

Kiberfizikai Rendszerek Intézet			Mintatanterv szerinti 4. félév 25-26-1			
Tantárgy neve:	Kódja:	Kredit:	Óraszám			
				ea	tg	lab
Bevezetés az adattudományba	NKXBATHBNE	5	nappali heti	2	0	2
Tárgyfelelős: Balázs Dr. Kail Eszter			Beosztás: egyetemi adjunktus			
Oktató(k): Dr. habil. Dineva Adrienn, Angyalné Dr. Alexy Márta, Dr. Alwahab Dhulfiqar Zoltán						
Előtanulmányi feltételek:	NIXAB0HBNE	Adatbázisok				
Számonkérés módja:	évközi jegy					
A tananyag						
Oktatási cél:	A tárgy célja az adattudomány alapfogalmainak és folyamatainak gyakorlati megközelítésű megismertetése. A hallgatók a gyakorlati életből vett valós alkalmazási példákon keresztül az ismereteket megtapasztalva, egyre mélyebbre haladva precíz elméleti és egyúttal praktikus gyakorlati ismeretekhez jutnak. Az elméleti ismeretek gerincét a gépi tanuló algoritmusok adják, a gyakorlati feladatok pedig az elmélet gyakorlati megjelenítését szolgálják Python nyelv használatán keresztül.					
Tematika:	Előadás: Történet, példák, esettanulmányok, az adattudományba sorolható területek. CRISP DM és a 6 fázisa. Adattisztítás. Felügyelt, felügyelet nélküli és megerősített tanulás. Túltanulás, alultanulás. Modellek validálása, tanulási/validációs/tesztelési halmaz, cross-validáció, Bias-Variance, ROC görbe. Legkisebb négyzetek módszere. Lineáris Regresszió. Gradiens módszer, maximum-likelihood becslés. Logisztikus regresszió, Tanulási/validációs/tesztelési halmaz, cross-validáció, Bias-Variance tradeoff, Precision-Recall, F1-score, ROC görbe. Neurális hálók. Döntési fák. Véletlen erdők. Boosting. Nem Ellenőrzött tanulás. Klaszterezés. K-means klaszterezés. Nagyobb esettanulmányok, kitekintés. Gyakorlat: Az adatmanipulálás, prediktív analízis, megjelenítés lépései valódi adatokkal elsősorban Python-csomagok (pandas, scikit-learn, matplotlib, ggplot) használatával.					

Féléves ütemezés	
Oktatási hét (konzultáció)	Témakör
1.	Bevezetés, az adatok jelentősége, az adattudomány gyakorlati alkalmazásai, Gyakorlat: Az adatok szerepének bemutatása a modern világban, az adattudomány alkalmazási területeinek áttekintése, valamint az alapvető programozási és adatelemzési eszközök megismerése.
2.	A CRISP-DM módszertan, a különböző adattípusok tulajdonságai, azok gyakorlati szerepe (numerikus, szöveges, képi és hangadatok, videók), Gyakorlat: Relációs adatbázisok kezelése és alapvető adatkísérletek.
3.	A digitális adatgyűjtésről, azok gyakorlati nehézségeiről, adatelőkészítés „garbage in, garbage out (GIGO)”, Gyakorlat: Adatgyűjtés problémái: hiányzó adatok, duplikátumok, hibás formátumok.
4.	Leíró statisztikai módszerek, a megfelelő adatbázis kiválasztásának szempontjai, a gépi tanulás alapjai (felügyelt, nem felügyelt és megerősítő), gyakorlati példák, Gyakorlat: Statisztikai mutatók számítása és a különböző ML-típusok illeszkedésének vizsgálata példafeladatokon.
5.	Adatelemzési modellek kiválasztásának szempontjai, leggyakoribb algoritmusok rövid jellemzése „No free lunch”, Gyakorlat: Egyszerű adatelemzési modellek (pl. lineáris regresszió, döntési fák, k-means) megvalósítása és eredményeik értelmezése.

6.	A természetes és a mesterséges neurális hálók jellemzői, gyakorlati alkalmazásának előnyei és hátrányai, Gyakorlat: Egyszerű neurális háló implementációja és kiértékelése képfeldolgozási példán (pl. MNIST)
7.	Asszociációs adatbányászat Gyakorlat:gyakori elemkombinációk („itemsetek”) és ezek közötti asszociációs szabályok származtatása tranzakciós adatokban (pl. piackosár-elemzés)
8.	Nem felügyelt gépi tanulás I: K-means és K-medoids algoritmus, távolságmértékek klaszterszám kiválasztás, hierarchikus klaszterezés, linkage-típusok, dendrogram, spektrális klaszterezés, gráf-Laplacian, sajátérték-alapú beágyazás, Silhouette-index, Davies–Bouldin-index. Gyakorlat: változatos adatkészleteken (Iris, Mall customers) K-means, hierarchikus és spektrális klaszterezés alkalmazása, klaszterminőség értékelése Silhouette- és DB-indexszel, dendrogram és spektrális beágyazás vizualizációja
9.	Nem felügyelt gépi tanulás II: Gauss-keverék modellek (GMM), klaszterek elmosódása, valószínűségi klasztertagság modellezése és vizualizálása, modellellenőrzés információelméleti kritériumokkal (BIC, AIC) Gyakorlat: GMM illesztése és klaszter-valószínűségek kiértékelése
10.	SZÜNET
11.	Modellértékelési szempontok felügyelt tanulásban, teljesítménymérő metrikák (accuracy, precision, recall, F1-score, ROC AUC), túlillesztés és alultanulás felismerése, cross-validáció szerepe, tanulási görbék elemzése, torzítás-variancia dilemma értelmezése Gyakorlat: különböző osztályozó és regressziós prediktív modellek teljesítményének kiértékelése valós adatkészleteken, modellek általánosító képességének vizsgálata
12.	Döntési fák működési elve, információnyereség és Gini-index alapú osztás, túlillesztés kezelése (metszés), valamint az egyes döntési struktúrák interpretálhatósága. Az ensemble (együttes) módszerek elméleti háttere, különös tekintettel a bagging (pl. Random Forest) és boosting algoritmusokra; a modellek stabilitása, torzítás–variancia csökkentésének mechanizmusai Gyakorlat: döntési fa, RF és boosting típusú ensemble modellek alkalmazása valós adatokon, teljesítményük összehasonlítása különböző értékelő metrikák mentén; a modellek viselkedésének vizsgálata tanulási görbéken keresztül, valamint a predikciók értelmezhetőségének bemutatása változók fontossága és döntési útvonalak segítségével.
13.	ZH (labor és elmélet)
14.	Pót ZH (labor és előadás)
Félévközi követelmények	
Évközi jegy / aláírás megszerzésének feltételei:	évközi jegy
Zárthelyi dolgozatok	
Oktatási hét	Témakör
13.	elméleti zárthelyi dolgozat (az előadás idejében), labor projektfeladat benyújtása
14.	pótló elméleti zárthelyi dolgozat és labor projekt-feladat pótleadása
Az évközi jegy kialakításának módszere (csak évközi jegyes tárgyak esetében töltendő ki)	
A félévzáró érdemjegy a zárthelyi dolgozatok (50) és a projektfeladat (50) pontszámai alapján kerül megállapításra. A maximálisan elérhető pontszám együttesen 100 pont.	
Pótlás módja	

A ZH / évközi jegy / aláírás pótlásának módja:	Pót zárthelyi írható az elmélet és a gyakorlat anyagából, minimum 50%-tól érvényes.
Vizsga módja (csak vizsgás tantárgy esetében töltendő ki)	
-	
Vizsgajegy kialakítása (csak vizsgás tantárgy esetében töltendő ki)	
-	
Az egyes érdemjegyek ponthatárai:	
A félévzáró érdemjegy a zárthelyi dolgozatok (50) és a projektfeladat (50) pontszámai alapján kerül megállapításra. A maximálisan elérhető pontszám együttesen 100 pont. Az érdemjegyek alsó ponthatárai a következők: elégséges – 51 pont, közepes – 63 pont, jó – 74 pont, jeles – 85 pont.	
Irodalom	
Kötelező:	A Moodle rendszerben közzétett jegyzet.
Ajánlott:	Richard S. Sutton and Andrew G. Barto: Reinforcement Learning: An Introduction J. Leskovec, A. Rajaraman, J. D. Ullman: Mining of Massive Datasets J. Gareth et al.: An Introduction to Statistical Learning
Egyéb segédletek:	