Teme za seminarski rad iz Istraživanja podataka 2

Student seminarski rad radi samostalno ili u paru. Jednu temu čini jedan skup podataka za obradu

- **za studente osnovnih studija smera Informatika**: i zadata metoda ili metode istraživanja podataka kojom se skup obrađuje (klasifikacija, klasterovanje ili pravila pridruživanja).
- **za studente master studija**: nad kojim je potrebno primeniti sve metode koje su pokrivene kursom (klasifikacija, klasterovanje i pravila pridruživanja).

Prijava teme za seminarski rad

- Tema za seminarski rad se prijavljuje popunjavanjem formulara. Student/tim može navesti do tri željene teme.
  Jednu temu može da radi samo jedan student osnovnih studija (ili tim ako je tema za dva studenta). Isti skup (ukoliko ispod teme stoji red sa tekstom *student master studija:* može da koristi jedan student master studija. Preciznije: jednu temu koju čine zadati skup + zadata tehnika/tehnike može da izabere jedan student osnovnih studija, a isti skup podataka može da izabere samo jedan student master studija. Prilikom biranja skupa podataka master student navodi broj teme, ali se podrazumeva da će primeniti klasifikaciju, klasterovanje i pravila pridruživanja. Student master studija može da izabere skup iz teme u okviru koje je navedena linija *student master studija.*

- Student može da
  - izabere jednu od predloženih tema koja nije zauzeta ili da
  - predloži skup podataka koji bi želeo da analizira i, ako je student osnovnih studija, metodu istraživanja podataka koju bi primenio (klasifikacija ili klasterovanje). Pri prijavi svog skupa podataka obavezno navesti izvor (adresu skupa).

Mogući izvori za izbor skupa:

1. https://www.kaggle.com/datasets
2. https://www.openml.org/
3. https://www.kdnuggets.com/datasets/index.html
4. http://archive.ics.uci.edu/
5. https://github.com/awesomedata/awesome-public-datasets
6. SPMF open-source data mining library
   http://www.philippe-fournier-viger. com/spmf/
7. Lista dodatnih izvora
   https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b/

Napomena: Birati samo skupove koji imaju bar 2000 slogova sa

- **za studente osnovnih studija smera Informatika:** bar 100 atributa.
- **za studente master studija**: bar 50 atributa.

Opšta struktura rada

- Obavezno izvršiti preprocesiranje podataka.
- Podatke obraditi traženom metodom koristeći minimum 5 algoritama.
- Izvršiti analizu dobijenih rezultata; uporediti rešenja dobijena za različite algoritme.
- Seminarski rad koji se predaje treba da sadrži
    - tekstualni deo (koji sadrži tekst zadatka, opis podataka, opis obrade podataka, opis i tumačenje rezultata). Tekstualni deo se predaje u pdf formatu.
    - podatke (početnu verziju, kao i verziju dobijenu preprocesiranjem),
    - konstruisane modele (u odogovarjućem obliku, zavisi od korišćenog alata).
- Materijal koji se šalje mora da sadrži sve što je potrebno za ponavljanje kompletnog postupka u lokalnom okruženju.
- Za rešavanje problema mogu se koristiti SPSS modeler, biblioteke programskog jezika python, ili neki drugi alat za istraživanje podataka, kao i programi i skripte koje je napisao sam student radi obrade podataka. Ako se koristi neki deo koda koji je preuzet sa mreže obavezno navesti njegov izvor.
- Obavezni koraci u istraživanju:
    - Vizuelno prikazati podatke sa 2D ili 3D.
    - Za klasifikaciju i klasterovanje napraviti modele sa svim atributima i sa različitim redukovanim skupovima atributa i uporediti modele.
    - Skupovi koji sadrže kolone sa tekstom (npr. kolona tags) ili tekst potrebno je obraditi za dobijanje informacija potrebnih za klasifikaciju ili klasterovanje
    - U skupovima gde ima smisla koristiti različite ciljne atribute za klasifikaciju potrebno je napraviti modele za svaki od njih.

Predložene teme

Klasifikacija

1. Job Dataset
   https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset
   student osnovnih studija: 110/2021 Nikola Radojičić
   student master studija: 1080/25 Natalija Lazić

2. Taiwanese Bankruptcy Prediction
   https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction
   student osnovnih studija: 97/2022 Maša Lazić
   student master studija: 1101/2024 Mina Velebit

3. FMA: A Dataset For Music Analysis (2 osobe)
   https://archive.ics.uci.edu/dataset/386/fma+a+dataset+for+music+analysis
   student osnovnih studija: 240/2019 Mina Živić, 86/2019 Ana Mišmaš
   student master studija:

4. TUANDROMD (Tezpur University Android Malware Dataset)

https://archive.ics.uci.edu/dataset/855/tuandromd+(tezpur+university+android+malware+dataset)
student osnovnih studija: 169/2020 Luka Arambašić
student master studija: 1094/2024 Nemanja Zivkovic

5. Musk (Version 2)
   https://archive.ics.uci.edu/dataset/75/musk+version+2
   student osnovnih studija: 231/2020 Masa Gacevic
   student master studija: 1064/2025 Милица Ђорђевић

6. Swarm Behaviour
   https://archive.ics.uci.edu/dataset/524/swarm+behaviour
   student osnovnih studija: 112/2021 Danilo Nikolaš
   student master studija:

7. Simulated Falls and Daily Living Activities Data Set
   https://archive.ics.uci.edu/dataset/455/simulated+falls+and+daily+living+activities+data+set
   student osnovnih studija: 137/2021 Andjela Jokovic
   student master studija: 1086/2025 Milica Rašula

8. Activity recognition using wearable physiological measurements
   https://archive.ics.uci.edu/dataset/552/activity+recognition+using+wearable+physiological+measurements
   Drugi link
   https://www.mdpi.com/1424-8220/19/24/5524/s1

   student osnovnih studija: 32/2022 Bogdan Micić
   student master studija: 1067/2025 Gala Posedi

9. Grammatical Facial Expressions
   https://archive.ics.uci.edu/dataset/317/grammatical+facial+expressions
   student osnovnih studija: 39/2020 Zarija Trtovic
   student master studija: 1084/2025 Petar Lukovic

10. Farm Ads
    https://archive.ics.uci.edu/dataset/218/farm+ads
    student osnovnih studija: 349/2021 Milica Mladenović
    student master studija: 1090/2025 Bogdan Mirovic

11. Character Font Images
    https://archive.ics.uci.edu/dataset/417/character+font+images
    student osnovnih studija: 71/2021 Nemanja Kelecevic
    student master studija: 1081/2025 Anja Đurić

12. p53 Mutants

https://archive.ics.uci.edu/dataset/188/p53+mutants
student osnovnih studija: 76/2021 Uros Kovacevic
student master studija:

13. Reuter_50_50
    https://archive.ics.uci.edu/dataset/217/reuter+50+50
    student osnovnih studija: 227/2020 Zagorka Pantovic
    student master studija:

14. MEx
    https://archive.ics.uci.edu/dataset/500/mex
    student osnovnih studija: 099/2021 Ana Arsić
    student master studija:

15. SpeedDating
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=40536
    student osnovnih studija: 160/2020 Ana Drobnjak
    student master studija: 1063/2025 Maja Saković

16. OVA_Breast
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1128
    student osnovnih studija: 168/2021 Sara Kabić
    student master studija: 1113/2025 Maša Miladinović

17. OVA_Ovary
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1166
    student osnovnih studija: 146/2020 Isidora Dukić

18. OVA_Uterus
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1138
    student osnovnih studija:

19. OVA_Kidney
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1134
    student osnovnih studija: 251/2020 Pavle Miličković

20. OVA_Lung
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1130

student osnovnih studija: 124/2020 Stefan Nikolic

21. OVA_Omentum
https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1139
student osnovnih studija:

22. OVA_Colon
https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1161
student osnovnih studija: 158/2022 Filip Djurkovic

23. **Klasifikacija ljudskih uzoraka** korišćenjem skupa Jednoćelijska transkriptomska mapa ljudskog i mišjeg pankreasa
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133
Kolone
- opis
- barcode
- assigned_cluster - tip ćelije : *ciljni atribut za kalsifikaciju*
- ostale kolone su geni, tj. broj prebrojanih gena

Uputstvo za preuzimanje skupa: klikom na Željeni uzorak u odeljku Samples (6) bićete preusmereni na stranicu uzorka. Na kraju stranice se nalazi link za preuzimanje skupa.
Npr. za uzorak GSM2230757 human pancreatic islets, sample 1, link je
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2230757
a skup podataka je GSM2230757_human1_umifm_counts.csv.gz na dnu stranice.
Obraditi uzorke
GSM2230757 human pancreatic islets, sample 1
GSM2230758 human pancreatic islets, sample 2
GSM2230759 human pancreatic islets, sample 3
GSM2230760 human pancreatic islets, sample 4

student osnovnih studija:
student master studija: 1057/2025 Jana Živković

24. **Klasifikacija mišjih uzoraka** korišćenjem skupa Jednoćelijska transkriptomska mapa ljudskog i mišjeg pankreasa
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133
Kolone
- opis
- barcode
- assigned_cluster - tip ćelije : *ciljni atribut za kalsifikaciju*
- ostale kolone su geni, tj. broj prebrojanih gena

Uputstvo za preuzimanje skupa: klikom na Željeni uzorak u odeljku Samples (6) bićete preusmereni na stranicu uzorka. Na kraju stranice se nalazi link za preuzimanje skupa. Npr. za uzorak GSM2230757 human pancreatic islets, sample 1, link je https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2230757 a skup podataka je GSM2230757_human1_umifm_counts.csv.gz na dnu stranice.

Obraditi uzorke

GSM2230761 mouse pancreatic islets, sample 1
GSM2230762 mouse pancreatic islets, sample 2

student osnovnih studija:
student master studija:

Klasterovanje

25. Brazilian E-Commerce Public Dataset
    https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_products_dataset.csv
    student osnovnih studija: 28/2022 Jovana Brkljac

26. FitBit Fitness Tracker Data
    https://www.kaggle.com/datasets/arashnic/fitbit/data
    student osnovnih studija: 42/2022 Dimitrije Spasojevic

27. NBA Database (2 osobe)
    https://www.kaggle.com/datasets/wyattowalsh/basketball
    student osnovnih studija: 264/2019 Marija Grekulović, 128/2021 Stepan Ignjatović
    student master studija: 1035/2025 Anđela Jovanović

28. Job Dataset
    https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset
    student osnovnih studija: 271/2021 Mateja Dabović

29. LinkedIn Job Postings (2023 - 2024)
    https://www.kaggle.com/datasets/arshkon/linkedin-job-postings?select=postings.csv
    student osnovnih studija: 26/2022 Marina Vračarić

30. Big Five Personality Test
    https://www.kaggle.com/datasets/tunguz/big-five-personality-test
    student osnovnih studija: 135/2019 Ksenija Ivanovic

31. Football Data from Transfermarkt
    https://www.kaggle.com/datasets/davidcariboo/player-scores?select=player_valuations.csv
    student osnovnih studija: 30/2022 Mateja Janic

32. Trending YouTube Video Statistics

https://www.kaggle.com/datasets/datasnaek/youtube-new?select=DEvideos.csv
student osnovnih studija: 103/2022 Aleksa Vukadinovic

33. Communities and Crime Unnormalized
https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized
student osnovnih studija: 19/2022 David Ćuruvija

34. FMA: A Dataset For Music Analysis (2 osobe)
https://archive.ics.uci.edu/dataset/386/fma+a+dataset+for+music+analysis
studenti osnovnih studija: 123/2021 Mihajlo Trifunovic, 227/2021 Jelena Đuric

35. TUANDROMD (Tezpur University Android Malware Dataset)
https://archive.ics.uci.edu/dataset/855/tuandromd+(tezpur+university+android+malware+dataset)
student osnovnih studija: 82/2020 Mihaela Filipović

36. Anonymous Microsoft Web Data
https://archive.ics.uci.edu/dataset/4/anonymous+microsoft+web+data
student osnovnih studija: 124/2021 Uroš Ivetić

37. Activity recognition using wearable physiological measurements
https://archive.ics.uci.edu/dataset/552/activity+recognition+using+wearable+physiological+measurements
student osnovnih studija: 355/2021 Martina Iricanin

38. Grammatical Facial Expressions
https://archive.ics.uci.edu/dataset/317/grammatical+facial+expressions
student osnovnih studija: 63/2021 Milos Krstic

39. Farm Ads
https://archive.ics.uci.edu/dataset/218/farm+ads
student osnovnih studija: 265/2021 Lazar Dunjić

40. Reuter_50_50
https://archive.ics.uci.edu/dataset/217/reuter+50+50
student osnovnih studija: 174/2020 Nikola Krstajic

41. MEx
https://archive.ics.uci.edu/dataset/500/mex
student osnovnih studija:

42. SpeedDating
https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=40536

student osnovnih studija: 150/2020 Viktor Danilovic

43. OVA_Breast
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=1128
    student osnovnih studija:

44. IMDB.drama
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=273
    student osnovnih studija: 17/2022 Marko Veljovic

45. Comprehensive-database-of-Minerals
    https://www.openml.org/search?type=data&status=active&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfFeatures=between_100_1000&id=43356

    student osnovnih studija: 13/2022 Nikola Trifunović

46. **Klasterovanje ljudskih uzoraka** korišćenjem skupa Jednoćelijska transkriptomska mapa ljudskog i mišjeg pankreasa
    https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133
    Kolone
    - opis
    - barcode
    - assigned_cluster - tip ćelije
    - ostale kolone su geni, tj. broj prebrojanih gena

    Uputstvo za preuzimanje skupa: klikom na željeni uzorak u odeljku Samples (6) bićete preusmereni na stranicu uzorka. Na kraju stranice se nalazi link za preuzimanje skupa. Npr. za uzorak GSM2230757 human pancreatic islets, sample 1, link je
    https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2230757
    a skup podataka je GSM2230757_human1_umifm_counts.csv.gz na dnu stranice.
    Obraditi uzorke
    GSM2230757 human pancreatic islets, sample 1
    GSM2230758 human pancreatic islets, sample 2
    GSM2230759 human pancreatic islets, sample 3
    GSM2230760 human pancreatic islets, sample 4

    student osnovnih studija:

47. **Klasterovanje mišjih uzoraka** korišćenjem skupa Jednoćelijska transkriptomska mapa ljudskog i mišjeg pankreasa
    https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133
    Kolone

- opis
- barcode
- assigned_cluster - tip ćelije
- ostale kolone su geni, tj. broj prebrojanih gena

Uputstvo za preuzimanje skupa: klikom na željeni uzorak u odeljku Samples (6) bićete preusmereni na stranicu uzorka. Na kraju stranice se nalazi link za preuzimanje skupa. Npr. za uzorak GSM2230757 human pancreatic islets, sample 1, link je https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2230757 a skup podataka je GSM2230757_human1_umifm_counts.csv.gz na dnu stranice.

Obraditi uzorke

GSM2230761 mouse pancreatic islets, sample 1
GSM2230762 mouse pancreatic islets, sample 2

student osnovnih studija:

Klasifikacija i klasterovanje

48. fake-and-real-news-dataset
    https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
    student osnovnih studija:
    student master studija: 1074/2025 Miloš Jovanov

49. Fake News Classification
    https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification
    student osnovnih studija: 359/2020 Maša Mitrović
    student master studija: mi251034, Nenad Dobrosavljevic

50. Food.com Recipes and Interactions (2 osobe)
    https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions?select=interactions_train.csv
    studenti osnovnih studija: 168/2022 Anja Petrovic, 184/2022 Marina Neskovic

Klasifikacija i pravila pridruživanja

51. REJAFADA (2 osobe)
    https://archive.ics.uci.edu/dataset/860/rejafada
    studenti osnovnih studija: 364/2022 Jovana Medenica, 366/2022 Mateja Miletic

52. DeliciousMIL: A Data Set for Multi-Label Multi-Instance Learning with Instance Labels
    https://archive.ics.uci.edu/dataset/418/deliciousmil+a+data+set+for+multi+label+multi+instance+learning+with+instance+labels
    student osnovnih studija:
    student master studija:

53. Analiza ćelijskog sastava ljudske jetre primenom single-cell RNA-seq metode
    https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115469 (2 osobe)
    Podaci:
    - GSE115469_CellClusterType.txt.gz
    Sadrži informacije o ćelijama i tipu. Za klasifikaciju se koristi CellType.
    - GSE115469_Data.csv.gz
    Sadrži podatke ekspresiji gena. Informacije o id gena su u prvoj koloni, a ostale kolone predstavljaju informaciju o jednoj ćeliji. Za analizu je potrebno transponovati matricu.

    studenti osnovnih studija: 47/2022 Milica Zubljić, 68/2022 Natalija Pavlićevič

Klasterovanje i pravila pridruživanja

54. Book Recommendation Dataset
    https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset?select=Books.csv
    student osnovnih studija: 173/2018 Sandra Milenković

55. European Soccer Database (2 osobe)
    https://www.kaggle.com/datasets/hugomathien/soccer
    studenti osnovnih studija: 173/2020 Luka Kizić, 75/2021 Petar Popović

56. Health News in Twitter
    https://archive.ics.uci.edu/dataset/438/health+news+in+twitter
    student osnovnih studija:

57. REJAFADA (2 osobe)
    https://archive.ics.uci.edu/dataset/860/rejafada
    studenti osnovnih studija: 144/2022 Aleksandar Đukić, 164/2022 Lazar Nikolić

58. DeliciousMIL: A Data Set for Multi-Label Multi-Instance Learning with Instance Labels
    https://archive.ics.uci.edu/dataset/418/deliciousmil+a+data+set+for+multi+label+multi+instance+learning+with+instance+labels
    student osnovnih studija:

59. Analiza ćelijskog sastava ljudske jetre primenom single-cell RNA-seq metode (2 osobe)
    https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115469
    Podaci:
    - GSE115469_CellClusterType.txt.gz
    Sadrži informacije o ćelijama i tipu.
    - GSE115469_Data.csv.gz
    Sadrži podatke ekspresiji gena. Informacije o id gena su u prvoj koloni, a ostale kolone predstavljaju informaciju o jednoj ćeliji. Za analizu je potrebno transponovati matricu.

    studenti osnovnih studija: 51/2021 Vuk Vujasinović, 36/2020 Bogdan Delić

Klasifikacija,. klasterovanje, pravila pridruživanja

60. Amazon Customer Reviews Dataset
    https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
    studenti osnovnih studija: 91/2021 Dragana Katic, 252/2021 Tamara Saponjic

Dodatne

RAG sistem za ucenje istrazivanja podataka
408/2021 Dusan Jevtovic

Klasterovanje i pravila pridruzivanja
skup college_scorecard
https://www.openml.org/search?type=data&sort=runs&status=active&qualities.NumberOfInstances=between_10000_100000&qualities.NumberOfFeatures=between_100_1000&id=46805

studenti: 146/2017 Viktor Gizdavic, 138/2019 Vojkan Panic

Klasterovanje
NASA's Confirmed Exoplanets: Regular Updated

https://www.kaggle.com/datasets/jameskychoi/confirmed-exoplanet-latest-update-dataset?

studenti: 60/2021 Katarina Šćekić, 26/2021 Boris Rajić

Klasterovanje
FlowSOM vs Standard Clustering Algorithms

Skup podataka: OpenML “gas-drift-different-concentrations”

https://www.openml.org/d/1477

studenti: 62/2021Jovan Rankovic, 303/2023 Filip Dramicanin

Klasifikacija
Home credit default risk
https://www.kaggle.com/competitions/home-credit-default-risk/data

student: 131/2020 Marko Petrovic

Klasifikacija i klasterovanje

ICU Patient Outcome Prediction:
https://www.kaggle.com/datasets/fdemoribajolin/death-classification-icu
studenti 80/2021 Aleksandar Ilić, 108/2021 Nikola Živković:

https://archive.ics.uci.edu/dataset/54/isolet
Angelina Jordanov 422/2019