

## Tárgytematika / Course Description

### Adatelemzés

**GKLM\_MSTM025****Tárgyfelelős neve /****Teacher's name:** dr. Harmati István**Félév / Semester:** 2021/22/2**Beszámolási forma /****Assesment:** Vizsga**Tárgy heti óraszám /****Teaching hours(week):** 0/0/0**Tárgy féléves óraszám /****Teaching hours(sem.):** 12/0/0

---

### OKTATÁS CÉLJA / AIM OF THE COURSE

A tárgy célja olyan módszerek, megoldások, szoftver rendszerek ismertetése, amelyek számítógépen megvalósítható adatelemzést tesznek lehetővé. Az erre vonatkozó gépi megoldások kiterjednek az üzleti döntési folyamatok támogatására, adatfeltárás, adatkutatás automatizálására.

---

### TANTÁRGY TARTALMA / DESCRIPTION

- Az adatelemzés ill. gépi tanulás fogalma, célja, folyamata. Néhány látványosabb alkalmazás. A gépi tanulás alapvető feladatai: Osztályozás, regresszió, klaszterezés. Matematikai alapok átismétlése.
- Python programozási alapok: A nyelv jellemzői, egyszerű adattípusok, kollekción, vezérlési szerkezetek.
- Python programozási alapok: Comprehension-ök, kicsomagolás, haladó indexelés és iterálás, függvények, fájlkezelés.
- A NumPy numerikus számítási csomag. Tömbök létrehozása, résztömbök, műveletek, broadcasting. Egyváltozós lineáris regresszió.
- A K legközelebbi szomszéd algoritmus. Tesztelés a Római Helyszínelők feladaton.
- A pandas adatelemző csomag. Series és DataFrame adatszerkezet. CSV fájlok betöltése. Légszennyezettségi adatok elemzése.
- Többváltozós lineáris regresszió. Tesztelés a Boston Housing adathalmazon. A scikit-learn alapjai. Keresztkiértékelés.
- Logisztikus regresszió. Egy- és többváltozós logisztikus eset. Tanítás Newton-módszerrel. Tesztelés a Wisconsin Breast Cancer adathalmazon.
- A túltanulás jelensége. L1 és L2 regularizáció. Ritka mátrixok. Regularizált ritka logisztikus regresszió tesztelése az SMS Spam adathalmazon.
- Neurális hálózatok. A többrétegű perceptron modell. Tanítás sztochasztikus gradiens módszerrel, tesztelés a Phishing Websites adathalmazon.
- Döntési fák. A döntési tönk és a döntési fa modell, tanítás "brute force" módszerrel, tesztelés a Boston Housing adathalmazon.
- Ensemble módszerek. Véletlen erdő, gradient boosting. Tesztelés a Boston Housing adathalmazon.
- Klaszterezés. A K-means algoritmus. Adatvizualizáció. A t-SNE algoritmus.

---

### SZÁMONKÉRÉSI ÉS ÉRTÉKELÉSI RENDSZERE / ASSESMENT'S METHOD

A tárgy számítógépes vizsgával zárul, ahol a hallgatóknak egyszerű adatelemzési feladatokat kell megoldaniuk Python nyelven. A vizsgán rendelkezésre álló idő 90 perc. Ponthatárok: 21-24: jeles, 18-20: jó, 15-17: közepes, 12-14: elégséges.

---

## **KÖTELEZŐ IRODALOM / OBLIGATORY MATERIAL**

- Ketskemény László-Izsó Lajos-Könyves Tóth Előd: Bevezetés az IBM SPSS Statistics program-rendszerbe, Artéria Stúdió Kft. Budapest 2011, ISBN 978-963-08-1100-2
- Jiawei Han-Micheline Kamber: Adatbányászat, Konceptiók és technikák PANEM 2004 ISBN: 963 545394
- Dr. Abonyi János: Adatbányászat COMPUTERBOOKS 2006 ISBN: 963 6183422
- Dr. Bodon Ferenc: Adatbányászati algoritmusok, BME jegyzet