>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



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, 2022. április 25.

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, 2022. április 25.

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. Magyarországon létezik egy népszerü hosszútávú futóverseny, ami a BSZM (másnéven Balaton Szupermaraton) névre hallgat. Ez az ultramarathon egy igazán érdekes kihívás minden hosszútávú futó számára. Kihívás, hiszen közel 200 kilóméternyi táv lefutását jelenti 4 nap lefolyása alatt. Emiatt találtam fontosnak, hogy elkészüljön a futóverseny eredményeinek kiértékelését támogató eszköztár és alkalmazás.

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. A predikciót különböző, gépi tanuló algoritmusok segítségével tudjuk megvalósítani, így a végfelhasználó számára többnyire pontos visszajelzést tudunk adni, a jövőbeli verseny kimeneteléről. A előrejelzés pontosításának az érdekében különböző mintavételezési és keresztvalidációs eljárások alkalmazására volt szükség.

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, és ez a nyelv a leginkább támogatottabb az adatbányászat és Data Science területén. Ezek mellett ,rengeteg keretrendszer és könyvtár áll rendelkezésre, ami segíti a fejlesztés menetét.

**Kulcsszavak:** [BSZM, Gépi tanulás, Data Science, Osztályozás, Python, Vizualizáció]

Abstract

Running is one of the most widely practiced types of training by people. It is no surprise, since it is easily accessible to everyone, does not require special equipment and has a fairly good effect on health as well. There exists a popular long-distance running competition in Hungary called BSZM (also known as Balaton Supermarathon). This ultramarathon is a real challenge for any long distance runner. It is difficult, as it means covering a distance of almost 200 kilometers in 4 days. For this reason, a toolkit and and application to support the evaluation of the results of the competition was created. The aim of the dissertation is to explore the results of such running marathons in more detail and to make them well visualizable.

The goal is to develop a complex user-friendly application that allows the evaluation of the results and can provide predictive predictions to the user based on historical data. Predictions can be implemented using various machine learning algorithms, so we can give the end user accurate feedback on the outcome of future competitions. Different sampling and cross-validation procedures were required to refine the prediction accuracy.

To create the application, I will use one the most common and popular programming language today, which is none other than Python. I chose this technology because of its popularity, it has a huge user base and community, it is easy to use, and it is the most supported language in Data Mining and Data Science. In addition, there are plenty of frameworks and libraries available to help with the development process.

**Keywords:** [BSZM, Machine Learning, Data Science, Classification, Python, Visualization]

**Tartalomjegyzék**

[1. Bevezetés 1](#_Toc102489698)

[2. Data Science 3](#_Toc102489699)

[2.1. Adatbányászat 4](#_Toc102489700)

[2.1.1. Főbb adatbányászati feladatok 4](#_Toc102489701)

[2.2. Web Scraping 6](#_Toc102489702)

[3. Futóversenyek és elemzéseik 8](#_Toc102489703)

[3.1. Maraton elemzések 8](#_Toc102489704)

[3.2. Ultramaraton elemzések 8](#_Toc102489705)

[3.3. Balaton Szupermaraton 8](#_Toc102489706)

[4. Technológiák 9](#_Toc102489707)

[4.1. Python 9](#_Toc102489708)

[4.2. Jupyter Notebook 9](#_Toc102489709)

[4.3. Visual Studio Code 10](#_Toc102489710)

[4.4. Pandas 10](#_Toc102489711)

[4.5. Plotting könyvtárak 11](#_Toc102489712)

[4.6. DataPrep 11](#_Toc102489713)

[4.7. Scikit-learn 11](#_Toc102489714)

[4.8. Streamlit 12](#_Toc102489715)

[5. Célkitűzés és irodalomelemzés 13](#_Toc102489716)

[6. Megvalósítás 15](#_Toc102489717)

[6.1. Első iteráció 15](#_Toc102489718)

[6.1.1. Adathalmaz 15](#_Toc102489719)

[6.1.2. Adatmigráció 16](#_Toc102489720)

[6.1.3. Adattisztítás és transzformáció 16](#_Toc102489721)

[6.1.4. Adatdiszkretizáció 17](#_Toc102489722)

[6.1.5. Feltáró elemzés 17](#_Toc102489723)

[6.1.6. Osztályozás 18](#_Toc102489724)

[6.1.7. Kiértékelés 19](#_Toc102489725)

[6.2. Probléma 19](#_Toc102489726)

[6.2.1. Kis adathalmaz 19](#_Toc102489727)

[6.2.2. Kiegyensúlyozatlan adathalmaz 19](#_Toc102489728)

[6.2.3. Zajos adathalmaz 19](#_Toc102489729)

[6.3. Második iteráció 19](#_Toc102489730)

[6.3.1. Adathalmazbövítés 19](#_Toc102489731)

[6.3.2. Adattisztítás és transzformáció 20](#_Toc102489732)

[6.3.3. Feature engineering 20](#_Toc102489733)

[6.3.4. Importance analysis 20](#_Toc102489734)

[6.3.5. Feltáró elemzés 20](#_Toc102489735)

[6.3.6. Osztályozás 20](#_Toc102489736)

[6.3.7. Kiértékelés 21](#_Toc102489737)

[7. Konklúzió 22](#_Toc102489738)

[8. Alkalmazás 23](#_Toc102489739)

[8.1. Telepítés 23](#_Toc102489740)

[8.2. Konfiguráció 23](#_Toc102489741)

[8.3. Streamlit 23](#_Toc102489742)

[8.4. Futtatás 23](#_Toc102489743)

[8.5. Felhö 24](#_Toc102489744)

[8.6. Felületek 24](#_Toc102489745)

[8.6.1. Statisztika 24](#_Toc102489746)

[8.6.2. Modellek 24](#_Toc102489747)

[8.6.3. Elörejelzés 24](#_Toc102489748)

[8.6.4. Adathalmaz feltöltés 24](#_Toc102489749)

[9. Továbbfejlesztési lehetöségek 25](#_Toc102489750)

[Irodalomjegyzék i](#_Toc102489751)

[Mellékletek ii](#_Toc102489752)

[Ábrajegyzék iii](#_Toc102489753)

[Táblázatjegyzék iv](#_Toc102489754)

# 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 az ultramarathon. 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 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.

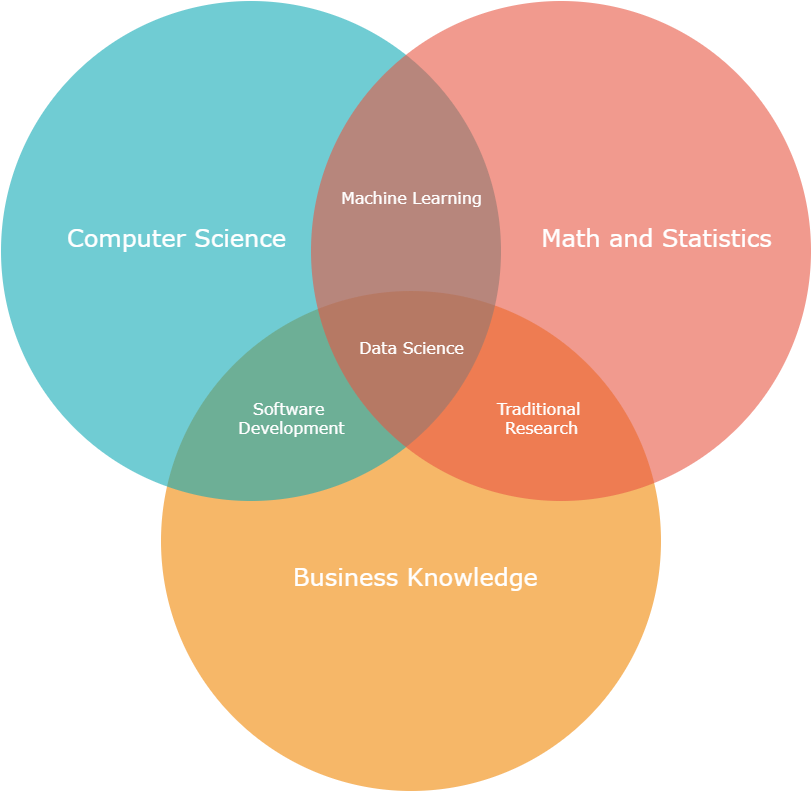
Ahhoz, hogy egy ilyen alkalmazást el lehessen készíteni, a manapság egyre népszerűbb és széles körben alkalmazott adatbányászati eljárásokat fogjuk felhasználni. Ez a technológia jelenlegi fénykorát éli az informatika világában, emiatt számos nagyvállalat számára is jó stratégiának számít. Alkalmazható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

### Főbb adatbányászati feladatok

Az adatbányászat egy több lépésből álló folyamat. Ahogy már említettem, 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.

# Futóversenyek és elemzéseik

## Maraton elemzések

## Ultramaraton elemzések

## Balaton Szupermaraton

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.

# 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 ezt alkalmazzam a fejlesztéshez.

## Python

A python manapság a 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
* Szkriptírás

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 Notebook

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.

## Visual Studio Code

A Visual Studio Code volt az a szövegszerkesztö amire a választásom esett, mivel nyílt forráskodúságának köszönhetöen, rengeteg hasznos pluginal bövíthetö és használható. Ilyen például a Jupyter kiegészítöje is, ami technikailag egy jupyter notebook environmentet imitál, csak a szövegszerkesztön belül. Azért preferálom jupyter VSCode-on belüli használatát, mivel így sokat tudnak segíteni a szövegszerkesztö hasznos funckiói is, mint például a kódkiegészítö használata vagy a kódsnippetek billentyüparancsra való beillesztése

## 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.

## Plotting könyvtárak

***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.

## DataPrep

Egy olyan könyvtár, ami a feltáró elemzés lépéseit teszi egyszübbé. Automatikusan készít kimutatást egy adathalmazról. Kifejezetten Jupyter notebook felhasználóknak készült, és ennek a használatára van a legjobban optimalizálva. A generált kimutatás beágyazható egyenes a notebook fájlba, vagy kiexportálható PDF, vagy egy weboldal formátumába is.

## 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 web alkalmazás 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.

Kifejezetten egyszerű a használata és a telepítése. 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 és irodalomelemzés

Adathalmaz feldolgozásánál az elsődleges célpont az, hogy lehetőség szerint a legrészletesebben feltárásra kerüljön az elérhető adat és hogy a lehetö legpontosabb elörejelzést tudjuk adni a felhasználó számára. 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.

A 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

A megvalósítás a tipikus adatbányászati lépésekkel történt meg. Mivel rendelkezésre állt az adathalmaz, nem volt szükség különbözö forrásból begyüjteni ezeket, ezért ez a lépés az elsö iteráció során kimaradt, de egy másik iteráció folyamán szükség volt adatbövítésre. Természetesen a korábban már ismertetett cél is kialakult már, emiatt nem kellett ezzel sem huzamosabb idöt tölteni. A tényleges munkafolyamatok lépéseinek részletesebb leírását a további fejezetekben lehet megismerni.

## Első iteráció

A munka kezdetével elöször az elérhetö adathalmazt kezdtem feldogozni, ami a 2015 és 2020 –as évek közötti egyéni eredményeket foglalta magába. A kiinduló adathalmaz magas dimenziójúnak tekinthetö, mivel számos tulajdonsággal rendelkezik. Tekintsünk is bele, mégis milyen oszlopok voltak elérhetöek a kiindulásnál.

### Adathalmaz

Amint már említettem, a kiinduló adathalmaz számos tulajdonsággal rendelkezett alapból. A tulajdonságok között voltak hasznos oszlopok, viszont számos olyat is tartalmazott, amire több mint biztos, hogy nincs szüksége az osztályozó modellnek a tanítás folyamán. Ettöl függetlenül, felhasználható adatról van szó így is, amennyiben diagramok segítségével szeretnénk a verseny eredményiröl kimutatást vagy egy vizualizációs alkalmazást készíteni.

1. táblázat kiinduló adathalmaz tulajdonságai típusmegjelöléssel

|  |  |  |  |
| --- | --- | --- | --- |
| **Kiinduló adathalmaz** | | | |
| **Nominális** | **Ordinális** | **Folytoson** | **Diszkrét** |
| Név | Helyezés | Rajtszám | Kategória |
| Ország | Születési év | Esemény éve | - |
| Csapat | Napi bontások idöadatai | - | - |
| Város | Napi összesített idöadatok | - | - |
| Nem | Végeredmény | - | - |
| Rajtszám | Megtett táv | - | - |

### Adatmigráció

Mivel külön CSV-fájlokban voltak eltárolva az adathalmazok, ezért ezeket egy közös szerkezetre kellett migrálni. Ehhez segítségünkre van a pandas könyvtár. Használatával könnyedén összeadhatóak az adathalmazok, amennyiben az oszlopok és a tulajdonságok megegyeznek a különbözö évi adatoknál. A mi esetünkben szerencsére megegyeznek, ezért ezt pár sornyi kóddal végre is tudjuk hajtani. Az így beolvasott CSV fájlokat egy DataFrame-re hallgató adattípusba olvassa be a pythonm amit pandas már megfelelöen tudja tudja kezelni. Ez az adattípus egy olyan 2 dimenziós címkézett adatstruktúra, ami képes különbözö típusú oszlopokat tárolni. Úgy is lehet rá tekinteni, mint egy egyszerü SQL táblára, vagy egy dictionary-re, ami sorozatobjetumokat tartalmaz.

### Adattisztítás és transzformáció

Sajnos a való életben nem gyakori az az eset, amikor a feldolgozandó adathalmaz olyan állapotban van, hogy hozzá sem kell nyúlni, hanem egyböl rá lehet engedni a gépi tanuló algoritmusokat. A valóság ennél egy picivel gyötrelmesebb. Elöfördulhat, hogy az adatok nem konzisztens módon lettek rögzítve. Ez történt a a használt adathalmaz esetében is. Sok hiányos adattal rendelkezett, amit így nem lehetett felhasználni az elsö iterációban. Nem tartalmazott olyan lényeges oszlopokat sem alapból az adathalmaz, amire mindenképpen szükségünk volna az osztályozás során, ezért ezt saját kezüleg kellett hozzádni. Ez a tulajdonság azt jelöli, hogy az egyén az adott évben bejefejezte-e a versenyt. Ezt olyan módon tudjuk utólagosan hozzádni az adathalmazhoz, hogy a megtett kilóméterböl következtetjük, hogy ki az aki lefutotta mind a 198 kilómétert és ki az aki nem.

A megtett kilóméter és az eredmény oszlopokat felhasználva kitudjuk számolni a futó átlagsebességét is, így ismét egy feltételezhetöen fontos tulajdonsággal tudjuk bövíteni az adathalmazt.

### Adatdiszkretizáció

Az adatdiszkretizáció során a cél az, hogy a folytonos adatokat valamilyen kategorikus adatszerkezetre tudjuk hozni.

***Vödrözés***

Egy olyan eljárás, amivel a folytonos értékeket tetszöleges *n* számú vödörbe osszuk fel. Két kiemelendö módja van ennek az eljárásnak.

* Azonos elemszámú vödrözés
* Azonos intervallumú vödrözés

Valami képet ide

Meg oda is

Legyen látványos

### Feltáró elemzés

Ahhoz, hogy nagyobb rálátásunk legyen arra, hogy mégis milyen információkat tartalmaz az adathalmaz, ahhoz a feltáró elemzésre lesz szükség. Fontos lépése az adatbányászati folyamatoknak, mivel így jobban értelmezni tudjuk azt, hogy az adatok hogyan viszonyulnak egymáshoz képest. Sok esetben ezen a ponton fényt lehet deríteni esetleges inkonzisztenciár vagy hibákra. A különbözö diagramok használatával outlier-ek felfedezésére is használni ezeket az eljárasokat. Rengeteg fajta feltáró elemzést segítö könyvtár áll rendelkezésre, ezekre példa:

* Matplotlib,
* Plotly,
* Seaborn,
* ggplot,
* Altair
* DataPrep

Valójában a plotting könyvtárak segítségével tudunk a legalkalmasabban feltáró elemzéseket készíteni. Alkalmazásuk egyértelmü, néhány függvény segítségével felparaméterezhetö a diagram, amit a python futtatva megjlelenít a jupyter notebook felületén, ami akár exportálható is egyböl a felületen.

### Osztályozás

Az elsö iterációs eljárás során két osztályoó eljárást alkalmaztam. Ezek a következök voltak:

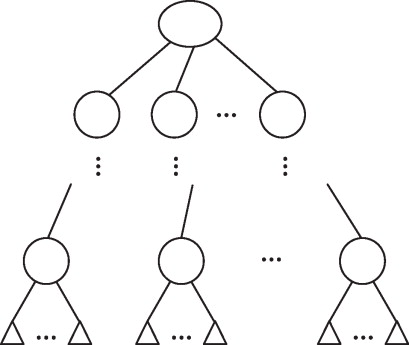
* Döntési fa osztályozó,
* Random forest osztályozó

Elsösorban azért ezekre az algoritmusokra esett a választásom, mivel alkalmazásuk gyakori és általában jó eredményeket tud hozni a megfelelö adathalmazokon.

#### Döntési fa

A döntési fa osztályozónak a müködése meglehetösen egyszerü. Fa szerüen lehet a segítségével tudást ábrázolni, amely akár átírható egy diszkrét szabályhalmazzá is, így könnyítve tovább a megértését. A fö elönye az, hogy képes különbözö részhalmazokat használni az osztályozás különbözö szakaszaiban. Egy általános döntési fa modell mindig egy csomópontból indul ki. Természetesen minden osztályozó modellnek vannak elönyei és hátrányai, ezért érdemes tudni azt, hogy milyen esetben melyik algoritmus alklamazása lesz a legoptimálisabb.

Figure 1 Döntési fa



***Elönyei***

* Könnyen értelmezhetö,
* Numerikus és kategorikus adattal is tud dolgozni
* Nincs sok elöfeltétele (pl.: nem szükséges dummy változókat létrehozni vagy vödrözést alkalmazni)

***Hátrányai***

* Könnyedén túlilleszt

#### Random forest

A random forest egy olyan továbbfejlesztése a döntési fa osztályozónak, ami növelni tudja a teljesítményt. Ezt oly módon teszi, hogy a modell létrehozásánál, több döntési fát hoz létre. Ezeknek a fáknak a számat általában *n* változóval szokták jelölni, és tetszöleges pozitív egész számot vehet fel. Az osztálycímkét végül ezen egyes döntési fáknak az eredménye fogja adni, úgy, hogy a többségben lévö osztálycímke lesz a jósolt érték.

### Kiértékelés

## Probléma

### Kis adathalmaz

### Kiegyensúlyozatlan adathalmaz

### Zajos **adathalmaz**

## Második iteráció

A

### Adathalmazbövítés

A

### Adattisztítás és transzformáció

A

### Feature engineering

A

### Importance analysis

A

### Feltáró elemzés

A

### Osztályozás

#### Döntési fa

A

#### Random forest

A

#### SVM

A ki

#### Neurális háló

A ki

#### Knn

A ki

#### Naive Bayesian

A ki

### Kiértékelés

A

# Konklúzió

# Alkalmazás

Miután sikerült a célul kitüzött feladatot megvalósítani, ezt valamilyen módon szeretnénk elérhetövé tenni egy alkalmazás, weboldal segítségével. Sokfajta megközelítése lehet egy ilyen webapplikáció fejlesztésének, de a dolgozat elkészítésénél, törekedtem az egyszerü alternatívák használatára. Emelett az is fontos volt, hogy egy modern, letisztult felület készüljön el ahol szépen, egyszerüen vizualizálva legyenek különbözö táblázatok és diagramok alkalmazásával a rendelkezésre álló adat. Természetesen elengedhetetlen manapság, a felhasználóbarát feületre való törekvés is, és a könnyü és gyors elérhetöség.

Ahhoz, hogy a fent említett elvárásoknak mind megfeleljen az alkalmazás, a streamlit névre hallgató python könyvtárt fogom felhasználni.

## Telepítés

A fejlesztés megkezdésénél szükséges volt elöször is a használt programozási nyelv telepítése az eszközre. Tetszölegesen választható a python verziójának a száma, de ajánlott mindenképpen a Python3 és ennél újabb verziók, mivel a 2-es verzió már jelentösen elavultnak tekinthetö és rengeteg hasznos funkciót nem tartalmaz, ami könnyíteni tudja a fejlesztések menetét.

Ahhoz, hogy a adatbányászati munkát tudjunk végezni, elöször is telepíteni kell a már az elözö fejezetben is említette könytárakat. Ezeket a python alapértelmezett csomagkezelöjével lehet telepíteni.

python3 -m pip install -r requirements.txt

Ahhoz, hogy a adatbányászati munkát tudjunk végezni, elöször is telepíteni kell a már az elözö fejezetben is említette könytárakat. Ezeket a python alapértelmezett csomagkezelöjével lehet telepíteni.

## Konfiguráció

Miután telepítettük a pythont, szükséges ennek a konfigurálása valamilyen kódszerkesztöben vagy valamilyen integrált fejlesztési környezetben. Ahhoz, hogy a python-t használni tudjuk a Visual Studio Code-on belül, ahhoz, telepíteni kell az ehhez készített bövítményt. Ha ez megvan, áll

## Streamlit

## Futtatás

## Felhö

## Felületek

### Statisztika

### Modellek

### Elörejelzés

### Adathalmaz feltöltés

### Információk

# Továbbfejlesztési lehetöségek

# Irodalomjegyzék

# Mellékletek

Ábrajegyzék

Táblázatjegyzék