

Kiberfizikai Rendszerek Intézet			Mintatanterv szerinti 2. félév 2024-25-2			
Tantárgy neve:	Kódja:	Kredit:	Óraszám			
				ea	tgy	lab
Adatelőkészítés és -feldolgozás	NKXAF1HMLF	5	levelező féléves	10	0	10
Tárgyfelelős: Balászné Dr. Kail Eszter			Beosztás: egyetemi adjunktus			
Oktató(k): Angyalné Dr. Alexy Márta, Gerse Ágnes						
Előtanulmányi feltételek:						
Számonkérés módja:	évközi jegy					
A tananyag						
Oktatási cél:	A tantárgy célja, hogy a hallgatók megismerjék az adatelőkészítési, -előfeldolgozási és adattranszformációs feladatok célját, módszereit, szükségességének okait az adatmodellek és adatkiértékelési folyamatok megfelelő működéséhez. A tárgy további célja az adatok gyűjtési, tárolási és betöltési kérdéseinek elemzése is. Az előadásokon bemutatott módszerek elméleti megközelítése és az adatelőkészítés gyakorlati jelentőségének ismertetése után a gyakorlatokon a hallgatók ezek Python programozási nyelven történő implementációját sajátíthatják el.					
Tematika:	A tárgy tematikája az adatok előfeldolgozásának négy fő lépése (adattisztítás, adattranszformáció, adatintegráció, és adatredukció) köré rendezve a következő témákat foglalja magába: CRISP-DM, adatok betöltése, adatok megismerése, hiányzó adatok, outlierek kezelése, adatmanipuláció, adattranszformáció, normalizálás, aggregáció, dimenziócsökkentés/-növelés, adattömörítés, diszkretizálás, adatminőség-biztosítás, valamint ezen feladatok megvalósításához használható eszközök ismertetése.					

Féléves ütemezés	
Oktatási hét (konzultáció)	Témakör
1.	Elmélet: CRISP-DM módszertan általános bemutatása, a probléma felismerésének és az ahhoz releváns adatok gyűjtésének szempontjai. Gyakorlat: Pandas modul, Pandas DataFrame, főbb adattípusok.
2.	Elmélet: A CRISP-DM módszertan adatelőkészítésre, modellezésre és a modellek értékelésének és alkalmazásának gyakorlati kérdései. Gyakorlat: adatbetöltés különféle adatforrásokból (relációs adatbázisok és egyéb gyakori forrástípusok).
3.	Elmélet: Leíró statisztika. Adattípusok, mérési szintek (nominális, ordinális, intervallum- és arányskála), idősorok jellemzői. Gyakorlat: Egyszerű leíró statisztikai feladatok Pandas környezetben. Matplotlib adatvizualizáció alapjai.
4.	Elmélet: Adattranszformáció (normalizálás, standardizálás, diszkretizálás, attribútumok származtatása, aggregáció). Gyakorlat: Adatátalakítások, adattranszformációs műveletek Pandas környezetben. One-hot encoding, skálázási műveletek.
5.	Elmélet: Adatok egyesítése, adatintegráció, adattömörítés (mintavételezés, dimenziócsökkentés). Gyakorlat: Adatintegrációs és adattömörítési feladatok Pandas környezetben.
6.	Elmélet: Adatok előfeldolgozása, adattisztítási szintek. 1. adattisztítási szint: elsődleges adatszerkezet és -formátumok kialakítása, indexek, oszlopnevek. Gyakorlat: Egyszerű adattisztítási feladatok Pandas környezetben.
7.	Elmélet: 2. adattisztítási szint: információk kinyerése, adatszerkezet, adatformátumok átalakítása. Gyakorlat: Adattranszformációk, adattisztítás SQL utasítások segítségével.
8.	Elmélet: 3. adattisztítási szint: hiányzó, kiugró és hibás értékek kezelése. Gyakorlat: Hiányzó adatok kezelése Pandas környezetben.
9.	Elmélet: Adatminőség biztosítása (adatminőségi dimenziók, mérőszámok). Gyakorlat: NumPy bevezetés, műveletek NumPy tömbökkel.

10.	Elmélet: Bevezetés a Big Data technológiákba. Gyakorlat: Adatelőkészítés és -feldolgozás Big Data környezetben, Big Data fájlformátumok.
11.	Elmélet: Batch/Stream feldolgozás Big Data környezetben. Gyakorlat: Databricks platform, Batch/Stream processing példák.
12.	Elmélet: Adatvédelmi kérdések, pszeudonimizáció és anonimizálás. Gyakorlat: Egyszerű adatelőkészítési feladatok Databricks platformon, modellezési példa bemutatása Big Data környezetben (Apache Spark MLlib).
13.	Elméleti ZH
14.	Elméleti ZH pótlás
Félévközi követelmények	
Évközi jegy / aláírás megszerzésének feltételei:	Az elméleti ZH-n legalább 36 pont elérése, valamint a féléves feladatra kapható pontszámból legalább 20 pont elérése.
Zárthelyi dolgozatok	
Oktatási hét	Témakör
13.	Elméleti ZH
14.	Elméleti ZH pótlás
Az évközi jegy kialakításának módszere (csak évközi jegyes tárgyak esetében töltendő ki)	
Az érdemjegyet meghatározó végső pontszám a következő két tétel összegéből áll össze a megadott ponthatárok szerint:	
1. Az elméleti ZH-n elért pontszám (max. 70 pont)	
2. A féléves feladatra kapott pontszám (max. 30 pont)	
Pótlás módja	
A ZH / évközi jegy / aláírás pótlásának módja:	Az elméleti ZH az utolsó héten és a vizsgaidőszak elején pótolható, a féléves feladat leadása a vizsgaidőszak elején pótolható.
Vizsga módja (csak vizsgás tantárgy esetében töltendő ki)	
Vizsgajegy kialakítása (csak vizsgás tantárgy esetében töltendő ki)	
Az egyes érdemjegyek ponthatárai:	
88 - 100 pont: jeles (5)	
76 - 87 pont: jó (4)	
64 - 75 pont: közepes (3)	
52 - 63 pont: elégséges (2)	
0 - 51 pont: elégtelen (1)	
Irodalom	
Kötelező:	Az előadásokhoz, laborgyakorlatokhoz kapcsoló prezentációk anyagai 1. Roy Jafari: Hands-On Data Preprocessing in Python, Packt Publishing, 2022, ISBN: 1801072132
Ajánlott:	1. Jake VanderPlas: Python Data Science Handbook, O'Reilly Media, 2022, ISBN: 1098121228 2. Ofer Mendelevitch, Casey Stella, Douglas Eadline: Practical Data Science with Hadoop and Spark, Addison-Wesley Professional, 2016, ISBN: 0134024141
Egyéb segédletek:	Egyéb Moodle-be feltöltött anyagok