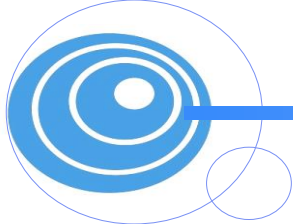


Big Data технологије у саобраћају, транспорту и логистици

Наставници (*e-mail*): др Слађана Јанковић, ванредни професор (s.jankovic@sf.bg.ac.rs)

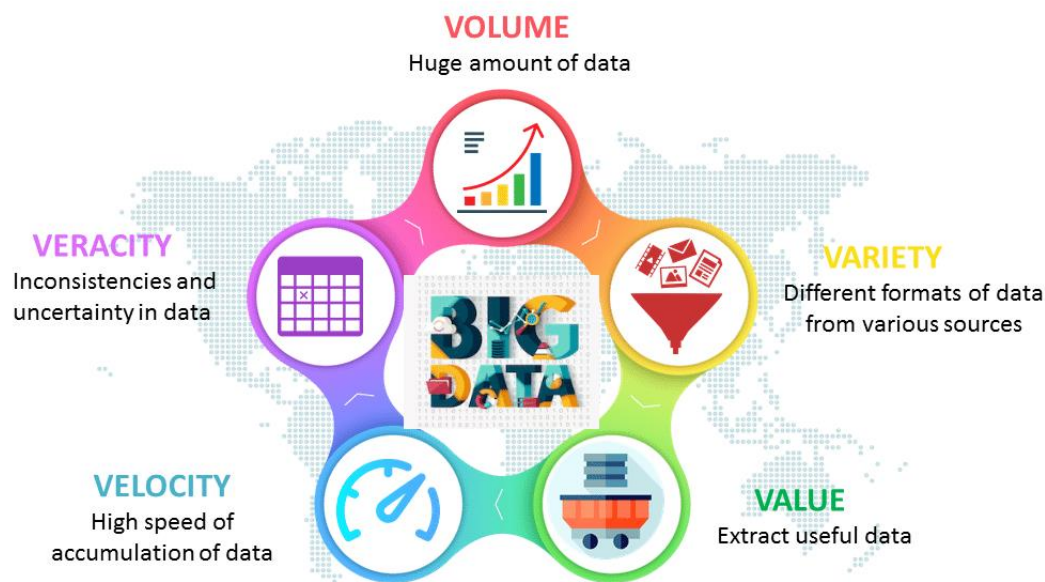
др Ана Узелац, доцент (ana.uzelac@sf.bg.ac.rs)

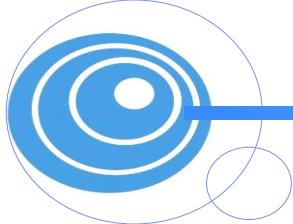
др Снежана Младеновић, редовни професор (snezanam@sf.bg.ac.rs)



Појам Big Data

- Под појмом Big Data подразумева се информациони ресурс велике количине, велике брзине увећавања и велике разноврсности података, који превазилази могућности уобичајено коришћеног софтвера за складиштење, обраду и управљање подацима.



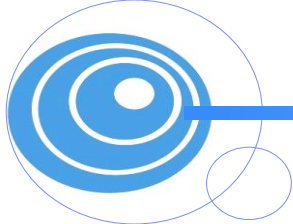


Зашто Big Data технологије?

- Са експлозијом сензора, паметних уређаја, технологија друштвеног умрежавања, подаци у једној организацији постају све комплекснији.
- Осим традиционалних релационих база података, све више су присутни сирови подаци, полуструктурирани и неструктурирани подаци са web страница, web log фајлови (укључујући click-stream податке), search indexes, social media forums, e-mail, документи, сензорски подаци са активних и пасивних система, и др.
- Процењује се да је свега око 20 процената података који се данас складиште у свету структурирано.

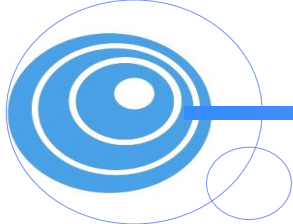


- Док количина података које организације складиште расте, проценат података које су организације способне да анализирају и употребе у свом пословању – опада! Та чињеница пружила је шансу Big Data технологијама.



Циљ предмета

- Оспособљавање студената за коришћење изабраних Big Data технологија, које омогућавају:
 - организовање и складиштење података који имају Big Data обележја,
 - примену алгоритама намењених за ефикасну обраду Big Data скупова података,
 - имплементацију метода Big Data аналитике (идентификовање односа, образаца и трендова у подацима).
- Упознавање студената са успешним случајевима коришћења Big Data технологија у саобраћајном инжењерству.



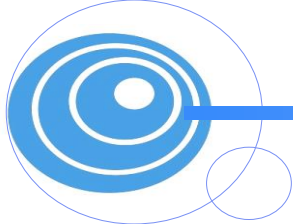
Исход предмета

- Студенти ће бити упознати са:
 - кључним технологијама које се примењују за складиштење Big Data скупова података (Apache Hadoop Distributed File System - HDFS, NoSQL базе података, језера података, ...);
 - Apache Hadoop MapReduce и Apache Spark системима за обраду Big Data;
 - библиотекама програмског језика Python које се користе за манипулацију и визуелизацију Big Data скупова података, као и Python модулом за машинско учење;
 - изабраним софтверским алатима који се користе за Big Data аналитику (Apache Hive, Weka, ...).
- По завршетку курса сваки студент ће бити способан да коришћењем одговарајућих технологија и алата самостално реализује бар два различита случаја коришћења Big Data технологија у саобраћају, транспорту или логистици.



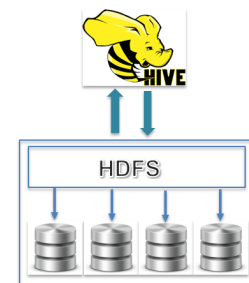
Садржај предмета – теоријски аспекти

- Појам Big Data.
- Big Data и IoT концепт.
- NoSQL базе података: кључ-вредност базе података, колонске базе података, базе докумената и графовске базе.
- Упитни језик HiveQL.
- Језера података.
- Методе Big Data интеграције.
- Виртуелизација података.
- Предиктивна анализа базирана на моделима машинског учења (алгоритми класификације, регресије и кластеровања).
- Библиотеке које служе за процесирање података и машинско учење доступне у Python-у.



Садржај предмета – практични аспекти

- Apache Hadoop алати за Big Data: HDFS, Apache Spark, Ambari.



- Нерелационе базе података: DynamoDB, Apache HBase, MongoDB, Neo4j.



- Apache HIVE складиште података.



Apache Ambari

- Denodo платформа за виртуелизацију података.

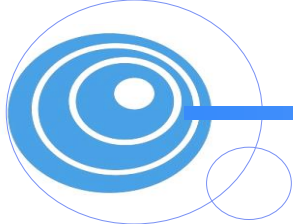


DynamoDB

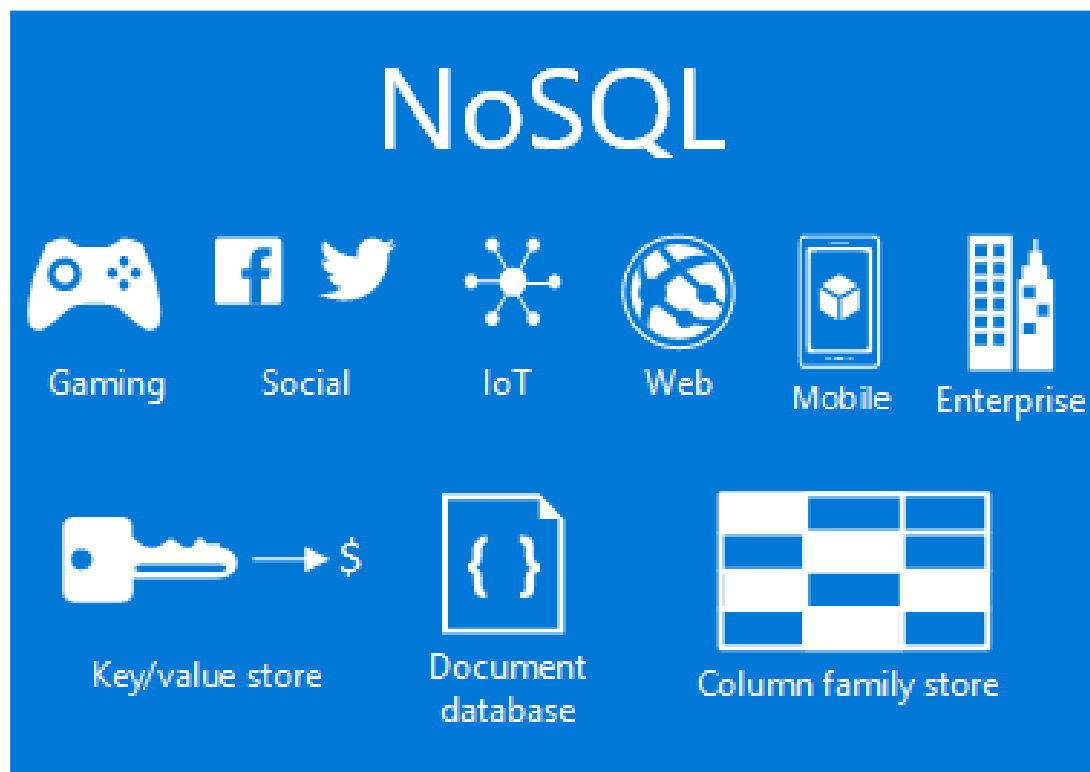
- Data mining софтвер Weka.

- Рад са Python модулом за машинско учење.

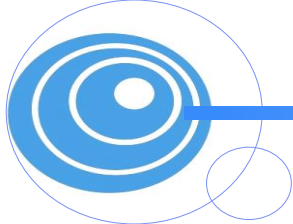




Садржај предмета

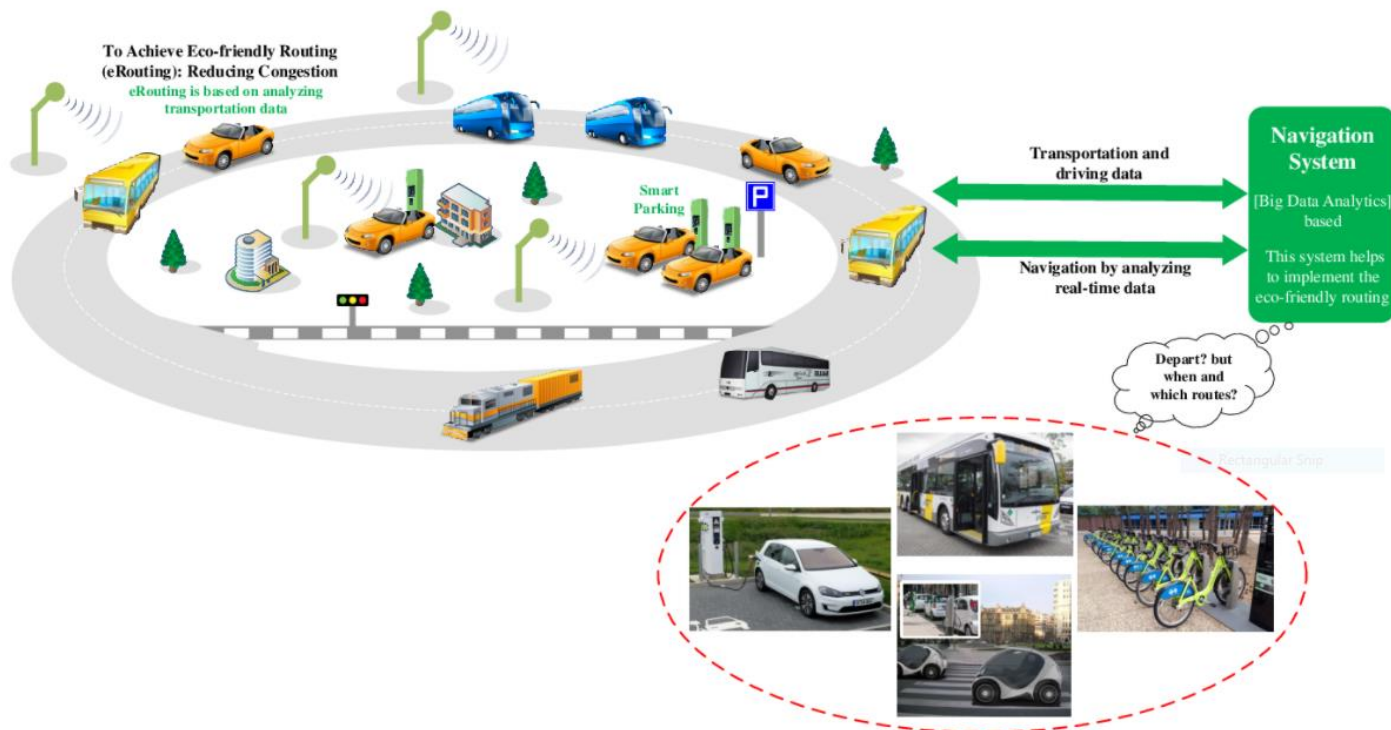


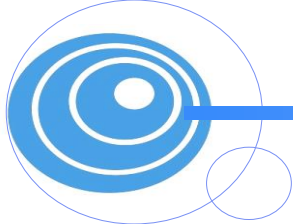
Нерелационе базе података (NoSQL)



Садржај предмета

- Улога Big Data аналитике у саобраћају





Пример нерелационе базе података у области друмског саобраћаја креиране на *Hadoop* платформи, упита над њом и визуелизације података из базе

Ambari - Sandbox

localhost:8080/#/main/views/HIVE/1.0.0/Hive

Ambari Sandbox 0 ops 2 alerts Dashboard Services Hosts Alerts Admin admin

Hive Query Saved Queries History UDFs

Database Explorer

brojanje_saobracaja

Search tables...

Databases

- brojanje_saobracaja
 - alibegovac
 - bocke
 - brojac_id STRING
 - naziv_brojaca STRING
 - x_koordinata STRING
 - y_koordinata STRING
 - redni_broj_vozila STRING
 - datum STRING
 - vreme STRING
 - kanal STRING
 - smer STRING
 - kategorija_vozila STRING
 - Lead more...
 - bulevar_evrope_3
 - kiisa
 - paragovo
 - sancevi
 - somborski_bulevar
 - tekije
 - tunel
 - vojvode_stepe

Query Editor

Worksheet

```
1 SELECT naziv_brojaca, x_koordinata, y_koordinata, smer,  
2 ROUND(AVG(brzina_vozila), 2) AS PROSEČNA_BRZINA, MAX(brzina_vozila) AS MAX_BRZINA,  
3 MIN(brzina_vozila) AS MIN_BRZINA, MAX(brzina_vozila)-MIN(brzina_vozila) AS RAZLIKA_MAX_I_MIN_BRZINA  
4 FROM bocke  
5 WHERE naziv_brojaca="Bocke" AND brzina_vozila > 10  
6 GROUP BY naziv_brojaca, x_koordinata, y_koordinata, smer;
```

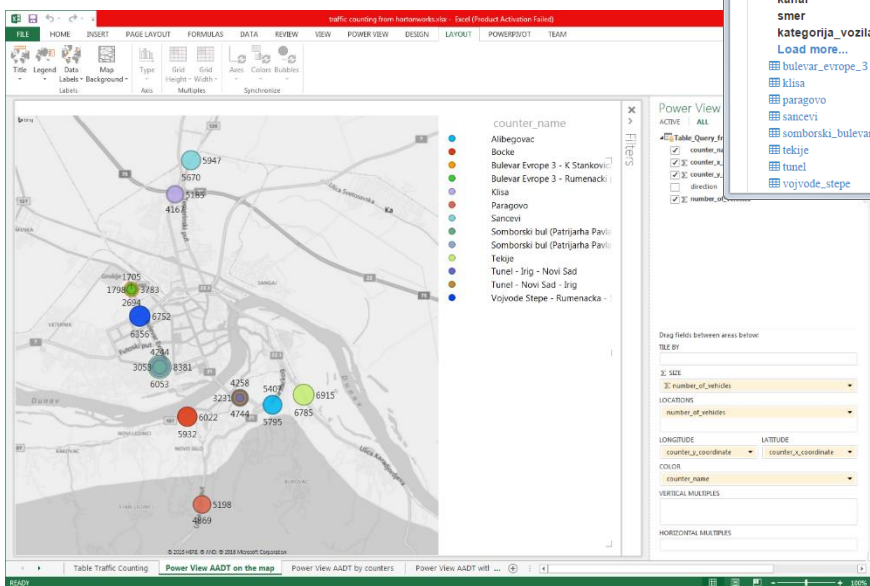
Stop execution Explain Save as... Kill Session New Worksheet

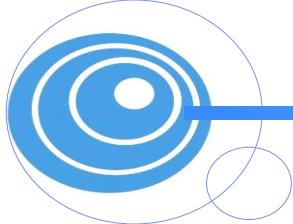
9%

Query Process Results (Status: RUNNING)

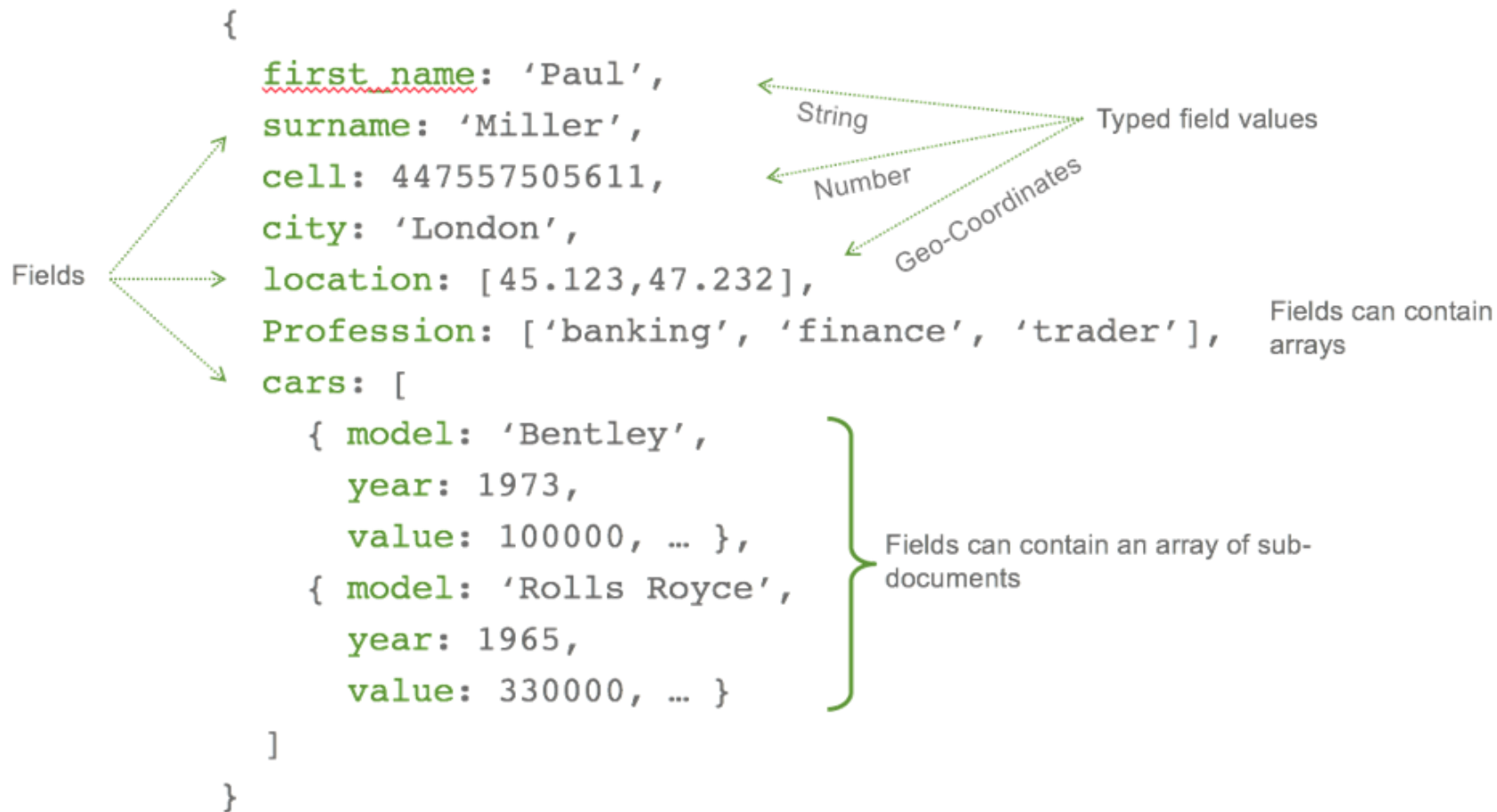
Logs Results

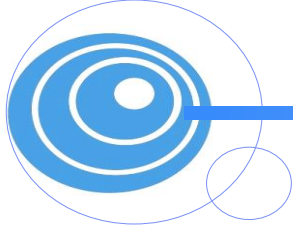
INFO: Tez session hasn't been created yet. Opening session





Пример MongoDB документа





Софтверски алат Weka – пример примене у области друмског саобраћаја

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose None | Apply | Stop

Current relation: Relation: Training dataset - PMDS, Instances: 1237, Attributes: 3, Sum of weights: 1237

Selected attribute: Name: PMDS, Missing: 0 (0%), Distinct: 1236, Type: Numeric, Unique: 1235 (100%)

Statistic	Value
Minimum	0
Maximum	28838.06
Mean	7399.954
StdDev	3823.469

Class: PMDS (Num) | Visualize All

Attributes: All | None | Invert | Pattern

No.	Name
1	brojac
2	mesec
3	PMDS

Status: OK | Log

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: Choose RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options: Use training set, Supplied test set, Cross-validation Folds 10, Percentage split % 66

Classifier output:

```
| mesec = oktobar : 15773.16 (5/320329.7)
| mesec = novembar : 14883.44 (5/391495.92)
| mesec = decembar : 14030.89 (5/771358.02)
```

Size of the tree : 483

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.9858
Mean absolute error	334.8252
Root mean squared error	643.0268
Relative absolute error	11.1024 %
Root relative squared error	16.8127 %
Total Number of Instances	1237

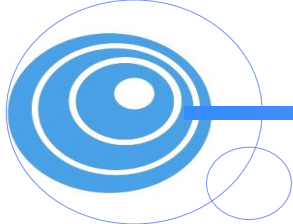
(Num) PMDS

Start | Stop

Result list (right-click for options):

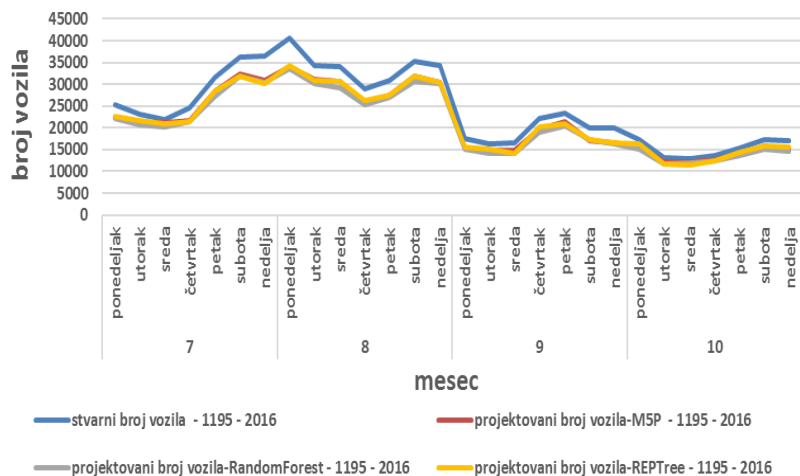
- 18:19:13 - trees.RandomForest
- 18:19:43 - trees.RandomTree
- 18:20:15 - trees.REPTree

Status: OK | Log

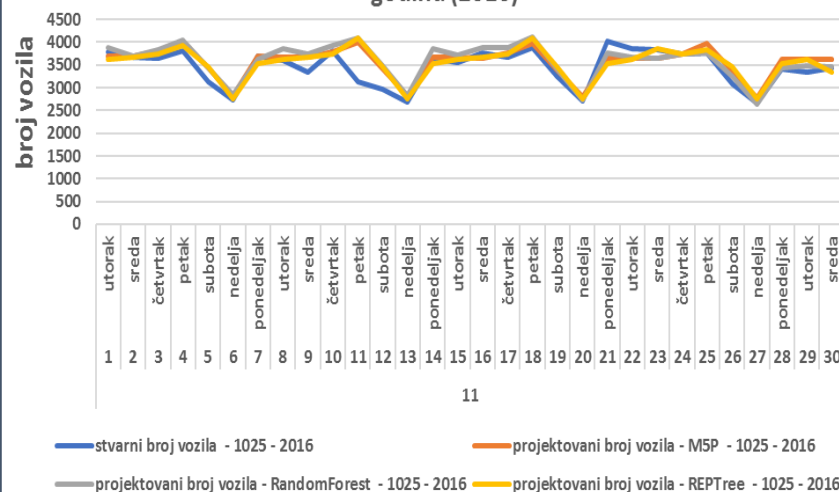


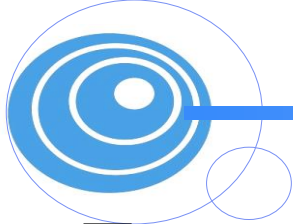
Примери резултата предиктивне анализе у области друмског саобраћаја

Укупан mesečni стварни и пројектовани проток возила по данима у недељи за изабрани бројач саобраћаја (ID: 1195), изабрану godinu (2016) и изабране mesece (jul, avgust, septembar i oktobar)



Стварни и пројектовани дневни проток возила за изабрани бројач саобраћаја (ID: 1025), изабрани месец (novembar) и изабрану godinu (2016)

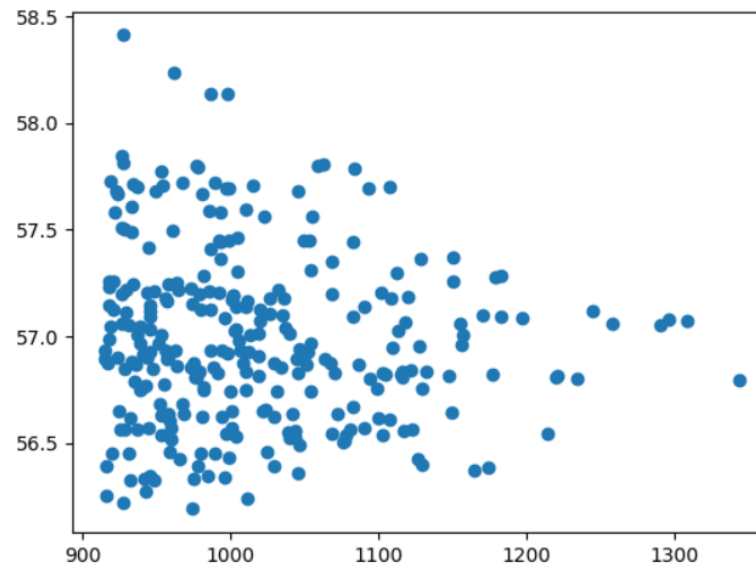


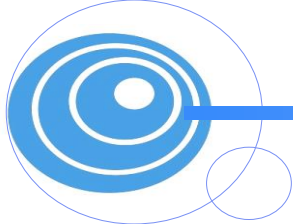


Пример употребе Matplotlib библиотеке у Python-у

```
import pandas as pd
import matplotlib.pyplot as plt

dataframe = pd.read_csv("scottish_hills.csv")
x = dataframe.Height
y = dataframe.Latitude
plt.scatter(x, y)
plt.show() # or plt.savefig("name.png")
```





Провера знања

1. Семинар (обавезан):

- креирање нерелационе базе података и упита над њом на платформи по избору студента,
- 30 бодова.

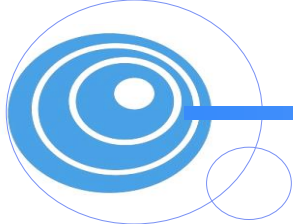
2. Пројектни задатак (обавезан):

- предиктивна анализа применом технике машинског учења у софтверском алату **Weka**,
- 40 бодова.

3. Завршни испит (није обавезан):

- обрада, анализа и визуелизација података коришћењем програмског језика **Python**.
- 30 бодова.

- Сва три дела испита подразумевају самосталан рад студента уз консултације са предметним наставницима, а сваки део може бити у форми семинарског/пројектног рада или у форми публикованог научно-стручног рада студента у сарадњи са предметним наставницима.



Литература

1. Sridhar Alla, *Big Data Analytics with Hadoop 3*, Packt, Birmingham – Mumbai 2018.
2. Tom White, *Hadoop: The Definitive Guide*, O'Reilly Media, 2015.
3. Ian Witten, Eibe Frank, Mark Hall, Christopher Pal, *Data Mining: Practical Machine Learning Tools and Techniques, 4th edition*, Morgan Kaufmann, 2016.
4. Doug Bierer, *MongoDB 4 Quick Start Guide*, Packt, Birmingham – Mumbai 2018.
5. Ivan Marin, Ankit Shukla, et al., *Big Data Analysis with Python*, Packt, Birmingham – Mumbai, 2019.
6. Научно-стручни радови и пројекти наставника носилаца предмета.