Futóverseny eredményeinek kiértékelését támogató eszköztár és alkalmazás fejlesztése

Témavezető: dr. Leitold Dániel

Csizmazia Máté - XI32IS

2022 Április 25.

Ultramarathon

BALATON SZUPERMARATON

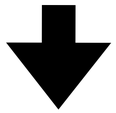
- Népszerű
- Nagy kihívás
- 4 napos esemény
- Többnyire publikus adathalmaz
- Data Science

Felépítés

- Adatösszegyűjtés / tisztítás
- Irodalomkutatás
- Kutatások alkalmazása az adathalmazon
- Modell-épités
- Kiértékelés
- Alkalmazás fejlesztése

Adathalmaz

- 2015 - 2020 közötti eredmények
- Számos tulajdonság
- Összesítve is alacsony részvételi arány



- ~ 1000 sornyi adat

Nominális	Folytonos	Diszkrét	Ordinális
Név	Helyezés	Rajtszám	Kategória
Ország	Születési év	Esemény éve	
Klub	Napi időadatok		
Város	Eredmény		
Nem	Futott kilóméter		
Rajtszám			

Első iteráció

Cél megfogalmazása

Teljesíthető?

Adat előkészítés

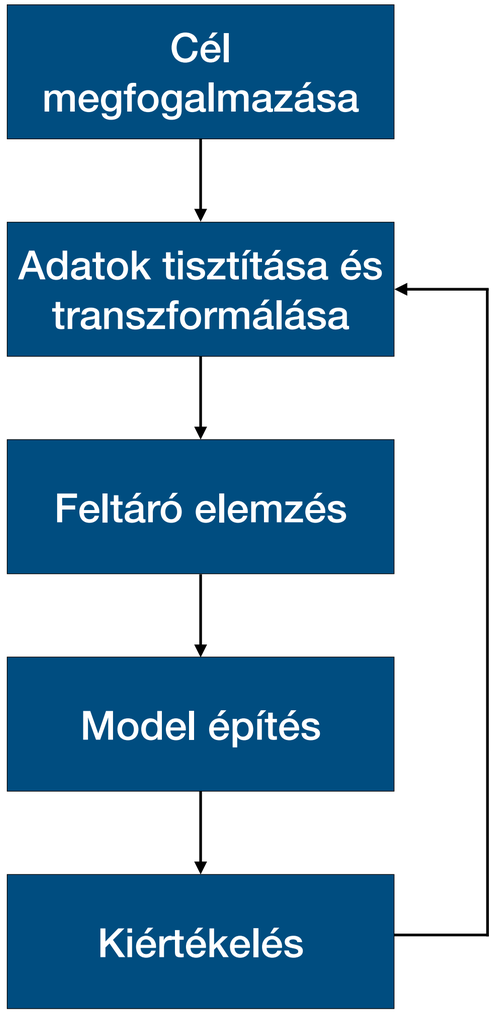
- Oszlopok törlése / hozzáadása
  (Név, Város, stb), (Befejezte, Tempo, stb)
- Értelmetlen adatok törlése
- Időadatok átalakítása
- Adat diszkretizáció (vödrözés)

Model építés

- Döntési fa osztályozó
- Random forest osztályozó

Kiértékelés

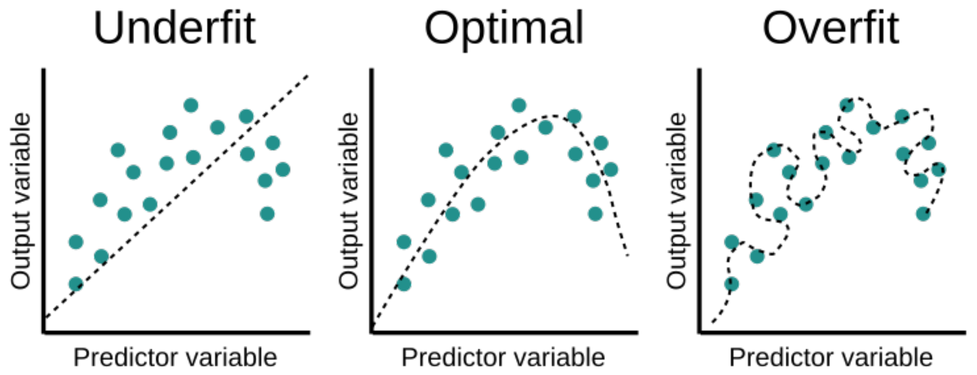
- Pontatlan modell (60-65%)
- Több adatra van szükség
- Szervezők bevétele, 2015 előtti eredmények feldolgozása
- Új iteráció



Problémák

- Kis adathalmaz, bővítve is kicsi
  2511 sor
- Imbalanced adathalmaz
  ~ 20% adta fel a versenyt
- Zajos adatok, nagyban eltérő információk
  Verseny hossza, stb
- Használhatatlan adatok
  2008 - 2011 közötti adatok
- 2/3 vágás nem elegendő
  Más megközelítésre van szükség

- Túlillesztés gyakori kicsi adathalmazok esetén
- High bias, mivel egyenlőtlen az adathalmaz
  ~ 95% pontos modellek, a bővített adathalmazokon
- Tisztítás során sok hamis információt generálhat



Megoldások

- Más gépi tanuló algoritmusok alkalmazása
- Irodalom kutatás
- Feature engineering
- Tulajdonságok bővítése
- Mintavételezési eljárások alkalmazása
- Keresztvalidálási eljárások
- Importance analysis

Nominális	Folytonos	Diszkrét	Ordinális
Név	Helyezés	Rajtszám	Kategória
Ország	Születési év	Esemény éve	Hőmérséklet
Klub	Napi időadatok		Tempó kategóriák
Város	Eredmény		Felhősség
Nem	Futott kilóméter		Légnyomás
Rajtszám	Napi szakaszok hossza		Eső mértéke
	Napi szakasz tempók		Szél
	Napi tempók		Széllökések
	Időjárás adatok		

Keresztvalidáció

- Olyan esetekben nagyban ajánlott az alkalmazása, amikor egy model pontossága bizonytalan (sokat ugrál a pontossága)
- Általános megközelítés a 2/3 -os vágás, ami általában működik alkalmas adathalmazokon
- Imbalanced adathalmazokon erősen high bias eredményt ad, ami sok pontatlan előrejelzést eredményez.
- Több fajta keresztvalidálási eljárás is létezik pl.: (Leave One Out CV), de a mi esetünkben kettő vizsgálatán lesz a hangsúly.

K-fold Cross Validation

- Pro
  - Teljesítményt növelheti
  - Ismeretlen adathalmazokat segít jobban megérteni
  - Kis adathalmazon optimális
- Cons
  - Imbalanced adathalmazon high bias
  - Nagy adathalmazon lassú lehet

Stratified K-fold CV

- Pro
  - Valós pontosságot tovább növeli speciális adathalmazokon (pl. Imbalanced).
  - Minden iterációban azonos eloszlású az osztálycímke
  - Kis adathalmazon optimális
- Cons
  - Lassabb

K-fold Cross Validation

Pl.: 1000 sor adatra 5 fold

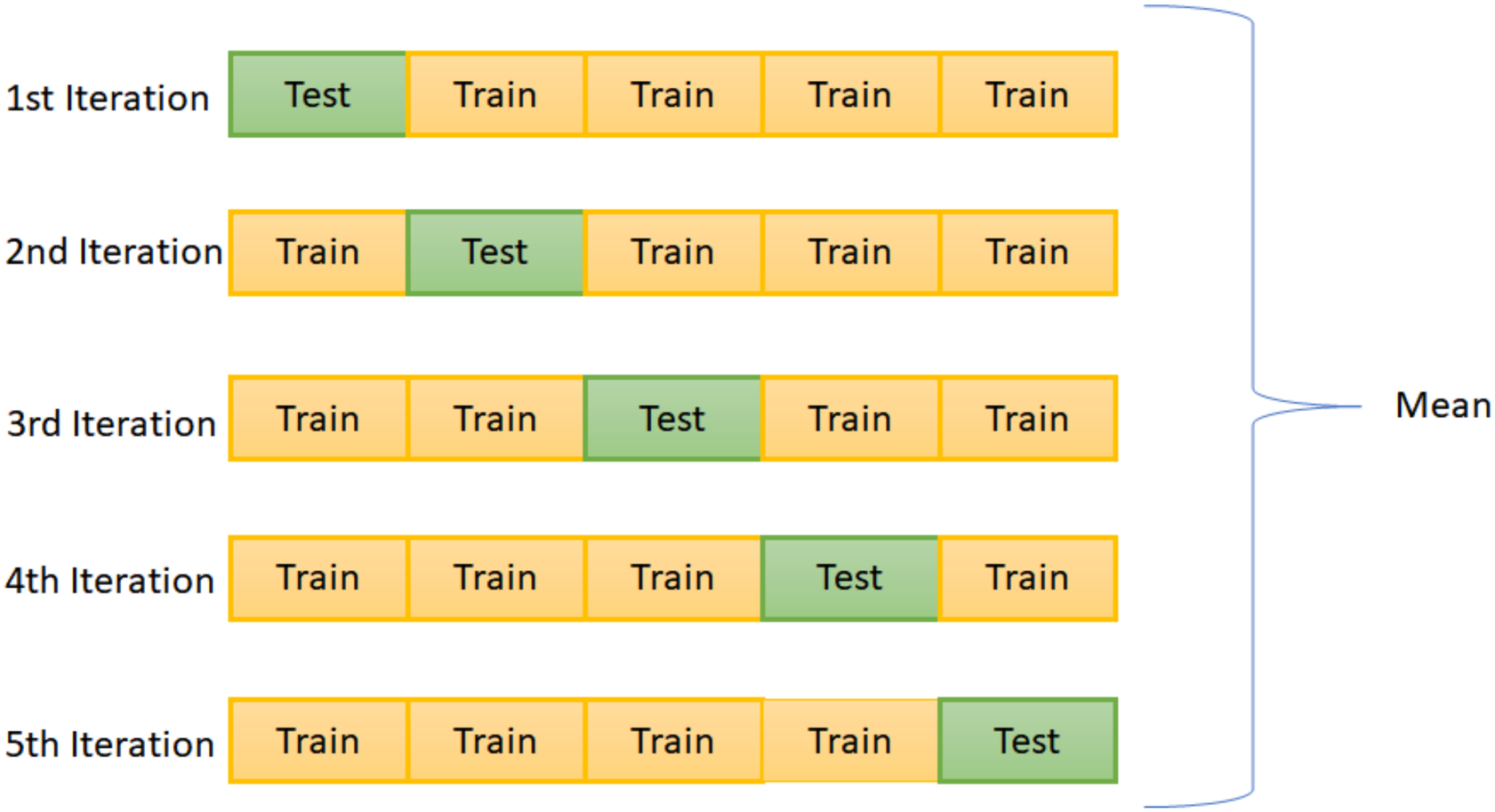
200 test, 800 train
180/20

200 test, 800 train
198/2

200 test, 800 train
120/80

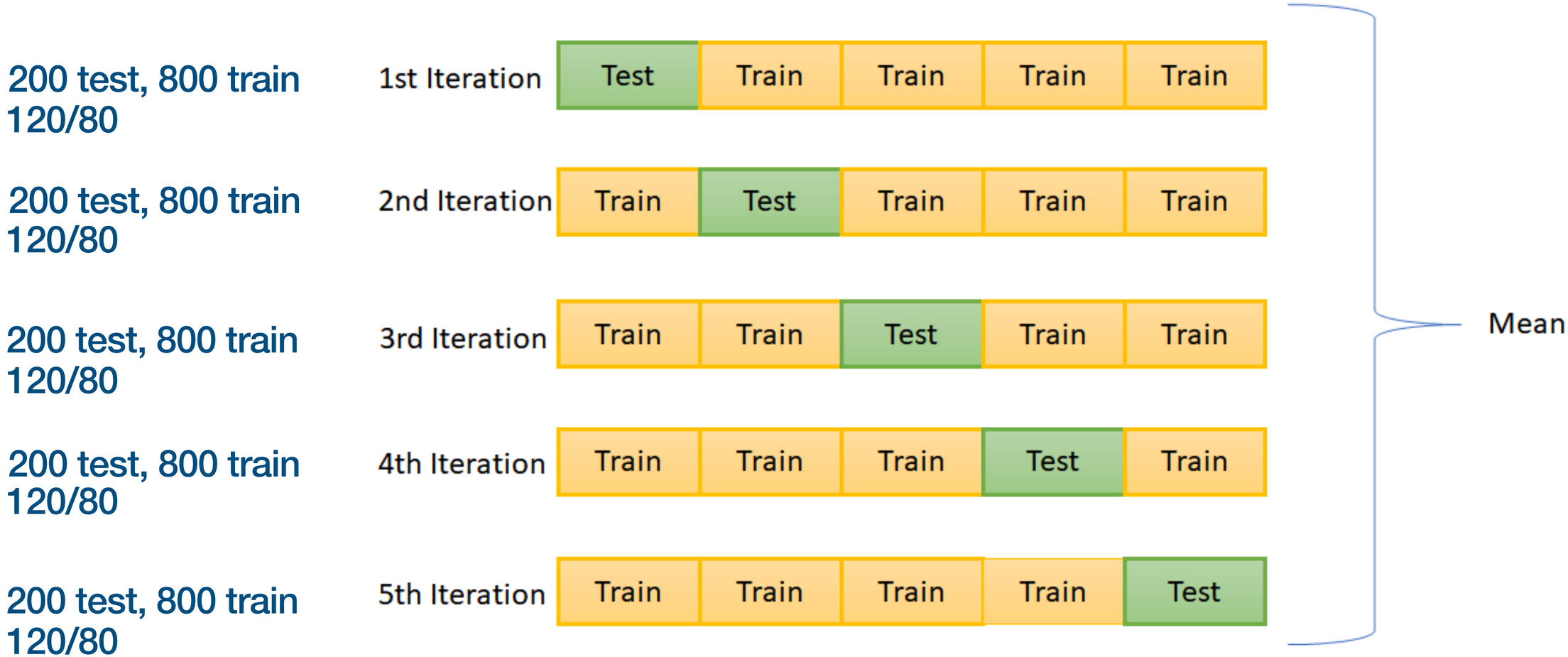
200 test, 800 train
110/90

200 test, 800 train
200/0



Stratified K-fold Cross Validation

Pl.: 1000 sor adatra 5 fold



Második iteráció

Adatok előkészítése

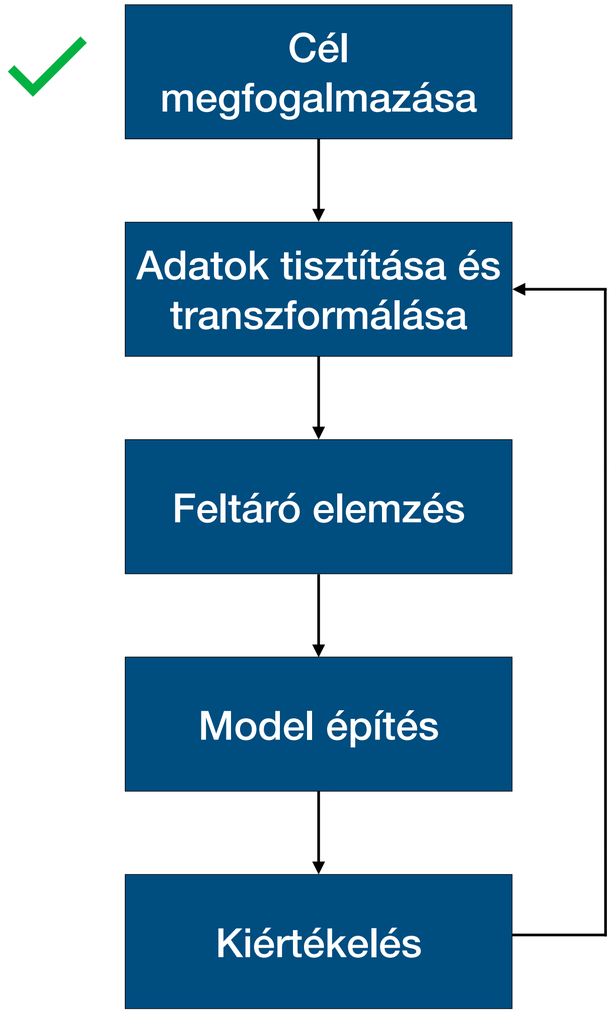
- Összes adathalmaz migrálása
- Normalizáció
- Teljesebb adattisztítás
- Outlier detection
- Feature Engineering

Feltáró elemzés

- Correlation analysis
- DataPrep alkalmazása
- Vizualizáció
- Adathalmaz részletes értelmezése

Model építés

- Random forest osztályozó
- SVM osztályozó (kis adathalmazokra alkalmas)
- Naive Bayesian osztályozó
- Neurális háló osztályozó
- Keresztvalidációs eljárások alkalmazása
- Over / Under sampling eljárások alkalmazása



Konklúzió

60% -> 95%

Nem valós teljesítmény növekedés, az adathalmaz bővült, de ezzel az adathalmaz kiegyensúlyozatlansága is növekedett. Emiatt a modell high bias módon osztályozott, ezért szükséges volt különböző mintavételezési eljárást alkalmazni.

Modellek

Az munka folyamán váltakozó pontosságú modelleket lehetett építeni a tisztított adathalmazból. Ezek közül a legalkalmasabbak a Support Vector Machine alapú osztályozók, és a Naive Bayesian osztályozók.

Cross Validation

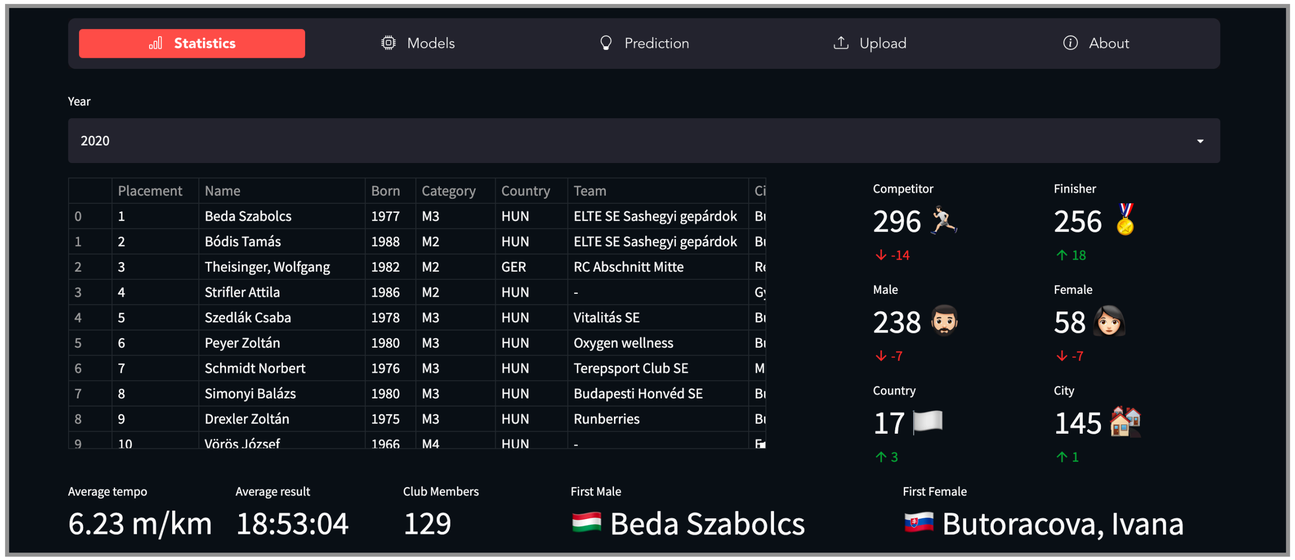
A keresztvalidálási eljárások alkalmazásával ténylegesen sikerült az épített modellek teljesítményét javítani. A legjobban 10 k-nál látszódott a javulás és a Naive Bayesian osztályozóval sikerült a legpontosabb modellt építeni, ami 78,23%

Eredmény

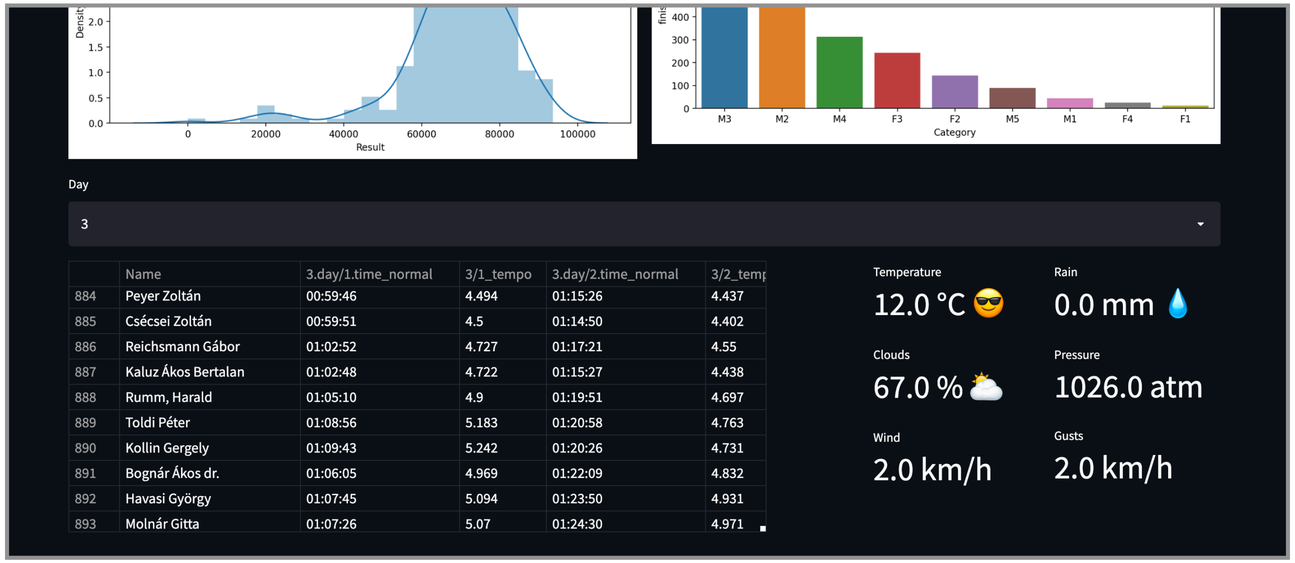
Az első iterációhoz képest nagyobb teljesítményű és pontosabb modellt sikerült építeni. Az adathalmazra és a felhasznált tulajdonságra a legalkalmasabban a Naive Bayesian osztályozó alkalmazható.

Alkalmazás / Vizualizáció

- Évi / összes eredmények kiértékelése
- Napi eredmények
- Időjárás adatok
- Osztályozó modellek és pontosságuk



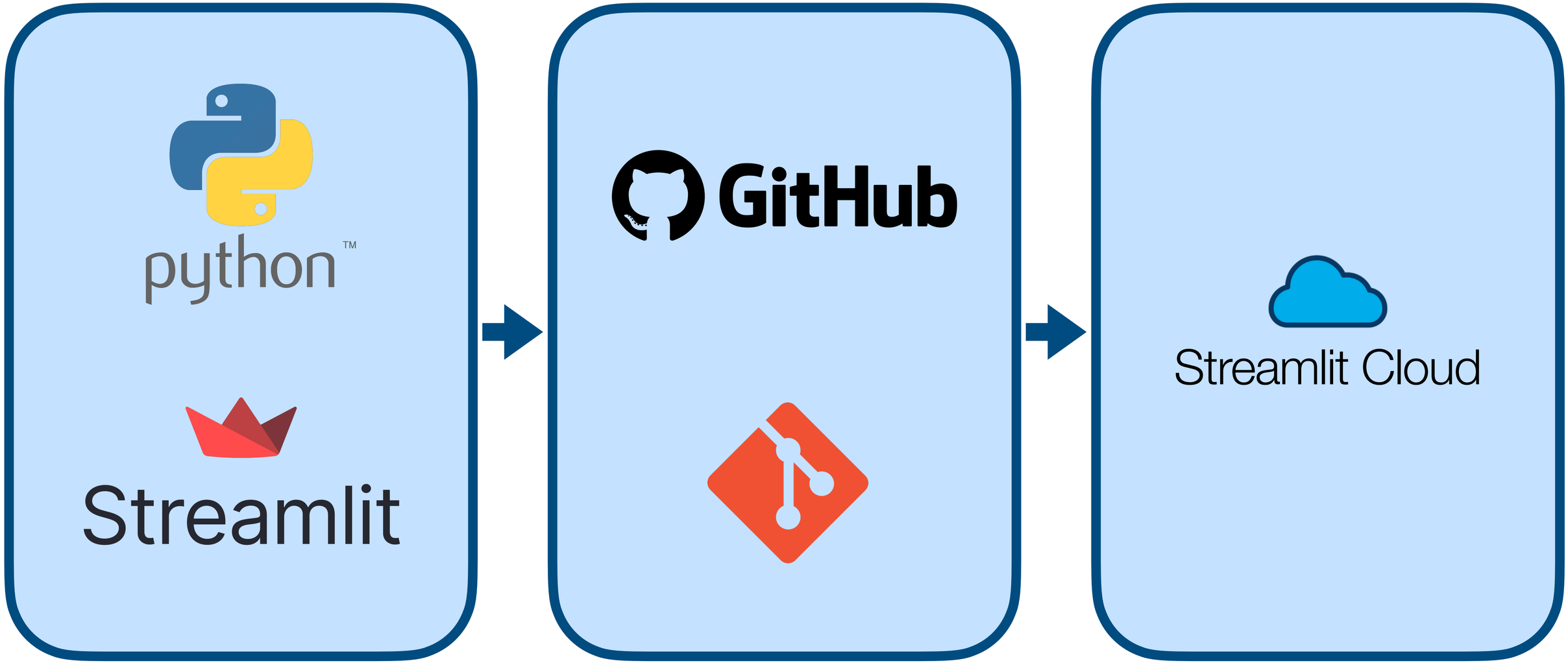
	Placement	Name	Born	Category	Country	Team
0	1	Beda Szabolcs	1977	M3	HUN	ELTE SE Sashegyi gepárdok
1	2	Bódis Tamás	1988	M2	HUN	ELTE SE Sashegyi gepárdok
2	3	Theisinger, Wolfgang	1982	M2	GER	RC Abschnitt Mitte
3	4	Strifler Attila	1986	M2	HUN	-
4	5	Szedlák Csaba	1978	M3	HUN	Vitalitás SE
5	6	Peyer Zoltán	1980	M3	HUN	Oxygen wellness
6	7	Schmidt Norbert	1976	M3	HUN	Terepsport Club SE
7	8	Simonyi Balázs	1980	M3	HUN	Budapesti Honvéd SE
8	9	Drexler Zoltán	1975	M3	HUN	Runberries
9	10	Vörös József	1966	M4	HUN	-



	Name	3.day/1.time_normal	3/1_tempo	3.day/2.time_normal	3/2_temp
884	Peyer Zoltán	00:59:46	4.494	01:15:26	4.437
885	Csécsei Zoltán	00:59:51	4.5	01:14:50	4.402
886	Reichsmann Gábor	01:02:52	4.727	01:17:21	4.55
887	Kaluz Ákos Bertalan	01:02:48	4.722	01:15:27	4.438
888	Rumm, Harald	01:05:10	4.9	01:19:51	4.697
889	Toldi Péter	01:08:56	5.183	01:20:58	4.763
890	Kollin Gergely	01:09:43	5.242	01:20:26	4.731
891	Bognár Ákos dr.	01:06:05	4.969	01:22:09	4.832
892	Havasi György	01:07:45	5.094	01:23:50	4.931
893	Molnár Gitta	01:07:26	5.07	01:24:30	4.971

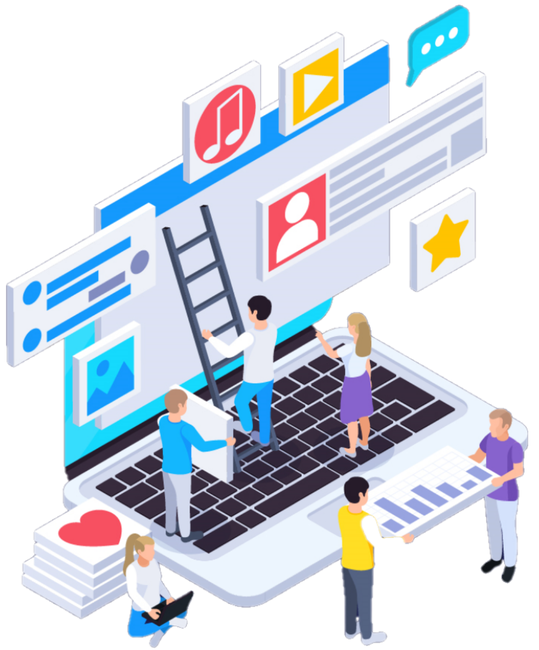
- Előrejelzés
- Új évi tesztadat feltöltése
- Információk (BSZM, elérhetőség, stb)

Alkalmazás elérhetősége



Továbbfejlesztési lehetőségek

- Mindig lehet tovább pontosítani, hiperparaméter hangolás
- Új célok -> pl.: Eredmény közelítő becslés
- További együttműködés a szervezőkkel, új kimutatások
- Alkalmazás deployment



Köszönöm a figyelmet!

Csizmazia Máté - XI32IS - Programtervező informatikus MSc

- Téma ismertetése
- Munkafolyamatok és problémák ismertetése
- Megoldások és eredmények ismertetése
- Kiértékelő alkalmazás bemutatása
- Továbbfejlesztési lehetőségek ismertetése





2022 Április 25.