Univerzitet u Kragujevcu
Fakultet inženjerskih nauka

Seminarski rad iz predmeta
Veštačka inteligencija

Tema:
Klasifikacija Leaf skupa podataka

Student:
Aleksandra Milosević 571/2016

Predmetni profesor:
Dr. Vesna Ranković, redovni profesor

Predmetni saradnik:
Tijana Šušteršič

Kragujevac, 2020.

Sadrzaj:

1. Postavka problema

Problem predstavlja kreiranje modela koji klasifikuje razlicite vrste biljaka. Ova baza podataka sadrzi 40 razlicitih vrsta biljaka. Skup podataka se sastoji od uzoraka svake vrste. Od svakog uzorka merene su odredjene karakteristike:

1. Class (Species) – predvidjamo klasu na osnovu atributa dole (2-16.)
2. Specimen Number (broj uzoraka)
3. Eccentricity (ekscentricnost)
4. Aspect Ratio (proporcija)
5. Elongation (izduzenje)
6. Solidity (cvrstoca)
7. Stochastic Convexity (stohasticka konveksnost)
8. Isoperimetric Factor (izoperimetrijski faktor)
9. Maximal Indentation Depth (maksimalna dubina zasecanja)
10. Lobedness (lisnatost)
11. Average Intensity (prosecni intenzitet)
12. Average Contrast (prosecni kontrast)
13. Smoothness (glatkoca)
14. Third moment (treci trenutak)
15. Uniformity (slicnost)
16. Entropy (entropija)

Zadatak je kreirati model koji na osnovu ovih karakteristika razlikuje kojoj vrsti biljka od porodice biljaka pripada. U ovom slucaju za model je koriscena viseslojna neuronska mreza koja je implementirana u programskom jeziku Python 3.5.2.

2. Opis i vizualizacija podataka

Ulazne podatke cine atributi 2-16, dok izlazni podatak predstavlja klasu kojoj pripada uzorak sa tim konkretnim atributima(1.). Uzorak moze da pripada jednoj od 40 klasa. Podaci se nalaze u fajlu leaf.csv. Prvu kolonu cine izlazni podaci broj klase, drugu cini broj uzoraka, trecu ekscentricnost, cetvrtu proporcija, petu izduzenje, sestu cvrstoca, sedmu stohasticka konveksnost, osmu izoperimetrijski faktor, devetu maksimalna dubina zasecanja, desetu lisnasnost, jedanaestu prosecni intenzitet, dvanaestu prosecni kontrast, trinaestu glatkoca, cetrnaestu treci trenutak, petnaestu slicnost, a sesnaestu entropija. Prvu kolonu cine izlazni podaci, a ostale kolone ulazni podaci. Svaki uzorak ima kompletne podatke.

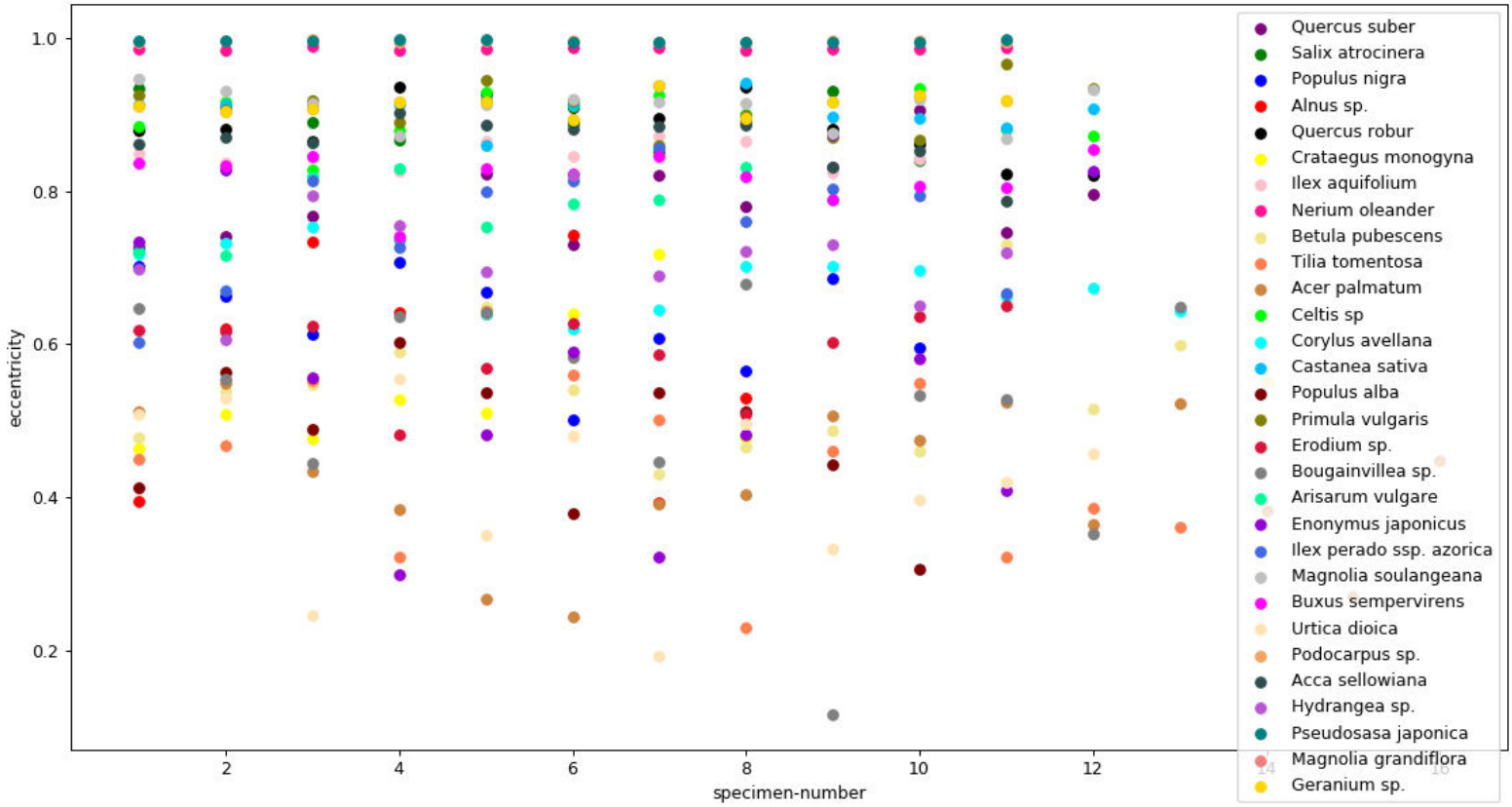
1	1	1	0.72694	1.4742	0.32396	0.98535	1	0.83592	0.0046566	0.0039465	0.04779	0.12795	0.016108	0.0052323	0.00027477	1.1756
2	1	2	0.74173	1.5257	0.36116	0.98152	0.99825	0.79867	0.0052423	0.0050016	0.02416	0.090476	0.0081195	0.002708	7.4846E-05	0.69659
3	1	3	0.76722	1.5725	0.38998	0.97755	1	0.80812	0.0074573	0.010121	0.011897	0.057445	0.0032891	0.00092068	3.7886E-05	0.44348
4	1	4	0.73797	1.4597	0.35376	0.97566	1	0.81697	0.0068768	0.0086068	0.01595	0.065491	0.0042707	0.0011544	6.6272E-05	0.58785
5	1	5	0.82301	1.7707	0.44462	0.97698	1	0.75493	0.007428	0.010042	0.0079379	0.045339	0.0020514	0.00055986	2.3504E-05	0.34214
6	1	6	0.72997	1.4892	0.34284	0.98755	1	0.84482	0.0049451	0.0044506	0.010487	0.058528	0.0034138	0.0011248	2.4798E-05	0.34068
7	1	7	0.82063	1.7529	0.44458	0.97964	0.99649	0.7677	0.0059279	0.0063954	0.018375	0.080587	0.0064523	0.0022713	4.1495E-05	0.53904
8	1	8	0.77982	1.6215	0.39222	0.98512	0.99825	0.80816	0.0050987	0.0047314	0.024875	0.089686	0.0079794	0.0024664	0.00014676	0.66975
9	1	9	0.83089	1.8199	0.45693	0.9824	1	0.77106	0.0060055	0.006564	0.0072447	0.040616	0.0016469	0.00038812	3.2863E-05	0.33696
10	1	10	0.90631	2.3906	0.58336	0.97683	0.99825	0.66419	0.0084019	0.012848	0.0070096	0.042347	0.0017901	0.00045889	2.8251E-05	0.28082

Slika 1 – Prvih 10 redova fajla leaf.csv

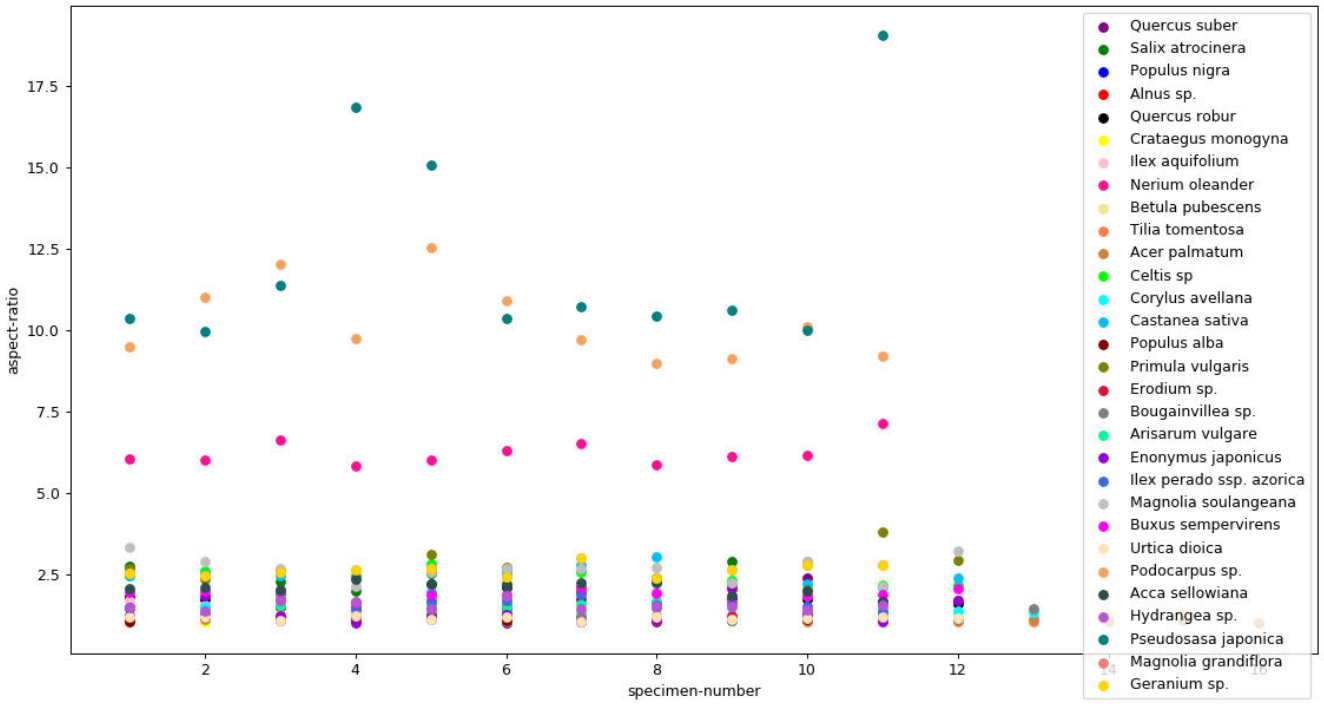
Class	Scientific Name	#	Class	Scientific Name	#
1	Quercus suber	12	21	Fraxinus sp.	10
2	Salix atrocinera	10	22	Primula vulgaris	12
3	Populus nigra	10	23	Erodium sp.	11
4	Alnus sp.	8	24	Bougainvillea sp.	13
5	Quercus robur	12	25	Arisarum vulgare	9
6	Crataegus monogyna	8	26	Euonymus japonicus	12
7	Ilex aquifolium	10	27	Ilex perado ssp. azorica	11
8	Nerium oleander	11	28	Magnolia soulangeana	12
9	Betula pubescens	14	29	Buxus sempervirens	12
10	Tilia tomentosa	13	30	Urtica dioica	12
11	Acer palmatum	16	31	Podocarpus sp.	11
12	Celtis sp.	12	32	Acca sellowiana	11
13	Corylus avellana	13	33	Hydrangea sp.	11
14	Castanea sativa	12	34	Pseudosasa japonica	11
15	Populus alba	10	35	Magnolia grandiflora	11
16	Acer negundo	10	36	Geranium sp.	10
17	Taxus bacatta	5	37	Aesculus californica	10
18	Papaver sp.	12	38	Chelidonium majus	10
19	Polypolium vulgare	13	39	Schinus terebinthifolius	10
20	Pinus sp.	12	40	Fragaria vesca	11

Slika 2 – Broj uzoraka po klasi

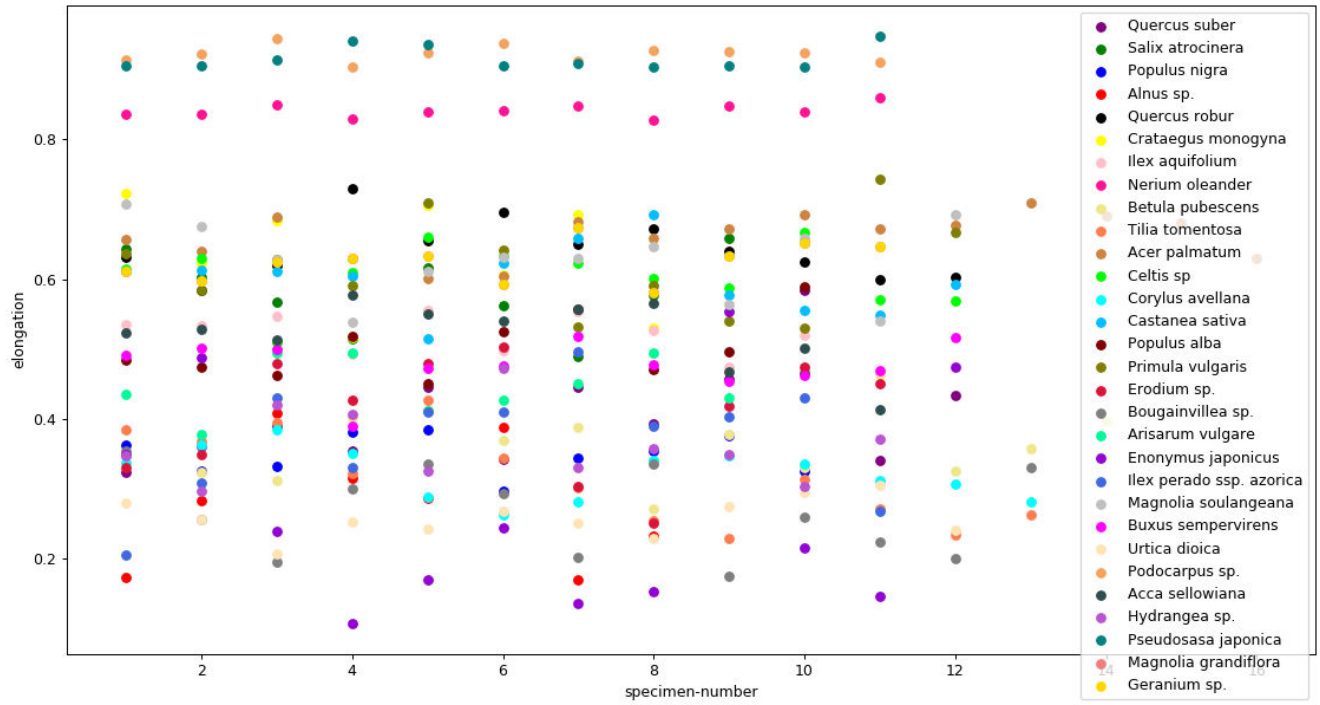
Vizualizacija podataka kombinovanjem ulaznih atributa prikazana je na sledecim slikama. Ukupno imamo 105 kombinacija, kombinovanjem svaka dva ulazna atributa. Ovde imamo prikazane primere tj. kombinacije prvog I drugog ulaznog atributa sa svakim atributom, da ne bismo prikazivali svih 105 kombinacija (svaki ulazni atribut sa svakim ulaznim atributom).



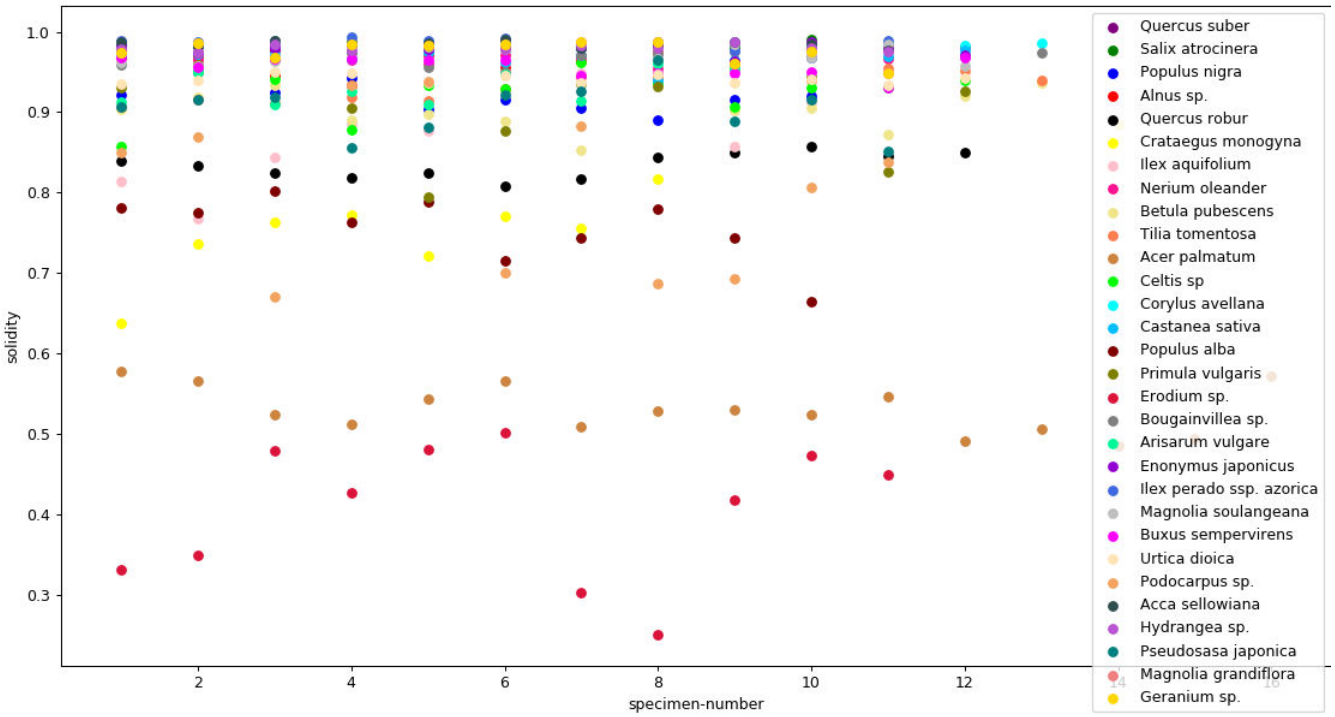
Slika 3 – Figure 1 (specimen-number I eccentricity)



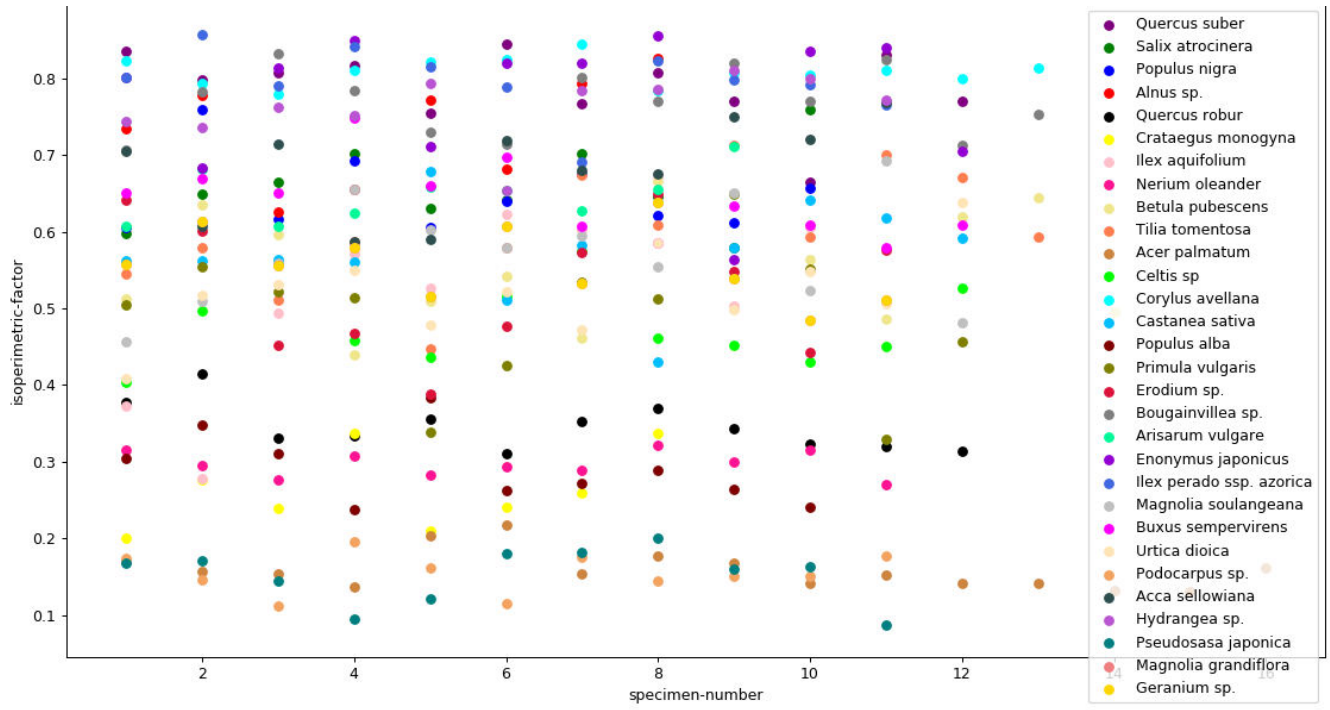
Slika 4 – Figure 2 (specimen-number I aspect-ratio)



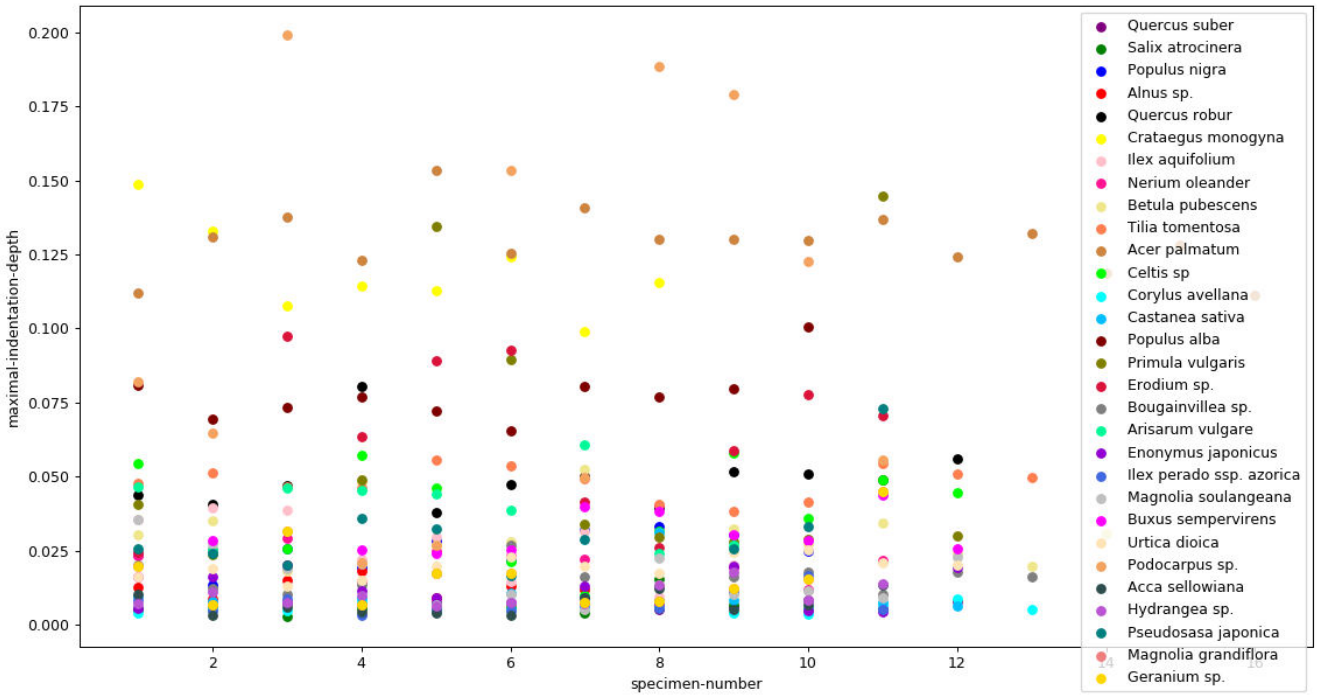
Slika 5 – Figure 3 (specimen-number I elongation)



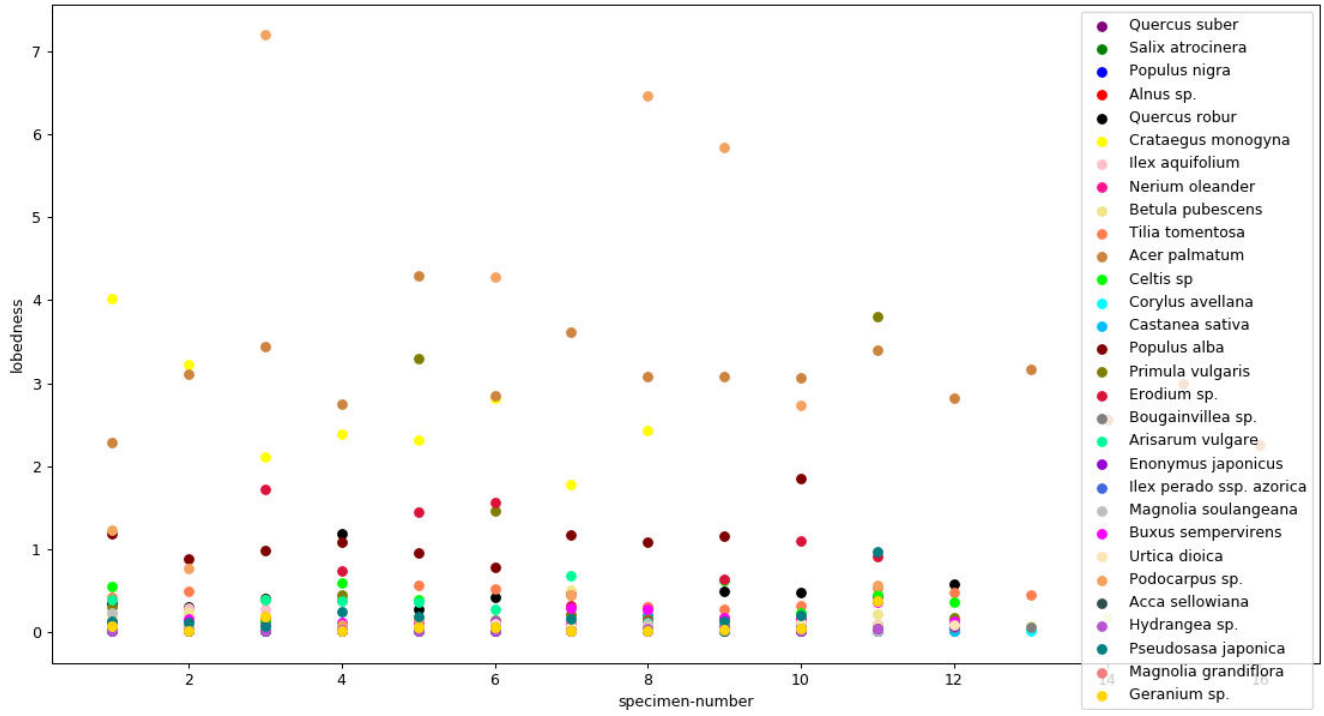
Slika 6 – Figure 4 (specimen-number I solidity)



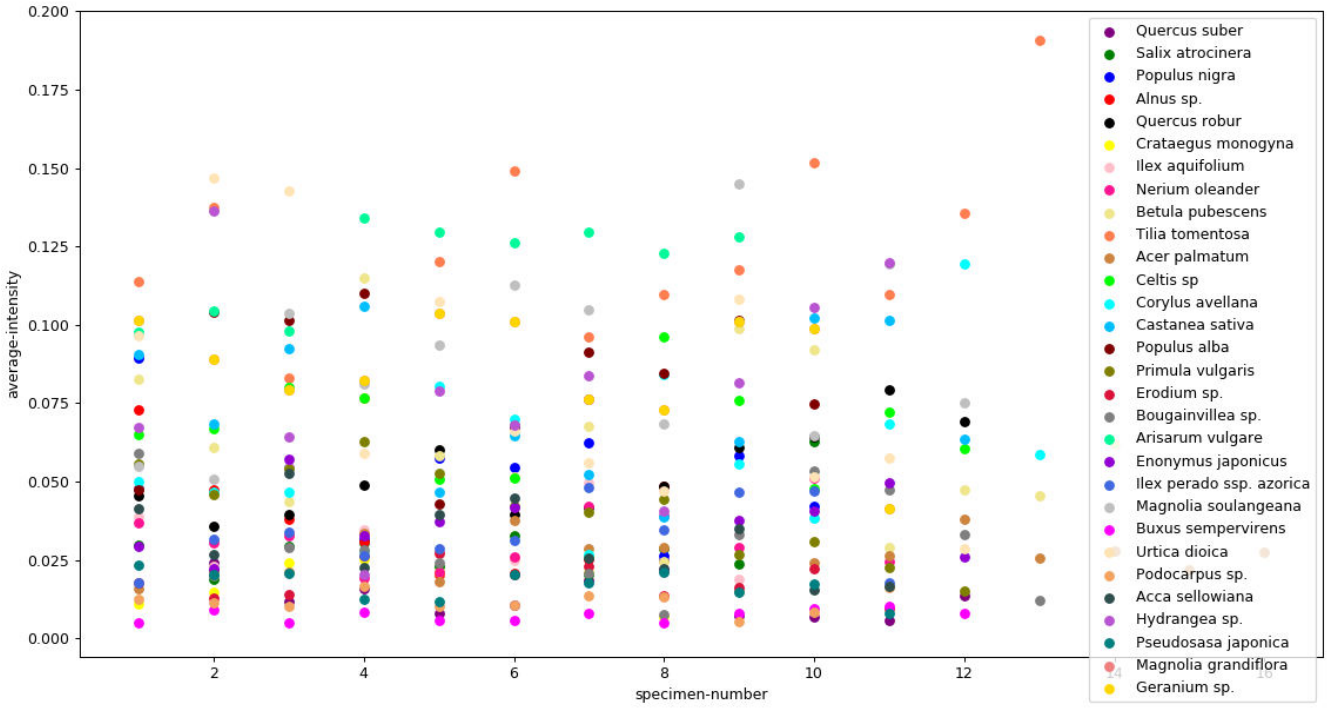
Slika 7 – Figure 5 (specimen-number I isoperimetric-factor)



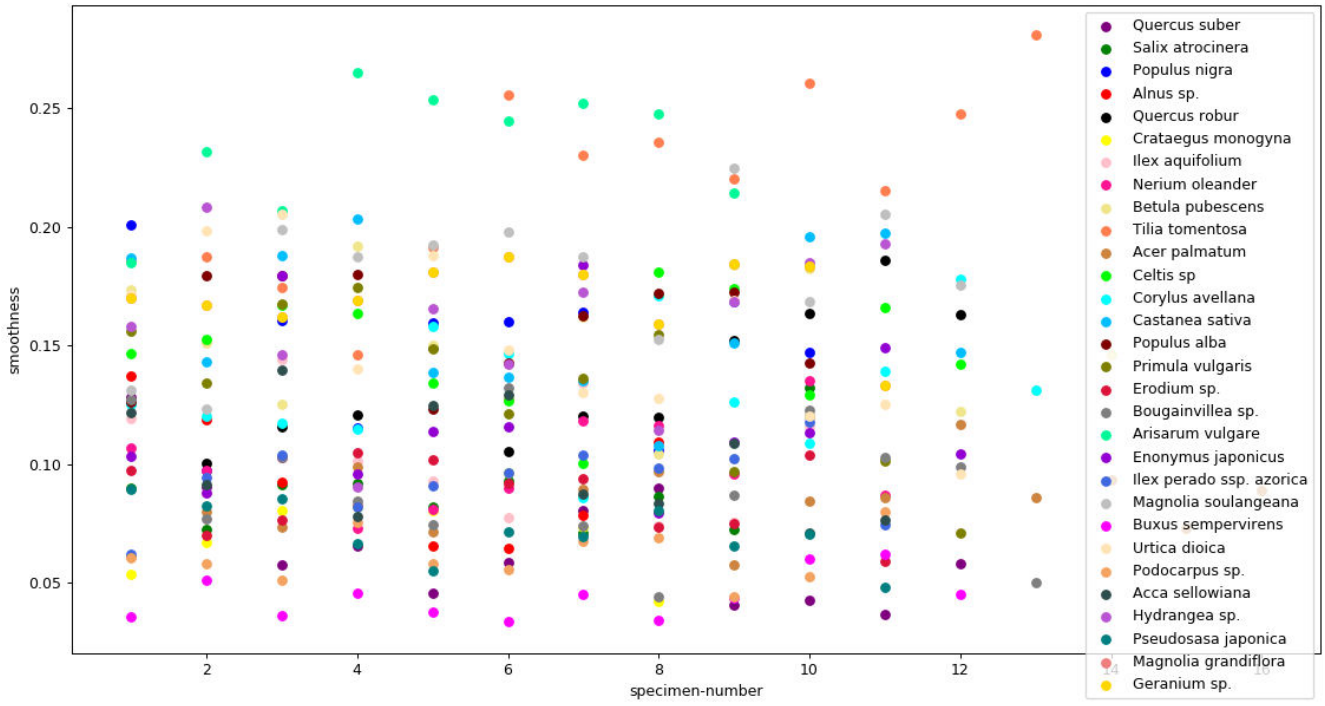
Slika 8 – Figure 6 (specimen-number I maximal-indentation-depth)



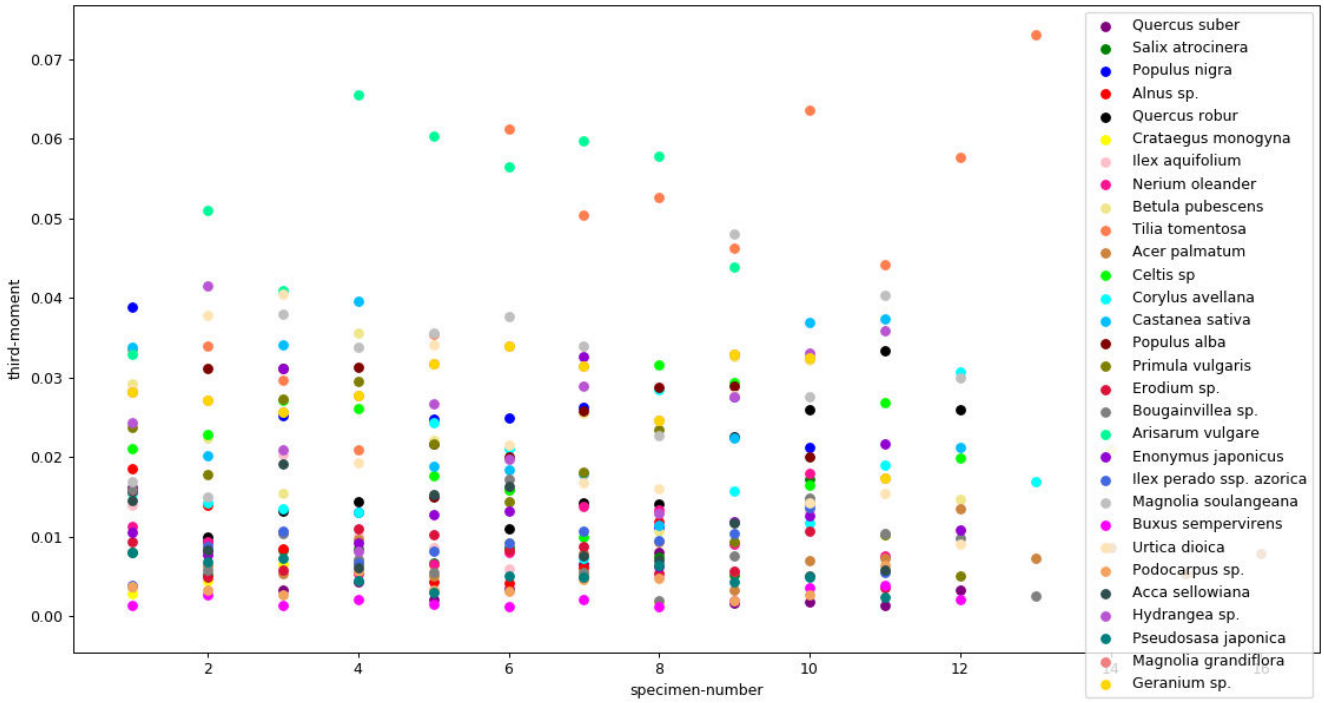
Slika 9 – Figure 7 (specimen-number I lobedness)



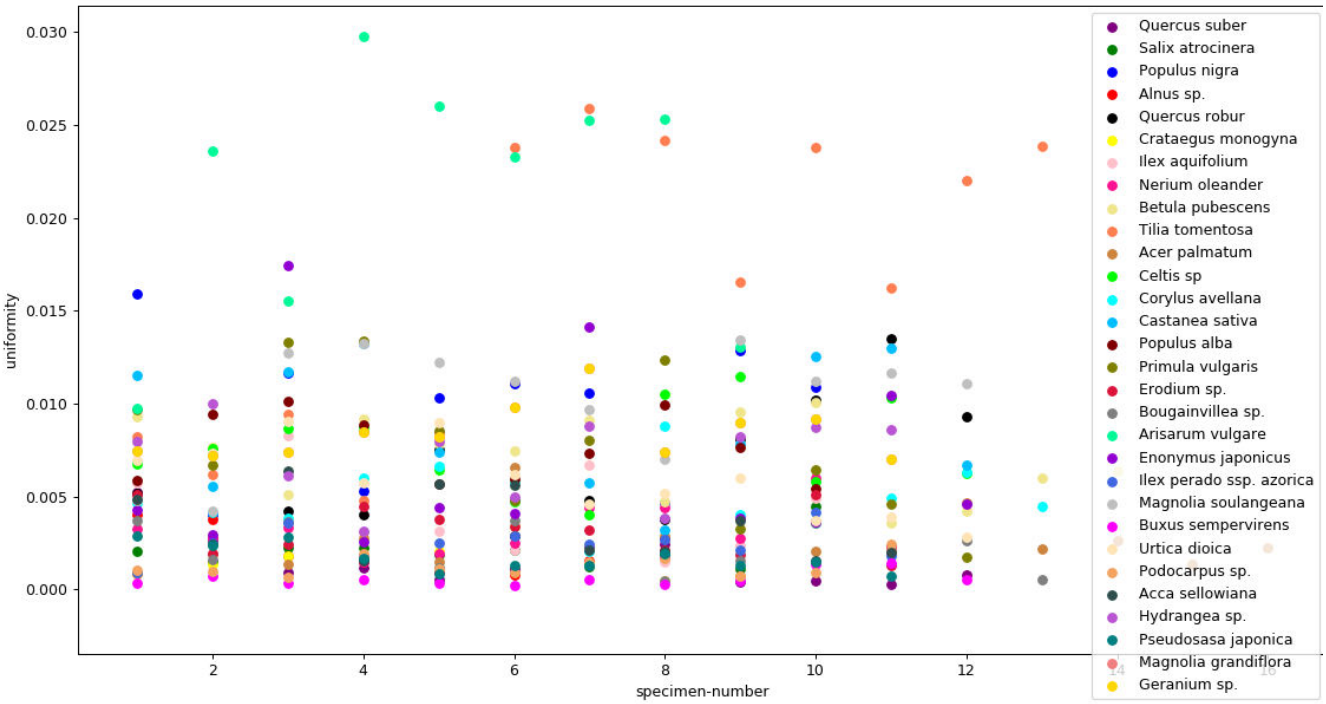
Slika 10 – Figure 8 (specimen-number I average-intensity)



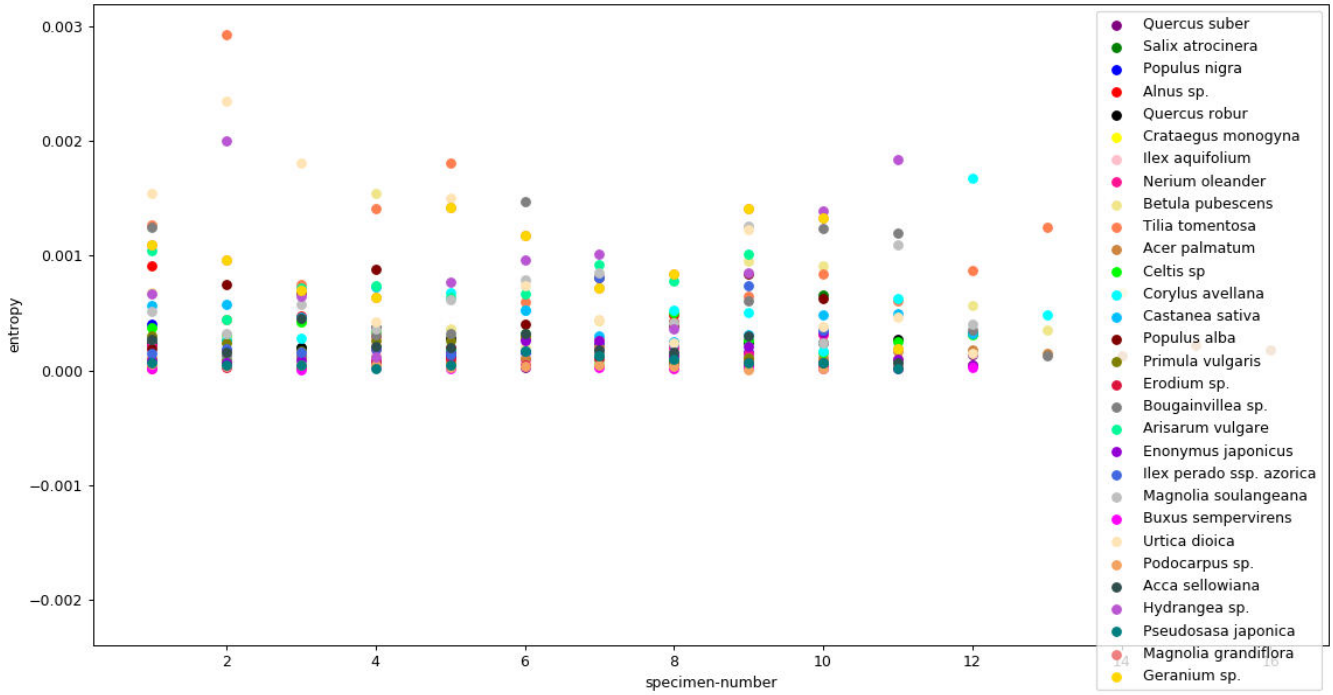
Slika 11 – Figure 9 (specimen-number I smoothness)



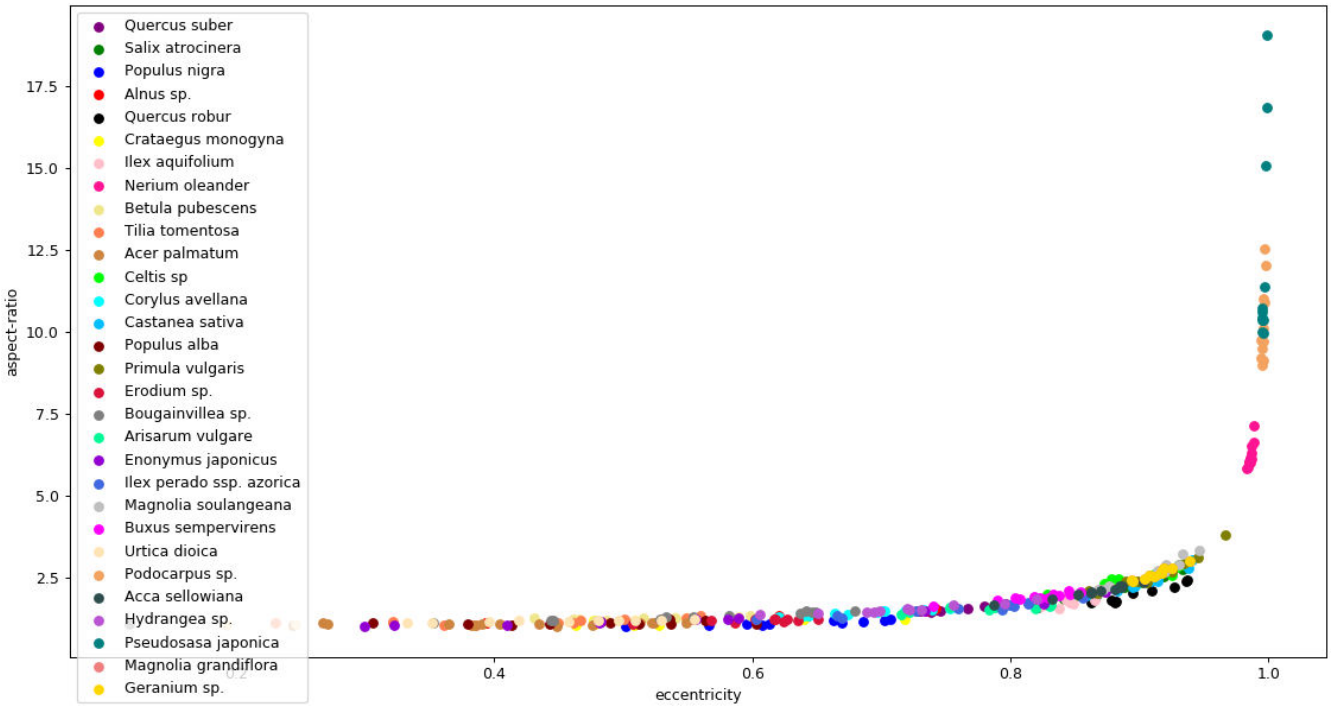
Slika 12 – Figure 10 (specimen-number I third-moment)



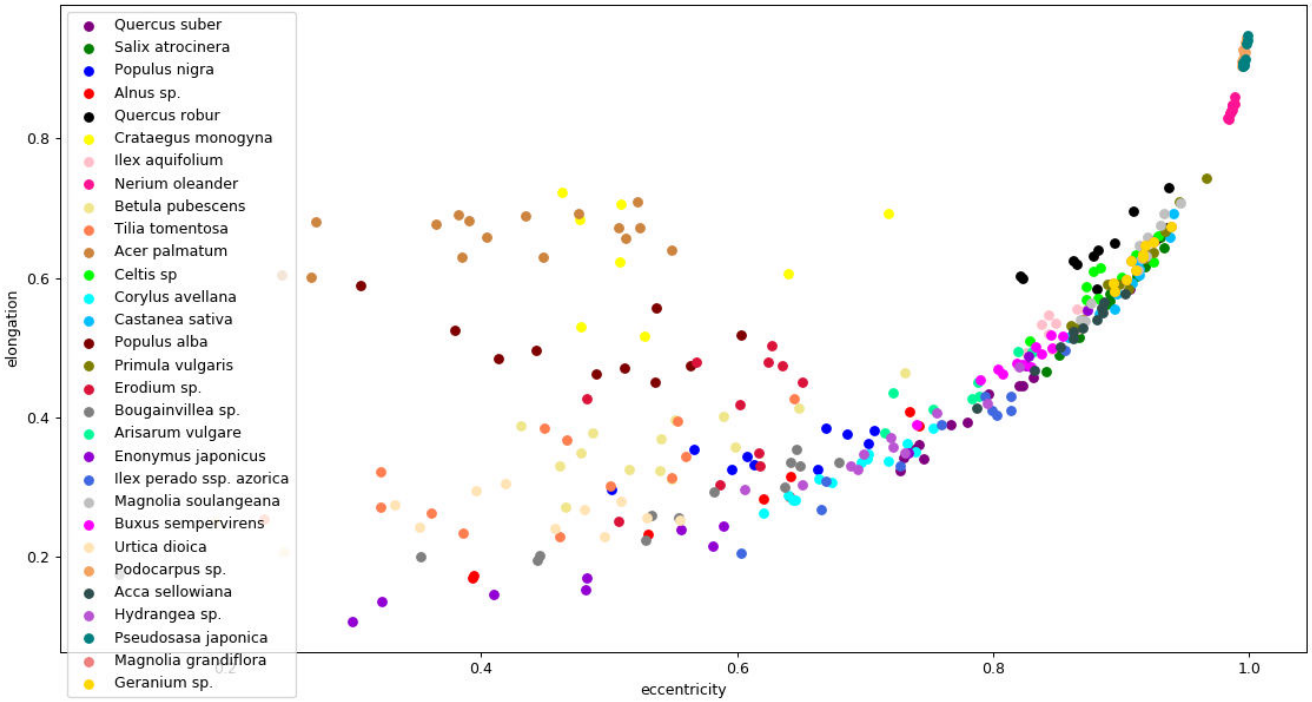
Slika 13 – Figure 11 (specimen-number I uniformity)



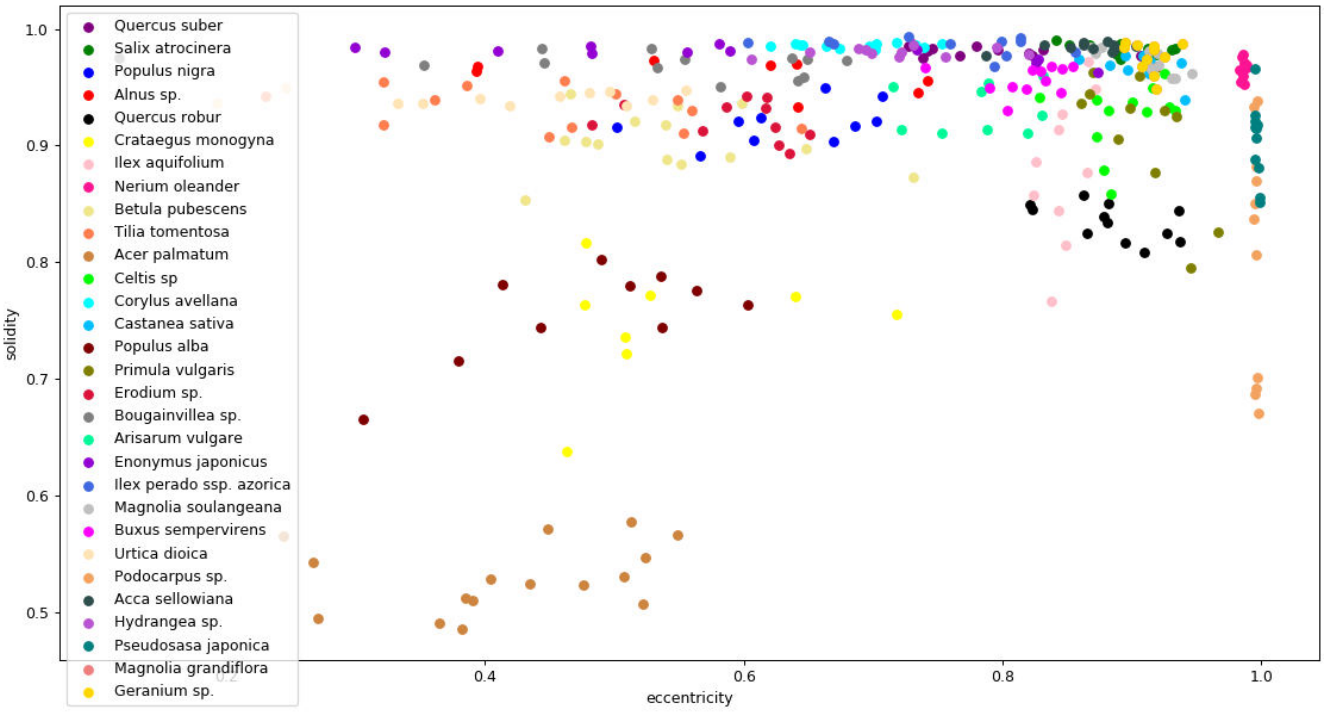
Slika 14 – Figure 12 (specimen-number I entropy)



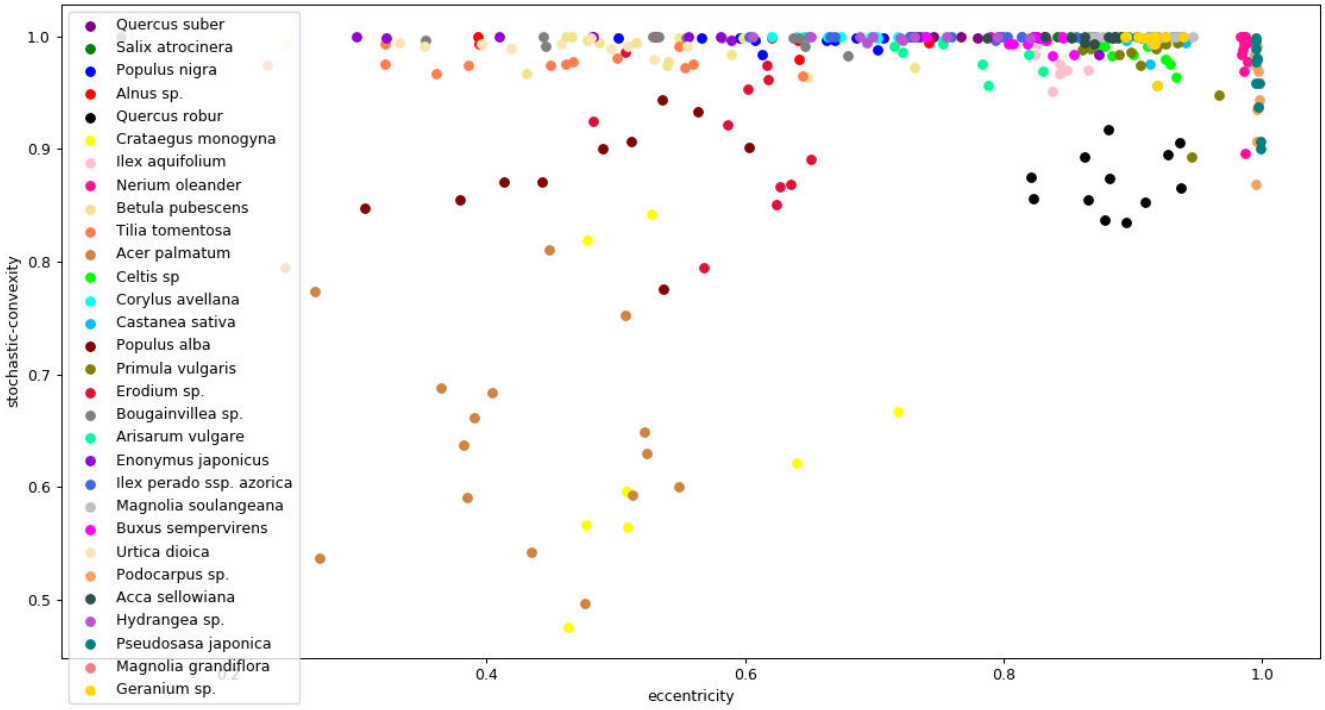
Slika 15 – Figure 13 (eccentricity I aspect-ratio)



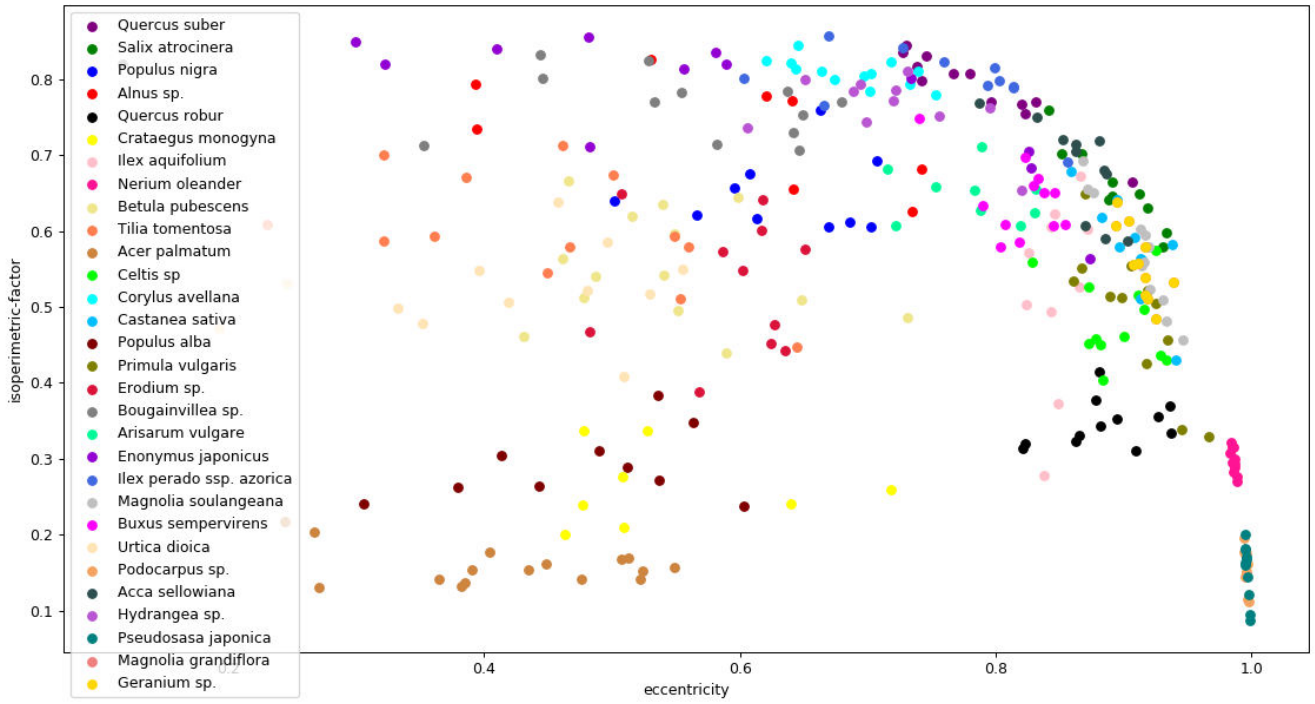
Slika 16 – Figure 14 (eccentricity I elongation)



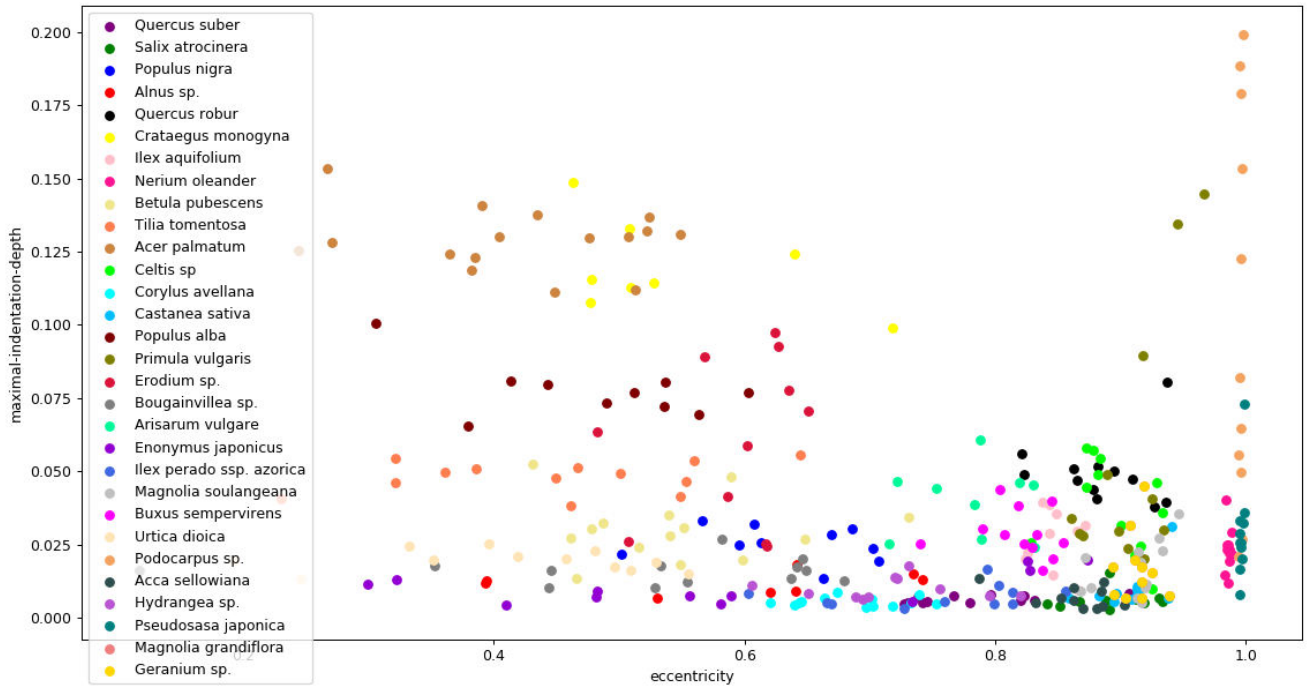
Slika 17 – Figure 15 (eccentricity I solidity)



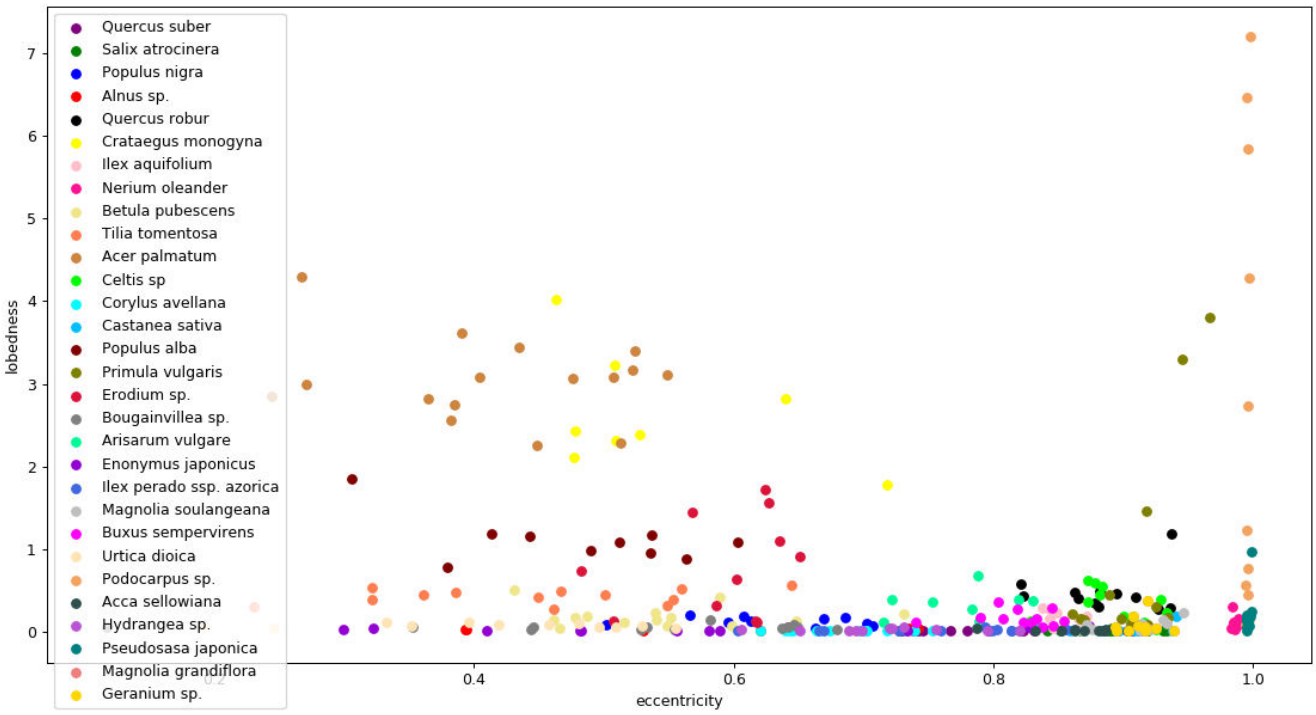
Slika 18 – Figure 16 (eccentricity I stochastic-convexity)



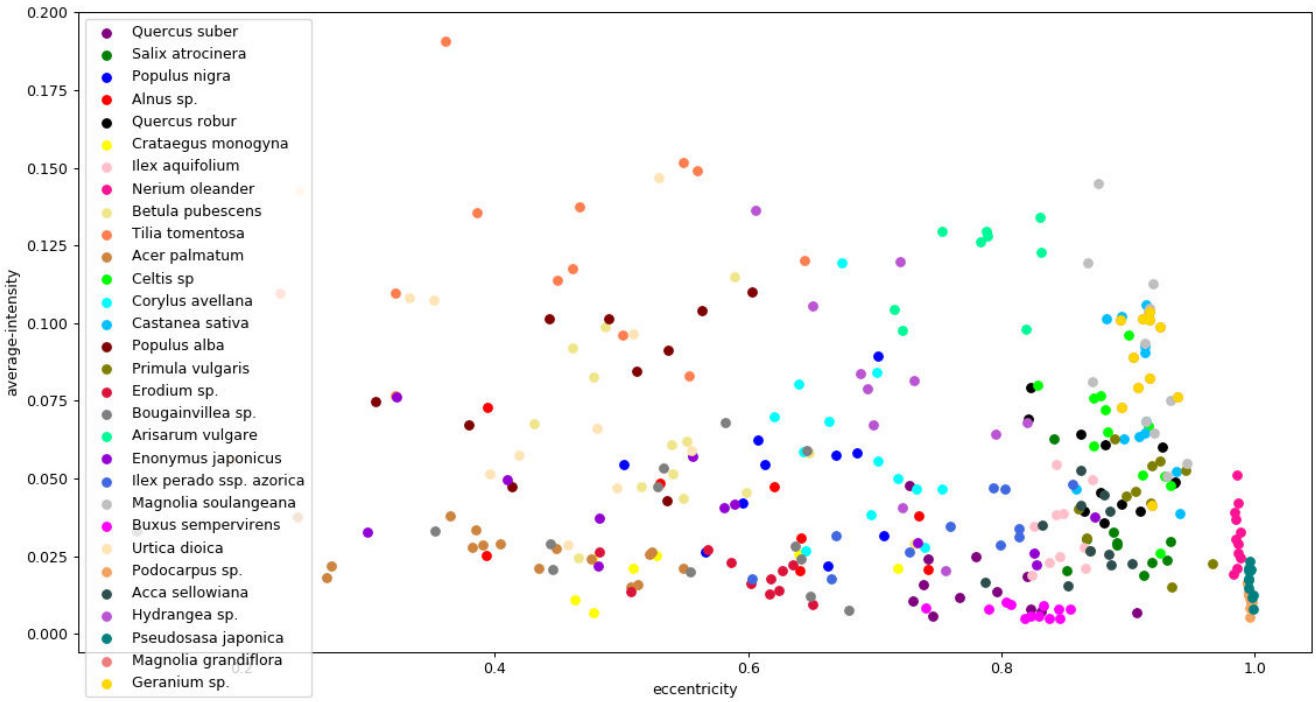
Slika 19 – Figure 17 (eccentricity I isoperimetric-factor)



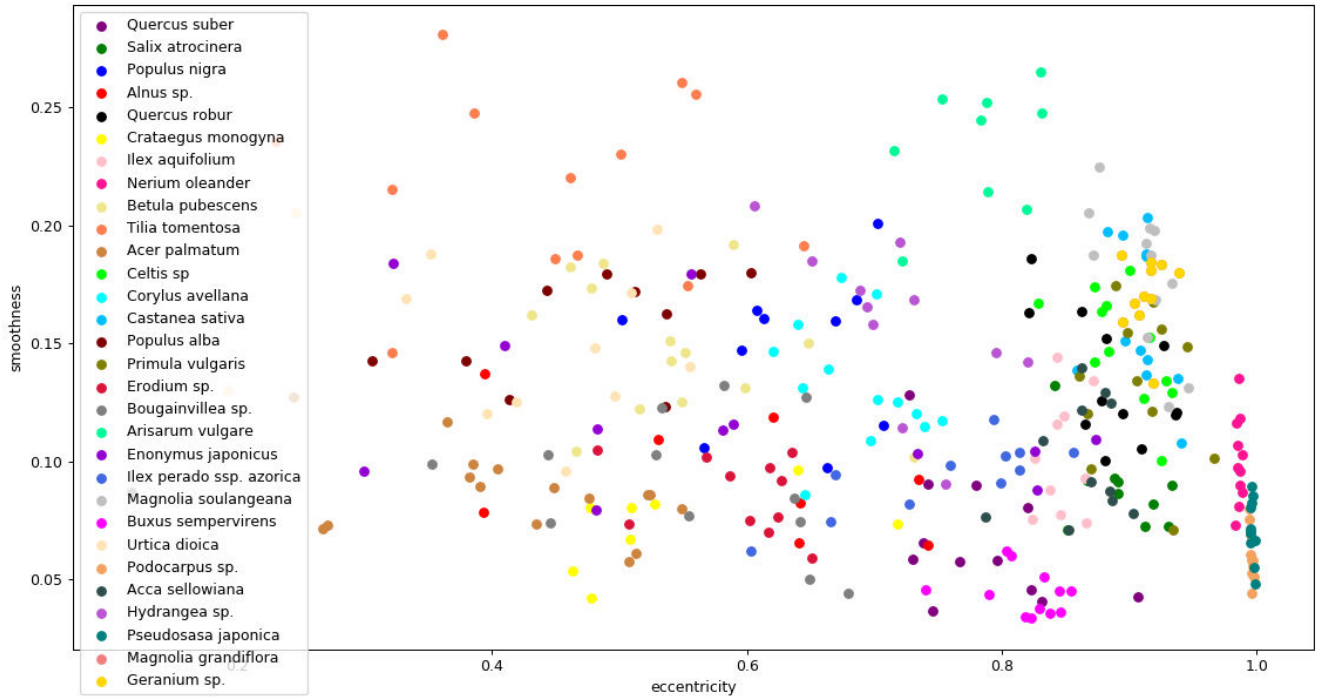
Slika 20 – Figure 18 (eccentricity I maximal-indentation-depth)



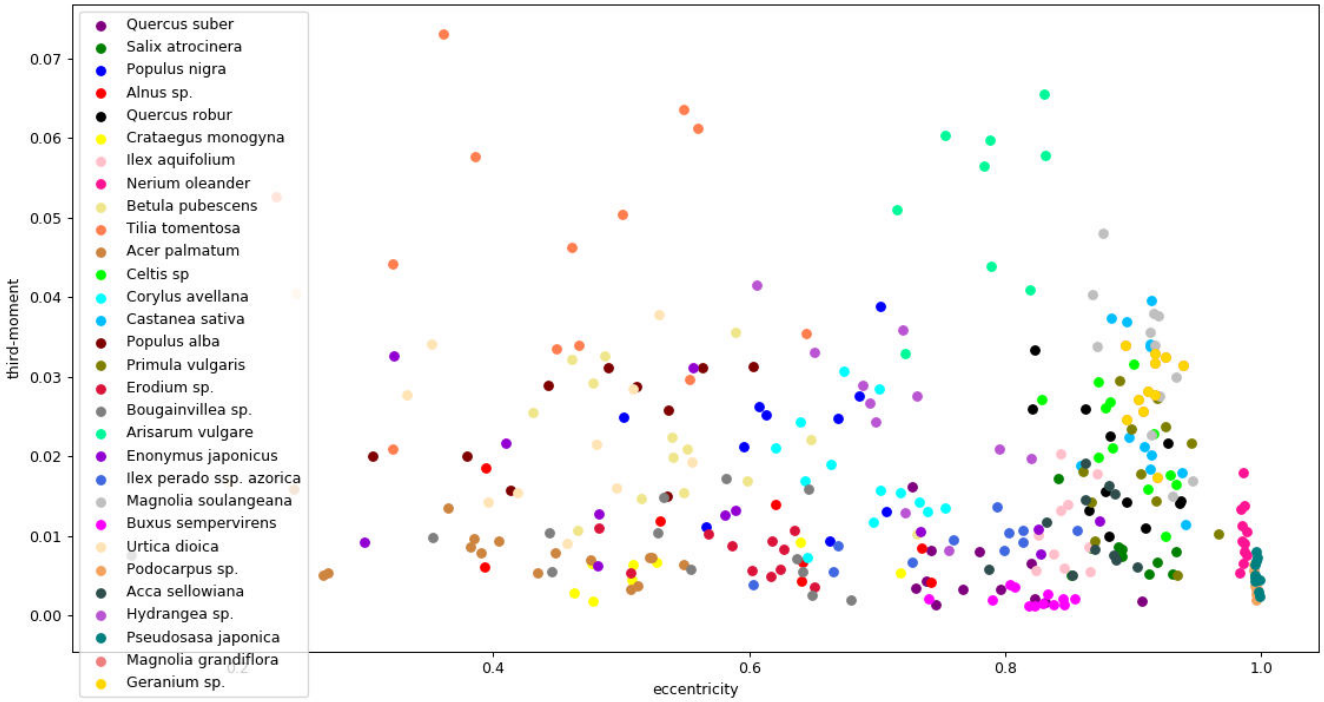
Slika 21 – Figure 19 (eccentricity I lobedness)



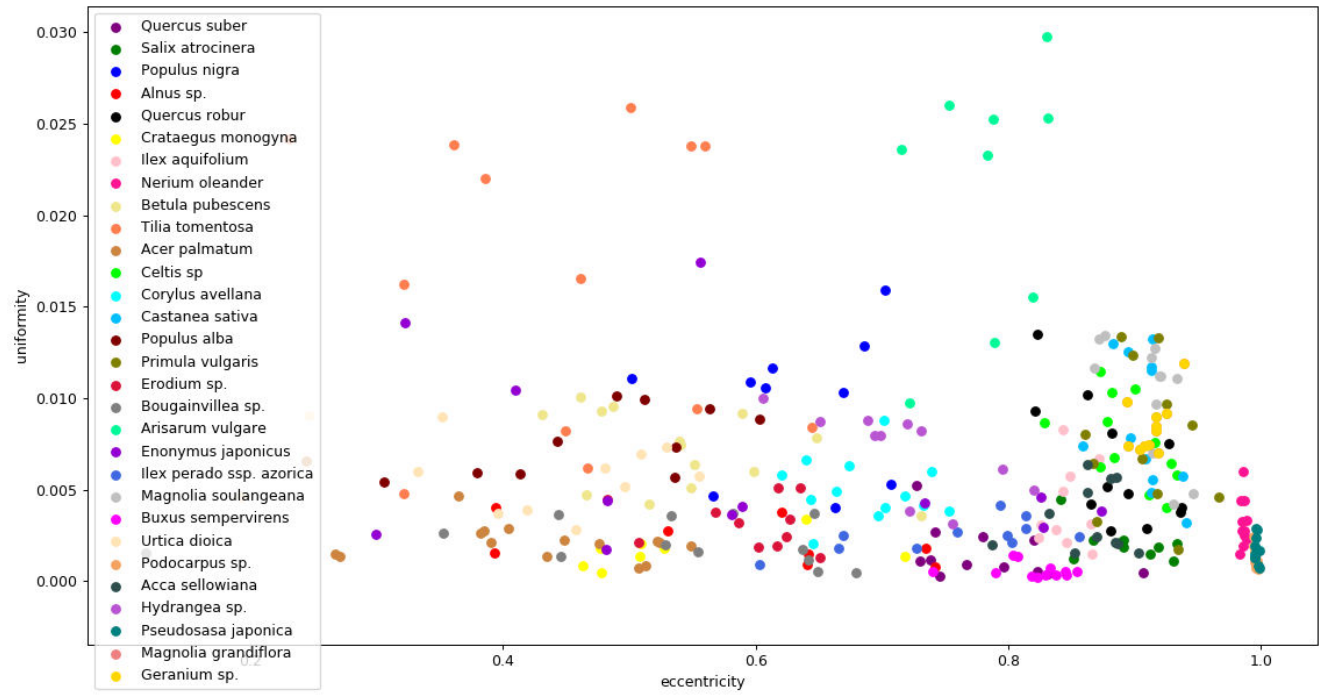
Slika 22 – Figure 20 (eccentricity I average-intensity)



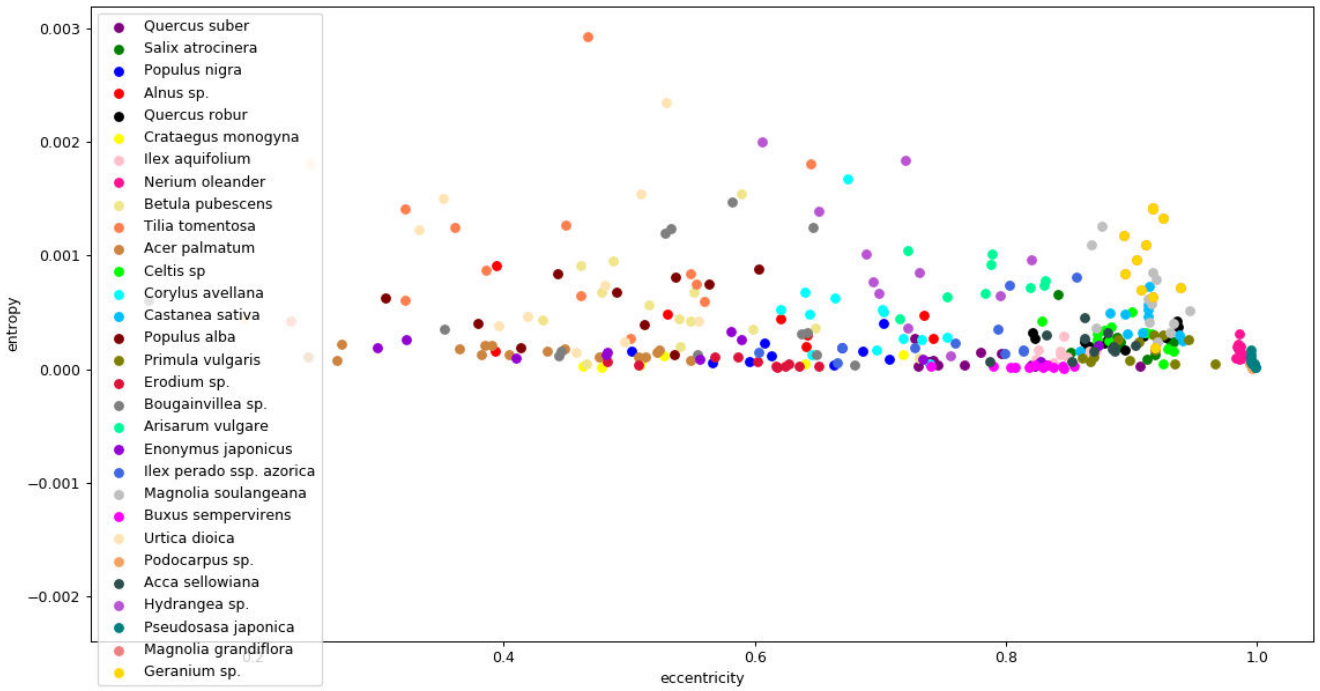
Slika 23 – Figure 21 (eccentricity I smoothness)



Slika 24 – Figure 22 (eccentricity I third-moment)



Slika 25 – Figure 23 (eccentricity I uniformity)



Slika 26 – Figure 24 (eccentricity I entropy)

3. Učitavanje i razdvajanje podataka

Implementacija je urađena u programskom jeziku*Python*i to verzija 3.5.2. Podaci koji se nalaze u*csv*fajlu učitani su pomoću funkcije *read_csv()* koja je deo *pandas* biblioteke. Prvu kolonu cini izlaz (class), dok sve ostale kolone cini ulaz (od druge do sesnaeste kolone fajla), pa su na taj način podeljeni na *X i Y*. Na početku je bilo potrebno da se podaci podele u dva skupa, jedan skup za treniranje, dok drugi skup je namenjen za testiranje modela. Trening skup podataka čini 80% podataka, dok skup za testiranje čini preostalih 20% podataka. Podela je napravljena pomoću funkcije *train_test_split()* koja za parametre prima X, Y i procenat podataka za test. Funkcija je deo biblioteke *sklearn i* vraća ulazni i izlazni skup podataka za treniranje, kao i ulazni i izlazni skup podataka za test. Kod za učitavanje podataka i podelu na trening skup i test skup prikazan je na slici 27.

```
data="leaf.csv"
names=['class', 'specimen-number', 'eccentricity', 'aspect-ratio', 'elongation', 'solidity', 'stochastic-convexity', 'isoperimetric-factor', 'maximal-
indentation-depth', 'lobedness', 'average-intensity', 'average-contrast', 'smoothness', 'third-moment', 'uniformity', 'entropy']
dataset=pd.read_csv(data, names=names)

#podela podataka na test i train
array=dataset.values
X=array[:,0:15]
Y=array[:,0]

X_train, X_test, Y_train, Y_test=train_test_split(X, Y, test_size=0.2)
print(dataset.values)
```

Slika 27 - Učitavanje podataka i podela na trening i test skup

4. Model, treniranje i testiranje

Problem klasifikacije je takav da ga je pogodno rešiti primenom višeslojne neuronske mreže. Za model višeslojne neuronske mreže moze se birati model tipa *Sequential* iz *Keras* biblioteke. Mreža ima 16 slojeva, od kojih su 15 skrivena, a jedan izlazni sloj. Pošto ima 15 ulaznih atributa, mreža ima 15 ulaza. Izlazni sloj ima 30 neurona zbog 30 mogućih izlaza u zavisnosti od klase. Za aktiavacionu funkciju skirvenih slojeva moze se birati funkcija *relu.* Funkcija izlaznog sloja moze biti *softmax,* jer na taj način je osigurano da su izlazne vrednosti između 0 i 1 i može se koristiti za predviđanje verovatnoće. Broj slojeva i neurona po slojevima, kao i aktivacione funkcije se biraju nakon ručnog ispitivanja šta daje najveću tačnost i najmanju vrednost funkcije gubitka. Za algoritam optimizacije odabran je *adam*, za *loss* funkciju *categorical_crossentropy* koja je prigodna za multiklasifikaciju. Broj epoha moze biti 100, dok za veličinu gradijentnog spusta mozemo uzeti vrednost 5. Kroz model se provuku trening podaci i na taj način je istrenirana mreža.

5. Literatura

[1] The UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Leaf (datum pristupa 17.06.2020.)
[2] Machine Learning Mastery, https://machinelearningmastery.com/machine-learning-in-python-step-by-step/(datum pristupa 17.06.2020.)
[3] TensorFlow,https://www.tensorflow.org/api_docs/python/tf/keras/utils/to_categorical(datum pristupa 17.06.2020.)
[4] Keras Documentation,https://keras.io/getting-started/sequential-model-guide/(datum pristupa 17.06.2020.)