

Kiberfizikai Rendszerek Intézet			Mintatanterv szerinti 2. félév 2023-24-2			
Tantárgy neve:	Kódja:	Kredit:	Óraszám			
				ea	tgy	lab
Adatelőkészítés és -feldolgozás	NKXAF1HMLF	5	levelező féléves	10	0	10
Tárgyfelelős: Balázsné Dr. Kail Eszter			Beosztás: egyetemi adjunktus			
Oktató(k): Gerse Ágnes, Szécsényi Gergely						
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 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, outlierok kezelése, adatmanipuláció, adattranszformáció, normalizálás, aggregáció, dimenzió csökkentés/növelés, adattömörítés, diszkretizálás, adatminőség biztosítás, valamint ezen feladatok megvalósításához használható eszközök ismertetése.					

Féléves ütemezés	
Oktatási hét (konzultáció)	Témakör
1.	Elmélet: CRISP-DM módszertan. Gyakorlat: Python bevezetés, Jupyter Notebook interaktív fejlesztői környezet, Python adattípusok, adatszerkezetek.
2.	Elmélet: Adattípusok, mérési szintek (nominális, ordinális, intervallum- és arányskála), idősorok jellemzői. Gyakorlat: adatátalakítások, -műveletek, Pandas modul, DataFrame objektumok, adatbetöltés.
3.	Elmélet: Leíró statisztika. Gyakorlat: Egyszerű leíró statisztikai feladatok Pandas környezetben, Matplotlib adatvizualáció alapjai.
4.	Elmélet: Adatok előfeldolgozása, adattisztítási szintek (1. szint – elsődleges adatszerkezet és -formátumok kialakítása, indexek, oszlopnevek, 2. szint – információk kinyerése, adatszerkezet, adatformátumok átalakítása, 3. szint – hiányzó, kiugró és hibás értékek kezelése). Gyakorlat: Egyszerű adattisztítási feladatok Pandas környezetben.
5.	Elmélet: Adatok egyesítése, adatintegráció, adattömörítés (mintavételezés, dimenziócsökkentés). Gyakorlat: Adatintegrációs és adattömörítési feladatok Pandas környezetben.
6.	Elmélet: Adattranszformáció (normalizálás, standardizálás, diszkretizálás, attribútumok származtatása, aggregáció). Gyakorlat: Adattranszformációs feladatok Pandas környezetben.
7.	Elmélet: Adatminőség biztosítása (adatminőségi dimenziók, mérőszámok), adatvédelem és anonimizálás. Gyakorlat: NumPy bevezetés, műveletek NumPy tömbökkel.
8.	Elmélet: Bevezetés a Big Data technológiába Gyakorlat: BigData fájlformátumok
9.	Elmélet: Adatfeldolgozás folyamata Gyakorlat: Spark keretrendszer
10.	Elmélet, Gyakorlat: Batch processing Spark keretrendszerrel
11.	Elmélet, Gyakorlat: Stream processing Kafka forrásból Spark keretrendszerekkel
12.	Elmélet, Gyakorlat: Databricks platform, Adattárolás Delta Lake-ben
13.	Előadás ZH, Labor ZH

14.	Előadás ZH pótlás, Labor ZH pótlás
Félévközi követelmények	
Évközi jegy / aláírás megszerzésének feltételei:	Mind az előadás, mind a labor ZH-n legalább 26 pont elérése.
Zárthelyi dolgozatok	
Oktatási hét	Témakör
13.	Előadás ZH, Labor ZH
14.	Előadás ZH pótlás, Labor 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 előadás ZH-n elért pontszám (max. 50 pont)	
2. A labor ZH-n elért pontszám (max. 50 pont)	
Pótlás módja	
A ZH / évközi jegy / aláírás pótlásának módja:	Az előadás ZH és a labor ZH az utolsó héten és a vizsgaidőszak elején pótolhatók.
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 - Roy Jafari: Hands-On Data Preprocessing in Python, Packt Publishing, 2022, ISBN: 1801072132
Ajánlott:	- Jake VanderPlas: Python Data Science Handbook, O'Reilly Media, 2022, ISBN: 1098121228 - 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