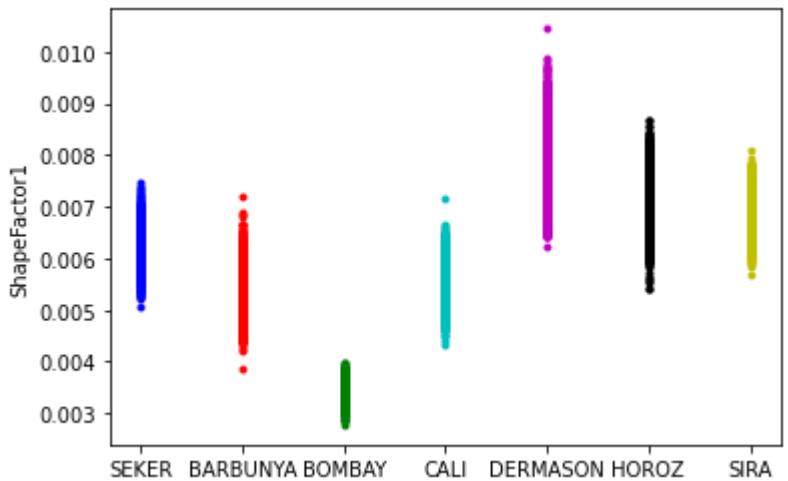
Area	Perimeter	Major Axis Length	Minor Axis Length	Aspect Ration	Eccentricity	Convex Area	Equiv Diameter	Extent	Solidity	Roundness	Compactness	Shape Factor 1	Shape Factor 2	Shape Factor 3	Shape Factor 4	Class
37564	736,706	236,4602	202,526	1,167555	0,516162	38184	218,6961	0,766299	0,983763	0,869748	0,924875	0,006295	0,002841	0,855393	0,998718	**SEKER**
155524	1523,825	559,7563	360,263	1,553744	0,765356	159126	444,9936	0,817506	0,977364	0,841661	0,794977	0,003599	0,000887	0,631989	0,981949	**BOMBAY**
56334	989,798	372,7248	193,539	1,925838	0,85462	57684	267,8184	0,674497	0,976597	0,722582	0,718542	0,006616	0,001088	0,516302	0,994315	**HOROZ**
47605	877,296	373,1439	163,5929	2,28093	0,898771	48131	246,1962	0,630396	0,989071	0,777267	0,659789	0,007838	0,000916	0,435321	0,992938	**HOROZ**
70563	1014,222	393,4801	229,3723	1,715465	0,812521	71437	299,7392	0,717352	0,987765	0,862027	0,761765	0,005576	0,001158	0,580285	0,995459	**CALI**
40526	760,728	287,5861	180,4546	1,593676	0,778632	41067	227,1548	0,68433	0,986826	0,880004	0,789867	0,007096	0,001704	0,62389	0,994279	**SIRA**
46815	822,48	322,0585	185,8291	1,73309	0,816741	47410	244,1449	0,698585	0,98745	0,869649	0,758076	0,006879	0,001401	0,57468	0,995971	**SIRA**
77007	1069,231	411,5527	240,3708	1,712158	0,811712	78077	313,1267	0,660381	0,986296	0,846442	0,760842	0,005344	0,001105	0,578881	0,991136	**CALI**
37832	720,476	263,0345	183,385	1,43433	0,716887	38289	219,4748	0,725168	0,988064	0,915862	0,834396	0,006953	0,002079	0,696216	0,998603	**DERMASON**
85890	1152,016	417,5364	262,7196	1,589285	0,777232	87188	330,694	0,714654	0,985113	0,813271	0,792012	0,004861	0,00118	0,627284	0,996933	**BARBUNYA**
51131	842,796	316,1856	207,029	1,527253	0,755828	51654	255,151	0,812506	0,989875	0,904585	0,806966	0,006184	0,001618	0,651194	0,994537	**SIRA**
27884	630,303	239,4054	148,4848	1,612322	0,784425	28196	188,4224	0,758253	0,988935	0,881995	0,787043	0,008586	0,002032	0,619437	0,998732	**DERMASON**
70344	1037,985	378,6511	237,9098	1,591574	0,777964	71521	299,2737	0,821354	0,983543	0,820455	0,790368	0,005383	0,001296	0,624682	0,994227	**BARBUNYA**

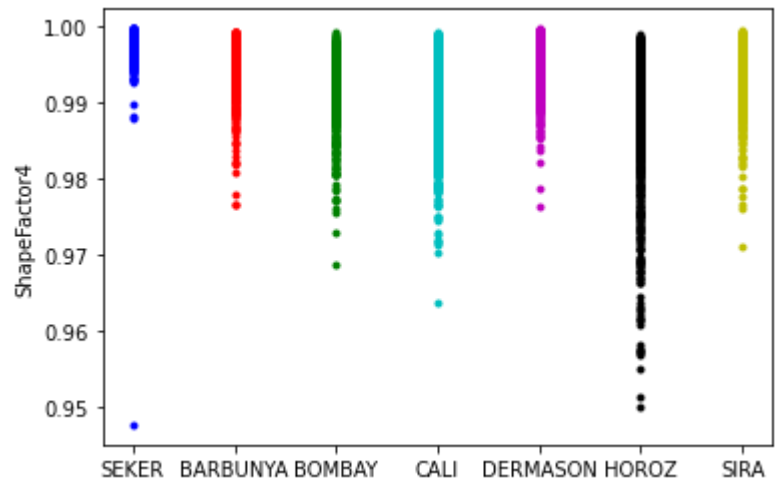
Liczebności poszczególnych klas

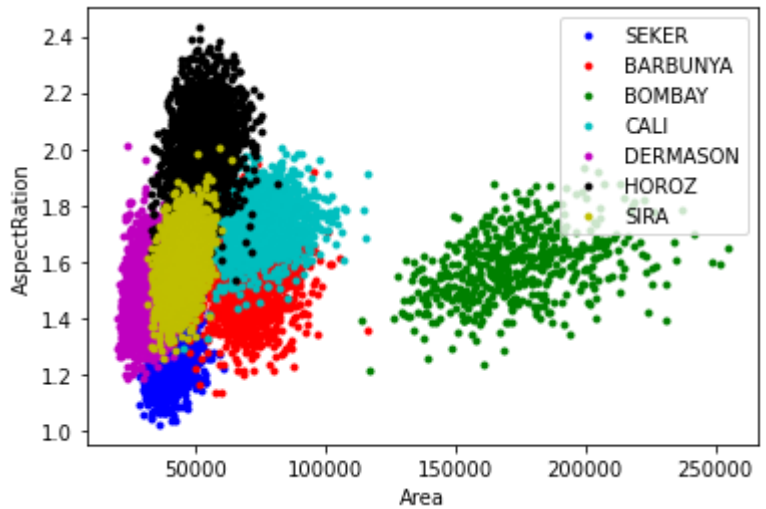
Odmiana	Liczebność
BARBUNYA	1322
BOMBAY	522
CALI	1630
DERMASON	3546
HOROZ	1928
SEKER	2027
SIRA	2636

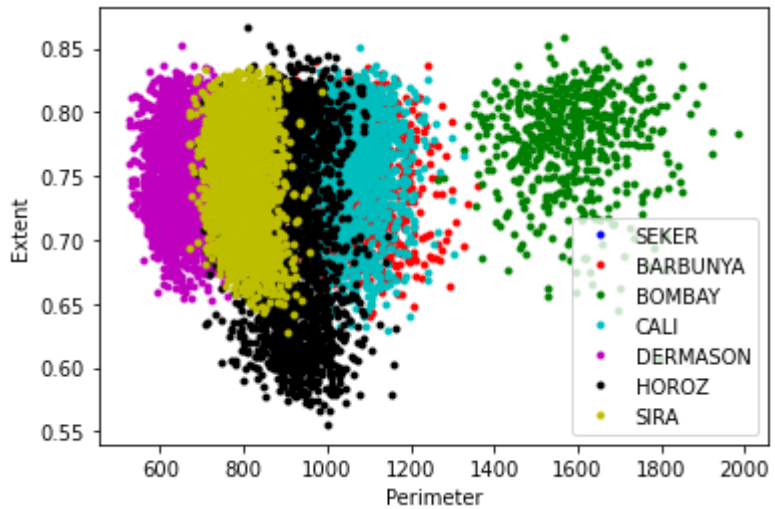

Area
250000
200000
150000
100000
50000
SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA

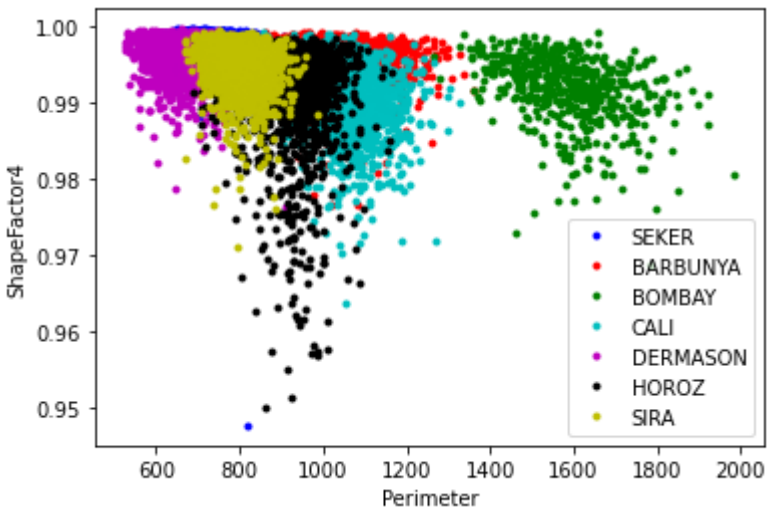

Eccentricity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA

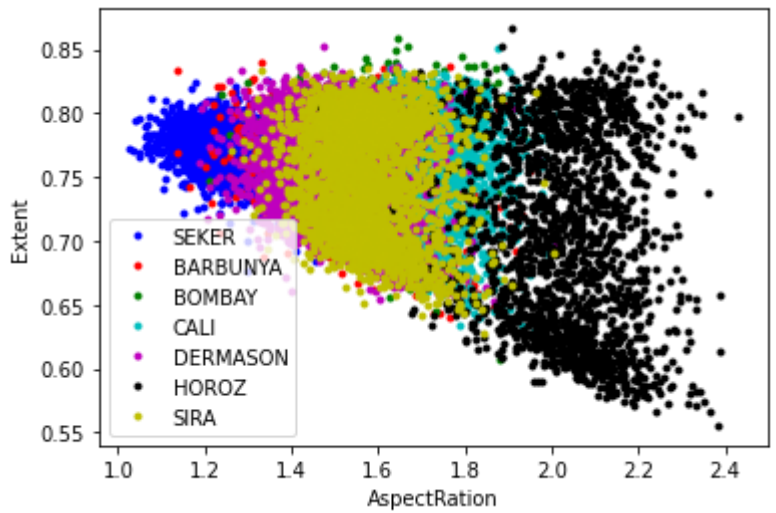

ShapeFactor1
0.010
0.009
0.008
0.007
0.006
0.005
0.004
0.003
SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA


ShapeFactor4
1.00
0.99
0.98
0.97
0.96
0.95
SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA


SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA
AspectRation
Area
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
50000
100000
150000
200000
250000


SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA
Extent
Perimeter
0.85
0.80
0.75
0.70
0.65
0.60
0.55
600
800
1000
1200
1400
1600
1800
2000

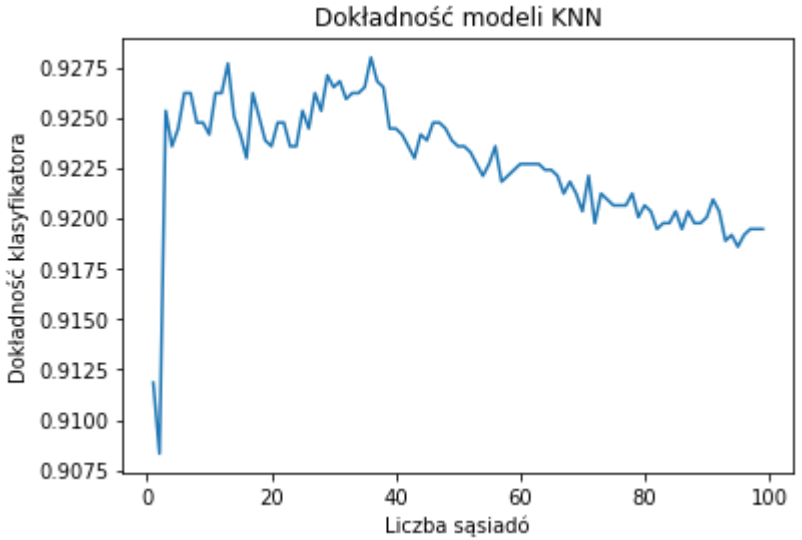

SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA
ShapeFactor4
Perimeter
1.00
0.99
0.98
0.97
0.96
0.95
600
800
1000
1200
1400
1600
1800
2000


SEKER
BARBUNYA
BOMBAY
CALI
DERMASON
HOROZ
SIRA
Extent
AspectRation
0.85
0.80
0.75
0.70
0.65
0.60
0.55
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4

Wybór klasyfikatorów

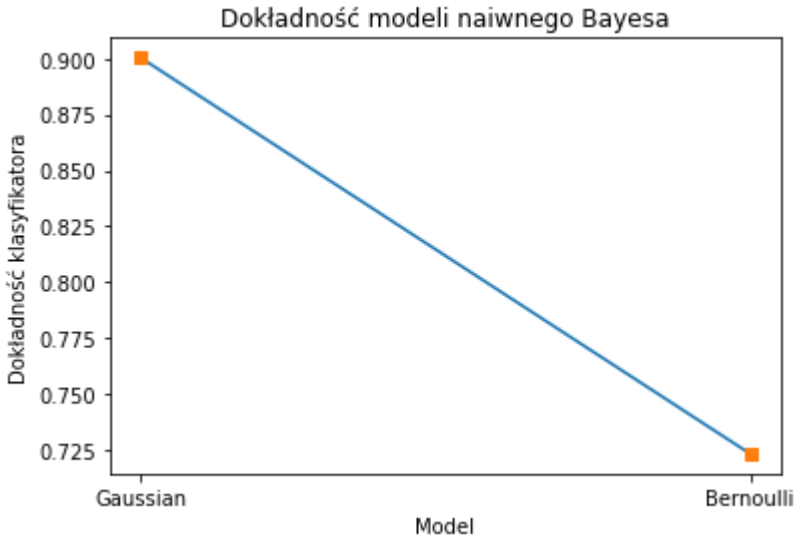
- Metoda K najbliższych sąsiadów
- Naiwny klasyfikator bayesowski
- Metoda wektorów nośnych (SVM, SVC)
- Regresja logistyczna
- Drzewo decyzyjne

Metoda K najbliższych sąsiadów



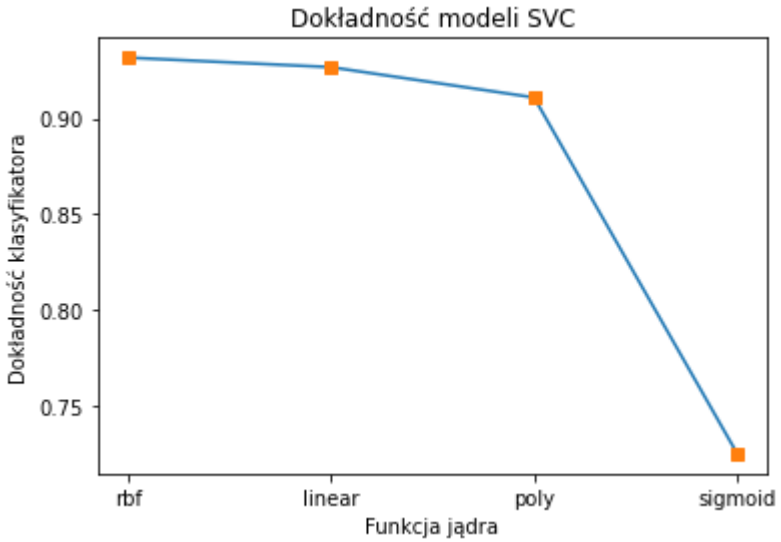
liczba sąsiadów	Dokładność
36	0,928005
13	0,927711
29	0,927123
37	0,926829
31	0,926829
…	…
96	0,919189
93	0,918895
95	0,918601
1	0,911842
2	0,908316

Naiwny klasyfikator bayesowski



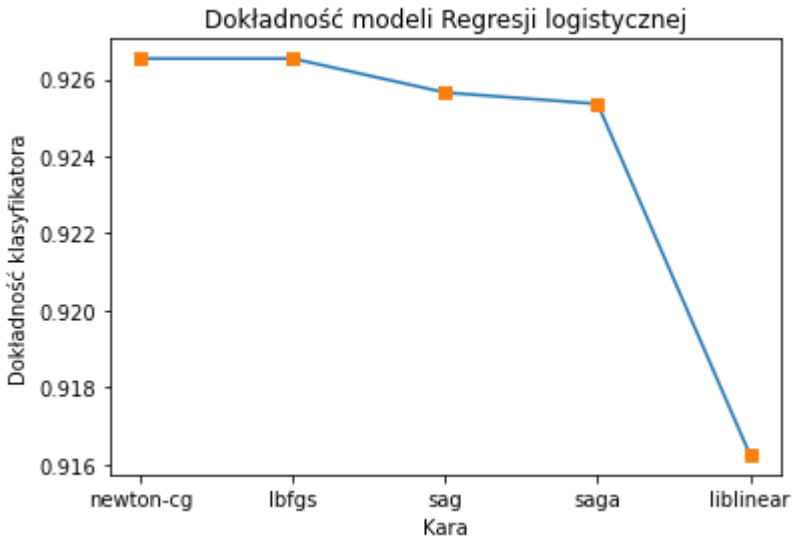
model	Dokładność
Gausowski	0,900382
Bernoulliego	0,723479

Metoda wektorów nośnych (SVC)



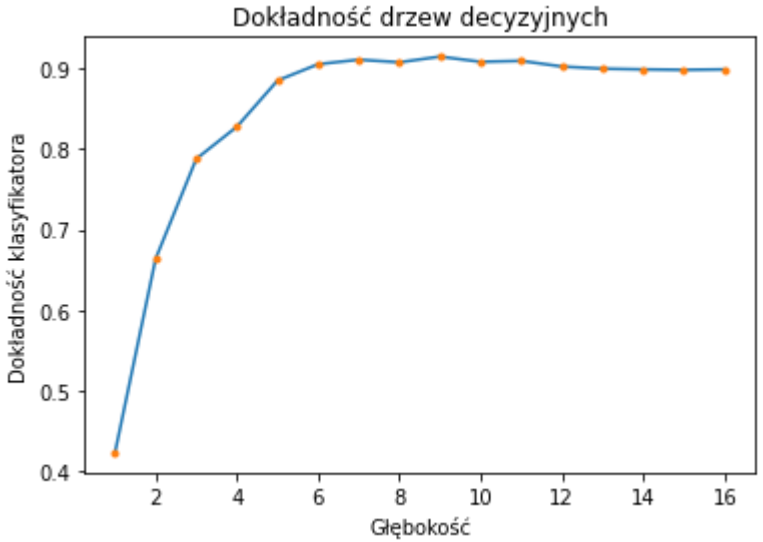
Funkcja jądra	Dokładność
radialna funkcja bazowa	0,931825
funkcja liniowa	0,926829
funkcja wielomianowa	0,910961
funkcja sigmoidalna	0,724655

Klasyfikator regresji logistycznej



Algorytm	Dokładność
Newton-CG	0,926535
L-BFGS	0,926535
SAG	0,925654
SAGA	0,92536
LIB LINEAR	0,91625

Drzewo decyzyjne



głębokość	dokładność
9	**0,914781**
7	0,910961
11	0,909492
10	0,908316
8	0,907728
6	0,905378
12	0,902439
13	0,899794
14	0,898913
16	0,898913
15	0,898325
5	0,885101
4	0,827799
3	0,787834
2	0,663826
1	0,423156

Ostatecznie wybrane modele

Model	Parametry
KNN	k=36 sąsiadów
Naiwny klasyfikator bayesowski	Gausowski
Metoda wektorów nośnych (SVC)	radialna funkcja bazowa
Klasyfikator regresji logistycznej	algorytm Newton-CG
Drzewo decyzyjne	głębokość d=9

Sprawdzian krzyżowy

Model	2	3	4	5	6	7	8	9	10
KNN	0,78	0,778	0,7778	0,77782	0,777825	**0,777825**	0,777825	0,777825	0,777825
Gaussowski NB	0,7	0,7	0,7005	0,70046	0,700457	**0,700457**	0,700457	0,700457	0,700457
SVC	0,69	0,692	0,6921	0,69208	0,692085	**0,692085**	0,692085	0,692085	0,692085
Regresja logistyczna	0,7	0,696	0,6963	0,69627	0,696273	**0,696273**	0,696273	0,696273	0,696273
Drzewo decyzyjne	0,34	0,337	0,3372	0,33723	0,337225	**0,337226**	0,337225	0,337225	0,337225

Wytrenowanie modeli.

Predykcja wartości.

Macierze błędów

KNN_n=36 - macierz błędów:

```
[[276   0  25   0   2   5  11]
 [  0 111   0   0   0   0   0]
 [  7   0 402   0   6   1   3]
 [  1   0   0 847   1  14  42]
 [  0   0  11   2 469   0   9]
 [  2   0   0   2   0 446  20]
 [  1   0   1  69   7   3 607]]
```

Gaussian_NB - macierz błędów:

```
[[259   0  40   0   2   3  15]
 [  0 111   0   0   0   0   0]
 [ 35   0 377   0   5   1   1]
 [  0   0   0 808   2  19  76]
 [  0   0  10   4 469   0   8]
 [  2   0   0   3   0 442  23]
 [  4   0   1  57  18  10 598]]
```

SVC_rbf - macierz błędów:

```
[[287   0  20   0   1   3   8]
 [  0 111   0   0   0   0   0]
 [  9   0 401   0   5   1   3]
 [  0   0   0 849   1  11  44]
 [  1   0   9   3 469   0   9]
 [  1   0   0   5   0 448  16]
 [  1   0   1  69   6   5 606]]
```

Log_Regression - macierz błędów:

```
[[282   0  21   0   1   4  11]
 [  0 111   0   0   0   0   0]
 [ 10   0 396   0   7   1   5]
 [  1   0   0 841   3  10  50]
 [  1   0   6   3 472   0   9]
 [  2   0   0   2   0 448  18]
 [  0   0   3  62  15   5 603]]
```

Decision_Tree - macierz błędów:

```
[[277   0  20   0   3   6  13]
 [  1 110   0   0   0   0   0]
 [ 19   0 387   0   8   1   4]
 [  0   0   0 842   1  13  49]
 [  0   0  12   3 456   0  20]
 [  2   0   0  14   0 436  18]
 [  1   0   0  71   6   5 605]]
```

BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ, SEKER, SIRA

Macierze błędów – podsumowanie

Model	Liczba poprawnych klasyfikacji	Liczba niepoprawnych klasyfikacji	Udział poprawnych klasyfikacji [%]
KNN	3158	245	92.80047017337644
Gausowski NB	3064	339	90.03820158683514
SVC	3171	232	93.18248604172788
Regresja logistyczna	3153	250	92.65354099324125
Drzewo decyzyjne	3113	290	91.47810755215986

Raporty klasyfikacji

	KNN	Gausowski NB	SVC	Regresja logistyczna	Drzewo decyzyjne
			PRECYZJA		
BARBUNYA	0.96	0.86	0.96	0.95	0.92
BOMBAY	1.00	1.00	1.00	1.00	1.00
CALI	0.92	0.88	0.93	0.93	0.92
DERMASON	0.92	0.93	0.92	0.93	0.91
HOROZ	0.97	0.95	0.97	0.95	0.96
SEKER	0.95	0.93	0.96	0.96	0.95
SIRA	0.88	0.83	0.88	0.87	0.85
			CZUŁOŚĆ		
BARBUNYA	0.87	0.81	0.90	0.88	0.87
BOMBAY	1.00	1.00	1.00	1.00	0.99
CALI	0.96	0.90	0.96	0.95	0.92
DERMASON	0.94	0.89	0.94	0.93	0.93
HOROZ	0.96	0.96	0.96	0.96	0.93
SEKER	0.95	0.94	0.95	0.95	0.93
SIRA	0.88	0.87	0.88	0.88	0.88

Dokładność modeli

Model	Dokładność
SVC	0.931825
KNN	0.928005
Regresja logistyczna	0.926535
Drzewo decyzyjne	0.914781
Gausowski NB	0.900382

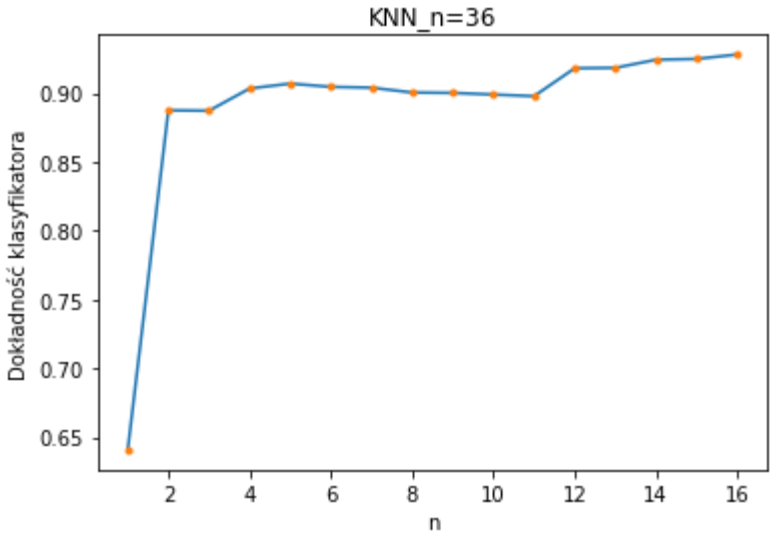
Wybór najistotniejszych cech

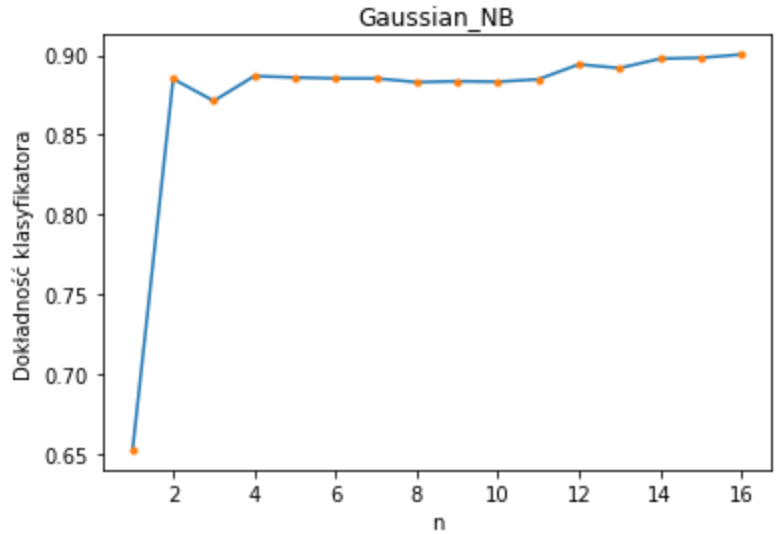
Badanie dokładności modeli w zależności od liczby wykorzystanych cech

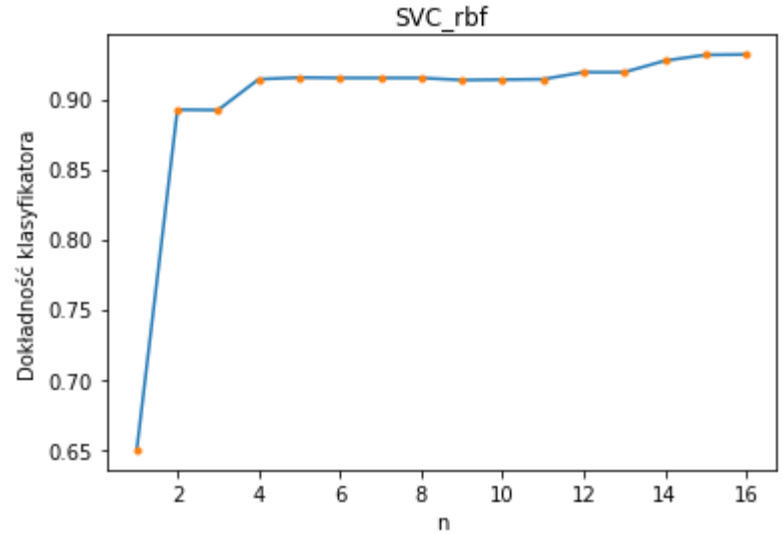
Początkowy indeks	Cecha	Istotność
1	Perimeter	0.096522
14	ShapeFactor3	0.094545
11	Compactness	0.094201
12	ShapeFactor1	0.090030
3	MinorAxisLength	0.081913
2	MajorAxisLength	0.077354
6	ConvexArea	0.071767
5	Eccentricity	0.061753
4	AspectRation	0.060185
7	EquivDiameter	0.057724
0	Area	0.055776
10	roundness	0.055103
13	ShapeFactor2	0.041672
15	ShapeFactor4	0.031570
9	Solidity	0.018447
8	Extent	0.011438

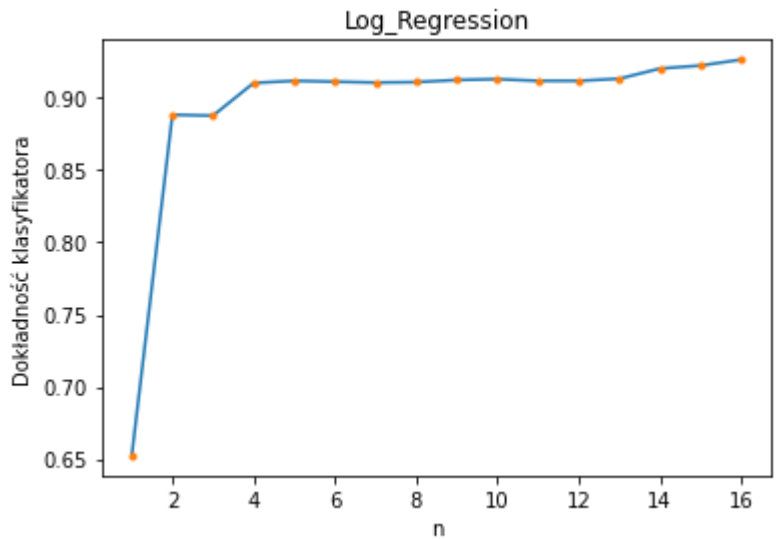
Dokładność modelu w zależności od liczby wykorzystanych cech

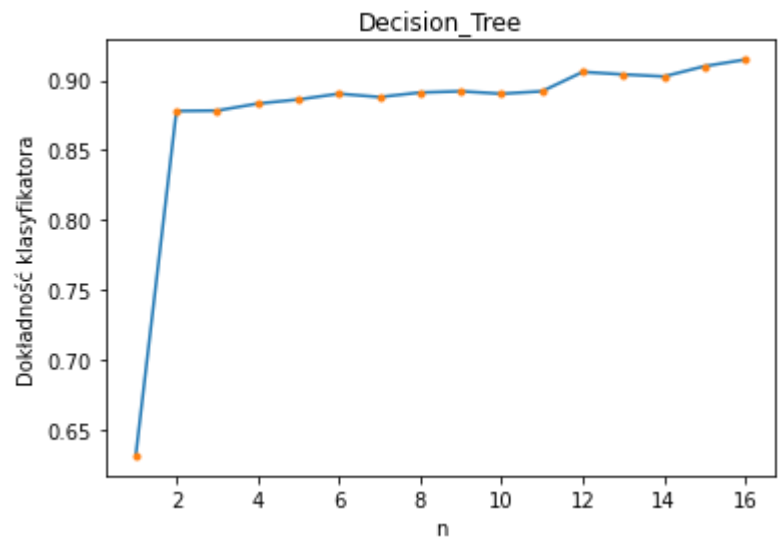
Model	1 cecha	2 cechy	3 cechy	4 cechy	5 cech	6 cech	7 cech	8 cech	9 cech	10 cech	11 cech	12 cech	13 cech	14 cech	15 cech	16 cech
KNN	0,640905	0,887452	0,887158	0,903321	0,906847	0,904496	0,903908	0,900382	0,900088	0,898913	0,897737	0,918014	0,918307	0,924185	0,924772	**0,928005**
Gausowski NB	0,652659	0,885101	0,87129	0,886865	0,885983	0,885395	0,885395	0,883044	0,883632	0,883338	0,884808	0,894211	0,89186	0,897737	0,898325	**0,900382**
SVC	0,649721	0,892448	0,892154	0,914193	0,915369	0,915075	0,915075	0,915075	0,913606	0,9139	0,914193	0,919189	0,919189	0,927417	0,931531	**0,931825**
Regresja logistyczna	0,652366	0,888334	0,887746	0,910373	0,911842	0,911255	0,910667	0,910961	0,91243	0,913018	0,911842	0,911842	0,913312	0,920364	0,922421	**0,926535**
Drzewo decyzyjne	0,631795	0,878049	0,878343	0,883338	0,886277	0,890391	0,88804	0,891272	0,892154	0,890391	0,892154	0,905965	0,904202	0,902733	0,910079	**0,914781**


KNN_n=36
Dokładność klasyfikatora
n


Gaussian_NB
Dokładność klasyfikatora
n


SVC_rbf
Dokładność klasyfikatora
n


Log_Regression
Dokładność klasyfikatora
n


Decision_Tree
Dokładność klasyfikatora
n

Maksymalna dokładność modelu

Model	Maksymalna dokładność	Liczba wykorzystanych cech	Kroswalidacja dla 7 podzbiorów
SVC	0,931825	16	0,692085
KNN	0,928005	16	0,777825
Regresja logistyczna	0,926535	16	0,696273
Drzewo decyzyjne	0,914781	16	0,337226
Gaussowski NB	0,900382	16	0,700457

Wnioski – jakość klasyfikatorów

- Najlepszy okazał się model wektorów nośnych, który uzyskał dokładność klasyfikacji na poziomie 93,18%.
- Wszystkie modele uzyskały dokładność ponad 90%.
- Sprawdzian krzyżowy oszacował dokładność modeli znacznie gorzej niż faktycznie uzyskane wyniki. Kroswalidacja wskazała model K najbliższych sąsiadów jako najskuteczniejszy klasyfikator ze średnią dokładnością na poziomie 77,78%.
- Sprawdzian krzyżowy dla modelu drzewa decyzyjnego wskazał dokładność na poziomie zaledwie 33,72%.

Podsumowując, najlepsze modele to model SVC i model KNN. Zwracają poprawne wyniki z wysoką dokładnością. Jednakże żaden klasyfikator nie przekroczył 94% poprawności.

Wnioski – klasyfikacja odmian

- Odmiana Bombay została zaklasyfikowana poprawnie w 100% przez niemal wszystkie algorytmy (pojedynczy błąd dla modelu drzewa decyzyjnego)
- Wszystkie modele najczęściej popełniały błąd przy klasyfikacji odmiany Sira. Co ciekawe, każdy algorytm najwięcej niepoprawnych klasyfikacji tej odmiany dokonał poprzez uznanie ziarna jako odmiana Dermason. Druga najliczniejsza grupa błędnych klasyfikacji dotyczyła odmiany Horoz.
- Dodatkowo, odmiana Dermason najczęściej błędnie była klasyfikowana jako odmiana Sira. Widać zatem, że odmiany te są bardzo zbliżone, biorąc pod uwagę cechy formy czy kształtu.
- Z kolei odmiana Horoz najczęściej błędnie była klasyfikowana jako odmiana Cali lub Sira.

Ogólne wyniki są zadowalające. Tego typu klasyfikator mógłby być wykorzystany w przemyśle. Automatyczna klasyfikacja ziaren byłaby optymalizacją procesu „ręcznej" identyfikacji, która jest czaso- i pracochłonna.

Dziękuję za uwagę!

MAKSIMOWICZ MARTYNA

WYDZIAŁ INFORMATYKI, POLITECHNIKA BIAŁOSTOCKA