

FEATURE ENGINEERING KATEGORIKUS ADATOKON

Radó Laci, Lead Data Scientist @ BT Group



FEATURE ENGINEERING KATEGORIKUS ADATOKON

Radó Laci, Lead Data Scientist @ BT Group

- Feature és Feature Engineering jelentése
- Mi a Feature Engineering szerepe
- Gyakorlati példák

FEATURE

Egyszerűen: **Egy modell bemeneti változója.**

A modellezett jelenség egy olyan **megkülönböztető jellemzőjét**, tulajdonságát leíró változó, melyet egy **modell felhasznál** a döntéshozatalhoz, vagy előrejelzéshez.

Lehet egy eredeti, nyers adatforrásból származó attribútum, vagy egy feldolgozott, átalakított változó, amit felhasználok a modellezés során.

FEATURE

Egyszerűen: **Egy modell bemeneti változója.**

A modellezett jelenség egy olyan **megkülönböztető jellemzőjét**, tulajdonságát leíró változó, melyet egy **modell felhasznál** a döntéshozatalhoz, vagy előrejelzéshez.

Lehet egy eredeti, nyers adatforrásból származó attribútum, vagy egy feldolgozott, átalakított változó, amit felhasználok a modellezés során.

Jellemző

Független / Magyarázó Változó

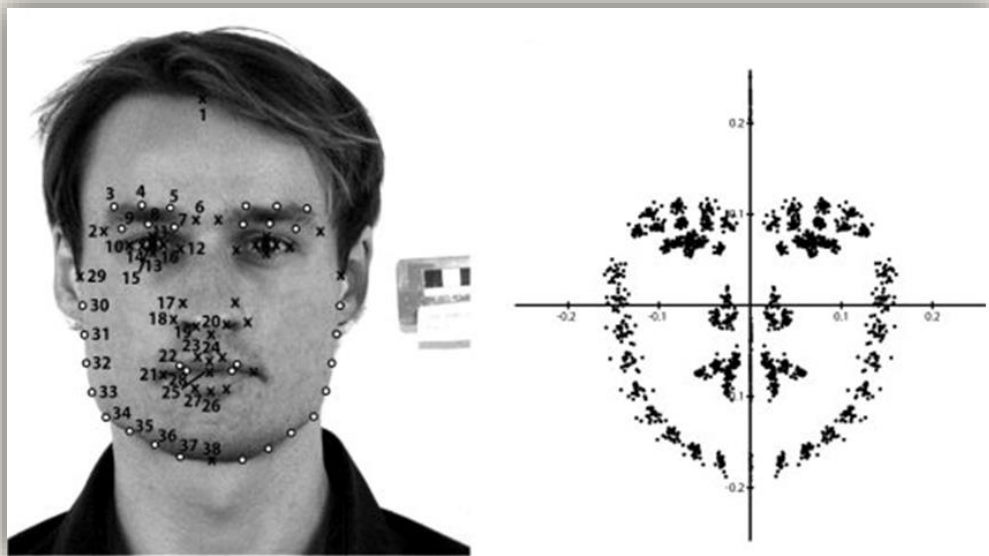
(Independent / Explanatory Variable)

FEATURE

Egyszerűen: **Egy modell bemeneti változója.**

A modellezett jelenség egy olyan **megkülönböztető jellemzőjét**, tulajdonságát leíró változó, melyet egy **algoritmus felhasznál** a döntéshozatalhoz, vagy előrejelzéshez.

Lehet egy eredeti, nyers adatforrásból származó attribútum, vagy egy feldolgozott, átalakított változó, amely a modell célját segíti.



Jellemző

Független / Magyarázó Változó

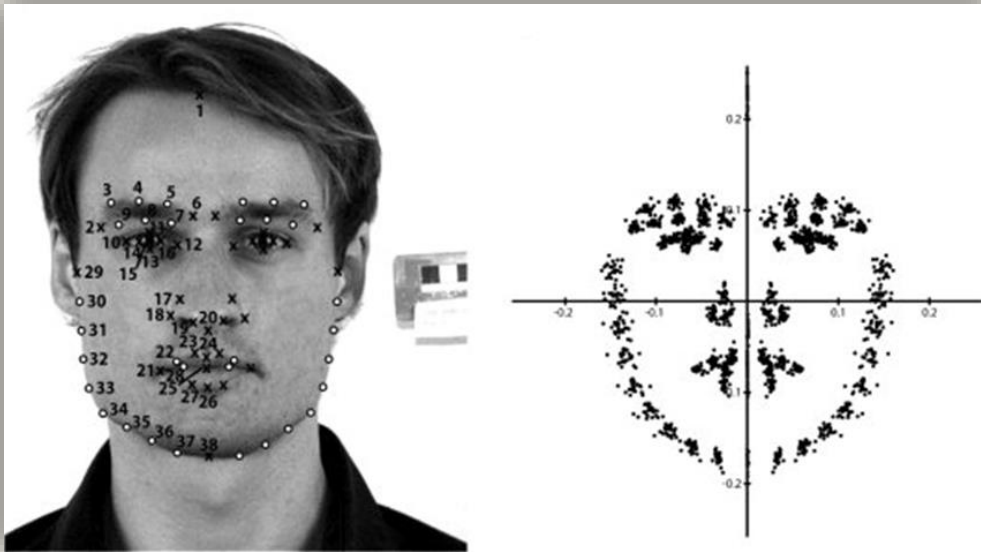
(Independent / Explanatory Variable)

A mai **meeting** során a **machine learning** algoritmusok **implementációjáról**, esett szó, különös tekintettel a **feature engineering** fontosságára és a modell **prediktív** **performanciájának** **optimalizálására**.

8x idegen kifejezés

A modellezett jelenség egy olyan megkülönböztető jellemzője, tulajdonságát leíró változó, melyet egy **algoritmus** felhasznál a döntéshozatalhoz, vagy előrejelzéshez.

Lehet egy eredeti, nyers adatforrásból származó attribútum, vagy egy feldolgozott, átalakított változó, amely a modell célját segíti.



FEATURE ENGINEERING

Egyszerűen: **Az a folyamat, ami során ezeket a jellemzőket kitaláljuk előállítjuk.**

Az a folyamat, amely során a nyers adatainkat úgy alakítjuk át hogy azok a modell számára **érelmezhető** formában, a döntéshez vagy előrejelzéshez hasznos információkat **könnyen hozzáférhetően** hordozzák.

Ennek során **új változókat** hozunk létre a meglévőkől, vagy a **meglévőket alakítunk át**, és kiemeljük azokat az információkat, amelyek a modellezett jelenség szempontjából lényegesek.

FEATURE ENGINEERING

Egyszerűen: **Az a folyamat, ami során ezeket a jellemzőket kitaláljuk előállítjuk.**

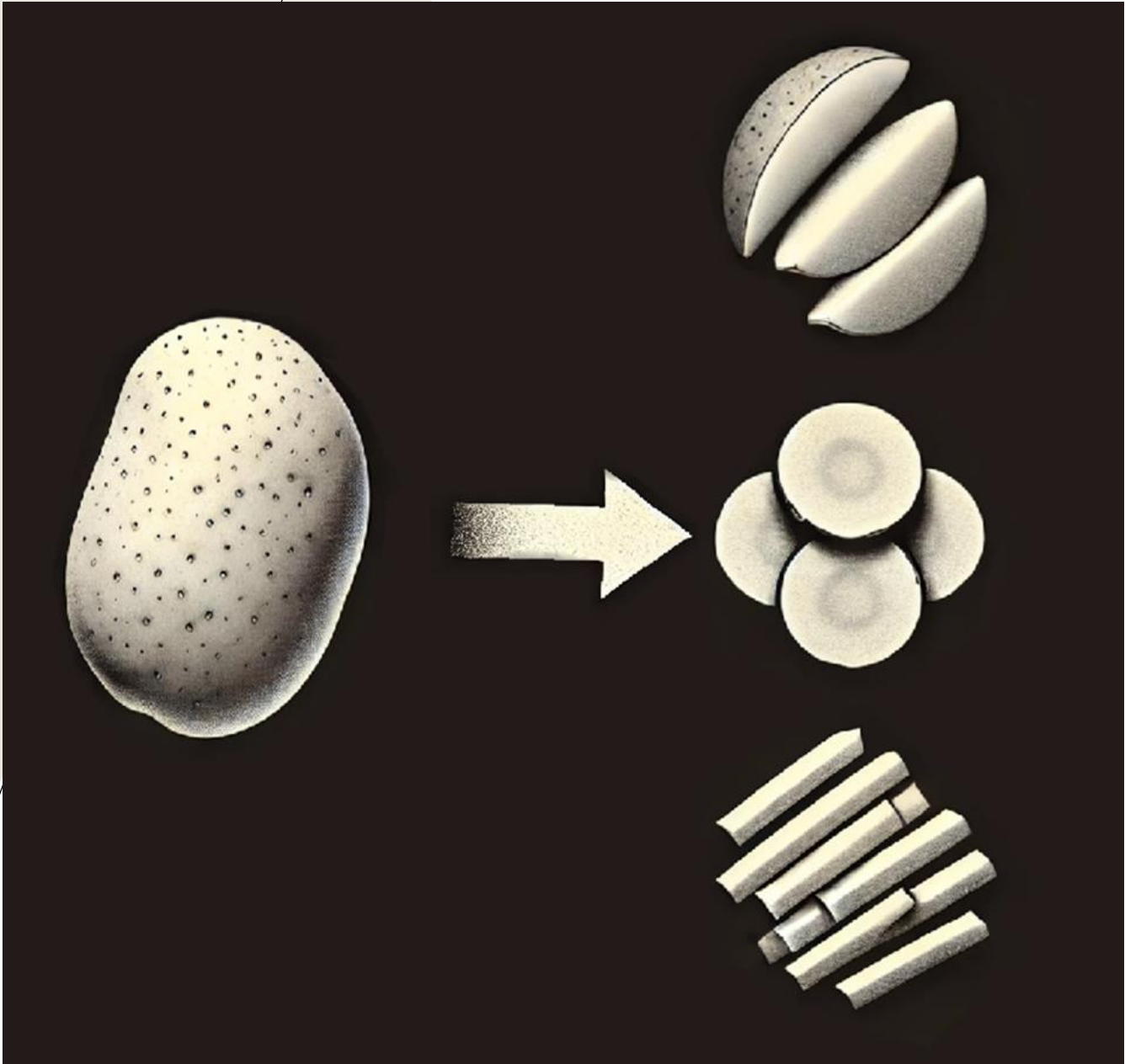
Az a folyamat, amely során a nyers adatainkat úgy alakítjuk át hogy azok a modell számára értelmezhető formában, a döntéshez vagy előrejelzéshez hasznos információkat könnyen hozzáférhetően hordozzák.

Ennek során új változókat hozunk létre, meglévőket alakítunk át, és kiemeljük azokat az információkat, amelyek a modellezett jelenség szempontjából lényegesek.

- Jellemzők leírása / kinyerése
- Számított változók tervezése
- Adatelőkészítés

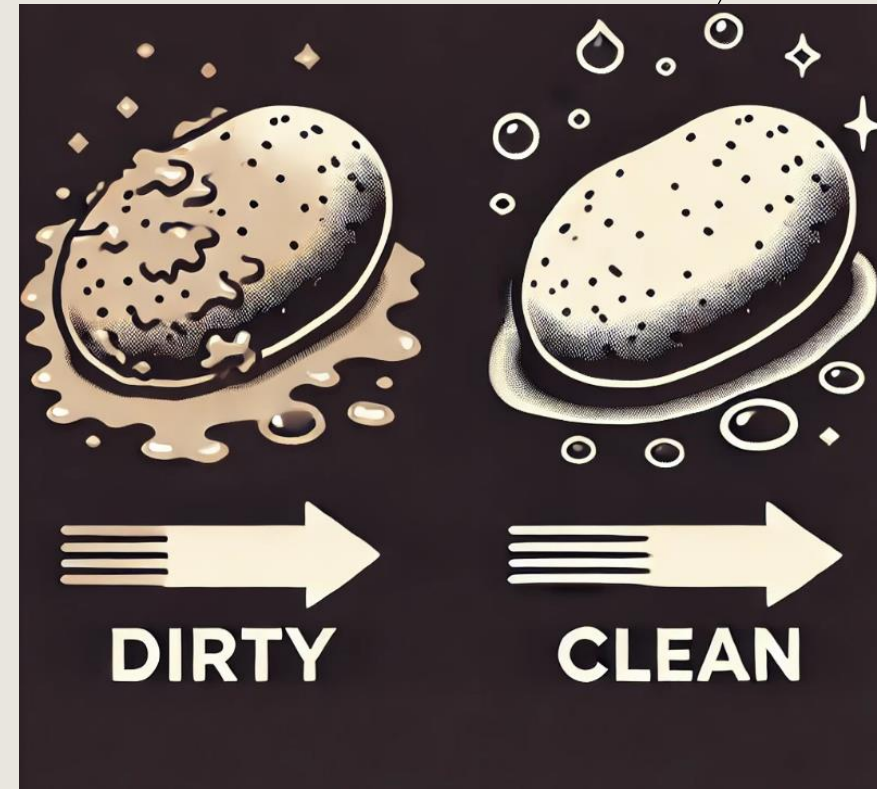
FEATURE ENGINEERING FELADATA

- Az adatok átalakítása a modell által feldolgozható formátumra.
 - A modell által nehezen hozzáférhető információk kiemelése, a hozzáférés megkönnyítése.
- Adatgazdagítás



NEM ADATTISZÍTÁS!!!

- NEM hiányzó értékek kezelése
- NEM duplikált adatok eltávolítása
- NEM kategóriák konszolidálása
- NEM helyesírási, adatrögzítési hibák javítása
- ...

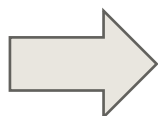


PÉLDA - 1

KATEGORIKUS VÁLTOZÓK KEZELÉSE

Sorszámozás
Label Encoding

Ország
Magyarország
Japán
Magyarország
Magyarország
Kanada
Brazília
Kenya

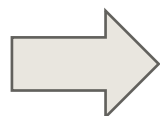


Ország
1
2
1
1
3
4
5

PÉLDA - 1

KATEGORIKUS VÁLTOZÓK KEZELÉSE

Ország
Magyarország
Japán
Magyarország
Magyarország
Kanada
Brazília
Kenya



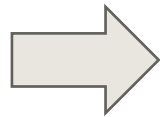
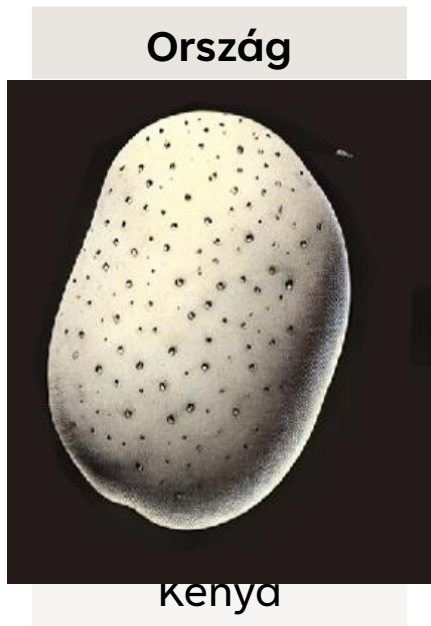
Sorszámozás Label Encoding	
Ország	
1	
2	
1	
1	
3	
4	
5	

Target Encoding	
Ország	
113.8	
223.1	
113.8	
113.8	
72.0	
459.9	
122.7	

One-Hot Encodinig		
Magyar	Japán	Brazília
1	0	0
0	1	0
1	0	0
1	0	0
0	0	0
0	0	1
0	0	0

PÉLDA - 1

KATEGORIKUS VÁLTOZÓK KEZELÉSE

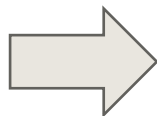


PÉLDA – 2

SZÁRMAZTATOTT VÁLTOZÓ LÉTREHOZÁSA

Eredeti Adat

Kezdő dátum	Vég dátum
2022.01.01	2023.05.01
2021.06.15	2022.03.20
2020.03.20	2021.06.20
2019.11.05	2020.10.01
2023.02.25	2024.07.30
2018.04.10	2022.12.31
2020.07.01	2021.09.15



Származtatott adat

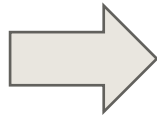
Napok száma
485
278
457
331
521
1726
441

PÉLDA – 2

SZÁRMAZTATOTT VÁLTOZÓ LÉTREHOZÁSA

Eredeti Adat

Kezdő dátum	Vég dátum
2022.01.01	2023.05.01
2021.06.15	2022.03.20
2020.03.20	2021.06.20
2019.11.05	2020.10.01
2023.02.25	2024.07.30
2018.04.10	2022.12.31
2020.07.01	2021.09.15



Származtatott adat

Napok száma
485
278
457
331
521
1726
441

Származtatott adat

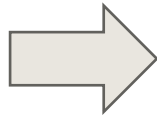
1 évnél rövidebb
Nem
Igen
Nem
Igen
Nem
Nem
Nem

PÉLDA – 2

SZÁRMAZTATOTT VÁLTOZÓ LÉTREHOZÁSA

Eredeti Adat

Kezdő dátum	Vég dátum
2022.01.01	2023.05.01
2021.06.15	2022.03.20
2020.03.20	2021.06.20
2019.11.05	2020.10.01
2023.02.25	2024.07.30
2018.04.10	2022.12.31
2020.07.01	2021.09.15



Szarmaztatott adat

Napok száma
485
278
457
331
521
1726
441

Szarmaztatott adat

1 évnél rövidebb
0
1
0
1
0
0
0

PÉLDA – 3

SZÁRMAZTATOTT VÁLTOZÓ LÉTREHOZÁSA

Felhasználók
Adatai

User_id	E-mail	Előfizeté	Regisztráci
Abc123	anna@abc.com	Alap	2023.05.01
Qwe987	bela@abc.com	Prémium	2021.03.08
Bnm456	csilla@abc.com	Alap	2019.12.30
Ghj678	david@abc.com	Prémium	2024.02.02

Felhasználási
adatok

User_id	Cím	Stílus	Lejátszás Dátuma
Abc123	Rambo	Akció	2023.05.01
Abc123	Süsü a sárkány	Mese	2021.03.08
Abc123	Star Wars	Sci-fi	2019.12.30
Abc123	Die Hard	Akció	2024.02.02

PÉLDA – 3

SZÁRMAZTATOTT VÁLTOZÓ LÉTREHOZÁSA

Felhasználók
Adatai

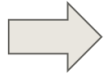
User_id	E-mail	Előfizeté	Regisztráci	Akció_pct	Múlt havi aktivitás
Abc123	anna@abc.com	Alap	2023.05.01	50%	3
Qwe987	bela@abc.com	Prémium	2021.03.08
Bnm456	csilla@abc.com	Alap	2019.12.30
Ghj678	david@abc.com	Prémium	2024.02.02

Felhasználási
adatok

User_id	Cím	Stílus	Lejátszás Dátuma
Abc123	Rambo	Akció	2025.03.01
Abc123	Süsü a sárkány	Mese	2025.03.02
Abc123	Star Wars	Sci-fi	2025.03.27
Abc123	Die Hard	Akció	2025.04.04

FEATURE ENGINEERING FELADATA

- Az adatok átalakítása a modell által feldolgozható formátumra.

Ország		Ország
Magyarország		1
Japán		2
Magyarország		1
Magyarország		1
Kanada		3
Brazília		4
Kenya		5

- A modell által nem, vagy nehezen hozzáférhető információk kiemelése, a hozzáférés megkönnyítése.

User_id	E-mail	Előfizeté	Regisztráci	Akció_pct
Abc123	anna@abc.com	Alap	2023.05.01	50%
Qwe987	bela@abc.com	Prémium	2021.03.08	...
Bnm456	csilla@abc.com	Alap	2019.12.30	...
Ghj678	david@abc.com	Prémium	2024.02.02	...