Adatelemzési Plaformok

(Data Analysis Platforms - DAP)

2018. tavaszi félév

Dmlab, BME-TMIT

[Tárgyinformációk](#_unbv6upox4mm)

[Órák időpontja](#_f8hzlnkiylds)

[Hivatalos tantárgyi adatlap](#_7rl1399ag8e2)

[Legutóbbi változások az oldalon](#_tn8iqs4179y7)

[Előadók és anyagokat kidolgozó oktatók](#_n5dew6ba037p)

[Követelmények](#_fchg25a4v1s)

[Számonkérés információk](#_47jjgtocg5hm)

[Eredmények](#_75d1kuyt8xv0)

[Előadások anyagai és időbeosztás](#_npwfc1gcncje)

[Telepítési útmutató](#_hhzjody71pq2)

[Nagyházi feladat](#_ttke6x66apiw)

[Kisházi feladatok](#_wgr8haguudmp)

[1. gyakorlat - 2017-02-13](#_vzng4w1pglyb)

[Python gyakorló anyagok](#_f2qm052e6wj4)

# **Tárgyinformációk**

## **Órák időpontja**

* Kedd 8:30-10:00 - QB104
* Szerda (**páros heteken**) 10:15-12:00 - QBF10

## **Hivatalos tantárgyi adatlap**

* [A BME VIK oldalán](https://portal.vik.bme.hu/kepzes/targyak/VITMMA05/)

## **Legutóbbi változások az oldalon**

* 2018. 02. 05. - tárgyinformációk és féléves ütemterv, telepítési útmutatók

## **Előadók és anyagokat kidolgozó oktatók**

Gáspár Csaba - [gaspar@tmit.bme.hu](mailto:gaspar@tmit.bme.hu) - előadó

Nagy-Rácz István - [nagy.istvan@tmit.bme.hu](mailto:nagy.istvan@tmit.bme.hu) - előadó

A tárgy slack csatornája: <http://dmlab-ds-courses.slack.com> / dap2018-as csatorna

# **Követelmények**

* utolsó héten zárthelyi
* félév közben egy házi
* 50 pont ZH és 50 pont nagyházi, a két pont összege adja meg a végső jegyet
* 6 gyakorlat lesz a félévben, mindegyik után kiadunk egy kis feladatot, aki megcsinálja kap a ZH pontszámához 4-4 pontot (így összesen 24 pont elérhető)
* aki jegyzetet ír, az is szerezhet pontokat a ZH-ban
* a nagyházira adott pontokat az alábbiak szerint osztjuk ki: vesszük a baseline megoldásunk és a legjobb megoldás visszamérési függvény szerinti értékének intervallumát
  + a legalsó felében lévők kapnak 10 pontot
  + a következő negyedben lévők 20 pontot
  + a következő nyolcadban lévők 30 pontot
  + a következő tizenhatodban lévők 40 pontot
  + végül a legfelső tizenhatodban lévők 50 pontot

# **Számonkérés információk**

|  |  |
| --- | --- |
| Zárthelyi |  |
| Pótló zárthelyi |  |
| Házi feladat |  |

# **Eredmények**

Az eredmények összefoglalása fokozatosan követhető lesz itt.

# **Előadások anyagai és időbeosztás**

* Több anyag együttes lelőhelye: [Google Drive](https://drive.google.com/open?id=1Da0-EE0hUspmyalg5Xq9LMCpXBHnIZzs) (közvetlenül az egyes előadásokhoz és gyakorlatokhoz linkelve alább)
* Opcionálisan feldolgozandó, ajánlott olvasmányok
  + Daniel Larose and Chantal Larose: Discovering Knowledge in Data: An Introduction to Data Mining, 2nd edition (Wiley, 2014) / Daniel Larose and Chantal Larose: Data Mining and Predictive Analytics, 2nd edition (Wiley, 2015)
  + [Pang-Ning Tan, Michael Steinbach és Vipin Kumar: Bevezetés az adatbányászatba (Panem Kiadó, 2011)](http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/index.html)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dátum** | **Nap** | **Téma** | **Előadó** | **Online anyagok** |
| 2018-02-06 | kedd | Felvezető, CRISP-DM, adattípusok, adatmodell, churn példa | István | [Az analitika 8 szintje](https://drive.google.com/open?id=10HN4Sawmjbuz3r2pY2iA_dCMy3_wiwuN) |
| 2018-02-07 | szerda |  |  |  |
| 2018-02-13 | kedd | RapidMiner és Python bevezető és felhozó feladatkiadás | István | [A gyakorlat anyaga (RM)](https://drive.google.com/open?id=1-WMLvu5ruteEzc8VGU-90E25_Fmeplju)  [Kaggle verseny linkje](https://www.kaggle.com/c/titanic) |
| 2018-02-14 | szerda | Hitelbírálati rendszer bemutatása, logisztikus regresszió mint scoring technika, mintavételezés. | István |  |
| 2018-02-20 | kedd | Hitelbírálati feladat adatbányászati megoldása (RapidMiner) | István |  |
| 2018-02-21 | szerda |  |  |  |
| 2018-02-27 | kedd | Ügyfélérték fogalma, lineáris regresszió | István |  |
| 2018-02-28 | szerda | Felügyelt és nem felügyelt feladatok szétválasztása, ügyfélszegmentálás, klaszterező eljárások, k-közép és k-medoid algoritmus | István |  |
| 2018-03-06 | kedd | Klaszterezés és outlier keresés: hasonlósági és távolsági mértékek, particionáló módszerek, hierarchikus klaszterezők, sűrűség alapú klaszterezők, outlier keresési technikák | István |  |
| 2018-03-07 | szerda |  |  |  |
| 2018-03-13 | kedd | Adatelőkészítési módszerek: adattisztítási módszerek, adatintegrációs és transzformációs technikák, adatredukciós módszerek, diszkretizációs technikák. | István |  |
| 2018-03-14 | szerda | Adattranszformációs és adatmanipulációs alapfogalmak és vizuális adatelemzés: adattípusok, adatelemzési problémák áttekintése, visszamérési módszerek. | István |  |
| 2018-03-20 | kedd | Távközlési cég ügyfeleinek elvándorlás (churn) előrejelzése (RapidMiner) | István |  |
| 2018-03-21 | szerda |  |  |  |
| 2018-03-27 | kedd | Osztályozási problémák megoldása: döntési fák, példányalapú mószerek. | István |  |
| 2018-03-28 | szerda | Kampányoptimalizáció biztosítási környezetben (Python) |  |  |
| 2018-04-10 | kedd | Kombinált adatbányászati eljárások, együttes osztályozók |  |  |
| 2018-04-11 | szerda |  |  |  |
| 2018-04-17 | kedd | Idősoros adatok feldolgozása: lineáris és nem-lineáris módszerek, regressziós fák |  |  |
| 2018-04-18 | szerda | Vásárlói kártya adatok adatbányászati feldolgozása (Python) |  |  |
| 2018-04-24 | kedd | A nagy adat (Big Data) jelensége és fogalma, szerepe. Az Apache Hadoop platform bemutatása. |  |  |
| 2018-04-25 | szerda |  |  |  |
| 2018-05-02 | szerda | Elosztott adattárolás és elemzések MapReduce alapokon. MapReduce programozási minták. Magasszintű nyelvek MapReduce felett (Hive, Pig) //// Alternatíva: SPARK direktíva |  |  |
| 2018-05-08 | kedd | Keresztértékesítés (Python) |  |  |
| 2018-05-09 | szerda |  |  |  |
| 2018-05-15 | kedd | Nagyházi feladat konzultáció |  |  |
| 2018-05-16 | szerda | ZH |  |  |
|  |  |  |  |  |

# Telepítési útmutató

## RapidMiner

1. <https://rapidminer.com/> Download
2. Operációs rendszer választás
3. A telepítő fájl letöltése és az instrukciók alapján feltelepítés
4. Egyetemi email címmel rendelkezők esetén: <https://rapidminer.com/educational-program/>

## Python

1. A Pythont futtató környezetet [az alábbi linken](https://notebook.dmlab.hu/tree/DAP2017) éritek el.
2. Jelszó: dataBook018
3. Mindenki a DAP2017 mappába és azon belül is a gyakorlathoz tartozó almappába dolgozzon. Az ezen kívül létrehozott notebookokat töröljük.
4. Amennyiben valaki mégis telepíteni szeretné saját gépre a pythont, úgy az [Anaconda disztribúcióját](https://www.continuum.io/downloads) javasoljuk.

# Nagyházi feladat

# Kisházi feladatok

## 1. gyakorlat - 2017-02-13

1. A feladat megoldásának határideje (eddig várjuk az e-maileket): **2018-02-20 12:00**
2. Csináld végig a RapidMiner - Basics tutorialokat
   1. Indítsd el a RapidMinert
   2. A kezdő képernyőn válaszd ki a Learn lehetőséget, vagy a szerkesztési nézetben a Questions? alatt a Tutorials menü pontot.
   3. Válaszd ki az első (Accessing Data) tutorialt és csináld végig: olvasd el a leírásokat és hozd létre a megfelelő adatelemzési folyamatot.
   4. Mentsd el, vagy exportáld a processzed.
   5. Csináld végig a maradék tutorialokat is.
   6. Az így létrejött rmp kiterjesztésű processzeket küldd el a következő [címre](mailto:nagy.istvan@tmit.bme.hu?subject=%27DAP%201.%20kisházi%27).
3. Csináld végig a következő [python tutorialt](https://learnxinyminutes.com/docs/python3/).
   1. Csinálj egy [notebookot](#_fpm9updpt79i).
   2. Kövesd végig a tutorialt és a kódokat próbáld ki.
   3. Küldd át a notebookod elérhetőségét a következő [címre](mailto:nagy.istvan@tmit.bme.hu?subject=%27DAP%201.%20kisházi%27).

# Python gyakorló anyagok

JAVASOLT ISMÉTLÉSHEZ, UTOLÉRÉSHEZ

Gyors túra python 3-ban, javaslom akár az órai anyag átnézésére, helyettesítésére. Lényegében maga a tutorial is egy python kód, a kommentben vannak a mondatok, utána csupa példa az egész. Nagyon hasonlít a szellemisége arra, ahogy mi haladtunk.

- <https://learnxinyminutes.com/docs/python3/>

HA MÉLYEBBEN BELE AKARSZ MENNI

Learn Python on the Hard Way - Nagyon részletes, lépésről lépésre haladó bevezető, annak javaslom, aki önállóan akar megtanulni mindent úgy, hogy a tutorial magyaráz is neki.

- <http://learnpythonthehardway.org/book/>

Lassan, az alapoktól felépülő, de nem mindent elmondani akaró tutorial:

- <https://en.wikibooks.org/wiki/Non-Programmer%27s_Tutorial_for_Python_3>

Egy igen részletes tutorial a Python 3 alapjaihoz, inkább referenciaként javasolt használni, vagy ha egy terület jobban érdekel (hirtelen szükség van rá), akkor érdemes lehet megnézni az adott fejezetet

- <https://docs.python.org/3/tutorial/>

- <http://www.python-course.eu/python3_course.php>

GYAKORLÁSI LEHETŐSÉGEK

Játékos formában ad lehetőséget a gyakorlásra a [CheckIO](https://checkio.org/): nagyon egyszerű és gyorsan megoldható feladatokkal, ellenőrzéssel, böngészőbe beépülő fejlesztő környezettel gyors sikerélményekhez juttat, azt kell mondjam, hogy egyenesen szórakozásból is érdemes lehet forgatni.

Egy játék, amiben egy-egy feladatot kell megoldani python-nal. Minden feladat valójában egy rejtvény, sokat kell törni a fejedet, hogy továbbjuss, de így a pythont lényegében a játék kedvéért tanulod meg, és próbálod ki.

<http://www.pythonchallenge.com/>

Egy másik gyakorlási lehetőség, ha egy-egy mini feladatot oldasz meg. Az alábbi platform kis függvényeket ad fel, lehet python-ban is kódolni. Itt az a lényeg, hogy egy függvényt kell fejleszteni, ami azt csinálja, amit a feladat leír, de úgy, hogy te írsz teszteket az adott függvényhez, és néha tesztelheted a megoldásod. Különösebb installálás nélkül használható.

<http://cyber-dojo.org/>

KÖNYVEK

Ha jobban szereted a könyveket, akkor az alábbiakat ajánljuk.

[Python for Data Analysis](http://shop.oreilly.com/product/0636920023784.do)

[Python Data Science Handbook](http://shop.oreilly.com/product/0636920034919.do), itt plusz poén az, hogy az összes [a könyvben lévő példához tartozó jupyter notebook](https://github.com/jakevdp/PythonDataScienceHandbook) ingyenesen elérhető. A notebookok letölthetőek, azokat nyugodtan töltsétek fel a mi szerverünkre és gyakoroljatok ott. Ehhez hozzatok létre egy saját mappát a DS 2 almappában. Amennyiben valaki mégis telepíteni szeretné saját gépre a pythont, úgy az [Anaconda](https://www.continuum.io/downloads) disztribúcióját javasoljuk.

Egyébként minden O'Reilly könyv jó lehet, ha szeretnénk tematikusan, témáról-témára elmélyülni a python lehetőségeiben