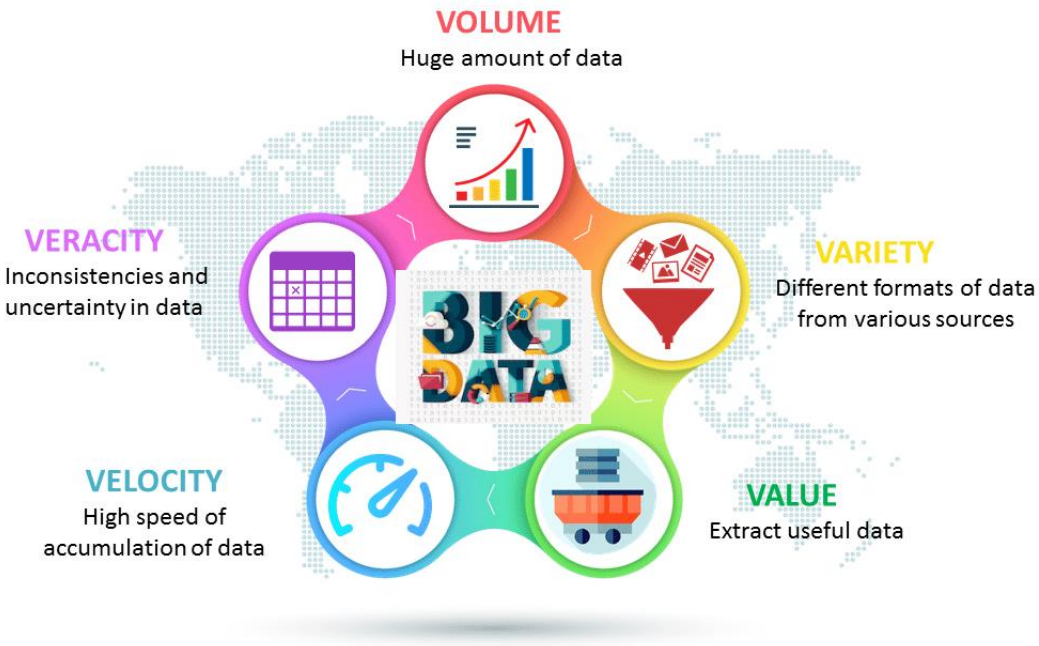
Big Data технологије у саобраћају, транспорту и логистици

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, полуструктуирани и неструктуирани 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 структуирано.

- 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

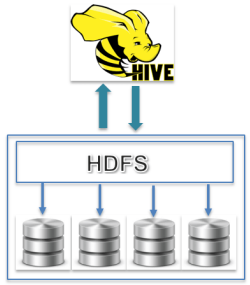
- Студенти ће бити упознати са:
  - кључним технологијама које се примењују за складиштење Big Data скупова података (Apache Hadoop Distributed File System - HDFS, NoSQL базе података, језера података, ...);
  - Apache Hadoop MapReduce и Apache Spark системима за обраду Big Data;
  - библиотекама програмског 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

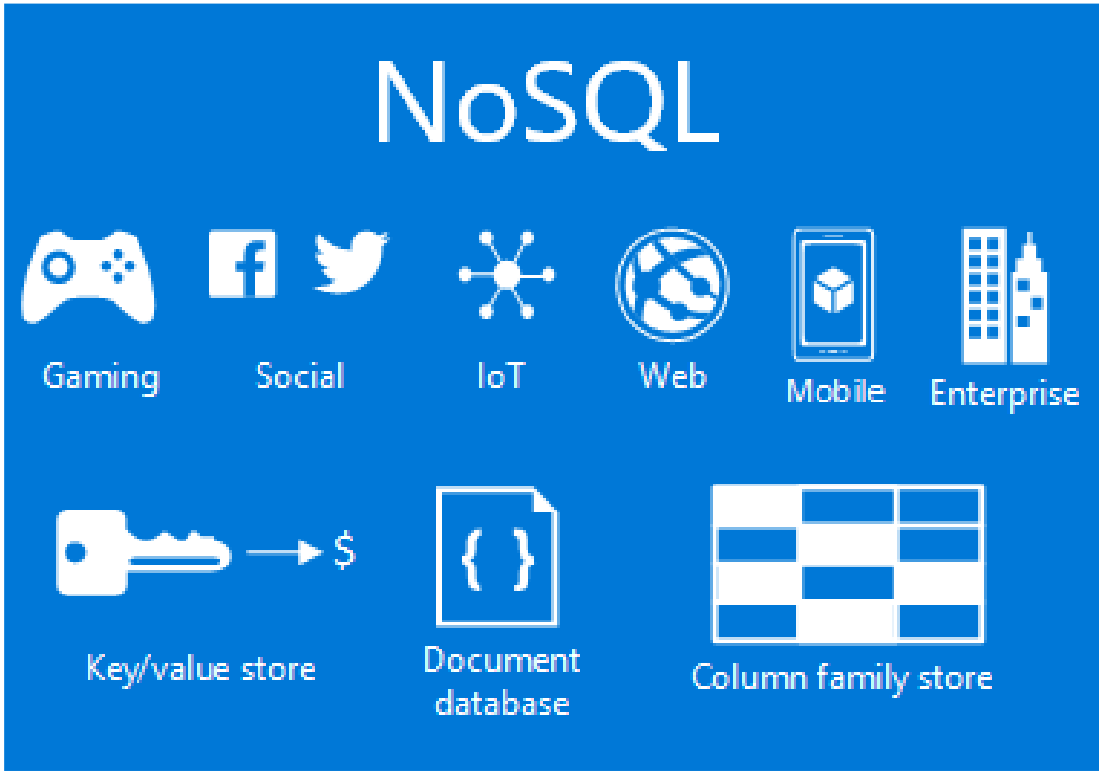
- Појам Big Data.
- Big Data и IoT концепт.
- NoSQL базе података: кључ-вредност базе података, колонске базе података, базе докумената и графовске базе.
- Упитни језик HiveQL.
- Језера података.
- Методе Big Data интеграције.
- Виртуелизација података.
- Предиктивна анализа базирана на моделима машинског 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.
- Denodo платформа за виртуелизацију података.
- Data mining софтвер Weka.
- Рад са Python модулом за машинско учење.



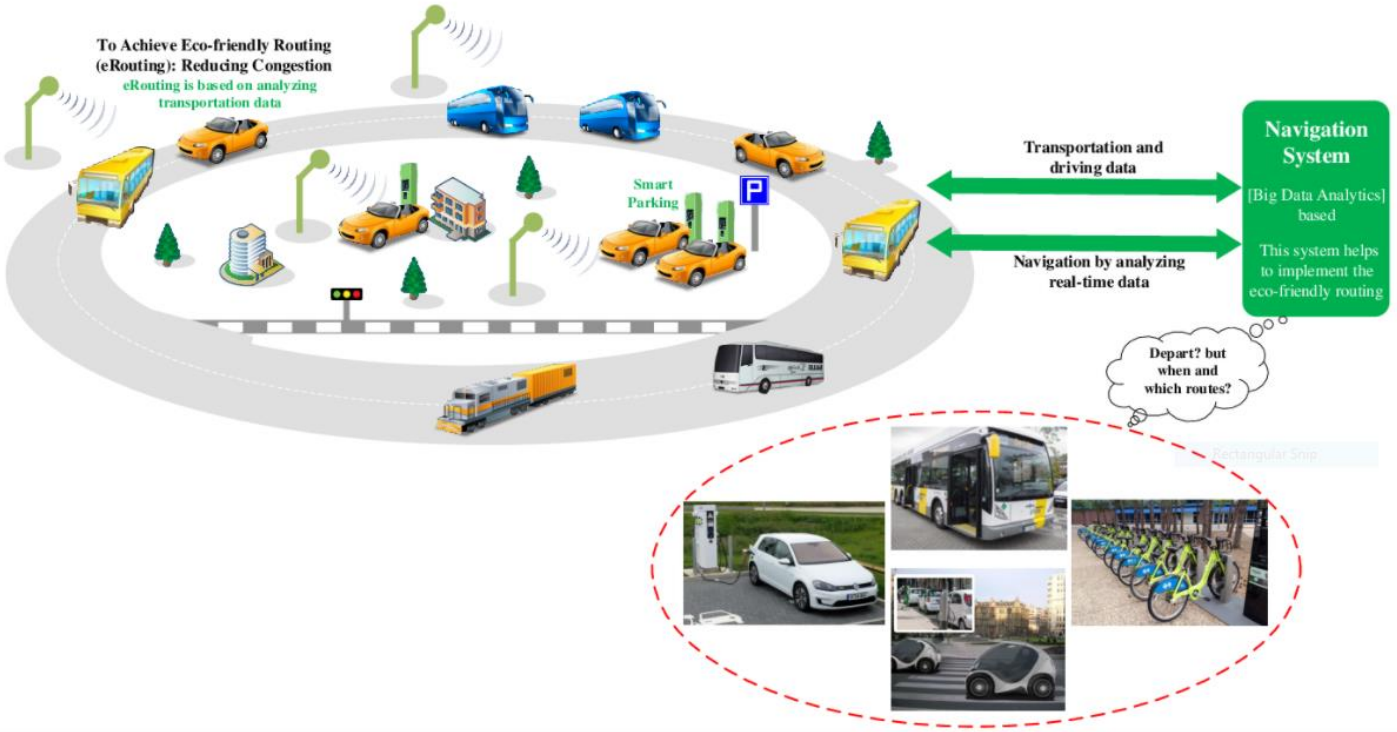
Садржај предмета



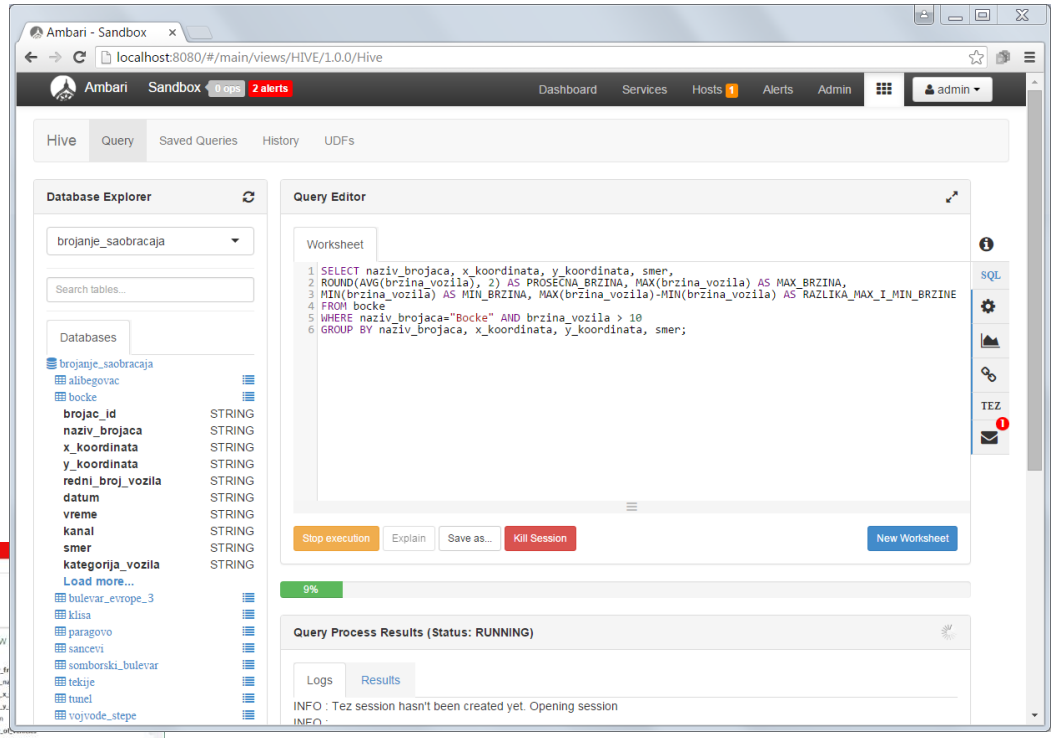
Нерелационе базе података (NoSQL)

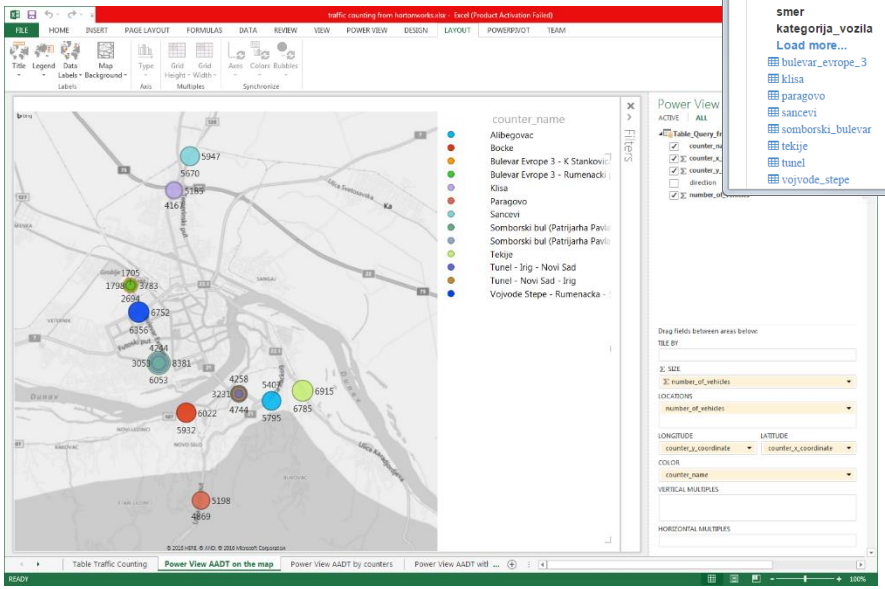
Садржај предмета

- Улога Big Data аналитике у саобраћају

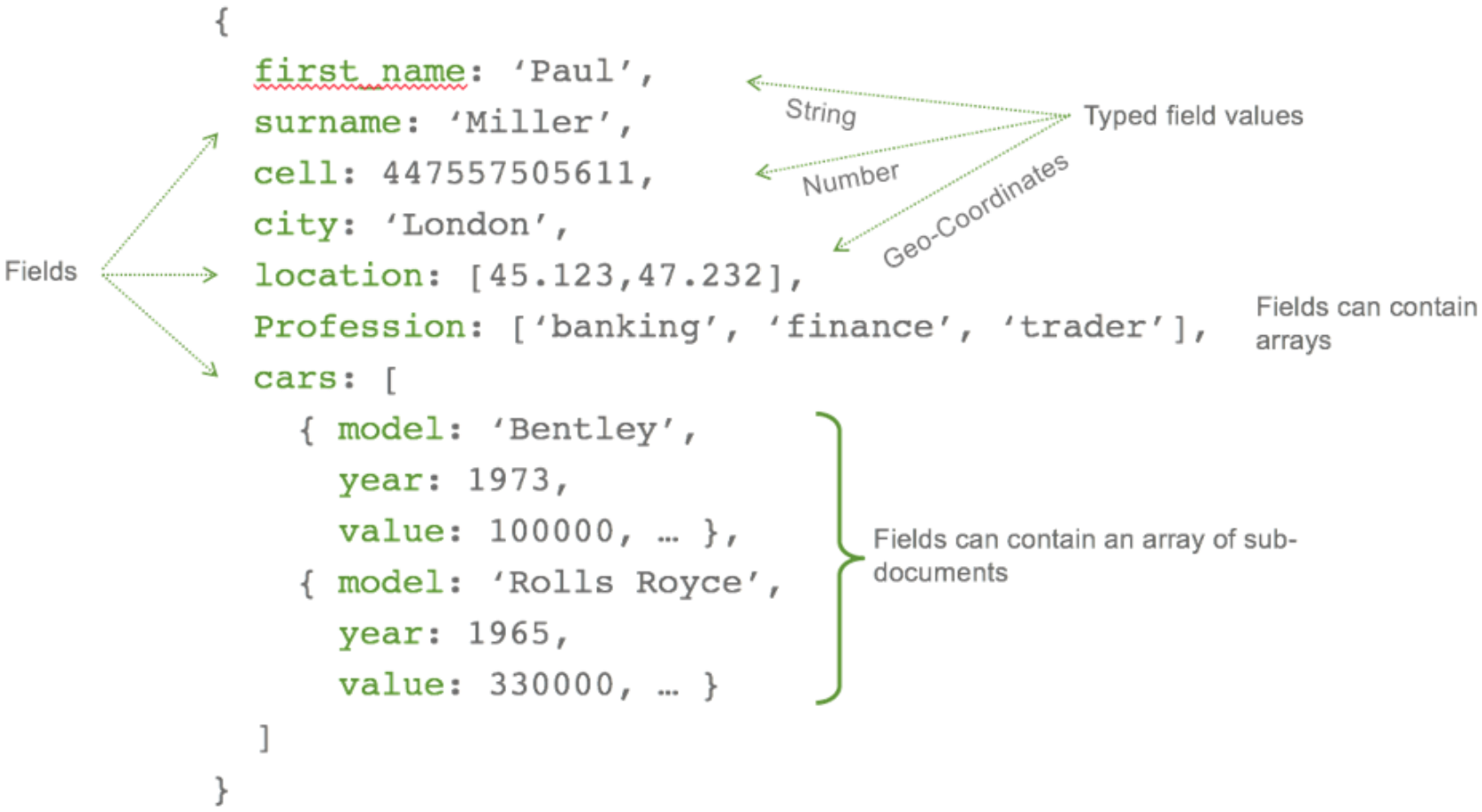


Пример нерелационе базе података у области друмског саобраћаја креиране на *Hadoop* платформи, упита над њом и визуелизације података из базе

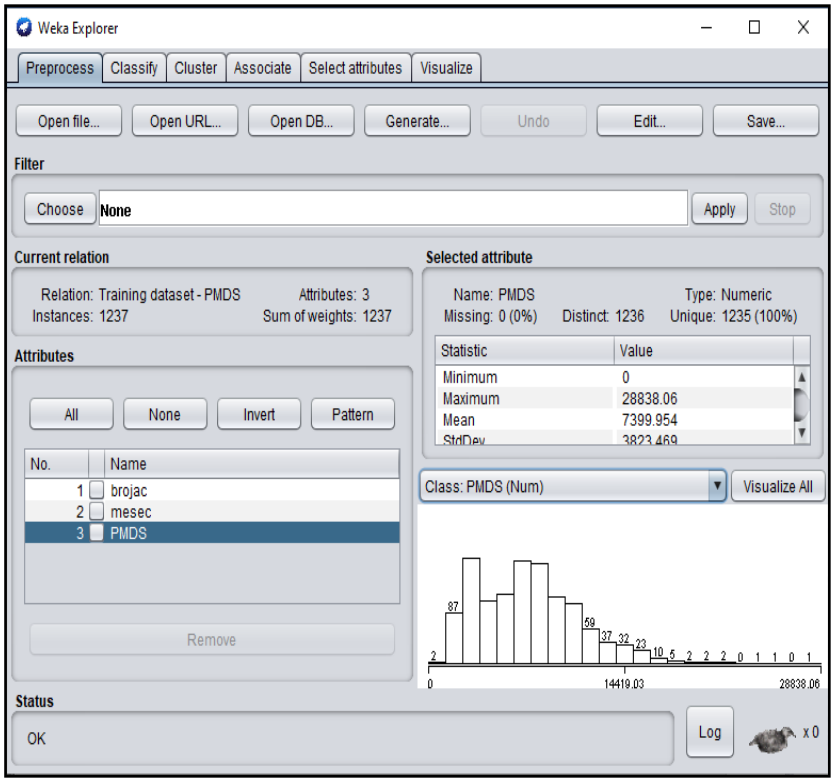


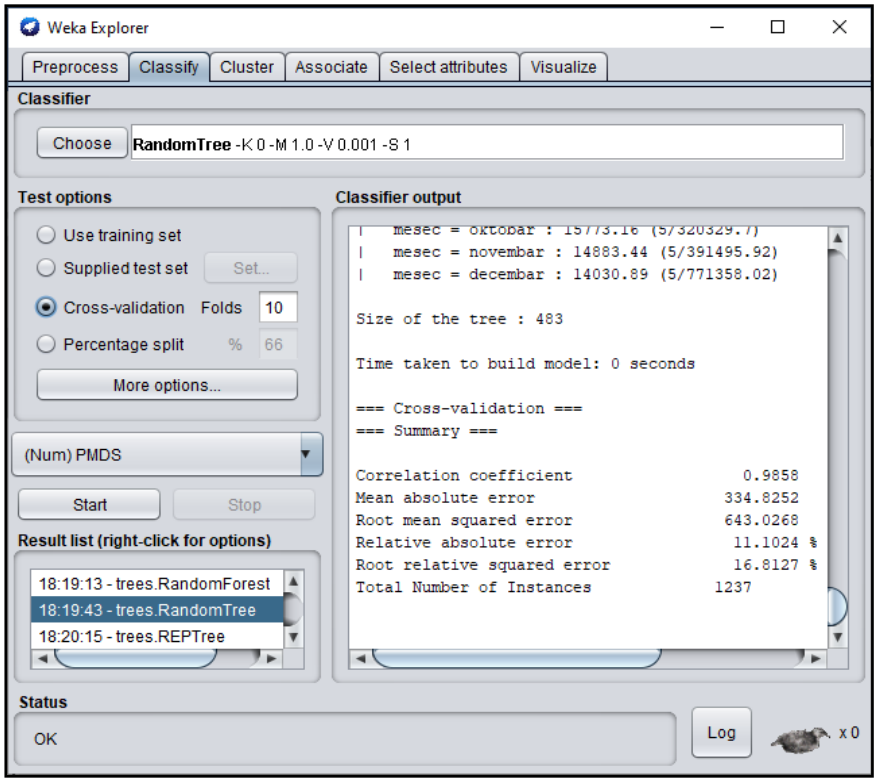


Пример MongoDB документа

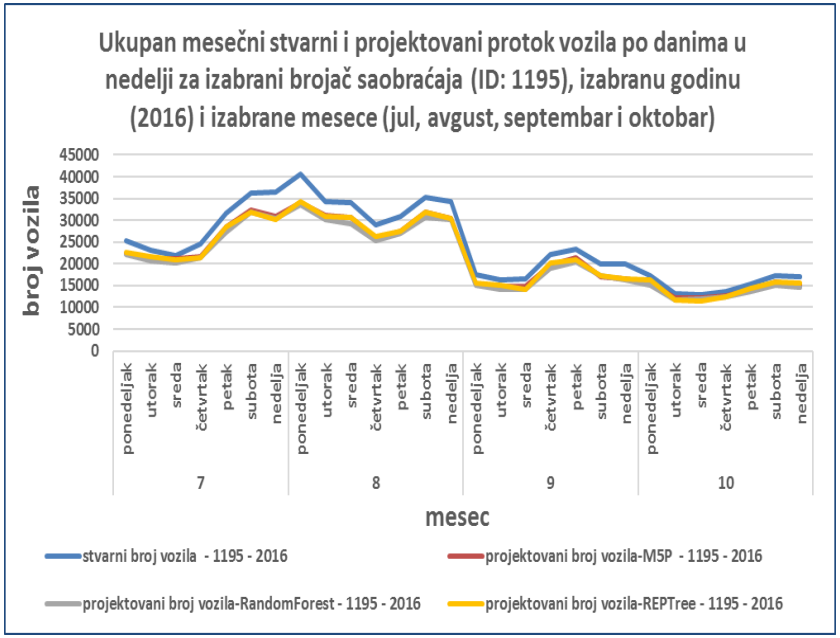


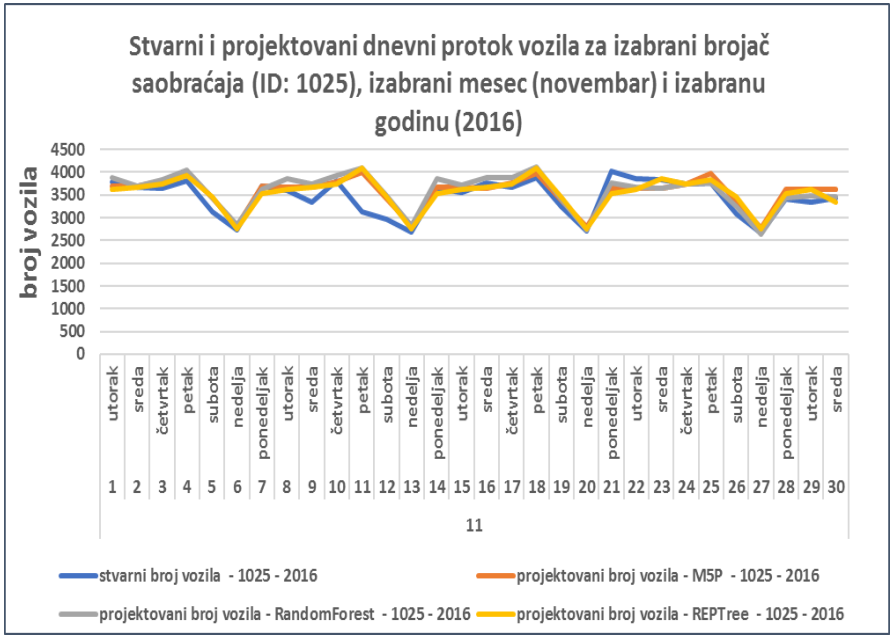
Софтверски алат Weka – пример примене у области друмског саобраћаја





Примери резултата предиктивне анализе у области друмског саобраћаја

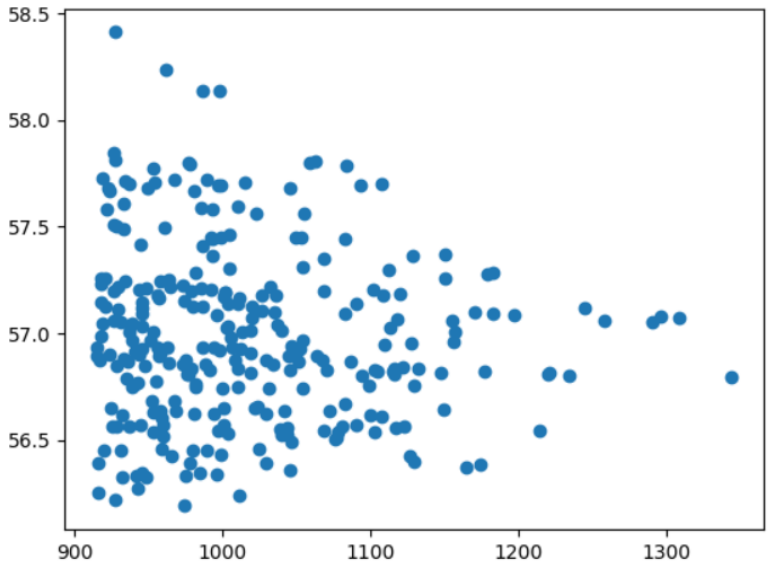




Пример употребе Matplotlib библиотеке у Python-у

```
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.