

DEBRECENI EGYETEM
GAZDASÁGTUDOMÁNYI KAR

TUDOMÁNYOS DIÁKKÖRI DOLGOZAT

MESTERSÉGES INTELLIGENCIA ALKALMAZÁSA A FUTBALL-
ADATELEMZÉSBEN

APPLYING ARTIFICIAL INTELLIGENCE TO FOOTBALL DATA ANALYSIS

DEBRECEN

2024.04.25.

Tartalomjegyzék

Tartalomjegyzék	2
Bevezetés	3
1. Témafelvetés.....	3
2. Szakirodalmi áttekintés	4
3. Anyag és módszer.....	6
3.1. Módszer	6
3.2. Anyag.....	6
3.2.1. A fejlesztés ismertetése.....	6
3.2.1.1. Microsoft Azure beállítása.....	6
3.2.1.2. Microsoft Visual Studio Code	7
3.2.2. A kérdőív ismertetése	9
4. Eredmények és azok értékelése	10
4.1. A fejlesztés értékelése.....	10
4.1.1. Hány labdát vesztett Szoboszlai Dominik az eddigi mérkőzések során a 2023/2024-es Premier League szezon során?	10
4.1.2. Melyik csatárnak volt a legmagasabb xG-je a Premier League-ben?.....	11
4.1.3. Ki a legjobb középpályás?	13
4.2. A kérdőív értékelése.....	14
5. Következtetések és javaslatok	17
Összefoglalás	18
Szakirodalmi jegyzék	18

Bevezetés

A labdarúgás kétségtelenül napjaink legnépszerűbb sportja, mely nem csak a szórakoztatóipar egyik kulcseleme, hanem az élsport tekintetében hatalmas gazdasági vonzatokkal járó sportág. A legmagasabb szinteken a nagy sportklubok többnyire profitorientált szemléletet követnek, amihez elengedhetetlen, hogy társuljon az eredménycentrikusság is. A legjobb teljesítmény elérésére több eszköz is rendelkezésre áll már a mai világban, de a technológiai fejlődésnek köszönhetően mindig vannak kiaknázatlan területek, ahol lehet még fejlődni. Az egyik ilyen terület a futballadat-elemzés, hiszen a mérőműszerek és a profi szakemberek segítségével számos teljesítménymutató kalkulálása lehetséges, melynek köszönhetően megalapozott és rentábilis döntéseket hozhatnak a klubok vezetői. Más iparágakban is egyre inkább elterjed az adatvezérelt döntéshozatal, mivel jelentős anyagi forrásokat lehet vele megtakarítani. Ezen futballadatok kielemezését megkönnyítheti egy új technológia, a mesterséges intelligencia elterjedése, amely nem csak a futball, de az élet más területein is nagy hasznára lehet az emberiségnek.

1. Témafelvetés

A dolgozat témájának kiválasztása során elképzelhetetlennek tartottam, hogy ne egy labdarúgással kapcsolatos kutatást végezzek, hiszen sportkarrierem során 15 évig futballoztam napi rendszerességgel. Emellett jelenleg is ezen a területen tanulok és dolgozok, ami lehetővé tette számomra a korábbi ismereteim és szakmai tapasztalatom alkalmazását a dolgozatom elkészítése során. A labdarúgáson túl mindig érdeklődő voltam a különböző informatikai megoldások és a modern, technológiai fejlesztések iránt, így szerettem volna megvizsgálni a két terület együttes alkalmazásának lehetőségeit és a benne rejlő potenciált. Egy fejlesztés keretein belül szerettem volna tudomást szerezni a futballadatok és a mesterséges intelligencia kombinálásának lehetőségeiről. Ezt egy MI alapú, strukturált lekérdezőnyelvet használó (SQL) ügynök létrehozásával valósítottam meg, melynek segítségével egyszerűbben lehet hozzájutni fontos adatokhoz, akár szöveges kérdések megfogalmazása során. A fejlesztés megerősítette az új technológia gyakorlatban való alkalmazásának létjogosultságát. További célkitűzésem során szerettem volna megtudni az átlagemberek véleményét és nyitottságát a futballadatokkal, a mesterséges intelligenciával és annak labdarúgásban való alkalmazásával kapcsolatban, így létrehoztam egy kérdőívet ennek felmérésére. A labdarúgás világában egyre elterjedtebbek a technológiai újítások, azonban itt nem kizárólag csak a játék során közvetlenül alkalmazott módszerekre kell gondolni. A pályán kívül is rengeteg kiaknázatlan lehetőség van még, melyek egyike lehet a futballadat-elemzést segítő, mesterséges intelligencia alapú szoftverek

használata. Bízom benne, hogy a fejlesztés egy futballelemzők, játékosmegfigyelők és sportkommentátorok által használt eszköz is lehet, valamint, hogy a dolgozat korábban fel nem tárt megállapításokat és összefüggéseket tartalmaz.

2. Szakirodalmi áttekintés

Az informatikai szolgáltatások használata során rengeteg adat generálódik. A mindennapi életben elég csak a banki, közigazgatási, egészségügyi vagy telekommunikációs rendszerekre gondolni, melyek működéséhez elengedhetetlen az adatok gyűjtésén túl a tárolásuk is. Ahhoz, hogy ezeknek az adatoknak a segítségével elérjük a kívánt célt, különböző elemzési módszereket kell elsajátítani. Ezen módszerek kiválasztása olykor nehezebb is lehet, mivel különböző esetekben más-más opciót kell választani. Sok adatbányászati, matematikai és statisztikai elemzési módszer létezik, a megfelelő opció kiválasztását pedig nehezíti a lehetőségek széles tárháza. Következésképpen a száraz adatok értékes információkká alakíthatók, majd olyan ismeretekre lehet szert tenni, amivel megalapozott gazdasági döntéseket is hozhatók (Fogarassyné & Starkné, 2011).

A futball sem maradhatott ki a már más üzleti ágazatokban és iparágakban kiteljesedett és bevált adatelemzésből (Bigcloud, n.d.). Ezeknek az új technológiáknak a bevezetése kihívások elé állítja az érintetteket, de nem csak a pénzügyi források felhasználásában, hanem a megfelelő személyi erőforrások és kompetenciák megtalálásában is (Herberger & Litke, 2021). A technológia fejlődésével elérhetővé vált a sportpiac számára, hogy számítógépes ismeretek alkalmazásával különböző elemzéseket végezzenek, és ezeket aztán alkalmazhassák a gyakorlatban, vagyis a mérkőzéseken. 2001-ben Billy Beane, az Oakland Athletics menedzsere felismerte a potenciált az iparágban. Forradalmasította az innovatív adatvezérelt menedzsmentet az amerikai profi baseball ligában. A statisztikai elemzések alapján hozott megalapozott döntésekből adódóan egyedülálló, húsz meccsig tartó veretlenségi szériát ért el (Lewis, 2004). Az olyan sportágak esetében, mint a baseball, a prediktív analitika jól működik a korlátozott mozgásokkal rendelkező, rögzített pozícióban lévő játékosokra vonatkozó mérőszámok alkalmazásakor. A futballban sokkal nehezebb statisztikai elemzést alkalmazni, mivel egy jóval dinamikusabb sportágról van szó (Bigcloud, n.d.). Az adattudósok és az adatelemzők nagy segítséget nyújthatnak a játékosmegfigyelők számára. Egy jó játékos leigazolásaért folytatott verseny egyre szorosabbá válik, ami miatt az adatalapú játékosmegfigyelés is egyre jelentősebb. Olyan minőségi, jó statisztikai adatokkal rendelkező játékosokat tudnak találni, amit ezen adatok hiányában nem lehetne megtenni, vagy csak nagyon nehéz és időigényes módon (Bigcloud, n.d.). Fontos azonban megjegyezni, hogy sosem

szabad csak és kizárólag ezekre az adatokra, mutatókra hagyatkozni; mindig teljes kontextusban kell értelmezni őket. Ez nem egy önálló eszköz a játékosmegfigyelők számára, hanem egy döntéstámogató rendszer, amit, ha jól és effektíven használnak, jelentősen megkönnyítheti a munkájukat. Régebben az adatelemzésben rejlő lehetőségek kihasználásának technikai korlátai voltak, gondolva itt a tárolható adat mennyiségére és a feldolgozási sebességre. Ezen korlátok mára megszűntek, ezzel új utakat megnyitva a gépi- és mélytanulási algoritmusok előtt, amelyek hatalmas mennyiségű adatok kezelésére alkalmasak. A korábban bevált leíró, előrejelző és előíró módszerek kiegészülnek az automatizálással és tanulással, ami a mesterséges intelligencia elterjedésének kezdetét is jelentette egyúttal (SAS, n.d.).

Az utóbbi években egyre jobban elterjedő mesterséges intelligencia térhódítása megkérdőjelezhetetlen. Sokan ebben látják a jövőt, hogy egy intelligens segítőeszközként szolgáljon az emberiség számára. Ennek ellenére nem a közelmúltban kezdtek el foglalkozni a mesterséges intelligenciával. Az első említés 1950-ben, Allan Turing nevéhez fűződik, ami egyben a legelterjedtebb fogalma a mesterséges intelligenciának. Egy imitációs/utánzó játékon keresztül próbálta meghatározni az AI fogalmát, aminek a neve „Turing teszt” volt. Ez egy olyan játék, aminek három résztvevője volt: egy ember, egy számítógép és egy emberi bíró, döntéshozó. A bíró el volt különítve az embertől és a számítógéptől és csakis szöveges üzenetekben kommunikálhattak egymással. A Turing teszt akkor volt sikeres, ha a döntéshozó nem tudta egyértelműen megkülönböztetni az embert a számítógéptől, nem tudta, hogy éppen kivel beszél. Így Turing (1950) szerint akkor beszélünk mesterséges intelligenciáról, ha az adott számítógép átmegy a Turing teszten. McCarthy (2004) következőképpen definiálta az MI-t: A mesterséges intelligencia nem más, mint az intelligens gépek megalkotásának tudománya, ahol az intelligencia a célok megvalósítására való alkalmasságot jelenti. Fitria (2021) véleménye alapján a mesterséges intelligencia olyan technológia, amely képes az emberi intelligenciához hasonló működésre, és várhatóan képes lesz olyan feladatok utánzására, amelyeket emberek is el tudnak végezni, legyen szó egy döntés meghozataláról vagy akár egy minta vagy kép felismeréséről. Mára az MI egyike azoknak a technológiai trendeknek, amelyeket bármilyen célra fel lehet használni (Fitria, 2023a). Népszerűsége folyamatosan növekszik mindegyik iparágban, többek között a chatbot technológiának köszönhetően (Wei et al., 2018; Luo et al., 2019). Ez egy olyan szoftver, amely természetes nyelvfeldolgozási (NLP) és mélytanulási technikákat használ szövegalapú online csevegések lebonyolítására (Dharani et al., 2020).

3. Anyag és módszer

3.1. Módszer

A kutatási módszer kiválasztásánál figyelembe vettem, hogy mely módszer használata esetén érhetem el a legpontosabb eredményeket. Kvantitatív módszerként a kérdőíves felmérést választottam. A kvantitatív kutatási módszerek legfőbb jellemzői közé sorolható, hogy mennyiségi alapon határoznak meg eredményeket és vonnak le következtetéseket, továbbá könnyen kivitelezhető és a többi kutatási módszerhez képest nem terheli le annyira a megkérdezetteket. Hátrányaihoz sorolható esetenként a bizalom és az őszinteség hiánya a megkérdezettek részéről (Lampek & Horváthné, 2015). Az MI alapú fejlesztés során egy futballadatokat tartalmazó adatbázisból dolgoztam, ezzel megteremtve a lehetőséget arra, hogy a technikai fejlesztést valós és értelmezhető keretek között valósítsam meg.

3.2. Anyag

3.2.1. A fejlesztés ismertetése

A fejlesztés középpontjában a mesterséges intelligencia alkalmazása áll egy futballadatokat tartalmazó adatbázis tartalma alapján. Ezen relációs adatbázist az xfb Analytics Kft. szolgáltatatta számomra, illetve ennek a cégnek az erőforrásait használva végeztem el a kutatást.

Az elemzésre az SQL nyelv szolgáltatja a megoldást. Alapvetően az adatelemzők és az adatbázisszakértők manuálisan szokták elemezni az adatbázisokat ennek a módszernek a segítségével. Megoldásomban szerettem volna egy automatizált felületet létrehozni, ahol nem szükséges az SQL által használt nyelvezetet használnia a felhasználónak, hanem értelmes, szöveges kérdések írásával tudhatja meg a várt válaszát. Ennek megvalósítására a LangChain névre keresztelt Python könyvtár használata mellett döntöttem, ahol egy úgynevezett SQL Agent-et, vagyis egy SQL ügynököt kellett létrehoznom. A LangChain egy olyan keretrendszer, ami olyan alkalmazások létrehozására szolgál, mely LLM-eket, másnéven nagy nyelvi modelleket használnak (Microsoft, 2024). Felhasznált szoftverek: Microsoft Azure, Microsoft Visual Studio Code, Microsoft SQL Server Management Studio. A következőkben bemutatom a fejlesztés folyamatát, a környezet felállítását, valamint a kutatás lépéseit.

3.2.1.1. Microsoft Azure beállítása

A Microsoft Azure egy felhőszolgáltatás, ahol technikai fejlesztéseket valósíthatnak meg a fejlesztők. Az xfb Analytics Kft. rendszerébe belépve az első feladatom az volt, hogy létrehozzak egy erőforrás csoportot, aminek a keretei között megvalósulhat a fejlesztés. Az erőforrás elkészítése után az Azure OpenAI szolgáltatást kellett üzembe helyezni, ami a

mesterséges intelligencia használatához volt szükséges, majd az Azure AI Studioban az általam használandó MI modellt kellett létrehozni, ami a természetes szöveg generálásához volt szükséges. A létrehozott modell paraméterei szerint az üzemelő példány neve „my-first-gpt”, a modell „gpt-4”, míg a maximum token száma 8192 lett. Ezen tényezők beállítása az összes szükséges Azure funkció meglétét eredményezte, így a következő lépéseket már a Microsoft Visual Studio Code szoftverben kellett elvégezni.

3.2.1.2. Microsoft Visual Studio Code

A Python programnyelv és az egyéb könyvtárak telepítése után létrehoztam egy `env_params` nevű fájlt, ami a környezeti változókat tartalmazza. Erre azért van szükség, hogy az érzékeny és titkos adatokat ne a fejlesztési kódba kelljen direkt módon beleírni, hanem ezek az adatok legyenek külön, egy kifejezetten erre a célra létrehozott fájlban. A tartalma a fájlban két részre osztható:

- „`OPENAI_API_KEY`”: Ide kellett beilleszteni az Azure-ban kapott kódot a OpenAI szolgáltatás létrehozását követően.
- „`uri`”: Itt kellett megadni az Microsoft SQL Szerver eléréséhez szükséges adatokat.

Az `env_params` fájl létrehozása után a következő feladat egy új Jupyter Notebook fájl létrehozása volt, aminek indításakor ki kellett választani a nyelvet, vagyis a Python 3.12.-es Kernel-t. Ezt követően kezdődhetett el a kódok megírása, ahol az első lépés a különböző könyvtárak és eszközök importálása volt. Az első részében a kódoknak öt könyvtárat importáltam, amelyek a következők voltak: a `create_sql_agent` az SQL ügynök létrehozásához, az `AgentType` az ügynök típusának meghatározásához, az `SQLDatabaseToolkit` a lekérdezések végrehajtásához, az `SQLDatabase` az SQL szerverhez való kapcsolódáshoz, a `ChatPromptTemplate` a párbeszéd sablonjának létrehozásához szükséges, míg az `env_params` importálása a teljes környezeti fájl kódba való integrálását jelentette. Ezt követően az Azure szolgáltatások csatlakoztatása következik a környezeti változók csatlakoztatásával, ahol három különböző paramétert kellett megadni, az `os` importálása után. Az `OPENAI_API_VERSION` a verziót, az `OPENAI_API_TYPE` a típust, míg az `OPENAI_API_KEY` az `env_params`-ban megadott kulcsot állítja be. A következő ablakban az 1. ábrán látható kódsorok kerültek meghatározásra.

```
uri = uri
db = SQLDatabase.from_uri(uri, include_tables=["..."])
```

1. ábra: A szerver és az adatbázis csatlakoztatása

Forrás: Saját szerkesztés (2024)

Az **uri** változóban egy tárolt elérési úton keresztül csatlakozik az SQL adatbázishoz, amit korábban már megadtam az `env_params` fájlban. A **db** során pedig magát az adatbázist csatlakoztatom, de nem a teljeset, hanem csak a fejlesztés elkészítéséhez szükséges táblákat. Ezekben a táblákban a meccsekre vonatkozó adatok, a játékosokhoz, csapatokhoz és bajnokságokhoz tartozó adatok, valamint a játékosok labdavesztésére és labdaszerzésére vonatkozó adatok szerepelnek.

A 2. ábrán a mesterséges intelligencia modell üzembe helyezésére kerül sor:

```
from langchain.chat_models import AzureChatOpenAI
llm = AzureChatOpenAI(deployment_name = 'my-first-gpt',
                      model_name = 'gpt-4')
```

2. ábra: A nyelvi modell üzembe helyezése

Forrás: Saját szerkesztés (2024)

Az `AzureChatOpenAI` modul kezeli az **llm**-ben meghatározott nagy nyelvi modellt, amit az Azure-ban elkészített, GPT-4-es modellt használó „my-first-gpt” jelöl. A program utolsó lépéseként kerül létrehozásra az SQL ügynök, aminek a beállításait a 3. ábrán mutatom be:

```
agent_executor = create_sql_agent(
    llm=llm,
    toolkit=SQLDatabaseToolkit(db=db, llm=llm),
    verbose=True,
    agent_type=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
    handle_parsing_errors=True,)
```

3. ábra: Az SQL ügynök beállításai

Forrás: Saját szerkesztés (2024)

Az **llm** és a **toolkit** esetében a már korábban létrehozott funkciók betöltésére kerül sor. A **verbose**, „Igaz” -ra állítva arra utal, hogy az MI a válasza esetén mutassa saját gondolatmenetét, látszódjon a logika a gondolkodásmódjában. Az **agent_type** a már korábban említett ügynök típus beállításához kapcsolódik, ahol a „ZERO_SHOT_REACT_DESCRIPTION” azt jelenti, hogy az MI minden egyes gondolatmenete után egy érvelő, logikailag megfelelő következtetést von le, majd csak ennek a cselekvésnek az elvégzése után lép tovább a következő gondolatmenetre. A **handle_parsing_errors** arra vonatkozik, hogy ha egy gondolatmenetnél elakad az MI, próbálja azt másképpen megoldani.

A fejlesztés további részét az Eredmény és azok értékelése című fejezetben folytatom tovább, ahol három különböző kérdést teszek majd fel, amit az SQL ügynök segítségével fogok megválaszolni. Mindezt az SQL Server Management Studio szoftver segítségével ellenőrzöm, manuálisan.

2.2.2. A kérdőív ismertetése

Kérdőívem központi témája a mesterséges intelligencia alkalmazása a futball-adatelemzésben volt, amin belül az emberek véleménye érdekelt arról, hogy hogyan vélekednek az MI elterjedéséről, véleményük szerint milyen hatásai lehetnek az MI használatának, illetve, hogy ennek milyen alkalmazási lehetőségei lehetnek a futballban. Emellett felmértem az emberek általános, futballadatokkal kapcsolatos tudását is, valamint kitértem a fejlesztés során használt LangChain könyvtár ismertségére is. Felmérésem feleletválasztós, likert típusú, valamint jelölőnégyzetes kérdéstípusokat tartalmaz. A kérdőív mintavétele során a hólabda módszert alkalmaztam, tehát megkértem a kitöltőket, hogy ők is osszák tovább a kérdőívet, így segítve a kutatást. Az elemzések során az Excelt, az SPSS statisztikai szoftvert és a Power BI adatvizualizációs szoftvert használtam. Összesen 154 fő töltötte ki az online kérdőívemet, melyet négy részre lehet osztani: demográfiai adatok, mesterséges intelligenciával kapcsolatos kérdések, futballadatokkal kapcsolatos kérdések és a mesterséges intelligencia és a futballadatok kapcsolatához fűződő kérdések. Ebben a fejezetben a demográfiai adatokat fogom elemezni.

A kérdőívet 78 nő és 76 férfi töltötte ki, aminek köszönhetően a kutatás rendkívül reprezentatív lett. A labdarúgás maszkulin jellegének ellenére számos nő is kifejezte érdeklődését a kérdőív kitöltése iránt. Az eredmények vizsgálata során több kérdést is részletesebben elemzek a nemek eloszlásának függvényében. Az életkor felmérése során a válaszadókat pontos életkor megadására kértem, hogy minél rugalmasabb elemezhetőséget biztosítsak a kutatás számára. A téma jellege miatt fontosnak tartottam, hogy elsősorban generációk alapján vizsgáljam meg a kitöltők életkorát. A Baby Boomers generációba tartozó kitöltők száma elég csekély volt, mindössze 1,3%-át tették ki a kitöltők számának. Az X generáció 10,39%-kal, míg az Y generáció 18,83%-kal képviseltette magát. A Z generáció abszolút többségben vett részt a kutatásban, több, mint kétharmados többségben voltak.

A következő kérdésben a lakhelyük alapján vizsgáltam a kitöltőket. A legkevesebben a községekben élő megkérdezettek töltötték ki a kérdőívet (11,69%), majd őket a fővárosiak követték 19,48%-kal. A megyeszékhelyeken élők száma a megkérdezettek csaknem

egyharmadát tették ki, míg a városokban élők száma 37,01%-ot. A következő kérdésben a válaszadók végzettségét vizsgáltam. Általános iskolában és szakmunkásképzőben végzettek száma alacsony, összesen két fő volt. Az érettségivel rendelkezők a kitöltők több, mint egyharmada, a főiskolai szakképzésben végzettek pedig 9,09%-a volt az összes kitöltő számának. Főiskolai végzettséggel a megkérdezettek csaknem 13%-a rendelkezett, míg az egyetemi diplomával rendelkezők kimagasló 42,86%-ban vettek részt a kutatásban. A következő kérdésben a munkavégzés típusát vizsgáltam, ahol arra voltam kíváncsi, hogy a megkérdezettek között milyen arányban oszlik el a fizikai, szellemi, fizikai- és szellemi munkavállalók, illetve a nyugdíjasok, tanulók és a tanulás mellett dolgozók száma. Megállapítható, hogy a megkérdezettek pontosan fele szellemi munkát végez, míg 15,58%-a fizikait. A kettőt egyszerre összesen mindössze 3,25%-uk gyakorolja. A tanulók 22,08%-ban, akik pedig mellette dolgoznak 15,58%-ban vannak jelen a mintában. A nyugdíjasok száma 1,3%, vagyis pontosan két fő volt.

A dolgozat folytatásában a fejlesztés, illetve a kérdőív további eredményeit részletezem.

4. Eredmények és azok értékelése

4.1. A fejlesztés értékelése

Az anyag és módszer fejezetben már ismertettem a fejlesztés technikai hátterét és a megvalósítás módját. A kódok a Python programozási nyelvben lettek megírva, amely mellett az SQL ügynökkel történő kommunikáció angol nyelven történt. Ezt szem előtt tartva a dolgozatban az ügynökkel történő párbeszédet magyar nyelven jelenítem meg. Fontos megjegyezni, hogy a szenzitív adatokat tartalmazó adatbázis használata miatt a végeredményeket nem jeleníthetem meg, csupán a folyamat működése és a rendszer használata a lényeges. A következő fejezetekben három kérdést tettem fel a mesterséges intelligenciának, melynek az volt a feladata, hogy a megfelelő SQL parancssorok legenerálása után szöveges üzenetben fogalmazza meg a feltett kérdésre vonatkozó választ.

4.1.1. Hány labdát vesztett Szoboszlai Dominik az eddigi mérkőzések során a 2023/2024-es Premier League szezon során?

Az SQL Ügynök futtatásához szükséges parancssort a 4. ábra mutatja be.

agent_executor.run („Hány labdát vesztett Szoboszlai Dominik az eddigi mérkőzések során a 2023/2024-es Premier League szezon során?”)
--

4. ábra: Szoboszlai Dominik labdavesztései számának meghatározására irányuló parancs

Forrás: Saját szerkesztés (2024)

A parancs futtatását követően az MI elkezdte kifejteni a saját gondolatmenetét, hogy milyen tényezők vizsgálata alapján kaphatja meg a kérdésre keresett választ. Az első lépése a következő volt: „Szoboszlai Dominik 2023/2024-es Premier League szezonban nyújtott teljesítményére vonatkozó információkat kell lekérdeznem az adatbázisból. De először is tudnom kell az adatbázis szerkezetét, hogy mely táblák tartalmazzák a vonatkozó információkat, és hogy ezek a táblák milyen mezőkkel rendelkeznek.” A második lépésben megvizsgálja a táblák neveit és tartalmát, majd kiválasztja a szükségeseket további elemzésre. A "report_ball_recorlost_fact" tábla tartalmazza részben a megoldást, de nincs benne minden szükséges információ, emiatt további táblákat is vizsgál a legjobb eredmény érdekében. Realizálja, hogy a „dim_players” táblában a játékosra, vagyis Szoboszlai Dominikra vonatkozó adatokat találja, emellett szüksége van még a bajnokság nevére és a szezon idejére, amiket a „dim_matches” táblában talál meg. Végül az információk megtalálása után létrehozza a válasz lekérdezéséhez szükséges parancssort, amely a 5. ábrán látható.

```
SELECT SUM(fact.ball_lost_cnt) AS total_balls_lost FROM report_ball_recorlost_fact AS fact
JOIN dim_players AS players ON fact.player_id = players.player_id JOIN dim_matches AS
matches ON fact.match_id = matches.match_id WHERE players.firstname = 'Dominik' AND
players.lastname = 'Szoboszlai' AND matches.season_name = '2023/2024' AND
matches.tournament_name = 'Premier League'
```

5. ábra: Az MI által létrehozott SQL parancs Szoboszlai Dominik labdavesztései számának meghatározására

Forrás: Saját szerkesztés (2024)

A kódban látható, hogy a különböző táblák összekötésre kerülnek, illetve a szűrők alkalmazása történik. Ezt követően az ügynök lefuttatja a generált SQL parancsot és szöveges üzenetben válaszol a feltett kérdésre: „Szoboszlai Dominik eddig „x” labdát veszített el a Premier League 2023/2024-es szezonjában.”; ahol „x” a konkrét számot jelöli. Ellenőrzésképpen felülvizsgáltam a kód helyességét, illetve manuálisan is lefuttattam a Microsoft SQL Server Management Studio szoftverben, ami igazolta az MI állítását.

4.1.2. Melyik csatárnak volt a legmagasabb xG-je a Premier League-ben?

Az előző kérdés feltevéséhez is szükséges parancssort lefuttatva – a kérdés cseréjét követően – szintén hasonló gondolkodásmódot lehetett felfedezni az MI-nél. Miután megtalálta a szükséges táblákat és azok oszlopait, ismét legenerálja az SQL kódot, ami a 6. ábrán látható.

```
SELECT TOP 1 p.firstname, p.lastname, MAX(r.xG_cnt) AS max_xG FROM
report_stat_players_matches_fact AS r JOIN dim_players AS p ON r.player_id = p.player_id
WHERE r.player_position_id = 'ST' AND r.tournament_id = 335 GROUP BY p.player_id,
p.firstname, p.lastname ORDER BY max_xG DESC
```

6. ábra: Az MI által létrehozott SQL parancs a legmagasabb xG értékkel rendelkező csatár meghatározására

Forrás: Saját szerkesztés (2024)

A készített kód működőképesnek tűnik, azonban mégis hibára fut: „Sajnos az adatok hiánya miatt nem tudtam megkeresni a konkrét játékos nevét.” Kielemezve a hibát megállapítható, hogy nem vizsgálta meg elég alaposan a játékos pozícióihoz tartozó értékeket. Az „ST”, vagyis a csatár szűrő lefuttatása után hiba keletkezett, mert az adatbázisban nem így szerepelnek az erre vonatkozó adatok. Az ilyen hibák kijavítására létezik egy olyan opció, ahol manuálisan lehet irányítani az ügynök gondolatmenetét. Ezt a 7. ábra szemlélteti.

```
question = " Melyik csatárnak volt a legmagasabb xG-je a Premier League-ben?"
final_prompt = ChatPromptTemplate.from_messages([
    ("system", """
        Te egy SQL-adatbázissal történő interakcióra tervezett ügynök vagy.
        ...
        Törekedj a lehető legkevesebb lekérdezésre a válasz generálása során.
        Ha a felhasználó a játékos egy adott pozíciójára kérdez rá, ellenőrizd az oszlop tartalmát, és
válaszd ki a lekérdezéshez a leginkább hasonló opciót. """),
    ("user", f'{question}\n ai: " ")])
```

7. ábra: Az ügynök finomhangolása

Forrás: Saját szerkesztés (2024)

Látható, hogy az ügynök gondolatmenetét különböző utasításokkal irányítom, ezzel biztosítva a jó válasz generálását. A kérdés szempontjából az aláhúzott mondat a lényeges, hiszen ott kap az MI utasítást, hogy ha nem talál olyan játékospozíciót, amit a felhasználó kér, keresse meg az ahhoz leghasonlóbb pozíciót. Ezzel kiküszöbölhető, hogy az ügynök csak és kizárólag a kérdésben feltett speciális eseteket nézze. Az egyik gondolatmenet során lehet látni a folyamatot: „A „csatár” pozíció nem létezik a „dim_players” táblázatban, helyette a lehetséges pozíciók a következők: „védő”, „támadó”, „középpályás”, „kapus” vagy „nincs adat”. Feltételezem, hogy a „csatár” kifejezés ebben a kontextusban a „támadó” kifejezéssel egyenértékű, így ennek megfelelően módosítom az SQL-lekérdezésemet.” A 8. ábrán látható SQL kódból látható, hogy miután lefutott a parancs, a válasz már jó volt és az MI sikeresen értelmezte a megadott paramétereket.

```
SELECT TOP 1 p.firstname, p.lastname, f.xG_cnt AS xG FROM
report_stat_players_matches_fact AS f JOIN dim_players AS p ON f.player_id = p.player_id JOIN
multilang_dim_tournaments AS t ON f.tournament_id = t.tournament_id WHERE
p.position1_name = 'Forward' AND t.tournament_nm = 'Premier League (England)' AND t.lang_id
= 'en' ORDER BY xG DESC
```

8. ábra: Az MI által kijavított SQL parancs a legmagasabb xG értékkel rendelkező csatár meghatározására

Forrás: Saját szerkesztés (2024)

Ezt követően az MI már meg tudta fogalmazni a helyes választ: „Játékos” volt a legmagasabb xG-vel rendelkező csatár a Premier League-ben, 2,53-as xG-vel.”; ahol a „játékos” az adott futballista nevét jelentette. Ennek a kérdésnek a feltevése során megállapítottam, hogy fontos az SQL ügynök finomhangolása a tökéletes eredmény és a hibalehetőségek redukálása érdekében.

4.1.3. Ki a legjobb középpályás?

Az előző kérdéshez hasonlóan adtam egy új utasítást az MI-nek, hogy milyen tényező alapján határozza meg a legjobb középpályás személyét. Ez a tényező a legtöbb pontos passzt jelentette, vagyis az MI-nek a megoldásában a legtöbb pontos passz alapján kellett meghatároznia a legjobb középpályást. A válasz a következőképpen nézett ki: „A legjobb középpályás „Játékos”; ahol a „játékos” az adott futballista nevét jelentette. Az SQL lekérdezés, amelyet lefuttatott a 9. ábrán látható.

```
SELECT TOP 1 p.firstname, p.lastname, SUM(f.passes_accurate_cnt) AS total_accurate_passes
FROM dim_players AS p JOIN report_stat_players_matches_fact AS f ON p.player_id =
f.player_id WHERE p.position1_name = 'Midfielder' GROUP BY p.firstname, p.lastname
ORDER BY total_accurate_passes DESC
```

9. ábra: Az MI által létrehozott SQL parancs a legjobb középpályás meghatározására

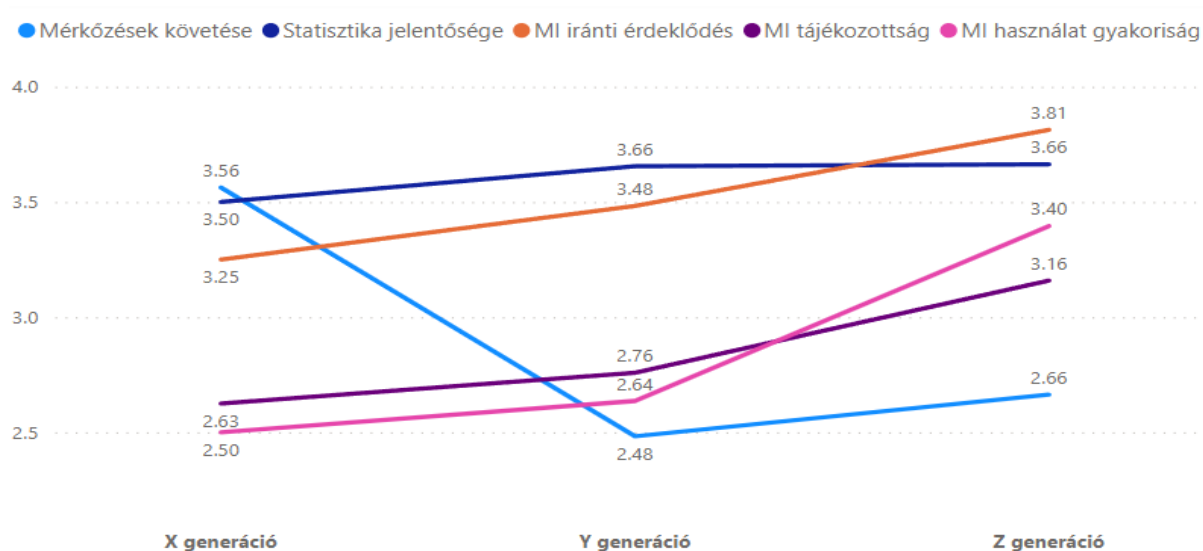
Forrás: Saját szerkesztés (2024)

Tudni kell, hogy ezek a lekérdezések anyagi vonzattal rendelkeznek, mivel minden egyes kérdés, úgynevezett „hívás”, pénzbe kerül. Az utolsó kérdés lefuttatása előtt egy módosítást végeztem a parancson, hogy szemléltessem, mennyibe került egy ilyen „hívás”, és megállapítottam, hogy a lekérdezés során 273 tokent használt fel az MI, ami a 2024. áprilisi árfolyam szerint körülbelül 3,77 Ft-nak felelt meg.

A fejlesztés sikerrel zárult, primer kutatásom további részében a kérdőívet elemzem.

4.2. A kérdőív értékelése

A kérdőívben öt likert típusú kérdés kapott helyet, amelyek felmérték a válaszadók futballmérkőzések követésével kapcsolatos szokásait, véleményüket a statisztikai elemzések jelenlétével kapcsolatban, az MI iránti érdeklődésüket, az abban való tájékozottságukat, valamint a mesterséges intelligencia használatának gyakoriságát. A válaszlehetőségeket egy 1-től 5-ig terjedő skálán adtam meg, ahol az 1-es a „nagyon ritkán” vagy „szinte soha”, a 2-es a „ritkán”, a 3-as a „néha”, a 4-es a „gyakran”, az 5-ös a „nagyon gyakran” vagy „szinte minden nap” értékeket vették fel. A 10. ábrán generációkra lebontott életkorok alapján vizsgáltam a megkérdezettek válaszait. A Baby Boomers generáció adatai a minta kis mérete miatt nem szerepelnek a vonaldiagramon.



10. ábra: Likert típusú kérdésekre adott válaszok megoszlása generációk tekintetében

Forrás: Saját szerkesztés, 2024 (N=154, N=128 MI használat gyakoriság esetén)

Az X tengelyen láthatók rendre az X, Y és Z generációk átlagos értékei, amelyekből megállapítható, hogy mely generáció tagjai hogyan vélekedtek a feltett kérdésekről. Az Y tengelyen a likert skálán megadott számok helyezkednek el, egy 2,4-től 4,0-ig terjedő skálán.

A mérkőzések követésének vizsgálata során látható, hogy az X generáció tagjai kimagasló, 3,56-os átlagos pontértéket adtak meg válaszként. Ezzel szemben az Y és Z generációk tagjai szinte egy egész pontnyival kevesebb átlagos pontértékkel rendelkeznek, ami azt jelenti, hogy jóval ritkábban néznek futballmérkőzéseket a televízióban, mint idősebb társaik. Ez ellentmond előzetes várakozásaimnak, mely szerint a Z generáció néz legtöbbet futball mérkőzéseket. A kérdésre adott összes válasz szórása 1,35 volt.

A statisztikai elemzések jelentőségének vizsgálata során megállapítható, hogy szinte alig van különbség a generációk véleménye között. Azon túl, hogy az Y és Z generációk értéke (3,66) pontosan megegyezik, az X generáció tagjainak válasza (3,50) is csak éppen, hogy alacsonyabb pontértéket képviselnek. Leszűrhető, hogy a statisztikai elemzésekről alkotott véleménye az embereknek nincs korhoz kötve, a fiatalok és az idősebbek is hasonlóan vélekednek ebben a kérdésben. A válaszok összesített szórása 1,21 volt.

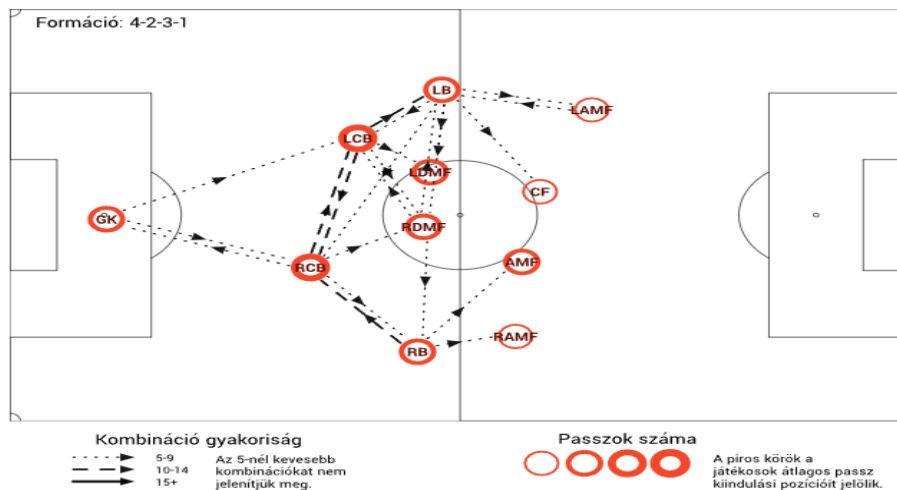
A mesterséges intelligencia utáni érdeklődés vizsgálata során megállapítható, hogy minél fiatalabbak a válaszadók, annál inkább érdeklődnek iránta. Míg az X generáció tagjainak válasza 3,25 és az Y generációé 3,48, addig a Z generáció 3,81-es átlagos pontértéke arra utal, hogy nagyobb mértékben foglalkoztatja őket a szakterület, mint idősebb társaikat. A szórás ebben az esetben 1,08 volt.

Az MI-ben való tájékozottságot és az MI használatának gyakoriságát felmérő kérdésekben hasonlóan alakultak a trendek az előző kérdés válaszaihoz. Míg az X generáció tagjai kevésbé tájékozottak és ritkábban használják az MI-t, addig a fiatalabbak jóval informáltabbak és sűrűbben használják a gyakorlatban az új technológiát. Előbbi kérdés esetén 1,05, utóbbi esetén 1,26 volt a válaszok szórása.

Összességében elmondható, hogy míg a futballmérkőzéseket az X generáció tagjai követik rendszeresebben, addig az Y és Z generációba tartozók az MI-vel kapcsolatos kérdésekben érték el magasabb átlagpontoszámot. A statisztikák jelentőségét vizsgáló kérdésben azonban nem volt felfedezhető számottevő különbség a generációk között.

Eredményeim elemzése során megállapítottam még, hogy minél inkább érdeklődik valaki a mesterséges intelligencia csevegőplatformjai iránt, annál valószínűbb, hogy elő fog fizetni ezen csevegőplatformok által kínált prémium funkciókra, fizetős tartalmakra.

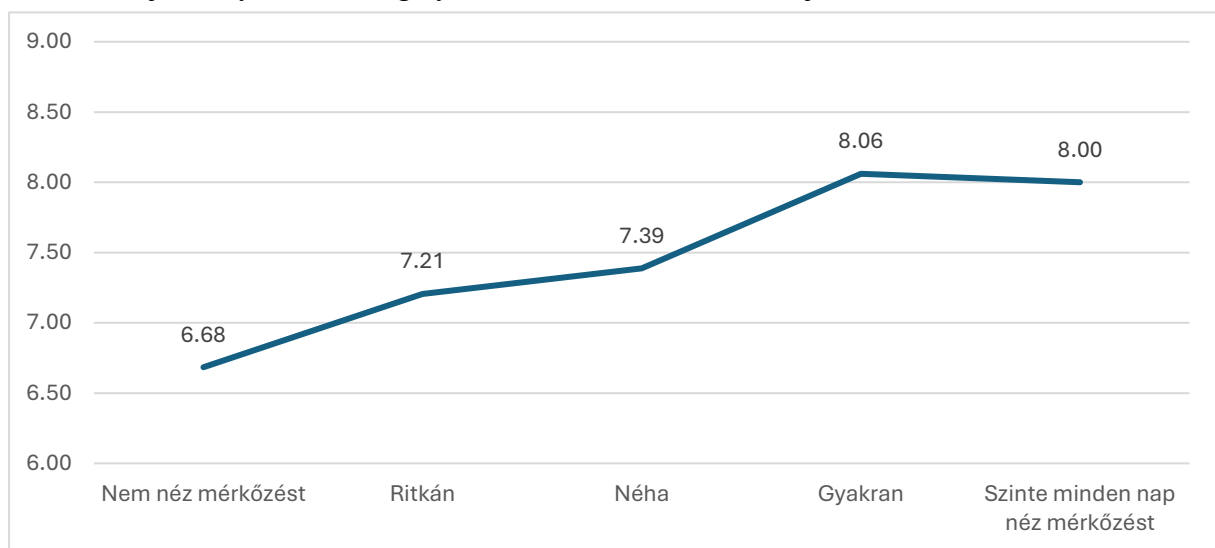
A következő kérdés során a kitöltőket az egyik NBI-es futballcsapat passzhálózatának vizualizációjáról (*11. ábra*) kérdeztem, mennyire tartják azt, illetve az ahhoz hasonló ábrákat hasznosnak, illetve, hogy milyen mértékben adhat ez segítséget az ellenfél csapatának feltérképezésekor.



11. ábra: Az egyik NBI-es csapat passzhálózata a 2023/2024-es szezonban

Forrás: xfb Analytics (2024), (N=154)

A válaszlehetőségek egy 1-től 10-ig terjedő likert skálán adtam meg, ahol az 1-es az „Egyáltalán nem ad segítséget”-et, míg a 10-es a „Nagy segítséget ad”-ot jelentette. A válaszok átlaga 7,39, szórása 1,97 lett. A vizsgálat során úgy gondoltam, érdemes megnézni, hogy a válaszok hogyan alakulnak az emberek mérkőzésekövetési szokásaival összefüggésben. Ennek eredményét a 12. ábra mutatja, melyen az Y tengely értékei 6,0 és 9,0 között terjednek.



12. ábra: A passzhálózathoz és ahhoz hasonló vizualizációk hasznossága a mérkőzésekövetéssel összefüggésben

Forrás: Saját szerkesztés, 2024 (N=154)

Az ábrán szépen kirajzolódik, hogy minél rendszeresebben nézik az emberek a mérkőzéseket, annál hasznosabbnak és fontosabbnak tartják az efféle vizualizációk meglétét az ellenfél csapatának feltérképezése céljából. Akik nem néznek futballmeccseket, 6,68-as átlagpontszámot jelöltek be a skálán. Ezt szem előtt tartva ez az érték így is magasnak mondható. A mérkőzéseket ritkán és néha követők rendre 7,21 és 7,39 átlagpontszámmal

rendelkeztek. Azok a megkérdezettek, akik gyakrabban tekintenek meg futballközvetítéseket, szinte egyenlő arányban, 8 pont körüli átlagot értek el, ami azt mutatja, hogy szükség van az efféle vizualizációkra.

Kérdőívem következő kérdésében a technológiai fejlesztéshez kapcsolódó kérdést tettem fel, ahol arra voltam kíváncsi, hogy hányan hallottak már az általam is használt LangChain könyvtárról korábban. Az eredmény nem volt meglepő számomra, hiszen igazolta, hogy nem túl széles körben ismert eszközről van szó. A 128 válaszadó közül csupán 9 fő, azaz a kitöltők 7,03%-a hallott már korábban erről a lehetőségről, amiből látszik, hogy egy egyáltalán nem elterjedt technológiáról van szó. Többek között emiatt is örömömre szolgál, hogy ennek a segítségével valósítottam meg a fejlesztést.

A dolgozat következő részében következtetéseket fogalmazok meg és javaslatokat teszek a kutatásokkal kapcsolatban.

5. Következtetések és javaslatok

A fejlesztés alappilléreinek számító LangChain könyvtárat a megkérdezettek mindössze 7,03%-a ismerte, így örültem, hogy ebben a környezetben valósítom meg az elképzelésemet. Ennek során megállapítottam, hogy a jövőben szükség lehet az efféle szoftverek használatára az emberi tevékenységek és a munka területén. Fontos az ügynök finomhangolása, hogy mindig megbízható és pontos eredményeket kapjunk. A fejlesztés a jövőben kivitelezhető lehet egy nagyobb adatbázis esetén is, biztosítva ezzel a széleskörűbb felhasználási módot. Mindazonáltal javaslataim között szerepel még az is, hogy az MI ne csak szöveges választ tudjon válaszként adni, hanem egy JavaScript Object Notation (JSON) formátumú kódot, ami által a végeredményben egyedi vizualizációk jöhetnek létre a hatékonyabb alkalmazás érdekében. A kérdőíves kutatás is ezt igazolja, hiszen az eredmények alapján elmondható, hogy nagy az igény a labdarúgás statisztikáit vizualizáló ábrákra. További kérdéseim során többek között szerettem volna felmérni, hogy melyik generáció használhatja nagyobb eséllyel a fejlesztést. Elmondható, hogy az 1965 és 1979 között született X generáció tagjai néznek a legsűrűbben futballmeccseket, ám a labdarúgó statisztikák jelentőségével foglalkozó kérdésben elmaradtak a fiatalabb generációktól. Emellett kevésbé érdeklődnek az MI iránt, kevésbé tájékozottak és ritkábban is használják azt. Ezzel szemben a Z generáció tagjai nagyon elkötelezettek az új technológia iránt, így véleményem szerint ők nagyobb eséllyel használhatják majd az általam kidolgozott fejlesztést. Bebizonyosodott az is, hogy minél sűrűbben néznek az emberek futballmérkőzéseket, annál inkább fontosabbnak tartják a

vizualizációkat. Javaslatom szerint a jövőben még nagyobb teret nyerhet a futballstatisztika és az MI, a kérdőív alapján pedig van is igény ezekre, így a két terület együttes alkalmazásával határtalan lehetőségek nyílnak meg a felhasználók előtt. Örömmre szolgálna még az adatalapú vizualizációk segítségével történő döntéshozatal a menedzsmenti szinteken, legyen szó futballról vagy akár az élet más területéről.

Összefoglalás

Tudományos diákköri dolgozatom első részében szakirodalmi források alapján ismertettem az adatelemzés alapjait. Ezt követően kifejezetten a futballadatok jelentőségével és hatásaival foglalkoztam. A szakirodalmi áttekintés utolsó részében a mesterséges intelligencia történetéről és elterjedéséről esett szó, majd minőségi szakirodalmi forrásokból vizsgáltam az MI fogalmát. Kutatásom egyik részében egy fejlesztést készítettem, amely a jövőben nagy szerephez juthat a technológia széleskörű elterjedésének köszönhetően. Ezek során egy átfogó képet kaptam a vizualizációk és a mesterséges intelligencia szükségességéről és hasznosságáról is. Ezt követően a vizsgálat második részében felmértem az emberek véleményét a futballal, futballstatistikákkal és a mesterséges intelligenciával kapcsolatban, mely során érdekes és megerősítő eredményeket kaptam. Javasolataim között szerepelt a vizualizációk létjogosultsága, az adatalapú döntéshozatal mielőbbi térnyerése, valamint az is, hogy a mesterséges intelligencia a válaszaiban jelenítsen meg ábrákat, diagramokat. Fontos még, hogy a növekvő adatmennyiségek miatt a fejlesztés működőképes legyen még nagyobb adatbázisok esetén is. Az emberek számára fontos a labdarúgás, hiszen az egyik legnépszerűbb sportágról van szó. Emellett elengedhetetlen haladni a sportban alkalmazandó modern fejlesztésekkel, hiszen a technológiai ipari forradalom ezt meg is követeli.

Összefoglalva elmondható, hogy egy hasznos, reprezentatív és átfogó kutatás készült, köszönhetően a kérdőív kitöltőinek és az xfb Analytics Kft. által nyújtott technikai háttérnek.

Szakirodalmi jegyzék

Bigcloud (n.d.). The Impact of Big Data Analytics in Football. <https://bigcloud.global/the-impact-of-big-data-analytics-in-football/>

Dharani, M., Jyostna, J. V. S. L., Sucharitha, E., Likitha, R. & Manne, S. (2020). Interactive transport enquiry with ai chatbot. *2020 4th international conference on intelligent computing and control systems (ICICCS)*, 1271-1276 IEEE.

Fitria, T. N. (2021). The use technology based on artificial intelligence in English teaching and learning. *ELT Echo: The Journal of English Language Teaching in Foreign Language Context*, 6(2), 213-223.

Fitria, T. N. (2023a). Augmented reality (AR) and virtual reality (VR) technology in education: Media of teaching and learning: A review. *International Journal of Computer and Information System (IJCIS)*, 4(1), 14-25.

Fogarassyné, Á. & Starkné, Á. (2011). *Intelligens adatelemzés*. Typotex.

Herberger, T. A. & Litke, C. (2021). The impact of Big Data and sports analytics on professional football: A systematic literature review. In: Herberger, T. A. & Dötsch, J. J. (Eds.), *Digitalization, digital transformation and sustainability in the global economy* (pp. 147–171).

Lampek, K. & Horváthné K., Zs. (2015): Általános kutatómódszertani alapok. In: Boncz, I. (Ed.) *Kutatómódszertani alapismeretek*. (pp. 9-59). Pécsi Tudományegyetem Egészségtudományi Kar, Pécs.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Luo, X., Tong, S., Fang, Z. & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, 38(6), 937-947.

McCarthy, J. (2004). What is artificial intelligence. Stanford University.

Microsoft. (2024). LangChain az Azure Databricksen LLM-fejlesztéshez. <https://learn.microsoft.com/hu-hu/azure/databricks/large-language-models/langchain>

SAS. (n.d.). Adatelemzés: Mit jelent és miért fontos? https://www.sas.com/hu_hu/insights/analytics/what-is-analytics.html

Turing, A. M. (2007). Computing machinery and intelligence. *Mind*, 59(239), 433-460 <https://doi.org/10.1093/mind/LIX.236.433>

Wei, C., Yu, Z. & Fong, S. (2018). How to build a chatbot: chatbot framework and its capabilities. *Proceedings of the 2018 10th international conference on machine learning and computing*, 369-373 <https://doi.org/10.1145/3195106.3195169>