>KÜLSŐ KÖTÉSTÁBLA<

A méret nem valós méret

DIPLOMADOLGOZAT

Csizmazia Máté   
2022

>CSERE ÜRES OLDALRA<

Pannon Egyetem   
Műszaki Informatikai Kar  
Rendszer- és Számítástudományi Tanszék   
Programtervező informatikus MSc

DIPLOMADOLGOZAT

Futóverseny eredményeinek kiértékelését támogató eszköztár és alkalmazás fejlesztése

Csizmazia Máté

Témavezető: Dr. Leitold Dániel

2022

>Feladatkiírás<

Csere az aláírt témakiírásra!

Nyilatkozat

Alulírott *Csizmazia Máté* hallgató, kijelentem, hogy a dolgozatot a Pannon Egyetem *Rendszer- és Számítástudományi* tanszékén készítettem az okleveles programtervező informatikus végzettség megszerzése érdekében.

Kijelentem, hogy a dolgozatban lévő érdemi rész saját munkám eredménye, az érdemi részen kívül csak a hivatkozott forrásokat (szakirodalom, eszközök stb.) használtam fel.

Tudomásul veszem, hogy a dolgozatban foglalt eredményeket a Pannon Egyetem, valamint a feladatot kiíró szervezeti egység saját céljaira szabadon felhasználhatja.

Veszprém, 2021. december 10.

Csizmazia Máté

Alulírott *Dr. Leitold Dániel* témavezető kijelentem, hogy a dolgozatot *Csizmazia Máté* a Pannon Egyetem *Rendszer- és Számítástudományi* tanszékén készítette okleveles programtervező informatikus végzettség megszerzése érdekében.

Kijelentem, hogy a dolgozat védésre bocsátását engedélyezem.

Veszprém, 2021. december 4.

Dr. Leitold Dániel

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani mindazoknak, akik szakmailag és erkölcsileg hozzájárultak a diplomadolgozatom megvalósításához. Kitartó munkájuknak, segítőkészségüknek, és szellemi támogatásuknak köszönhetően jöhetett létre a dolgozat.

Köszönöm témavezetőmnek, Dr. Leitold Dánielnek a sok segítséget, támogatást és építő kritikát. Munkája és szakmai hozzáértése nélkül nem születhetett volna meg ez a dolgozat és a hozzá tartozó szoftver.

Továbbá hálával tartozom családomnak a kitartó szellemi támogatásért és azért, hogy lehetővé tették tanulmányaimban való előre haladásom.

Tartalmi összefoglaló

A futás az emberek által egyik legszélesebb körben űzött edzésfajtája. Nem véletlen, hiszen mindenki számára könnyedén elérhető, nincs szükség speciális felszerelésre és meglehetősen jó hatással van az egészségre is. A sportágat űzők közül a többség természetesen csak amatőr szinten atletizál, de az amatőrök között is lehet találni olyan egyéneket, akik szeretnének szintet lépni a sportágban. Ilyen feljebb lépés esetében nagyon fontos, hogy az egyén fizikailag készen álljon erre, mivel, ha nem elég edzett, az kudarchoz vagy esetleges sérülésekhez is vezethet. Az atletizálásnak rengeteg fajtája van, de a dolgozatban főképp a maratonon és az ennél hosszabb futószámokon (szupermaraton) lesz a hangsúly.

A dolgozat célja, hogy az ilyen futómaratonok eredményeit részletesebben feltárja és jól vizualizálhatóvá alakítsa. A dolgozathoz egy olyan komplex felhasználóbarát alkalmazás fejlesztése a cél, amely lehetővé teszi az eredmények kiértékelését és historikus adatok alapján prediktív előrejelzéseket tud adni a felhasználó számára.

Az alkalmazás elkészítéséhez a manapság leginkább elterjedt és legnépszerűbb programozási nyelvet fogom igénybe venni, ami nem más, mint a Python. A választásom azért esett erre a technológiára, mivel népszerűségének köszönhetően hatalmas felhasználó bázissal rendelkezik, egyszerű a használata, rendelkezem előzetes tapasztalattal az alkalmazásában és nem utolsó sorban, ez a nyelv a leginkább támogatott az adatbányászat és Data Science területén. Rengeteg könyvtár áll rendelkezésre, ami megkönnyíti a feladatokat, ilyen például a pandas , aminek a segítségével gyorsan és könnyedén lehet adattisztítást végezni, a matplotlib adatvizualizációs feladatokat lát el, a scikit-learn pedig gépi tanuló modellek konstruálását látja el. Ezeken felül számos más hasznos könyvtár is létezik, amikről a későbbi fejezetekben bővebben lehet olvasni.

Abstract

Running is one the most broadly practiced sorts of excercise. It is no surprise, since it is easily accessible to everyone, does not require special equipment and has a rather good impact on health as well. Needless to say, majority of the runners do it as a hobby, yet there are others who want to step up and improve in this sport. In case of such a move, it is vital for the person to be truly prepared for it, if not, it can lead to disappointment or conceivable injury. There are many types of runs, however in this thesis, I will be focusing on the longer runs (marathons / supermarathons).

The point of the diplomawork is to better understand the results of running competitions and to easily visualize the data. The goal is to develop an easy to use application which helps in the evaluation of the data and can predict future outcomes for the user.

To create the application, I will utilize the most well-known and popular programming language today, which is Python. I sided with this technology since it has a massive user-base, it s fairly easy to use, I have previous experience in its application and to wrap things up, this language is the most supported in the field of Data Mining and Data Science. There are several libraries accessible to make work easier, such as pandas, which helps you with data cleaning, matplotlib to help you easily visualize data with diagrams, and scikit-learn to build machine learning models. There are various other helpful libraries which I have not mentioned now, but they will be discussed later on in this paper.

Tartalomjegyzék

[1. Bevezetés 1](#_Toc89112324)

[2. Data Science 3](#_Toc89112325)

[2.1. Adatbányászat 4](#_Toc89112326)

[2.1.1. Jelentős adatbányászati feladatok 4](#_Toc89112327)

[2.2. Web Scraping 6](#_Toc89112328)

[3. Adathalmazok 8](#_Toc89112329)

[3.1. Spuri Balaton Szupermaraton egyéni (BSZM) 8](#_Toc89112330)

[3.2. Boston maraton 9](#_Toc89112331)

[4. Technológiák 10](#_Toc89112332)

[5. Célkitűzés 14](#_Toc89112333)

[6. Megvalósítás 16](#_Toc89112334)

# Bevezetés

A futás, mint sport a világ egyik legszélesebb körben űzött edzésfajtája. Vannak, akik csak hobbiként tekintenek a futásra vagy ezt választják az általános edzésüknek, és vannak olyanok is, akik professzionális szinten űzik az atletizálást. Természetesen a futásnak is számos típusa létezik, ezekre néhány példa a könnyű futás, hosszú(táv) futás, fartlet, dombfutás. A dolgozatban főképp a hosszabb távú futószámokon lesz a hangsúly, ami nevén nevezve a maraton.

A maraton talán a legismertebb versenyfajtája a futásnak. Eredete egészen az ókori Görögországig visszavezethető, és hírnevét a futószám távja miatt érdemelte ki magának. Egy ilyen hosszúságú táv teljesítése többek szerint közel emberfeletti teljesítményt szimbolizál, emiatt is vált ez a sportág a kitartás egyik fő szimbólumává.

Létezik a maratonnak egy ennél is megerőltetőbb verziója, ami nem más, mint a szupermaraton. Erről azt érdemes tudni, hogy egy ilyen esemény általában egy többnapos versenyszám, amit lehet csoportosan váltóban is végezni, de a legedzettebbek egyénileg is képesek teljesíteni. A szupermaraton alatt megtett kilométer összesen meghaladhatja a 200-at is néhány esetben, ami egy 4 napos rendezvényen is napi legalább egy maratonnyi távot jelent. Érezhető, hogy egy ilyen feladat teljesítése talán még a legedzettebb futóknak is nagy kihívást jelenthet.

A dolgozat kiindulásához egy olyan adathalmaz áll rendelkezésre, ami egy ilyen szupermaraton eredményeit tartalmazza. Ez a futóverseny a BSZM (Balatoni Szupermaraton), ami lényegében a Balaton körbefutásáról szól. Ez egy négynapos esemény és távban körülbelül 180-200 kilométert jelent. A táv kissé eltérhet a különböző évek adatai között, mivel az útvonalat az évek során több esetben is megváltoztatták, ami a táv növekedését vagy csökkenését eredményezte.

A dolgozat elkészítésének az a fő célja, hogy ezen adathalmazok feldolgozásával nagyobb rálátást nyújtson az eredményekre, jól vizualizálhatóvá alakítsa ezt és hasznos vagy érdekes új információkat tárjon fel. Ehhez, egy olyan komplex alkalmazás elkészítése társul, ami lehetővé teszi az eredmények felhasználó által állítható kiértékelését, és historikus adatok alapján prediktív előrejelzéseket is tud adni a felhasználó számára.

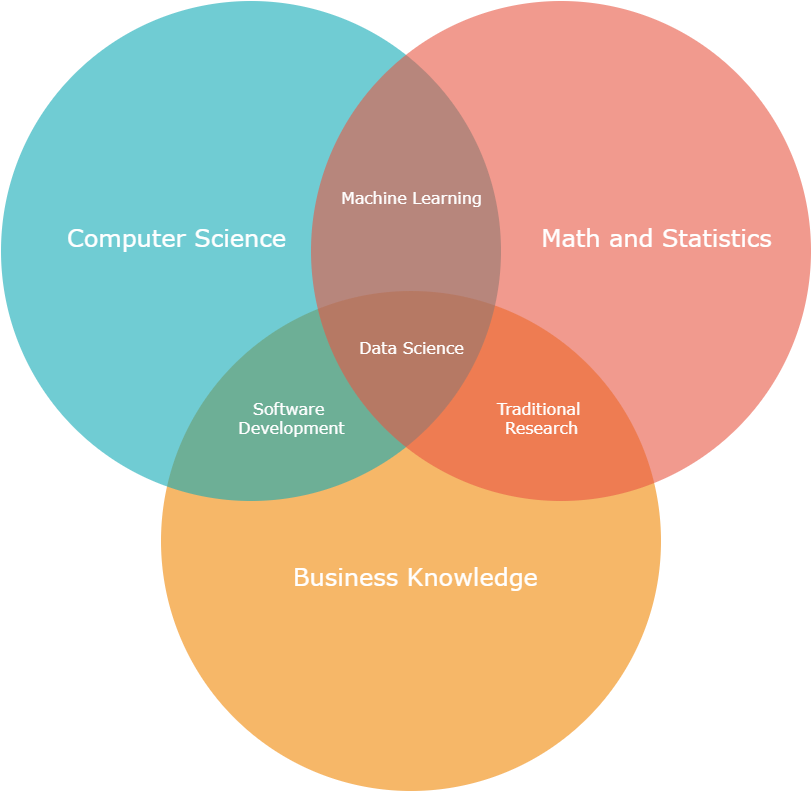
Ahhoz, hogy egy ilyen alkalmazást el lehessen készíteni, a manapság egyre népszerűbb és széles körben alkalmazott adatbányászati eljárásokat fogjuk felhasználni. Ez a technológia jelenlegi fénykorát éli az informatika világában, emiatt számos nagyvállalat számára is jó stratégiának számít. Alkalmazható kereskedelem, orvostudomány, autóipar, sport vagy számos más területeken is egyaránt.

Az alkalmazás elkészítéséhez manapság az egyik legnépszerűbb programozási nyelvet fogom használni, ami nem más, mint a Python. Egyértelmű volt, hogy ezzel a technológiával valósítsam meg az alkalmazást, mivel előzetesen már volt lehetőségem alkalmazni egyéni projektjeim során és hatalmas felhasználói bázissal rendelkezik, ami nagy segítséget tud jelenteni a fejlesztésben. Ezek mellett a használata és szintaxisa a nyelvnek meglehetősen egyszerű. A Data Science és az adatbányászat területén is kiemelkedő támogatásnak örvend a Python, de természetesen nem ez az egyetlen programozási nyelv, amin alkalmazni lehet.

Számos könyvtár áll rendelkezésre a Python nyelvhez, amik megkönnyítik a munkát. Ilyen például a pandas ami főképp olyan feladatokat lát el, aminek a segítségével könnyedén tudjuk az adathalmazt manipulálni, transzformálni és tisztítani. Másik példa a matplotlib, ami kifejezetten abból a célból jött létre, hogy az adatvizualizációt segítse rengeteg féle diagram segítségével. A scikit-learn egy másik gyakran használt könyvtár, aminek a segítségével gépi tanuló modelleket lehet konstruálni. Ezeken felül természetesen számos más hasznos könyvtár is létezik, amikről a későbbi fejezetekben lehet olvasni.

# Data Science

A Data Science egy olyan tudományos ágazat, ami kifejezetten a nagy mennyiségű adatok feldolgozásával és értelmezésével foglalkozik. Ez a rövid megfogalmazása, de a valóságban sok apróbb témát is lefed a kifejezés. Ha pontosabban szeretnénk vizualizálni, hogy mégis milyen fő részekből épül össze ez a tudomány, ezt legegyszerűbben egy Venn diagram segítségével tudjuk megtenni.



1. ábra Data Science Venn diagram

Számítástechnika

Ez talán a legtriviálisabb alkotóeleme ennek a tudománynak. Nélkülözhetetlen az alkalmazása, mivel ekkora adathalmazok feldolgozását csak számítógépek segítségével lehet optimálisan és könnyedén végrehajtani.

Matematika és statisztika

A matematika adja meg az alapját a Data Science-nek. A matematikának 3 fő részét érdemes kiemelni, ami a kalkulus, a lineáris algebra és a valószínűségszámítás / statisztika. Azért elengedhetetlen a használata, mivel a matematikának léteznek olyan eszközei, amivel kiszámolhatjuk a nagy adatok közötti összefüggéseket, mintákat találhassunk az adatok között vagy következtetéseket vonjunk le a feldolgozott adatok alapján. Olyan fontos matematikai fogalmak fognak előkerülni a későbbi fejezetekben, mint az átlag, medián, szórás, korreláció analízis, valószínűségszámítás és még rengeteg más is.

Üzleti tudás

Ahhoz, hogy a lehető legeffektívebben lehessen alkalmazni a technológiát, elhanyagolhatatlan a területi vagy üzleti ismereteknek a megléte.

Az ábrán látható metszetek többnyire maguktól értetődőek, hiszen szoftverfejlesztéshez az informatikai és a területi ismeretek elegendőek ahhoz, hogy egy üzleti alkalmazás elkészüljön, gépi tanuló algoritmusok elkészítéséhez matematikára és informatikára van szükség, és nem utolsó sorban, tradicionális módon is lehet kutatást végezni, de ez nem a legoptimálisabb eljárás manapság. A három főtémakör metszete pedig nem más, mint maga a Data Science.

## Adatbányászat

Az adatbányászati algoritmusokat az adathalmaz vagy adatbázisból való tudásfeltárás során alkalmazzák. Ez egy olyan folyamat, amelynek során valamilyen újszerű, hasznos információt vagy mintát sikerül felfedezni. Napjainkban leginkább hasznosított területe az üzleti szférában érzékelhető, de előszeretettel használják minden területen, legyen szó genetikáról, orvostudományról vagy sportról.

### Jelentős adatbányászati feladatok

Az adatbányászat egy több lépésből álló munka. Ahogy már említettük, nagy adathalmazokat feldolgozó feladatokat végez. Ekkora mennyiségű adat összegyűjtése, nem minden esetben áll rendelkezésre, ezért sokszor a munka az adatok összegyűjtésével kezdőik.

Adatelőkészítés

Amennyiben megvan az adathalmaz, ezt a további lépésekben elő kell készíteni mielőtt adatbányászati algoritmusokat alkalmaznánk rajtuk, mivel ezek csak az adatokban ténylegesen jelenlévő mintákat képesek feltárni. Az előfeldolgozás (pre-processing) elengedhetetlen lépése a folyamatnak, ahol az adatkészletet megtisztítják és transzformálják.

Gyakori minták kinyerése

Cél megtalálni gyakran előforduló objektumokat. Ezt a legegyszerűbben úgy lehetséges, hogy vizualizáljuk az adatokat diagramok segítségével.

Attribútumok közötti összefüggések

Attribútumok, másnéven tulajdonságok között azért érdemes megvizsgálni a kapcsolatokat, mivel ezáltal lehetőségünk van pótolni a hiányos adatokat és lehet javítani a zajokat. Többféle típusa is létezik a folyamatnak, általában az adattípusnak megfelelő eljárást érdemes használni. Skála típusú változók esetében az összefüggést korreláció és regresszió analízissel tudjuk kimutatni. Ordinális típusú változóknál rangkorrelációs mutatók mérik ezt.

Csoportosítás

Csoportosítás folyamán az adatokat külön klaszterekbe kell elválasztanunk úgy, hogy az egy csoportba tartozó elemek mind hasonlóak legyenek, de a különböző csoportokban lévők pedig különbözzenek egymástól.

Osztályozás

Osztályozás eljárás során arra törekszünk, hogy egy rekordot, előzetes adatok alapján, valamilyen osztályba tudjunk sorolni. Erről a későbbi fejezetekben bővebben lehet olvasni.

Regresszió

A regresszióanalízis segítségével egy függő változó és vagy több független változó közötti kapcsolatok becslése végezhető. Több változata is létezik, de a legelterjedtebb formája a lineáris regresszió, amelynek során egy olyan egyenes megtalálása a cél, amely a legjobban illeszthető az adatokhoz.

Dimenziócsökkentés

A dimenziócsökkentés az adathalmaz olyan átalakítását jelenti, amely során a nagy dimenziójú térből egy alacsony dimenziós teret hozunk létre úgy, hogy az alacsony dimenziós ábrázolás megtartja az eredeti adatok fontos tulajdonságait, ideálisan a belső dimenziójához közel.

Szövegbányászat

A szövegbányászat manapság egyre inkább elterjedt módszer. Lényege, hogy a strukturáltalan szöveget strukturált formátumúvá alakítsa, hogy értelmes mintákat tudjon felfedezni bennük. Számos algoritmus segíti a vállalatokat, mint a Naive Bayes vagy a Support Vector Machines, hogy felfedezhessenek strukturált adataikon belüli rejtett kapcsolatokat.

## Web Scraping

Amint az előző fejezetben említettem, az adatok beszerzése sokszor bonyolultabb lehet, mint gondolnánk. Amennyiben szerencsénk van, ezt készen is elérhetjük csv formátumban, de sajnos a legtöbb esetben ez nem így van. Ilyenkor segítségünkre vannak olyan technológiák, amelyek segítségével automatizált módon összegyűjthető az adat, legyen szó bármilyen weboldalról. Ez a technológia nem más, mint a webscraping. Ahogy a neve is utal rá, itt ténylegesen a weboldal forráskódjából „kaparjuk” le az adatokat, valamilyen automatizált algoritmus segítségével.

Selenium

A selenium talál a legelterjedtebb könyvtár, amivel lehet webscraping-et alkalmazni. Több programozási nyelvet is támogat, mint a Java, Python, C#, stb. Ez egy ingyenes nyílt forráskódú keretrendszer, amit több céllal is lehet használni, de a mi esetünkben, csak a webscraping-re lesz alkalmazva. Működése meglehetősen egyszerű, mivel kódon belül, egyszerűen manipulálható teljes mértékben a működése. Néhány böngésző esetében engedélyezni kell, hogy automatikusan tudjon műveleteket végrehajtani, erre példa a Safari, de ez is pár kattintással végrehajtható. Többnyire az összes böngészőt támogatja, tehát bárki számára elérhető.

A seleniumot alkalmazva gyakorlatilag bármilyen weboldalról tudunk automatizálva adatokat begyűjteni, ahol táblázatosan kilistázható az információ. Innentől csak annyi dolgunk van, hogy megírjuk az algoritmust ami a HTML-ből kinyeri az adatokat, egy közös szerkezetre hozza, és végül kiexportálja egy olyan formátumba, amit már fel tudunk dolgozni egyenesen pythonnal.

# Adathalmazok

A dolgozat során többféle adathalmaz is felhasználásra fog kerülni a legrészletesebb elemzések érdekében. Erre azért van szükség, mivel az adathalmaz, amivel eleinte kiindultam, sajnos nem tartalmaz adatokat olyan számosságban, ami optimálisnak tekinthető. Ezért lesz számunkra fontos, hogy olyan nevezetes maratoni futóeredményeket is bevethessünk a dolgozat során, ami lényegesen több rekordot (akár tízszeres vagy húszszoros mennyiségű) tartalmaz.

## Spuri Balaton Szupermaraton egyéni (BSZM)

Azért esett a választásom kifejezetten erre az adathalmazra, mivel egy érdeklődő személy kifejezetten ennek a versenyszámnak a részleteire volt kiváncsi. Egy olyan részletes elemzésre volt szüksége, amivel tisztán láthatóvá válhat, hogy a jelenlegi teljesítménye elegendő-e egy ilyen nagy kihívás teljesítésére. Emellett érdekelték olyan érdekességek is mint, hogy milyen eloszlású volt a futók sebessége vagy milyen arányban teljesítették a hasonló korabeli futók a versenyt. Ezen felül számos érdekesnek tekinthető információ is kinyerhető akár ebből a kisebbnek nevezhető adathalmazból.

Egy magyar weboldalról sikeresen összegyűjthető volt a Balatoni szupermaraton előző 6 év eredményeinek az adathalmazai. A régi weboldaluk, ahol volt lehetőség az adatokat xlsx formátumban letölteni sajnálatos módon jelenleg már nem elérhető, ezért csak ezt a 6 évnyi adatot kerül megvizsgálásra. A vizsgálat során hamar fény derül arra, hogy még rengeteg adattisztítást és adattranszformációt lesz szükséges elvégezni az adathalmazon ahhoz, hogy érdembeli eredményeket lehessen kapni az adatfeltárás során.

A Futanet weboldalán természetesen több más futóverseny adatai is megtalálhatóak, amiket a későbbiekben valamilyen webscraping megoldással lehet majd összegyűjteni, amennyiben nem tér vissza a régi weboldal, ahol volt lehetőség ezen adatokat exportálni egyenesen csv formátumba.

## Boston maraton

Kisebb kutatás után sikerült megtalálnom talán a világ leghíresebb futóversenyének 2019-es évi adathalmazát. Ebben az évben körülbelül 28 ezer sornyi értéket vittek fel, ami sokszorosa a BSZM előző 6 évi adatának, így lényegesen hasznosabb lesz ennek a feldolgozása a későbbiekben.

# Technológiák

Ahhoz, hogy a már meglévő adathalmaz feldolgozásra kerülhessen, alkalmaznunk kell különböző technológiákat, könyvtárakat és keretrendszereket. Mivel a python a legtámogatottabb nyelv a Data Science világában, ezért egyértelmű volt, hogy ebből érdemes kiindulni.

***Python***

A python a mai világ egyik legismertebb és legtöbbet használt programozási nyelve. Nem véletlen, hiszen alkalmazása és szintaxisa meglehetősen egyszerű, így vonzó lehet a tapasztalatlan, kezdő fejlesztőknek. Természetesen nem csak kezdő szintű fejlesztők alkalmazzák, mivel így is egy egyszerű, de magas szintű, nyílt forráskódú nyelvről beszélünk.

A python már számos területen bizonyította hasznosságát:

* Webfejlesztés,
* Játékfejlesztés,
* Tudományos számításokat végző alkalmazások,
* Mesterséges intelligenciához köthető tudományok,
* Asztali alkalmazások
* Nagyvállalati alkalmazások
* Operációs rendszerek

Jól látható, hogy valóban számos területen alkalmazzák, így népszerűsége nem is kérdéses.

Amint említettem, ez a nyelv rendelkezik a legtöbb olyan könyvtárral és keretrendszerrel, ami megkönnyíti az adatbányászati feladatokat.

***Jupyter Notebooks***

A Jupyter Notebook egy olyan nyílt forráskódú interaktív webalkalmazás, ami az adatbányászok számára lehetővé teszi olyan dokumentumok létrehozását, ami könnyedén megosztható és python kódot tud futtatni. Ez kiegészül egy markdown rendszerrel, aminek a segítségével könnyedén és jól láthatóan dokumentálható integráltan az egész munkamenet. Ennek a segítségével, könnyedén megjeleníthetővé válnak egy közös környezetben egyenletek, számítási eredmények, vizualizációk vagy képek.

Alkalmazása az összes adatbányászati feladatban hasznos, beleértve az adattisztítást, adattranszformációt, feltáró adatelemzést, adatvizualizációt, modellezést, gépi tanulást és még sok mást.

Tehát összefoglalva, egy olyan adatbányászati környezetet biztosít a felhasználó számára, ami nem csak egy fejlesztői környezetként működik, hanem egyben egy bemutatót, prezentációt segítő eszköz is.

A Notebook 2 fő összetevőkből áll össze, kódcellákból és markdown cellákból. A kódcellákba írt programot a weboldal továbbítja egy háttérkernelnek, amely lefuttatja a kódot és visszaadja az eredményeket a weboldal számára. Ezeknek a kerneleknek nem feltétlenül szükséges a használt eszközön futniuk, erre tökéletes példa a Google saját hasonló működést biztosító eszköztára, ami a Google Colaboratory projektnévre hallgat. Amennyiben saját eszközről kivánja a felhasználó futtatni a Jupytert, ezt megteheti akár hálózati hozzáférés nélkül is, és helyileg tudja végezni a munkát.

***Pandas***

A pandas a legelterjedtebb szoftverkönyvtár, amit a python nyelven belül adatmanipulációra és analízisre használnak. Számos különféle forrásból képes beolvasni adatokat, legyen szó csv-ről, xlsx-ről, vagy akár html-ről. A beolvasott adatokat egy saját DataFrame elnevézsre hallgató objektumban tárolja el, amin már a beépített műveleteket el lehet végezni. Ilyen műveletek például a több adatforrás összefűzése, adatmanipuláció, logikai vizsgálatok, adattisztítás vagy adatdiszkretizáció. Könyvtár rendkívül jól optimalizált teljesítmény szempontjából, ami miatt rendkívül jól alkalmazható nagy adathalmazok feldolgozására és módosítására.

***Numpy***

Egy olyan python könyvtár, amely támogat nagy, többdimenziós tömböket és mátrixokat. Rengeteg féle magas szintű matematikai függvényt tartalmaz.

***Matplotlib***

Alkalmazásával gyorsan és könnyedén lehet diagramokon megjeleníteni a DataFrame-ben tárolt adatokat. Rengeteg diagram áll rendelkezésre, legyen akár szó histogramokról, kör vagy vonaldiagramokról.

***Seaborn***

Hasonlóan a matplotlib-hez ez is egy diagramokat készítő és megjelenítő könyvtár. Főképp csak kinézetben különbözik a másik könyvtártól.

***Scikit-learn***

A scikit-learn egy ingyenesen elérhető gépi tanulást támogató könyvtár, ami a python programozási nyelvhez íródott. Többféle eljárást is támogat, mint osztályozás, regresszió, csoportosítás. Ezen eljárásoknak többféle fajtáját is támogatja.

***Streamlit***

A streamlit névre hallgató keretrendszer nemrégiben kezdett egyre inkább nagyobb népszerűségre szert tenni. A könyvtár segítségével gyorsan és könnyedén lehet adatbányászati projekteket egy weboldal formájára átalakítani. Számos beépített komponensének köszönhetően szinte bármilyen feladatot látványosan meg lehet jeleníteni. Szinte minden könyvtárat támogat amit adatbányászati feladatokhoz a leginkább alkalmaznak a felhasználók.

Rettentően egyszerű használni és telepíteni. A python package manager-ének segítségével (pip) telepítés után, pár parancs kiadásával bele is lehet tekinteni milyen szolgáltatásokat nyújt a streamlit. Interaktív módon akár egyből a felületen módosíthatóak a bemeneti értékek, szűrhetőek az adatok, vagy újra futtathatóak a parancsok. Az alkalmazás futtatása után valójában egy webszervert indít a számítógépen, amit innentől kezdve bármilyen lokális webböngésző segítségével el lehet érni. Nagy előny, hogy amikor a script file módosításra kerül, ilyenkor a rendszer automatikusan betölti a változtatásokat az oldalon, így nem kell folyamatosan újraindítani a webszervert vagy külön bővítményeket telepíteni, amik megteszik ezt helyettünk.

Ezek mellett a streamlit szolgáltat egy külön felhő szolgáltatást is a regisztrált felhasználóknak, ahová pillanatok alatt telepíthetjük az alkalmazást. Természetesen ez csak egy bizonyos fokig ingyenes, de így is három projekt publikálására ad lehetőséget költségek nélkül. Ahhoz, hogy az alkalmazás futhasson a felhő rendszerben, nincs másra szükség, hogy a publikálandó projektet egy GitHub repository-ban tároljuk el. Ennek a repository-nak publikusnak kell lennie ahhoz, hogy a streamlit felhő rendszere is el tudja érni. Miután megadtuk a repository elérését a streamlit felhő kezelő oldalán, a rendszer elkezdi telepíteni az alkalmazást. Amint végzett a telepítéssel, elérhetővé válik egy URL segítségével, amit a világ minden pontjáról el lehet majd érni.

# Célkitűzés

Adathalmaz feldolgozásánál az elsődleges célpont az, hogy lehetőség szerint a legrészletesebben feltárásra kerüljön az elérhető adat. Ennek a célnak az elérése egy többlépéses folyamat.

**1. lépés:** Első lépésben szükséges meghatározni a feladatot. Ehhez arra van szükség, hogy az adott témát részletesebben megismerjük. Ha kialakult egy elképzelés, hogy milyen adatok lehetnek érdekesek a témában, akkor folytatódhatnak a munkálatok.

**2. lépés:** Ebben a lépésben szükséges megtalálni azokat az adatforrásokat, ahonnan az adatok fognak származni. Ezekből a forrásokból az adatokat össze kell gyűjteni és egy közös szerkezetre hozni. Ezek után érdemes ezt egy file-ban eltárolni (csv)

**3. lépés:** Az összegyűjtött adatokat ebben a lépésben rendkívül fontos megtisztítani, mivel, ha ez a lépés kimarad, a későbbiekben gyakran hamis kiértékelésekhez vezethet.

**4. lépés:** A megtisztított adatokat amennyiben lehetséges, érdemes bővíteni. Itt olyan transzformációkat lehet elvégezni, amivel plusz olyan oszlopokat hozunk létre, amik hasznosak lehetnek a bányászat során. Ezek lehetnek egy másik adatforrásból érkező információk, vagy a meglévő oszlopok átalakításával vagy módosításával létrehozott új attribútumok.

**5. lépés:** Ebben a lépésben egy olyan adathalmaz áll rendelkezésre, amiben nagy valószínűséggel már lehet találni valamilyen mintázatot vagy érdekességet a diagramok segítségével. Ezt a lépést a feltáró adatelemzésnek nevezik (Exploratory Data Analysis)

**6. lépés:** Ebben a lépésben már nincs akadálya annak, hogy valamilyen gépi tanuló algoritmus alkalmazzunk modellépítéshez. Az első lépésben meghatározott feladatokhoz mérten szükséges választani a rendelkezésre álló algoritmusok közül. A gépi tanuló modell létrehozása után, alkalmazhatóak a kiértékelésnél.

**7. lépés:** Az adatbányászat esetében nem minden esetben sikeres első alkalommal az adatfeltárás vagy vizsgálat. Ezesetben többszöri iterációra van szükség.

Elsődleges cél a munka során egy olyan gépi tanuló modell építése és alkalmazása, ami képes megjósolni egy futó nevezésének a kimenetelét. A kérdés egyszerű. Képes-e teljesíteni egyénileg egy ilyen hosszúságú szupermaraton-t, vagy felkényszerül adni a versenyt valamelyik nap. Az is vizsgálható, hogy melyik napokon adták fel a legtöbben, esetleg melyik szakasz tekinthető a legnehezebbnek. Ehhez további kutatásra lesz szükség, mivel az elérhető adathalmazokon belül jelenleg ilyen információk nincsenek jelen, tehát valamilyen külső segítséggel adatbővítésre lesz szükség.

Amennyiben sikerült feltárni azokat az információkat, amiket célul kitűztünk, lehet átültetni a kapott eredményt egy interaktív webes felületre, hogy ezt a kutatást mások számára is elérhetővé tegyük.

# Megvalósítás