Óbudai Egyetem
Keleti Károly Gazdasági Kar

Statisztika az oktatásban - Hogyan szerettethető meg?

Adatok legyünk vagy szabadok? - Klasszikus statisztika és a Big Data elemzési eszközök

Előadó:
Szabó Zsolt Mihály
doktorandusz, ÓE-BDI

Konzulens:
Dr. habil. Csiszárik-Kocsir Ágnes
Egyetemi docens, Intézetigazgató

Budapest, 2018. november 13.
1084 Budapest, Tavaszmező u. 15-17. TA 122.

Téma aktualitása

Az adat az új olaj

- **Ez az analógia jól szimbolizálja**, hogy az adatok milyen mértékben lesznek képesek az üzleti élet megváltoztatására. Napjainkban szinte minden területen zajlik az a **digitalizációs forradalom**, amelynek alapját az adat és az ahhoz kapcsolódó lehetőségek adják.
- **Napjainkra az információ számít az egyik legfontosabb erőforrásnak.** Az a vállalat szerez versenyelőnyt, amelyik jobban tudja az adatokat hasznosítani, tárolni és kezelni, illetve megosztani az illetékesekkel és megvédeni az illetéktelenektől.
- Az **adat jelentősége folyamatosan nő** és még messze vagyunk a jelenlegi, adatokkal kapcsolatos képességeink kihasználásától.
- **Az adat, az adatvagyonnal való tudatos gazdálkodás a vállalati stratégiák, innovációk egyik legfontosabb központi témájává vált.**
- **Big Data korát éljük** - Elég bármely szervezet szerteágazó informatikai rendszereire gondolni, amely óriási – strukturált és egyre több nem strukturált – adattömeget keletkeztet és tárol nap mint nap.
- **Online digitális világ vállalatai (**Facebook, LinkedIn, Google, Amazon**)** már ma is mennyire meghatározó szereplői a piacnak és üzleti modelljüknek mennyire integráns alapkövei az adatalapú szolgáltatások.
- A **nyers adatokat értékesítésével** nem elégszenek meg a hazai és nemzetközi szereplők jellemzően nem elégszenek meg . Ehelyett meglevő adataik értékét további specifikus feldolgozással, elemzésekkel vagy publikusan elérhető adatokkal való összekapcsolással, esetleg egyszerre mindkettővel növelik. Így az adatkereskedelem helyett **értéknövelt szolgáltatással jelennek meg a piacon**.
- **A bizalom a legújabb olaj.** Az üzleti szempontok mellett az **új adatvédelmi rendelet (GDPR)** bevezetésének évében a jogi vonatkozásokat sem lehet figyelmen kívül hagyni. Magyarországra ebben az esetben is inkább a követő magatartás jellemző.

A vacak adat az új olaj

- **A Cork University három adattudósának ipari felmérése szerint csak a cégek elenyésző kis része rögzít jó adatokat, a többiek vacak alapanyagból dolgoznak drágán.**
- *„A rossz adatok pazarolják az időt, növelik a költségeket, gyengítik a döntéshozatalt, idegesítik az ügyfeleket és bonyolultabbá teszik az adatstratégia végrehajtását" – Harvard Business Review (September 11, 2017)*

Adathasználat, mire is jó a gazdaságstatisztika?

- Adat (vs. ember): **az információ hordozója**, a tények, fogalmak vagy utasítások formalizált ábrázolása, amely az emberek vagy automatikus eszközök számára közlésre, megjelenítésre vagy feldolgozásra alkalmas. A gazdaság, a társadalom egységeiről valamilyen információt hordozó jelek.
- Adathasználó/adatkezelő (vs. etika): az a természetes vagy jogi személy, valamint jogi személyiséggel nem rendelkező szervezet, aki vagy amely önállóan vagy másokkal együtt az **adatok kezelésének célját meghatározza**, az adatkezelésre (beleértve a felhasznált eszközt) vonatkozó döntéseket meghozza és végrehajtja, vagy az adatfeldolgozóval végrehajtatja.
- Statisztika (vs. fejlődés): **a valóság számszerű információinak megfigyelésére, összegzésére, elemzésére és modellezésére irányuló gyakorlati tevékenység és tudomány.** A tömegjelenségek jellemzőinek tömör, számszerű megismertetését szolgáló módszertana.
- Gazdaságstatisztika (vs. biztonság): **a gazdasági tömegjelenségek vizsgálatának tudománya, gyakorlati eljárása.**
  - Tárgya: A gazdaság alapegységei és folyamatai.
  - Tevékenységének köre: Mikro-, makro- és nemzetközi szintű (termelés (ipar, mezőgazdaság), szállítás, hírközlés, város- és községfejlesztés (kommunális gazdálkodás), életszínvonal, nemzeti jövedelem (GDP) kérdéseivel foglalkozik).
  - **Feladata: Segíteni a gazdasági élet döntéshozóit abban, hogy a rendelkezésre álló adatok alapján a statisztika módszertanának segítségével helyes döntéseket hozzanak** (a statisztika szó a latin Status szóból származik, államot jelent. Ebből képezték a az államtudományokkal foglalkozó egyén megjelelölésére olasz nyelven a statista (államférfi) szót. Ebből ered a statistika, mely a gyakorlati politikusok számára szükséges ismereteket jelentette.).

Adatbányászat

- Egyre több adat és hatalmas információmennyiség keletkezik a világ nagyvállalatai üzleti folyamatainak során.
- **Gyakorlatilag minden műszaki eszköz, ami környezetünkben található, ontja magából az adatokat.**
- **Információk keletkeznek**, ha telefonálunk, ha rákeresünk valamire az interneten, ha rendelünk valamit, ha bemegyünk egy üzletbe, ha bankkártyával fizetünk, vagy ha felkapcsoljuk a villanyt a konyhában. Olyan korban élünk, amikor adatforrásként tekinthetünk testünkre, mozgásunkra, viselkedésünkre és döntéseinkre. Bármi, amit teszünk, digitális nyomot hagy, és számok, képek, szövegek vagy navigációs adatok formájában tárolódik valahol.
- **Az óriási adattengerben izgalmas összefüggések, szabályszerűségek és minták rejlenek,** amelyek feltárására, vagyis a big data hasznosítására egyre nagyobb az igény az élet szinte minden területén az egészségügytől kezdve a közlekedésen át az időjárás-előrejelzésig.
- **A felhalmozott adatok hatalmas üzleti értékkel bírnak a vállalatok számára. Aki képes tárolni, strukturálni és gyorsan elemezni ezeket, komoly versenyelőnyre tehet szert.**

Big data és adatbányászat: mi közük a szerzői joghoz?

- **A 21. század társadalmának az adat szolgáltatja az üzemanyagot.**
- Az elmúlt pár évben minden területen megkerülhetetlenné vált a big data analitika, és az abból kinyerhető adatok, információk, új összefüggések feldolgozása.
- **A globális akadémiai és kutató közösség például mintegy másfél millió új tudományos cikket hoz létre évente**, nem beszélve az IoT (Internet of Things, a dolgok internete) által generált hatalmas mennyiségű adathalmazról.
- **A kutatók pusztán emberi erőforrással képtelenek lennének ezt az adathalmazt hatékonyan feldolgozni, ezért automatizált technológiai eszközöket, szoftvereket használnak az elemzésre.**
- A szöveg- és adatbányászati technológiák használata azonban nagy mennyiségű, szerzői jogilag védett tartalmak másolásához, többszörözéséhez, tárolásához vezethet.
- A **szöveg- és adatbányászat sértheti** a jelentős ráfordítással létrejött, úgynevezett kapcsolódó jogi, **sui generis oltalom**ban részesülő adatbázisok előállítóinak jogait és jogos érdekeit is, mivel ezen adatbázisok részeit rendszeresen kimásolják és újrahasznosítják a kutatók az adatbányászat során.

ADATBÁNYÁSZAT A GAZDASÁGI ÉLETBEN

- Az **adatbányászat egy döntéstámogatási módszer**, olyan üzleti intelligencia megoldás, amely új üzleti lehetőségeket segít megtalálni és kiaknázni a nagytömegű adathalmazokban rejlő, nem ismert összefüggések feltárásával.
- **Egyesíti az adatbáziskezelés, a statisztika és a mesterséges intelligencia kutatások eredményeit.**
- Az adatbányászat mint az adatelemzés eszköze és lehetősége – a statisztika hasonló kategóriáinak megfelelően – két nagyobb kategóriába sorolható:
  - a **leíró adatbányászat** az adatok alap jellemzőinek meghatározását jelenti.
  - a **következtetéses adatbányászat** alapvetően összefüggések feltárásával foglalkozik.
- A **statisztikai eszközökkel** „kis és közepes" adatmennyiségek esetén meg lehet találni bizonyos szabályszerűségeket és korrelációkat, de ezek az eszközök igazán nagy mennyiségű adattal már nem képesek megbirkózni.
- **Az adatbányászatnak nincsenek ilyen korlátai.** Az adatbányászat az adatok mélyére hatol. (Gáspár, 2006) Az adatbányászat alapadatai egyaránt lehetnek üzleti, kutatási, mérési adatok. Lényegében bármilyen nagytömegű adathalmaz elemei képezhetik az adatbányászat alapadat-állományait. Az adatok két köréhez kötődik az adatbányászat speciális területe: a szövegbányászat (textmining) és a webbányászat (webmining).

Elemző algoritmusok

- A **data science** (adat tudomány) területén, ahol olyan **elemző algoritmusok** kifejlesztése a cél, amikkel – a statisztikára, a gépi tanulásra és a nagy mennyiségű adatok feldolgozására támaszkodva – bonyolult összefüggések tárhatók fel.
- A számítógépeket nagy mennyiségű adat segítségével egyszerűen megtanítjuk a múlt eseményeire, vagyis arra, hogy mik voltak a normális működés jellemzői, és milyen körülmények között következtek be az anomáliák.
- A **data science** elsősorban az ipari internetre fókuszál, például:
  - Azt vizsgálják, hogy milyen tényezők hatnak az USA területén található szélerőművek teljesítményére. Ennek részeként több évtizedre visszamenőleg dolgozzák fel a különböző időjárás előrejelzők adatait, köztük a Nemzeti Óceán- és Légköri Hivatal (NOAA) jelentéseit, hogy megnézzék, az adott területre ki adta a **pontosabb becslést**.
  - Egy gázturbina gyártó számára fejlesztettek ki egy olyan hőkamerás képeket elemző algoritmust, aminek segítségével a mérnökök **előre tudják jelezni a meghibásodásokat**. Ezzel töredékére csökkenthető a szervízelési idő és a javítás költsége. A komplex rendszereknél, mint például egy gázturbinákat használó erőmű, egy nem tervezett leállás hatalmas bevételkiesést jelent.

Információk valós időben

- A **data engineering** (adatkezelés) ott kezdődik, ahol a **data science** véget ér. Itt már nem az elemző algoritmusok megalkotása a feladat, hanem az, hogy **az adatelemzés és az adatvizualizáció hatalmas mennyiségű adatok esetén is gyorsan, akár valós időben elvégezhető legyen**.
- A **data engineering** elsősorban arra fókuszál, ha a vállalat gyorsítani akarja adatfeldolgozási folyamatait, például:
  - A feladat az volt, hogy a közel 90 millió előfizető által generált több százmillió adatot felhasználva a világ legnagyobb streamingszolgáltatójának (Netflix) elemzői másodpercek alatt hozzáférjenek a legfontosabb mutatószámokhoz, mint például a nézettség, a sorozat öregedés vagy a folytatási hajlandóság.
  - A Facebook az adatvizualizációs szervereit egy ideje már egy magyar IT-tanácsadó cég (https://starschema.com/) programjaival felügyeli és frissíti, jövőre pedig a magyar cég fogja segíteni a világ legnagyobb közösségi hálózatát egy olyan algoritmus elterjesztésében, amivel könnyebb és gyorsabb lesz felderíteni a gyanús felhasználói aktivitásokat.

Az adatok vizualizációja

- **A puszta számok, főleg ha nagyok, nem sokat mondanak az embereknek. Ahhoz, hogy valóban érzékeltessük őket, találó módon kell őket képi formába önteni.**
- Az adatok (szakítások száma az év napjain) a Facebookról származnak, a megfelelő kifejezésekre történő keresésből.
- Egy-egy népszerű webes alkalmazás soktíz- és százezer, akár sok millió felhasználó adataival rendelkezik, jó esetben örködik felettük. A **magánszférára védelmétől** és a **személyiségi jogoktól** az online szolgáltatások kapcsán rendszeresen hangos a média, azonban kevesebbet foglalkoznak az aggregált adathalmok jelentőségével.
- Eddig a felhasználói profilok és látogatottsági adatok felhasználása nagyjából egészében a házon belüli, fejlesztési célú elemzésekre, illetve a reklámcélcsoport azonosítására korlátozódott. Ennél azonban sokkal több rejlik a rekordok között.
- **Eljön-e az az idő, amikor a webszolgáltatások együttal elemzőintézetekké válnak, profilfá válhat-e a weben a statisztikák kereskedelme?**

A mesterséges intelligencia és az emberek kapcsolata

- **Az adatokat sokszor meg sem kell érteni, elég ha az összefüggéseket látjuk.** Az adatok felhasználásával még a folyamatok megértése nélkül is eredményeket érhetünk el.
- A **mesterséges intelligencia:** megvizsgálja a siker és a bukás bemeneti adathalmazát, majd eldönti, hogy milyen előzmények milyen végkimenetelt okoztak.
- **Ez a gondolkodás, ami az adatokra épít, nagyon más, mint amit az eddigiekben megszokhattunk.**
- A digitalizáció nem csak a technológiáról szól. **A digitalizáció új megközelítéseket követel meg a gyártástól, a stratégiai tervezéstől és a gondolkodásmódunktól.**
- Az **adatvezérelt gondolkozás** ma versenyelőny, holnap, aki nem csinálja, az hátrányban lesz.
- Ma már sokkal nagyobb a kockázata annak, ha eldobjuk az adatainkat annál, mintha gyűjtenénk őket abban a reményben, hogy hasznos információval szolgálhatnak a jövőben.
- **Az adatokat nem feltétlenül kell megértenünk, sokszor elég az is, ha felismerjük, hogy mely körülmények játszanak közre a sikereinkben és kudarcainkban.**

Humán analitika

- **Aki (jól) keres, talál:** Ha jól használjuk, a **humán analitika** évekre előre megkönnyítheti a HR-esek munkáját. Ha megvan az ideális jelölt egy pozícióra, mindenki boldog. De ha nem, arra könnyen ráfizethetünk, a szó legszorosabb értelmében: az amerikai munkaügyi minisztérium becslése szerint egy elhibázott felvételen átlagosan annyit bukik egy cég, mint a dolgozó első éves fizetésének 30%-a.
- **A megfelelő adatok birtokában viszont szinte tökélyre fejleszthetjük a toborzási stratégiánkat, hogy már a kezdetektől biztosan a megfelelő tehetségeket vonzzuk be.**
- **Adatalapúvá tesszük:** Érdemes körbejárni például, milyen csatornából jutnak el hozzánk a legjobb jelöltek. LinkedInen, a kollégák ajánlásán vagy a karrieroldalon keresztül? Melyik országból, városból vagy egyetemről érkeznek? És milyen pozícióból? Minél többet tudunk meg a tuti befutókról, annál könnyebb lesz a dolgunk, amikor hozzájuk hasonlókat keresünk.

Amikor a mesterséges intelligencia támad bennünket

- **Alan Turing**, a modern számítógép-tudomány néhai atyja sem gondolta komolyan, hogy egy gép valaha is teljesíteni fogja a Turing-tesztet. Igaz, **1950 körül mondta, hogy az ezredfordulóra lesz olyan MI, amely képes lesz embernek kiadni magát.**
- **A jóslata valóra vált**, 64 év után tényleg „eladta" magát egy gép a tesztelő szakemberek 33 százalékának.
- Mi pedig most döbbenten vakarjuk a fejünket, hogy az **IBM mérnökei laborkörülmények között létrehoztak egy olyan kártevőt, amit már mesterséges intelligencia irányít** – jó ötletnek tűnik, biztos nem néztek elég olyan filmet, amikor valami elszabadul a laborokból és rendet tesz a világban.
- Percenként kilenc új támadás. Személyre szabott vírusok.
- **Támadhatóak a személyes és pénzügyi adatok.**
- **Még mindig a humánfaktor a legveszélyesebb a gépeinkre?**
- **Új játékszabályok kellenek.**

Valósidejű gazdaság

- ,,Digitalizációból üzleti érték" - rengeteg lehetőséget rejt és komoly felelősséggel is jár.
- A fogyasztók - elvárják, hogy az új technológiák minden előnyét, kényelmét, gyorsaságát, rugalmasságát élvezhessék.
- Változó környezet
- Hasznosítani a sötét adatot
- Újragondolt termelés
- Folyamatos újítások, nehezen mozduló KKV-k
- **A most gazdasága:** megváltoznak a klasszikus megrendelői és szállítói szerepek. Nem lesz idő arra, mint régen, hogy egy szerződésen akár hetekig, több körös egyeztetéssel dolgozzanak. A piac perceken belül bünteti, ha nincs megoldás, szerződés, és B2B (Business-to-business) oldalról kell jönnie a megoldásnak.
- Fel kell ismerniük a cégeknek, hogyan változzanak a digitalizációban, hogyan egyszerűsödjenek, ami szerint az előfeltétele az érdemi átalakulásnak.
- Mi magunknak is változnunk kell!

Számítógépes modell jelzi előre kiskereskedések bukását

- **Cambridge Egyetem által vezetett kutatócsoport a világ tíz városából származó adatokat használva fejlesztett új üzletek fél éven belüli bukását 80 százalékos pontossággal előrejelző modellt.**
- **74 milliónál több** chicagói, helsinkii, jakartai, londoni, los angelesi, new yorki, san franciscói, párizsi, szingapúri és tokiói Foursquare (egy helykeresésen és felfedezésen alapuló mobilos alkalmazás) bejelentkezést, valamint **181 millió** new yorki és szingapúri taxizás adatait használták.
- **A helyeket a környék jellemzői, környékek kapcsolatai és a különböző napszakok látogatási mintázatai együttes alkalmazásával osztályozták.**
- „Bármelyik új kereskedés számára az egyik legfontosabb kérdés, hogy mekkora a kereslet, mert közvetlenül kapcsolódik a siker valószínűségéhez. **De milyen mértékegység használható ezekhez az előrejelzésekhez?"**
- Egy kereskedés bukása számos **kontrollálható** (termékminőség, ár, nyitvatartási időpontok, vásárlók elégedettsége) és **kontrollálhatatlan** (munkanélküliek aránya, általános gazdasági feltételek, városvezetési stratégia) **tényező** összjátéka.
- „Kiderült, hogy a kontrollálhatatlan tényezőkre vonatkozó információk nélkül is tudjuk használni a helyszín-specifikus, a környékkel kapcsolatos és a mobilitás-alapú adatokat egy vállalkozás valószínű bukásának előrejelzéséhez"
- A modell rávilágított, hogy egy üzlet nyitási időpontjához és a helyszín kiválasztásához a környék statikus jellemzőin kívül más tényezőket is figyelembe kell venni, például hogy különböző napszakokban hogyan közelíthető meg stb.

Adattudós képzés

- **A digitalizációt nagymértékben tudja hátráltatni, ha nincsenek szakemberek, akik a szükséges rendszereket megtervezik, telepítik és működtetik.** Többféle szám is kering arról, hány informatikus hiányzik a magyar munkaerőpiacról, de az biztos, hogy több tízezerre tehető a számuk.
- A világ többi részéhez hasonlóan tehát Magyarországon sem tesz rossz lóra az, aki az **data scientist** (adattudós) pályát választja.
- **Nemzetközi trendet követve a magyar vállalatok is egyre nagyobb arányban kezdik felismerni a működésük során keletkező adatokban rejlő lehetőségeket**.
- Egyetemmel és főiskolával való együttműködés: szakdolgozati esettanulmányok kidolgozásának támogatása és gyakornoki programok támogatása.
- **Kompetenciák: egyszerre van szükség erős informatikai, matematikai és statisztikai háttérre, de az elemző gondolkodás és az új ismeretek elsajátításának képessége is alapvető követelmény.**
- Hornyák Miklós, a Pécsi Tudományegyetem Közgazdaságtudományi Karán belül működő Kvantitatív Intézet tanársegédje szerint: **„Látni kell azonban, hogy ez egy viszonylag fiatal és dinamikusan fejlődő ágazat, ahol rövid időn belül is hatalmas lemaradásba kerülhetünk. Ahhoz, hogy pozíciónkat és versenyképességünket megőrizzük, az iparági szereplők erőteljesebb aktivitására és támogatására van szükség, különösen a Budapesten kívüli régiókban".**

Köszönöm a figyelmet!