Matematikari Euskaldunen
VI. Topaketak

behategia
EUSKAL HEDABIDEEN BEHATOKIA

EUSKARAZKO HEDABIDEEN KONTSUMOAN ERAGITEN DUTEN ERABILTZAILE ALDAGAIAK

Naroa Burreso Pardo
naroa.burreso@ehu.eus

2024ko ekainak 17

BEHATEGIA

- Behategia (Euskal Hedabideen Behatokia) NOR ikerketa taldearen barruan kokatzen den Hekimen, EHU, MU, UD eta UEUren arteko hitzarmen bat da.

- Euskarazko komunikabideen ikerketarako espazioa da.
  - Euskarazko hedabideen **kontsumoa** eta **erabiltzaileak** ikertzen dira.

HELBURUA

Helburua: Euskarazko hedabide tradizionalen kontsumoan eragina duten **erabiltzaile aldagaiak** identifikatzea.

Zertarako? Pertsona batek euskarazko hedabide tradizionalak kontsumituko ote dituen **aurresateko.**

Nola? Euskarazko hedabide tradizionalen kontsumoa aurresaten duen **modelo bat eraikiz.**

DATU-BASEA

CIESen Komunikabideen Audientzien Azterketa (2021)

- **Helburua:** Komunikabideen audientzien kontsumoen neurketa egitea.
- **Xede-populazioa:** Hego Euskal Herriko (EAE+Nafarroa) 14 urtetik gorakoak.
- **Metodoa:** Laginketa
  - Urtero 8.600 inkestatu (4.300 martxoan eta 4.300 urrian).
- **Tresna:** Inkesta
  - 200 galdera inguru, 7 taldetan banatuta:
    - Soziodemografikoak
    - Prentsa
    - Irratia
    - Telebista
    - Aldizkako argitalpenak
    - Internet
    - Ekipamendua

8.600 erregistro
5.963 aldagai

DATU-BASEAREN TRANSFORMAZIOA

- **Aldagaien sorkuntza eta hautaketa:**
  - Erantzun aldagaia: Bezperan euskarazko hedabideren bat kontsumitu bai/ez → Aldagai dikotomikoa.
  - Aldagai azaltzaileak: Indibiduoaren soziodemografikoak, sostengatzaile nagusiaren datuak eta bizi den etxeari buruzko datuak → 23 aldagai.

- **Datu-basearen garbiketa:**
  - Balio galduen, *outlier*-en eta inkongruentzien azterketa (11 erregistro ezabatu).
  - Bikoiztutako erregistroen (177) ezabatzea.

9.858 erregistro
65 aldagai

- **Aldagai kategoriko nominalen dikotomizazioa** → 65 zenbakizko aldagai.

- **Aldagai guztien normalizazioa** (0 eta 1 arteko balioak har ditzaten).

- **Entrenamendu, baliozkotze eta testatze multzoen banaketa:**
  - Erregistroen %70 (5887) entrenamendua, %20 (1683) testatzea eta %10 (842) baliozkotzea.

- **Menpeko aldagaiaren kategoriak orekatu entrenamendu-multzoan:**
  - %83,73 (4929) Ez; %16,27 (958) Bai → *Oversampling*: 9858 erregistro entrenamendu-multzoan.

MODELOEN ERAIKUNTZA-PROZESUA

- 5 klasifikazio-algoritmo erabili dira: *K-Nearest Neighbour* (*KNN*), *Decision Tree*, *Support Vector Machine*, *Random Forest* eta *Logistic Regression*.
- Algoritmo bakoitzarekin emandako urratsak:
  1. Entrenamendu-multzoan **sentikortasuna** optimizatzen duten hiper parametroen bilaketa egin (*Random Search*, *Grid Search*).
  2. Hiper parametro horiekin modeloa eraiki eta baliozkotze-multzoan duen sentikortasuna kalkulatu.
  3. Baliozkotze-multzoan sentikortasun altuena duen modeloa lortu.
  4. Modelo horrek testatze-multzoan duen sentikortasuna kalkulatu.

		Aurresandakoa	
		Positiboak	**Negatiboak**
Behatutakoa	**Positiboak**	TP	FN
	Negatiboak	FP	TN

Sentikortasuna: behatutako positiboen zein ehuneko dagoen ondo sailkatuta = TP/ (TP+FN)

ALDAGAI AZALTZAILEEN HAUTAKETA

1. Menpeko aldagaiaren eta aldagai azaltzaileen arteko **korrelazioak** kalkulatu dira.
   - Orokorrean ez dute erlaziorik edo erlazioa ahula da.
   - Erlazio **sendoena** duten aldagaiak:
     - Euskara lehen hizkuntza: **0,43**
     - Euskaraz hitz egin: **0,41**
     - Euskaraz irakurri: **0,40**
     - Euskaraz idatzi: **0,40**
     - Euskaraz ulertu: **0,38**
     - Gaztelania lehen hizkuntza: **-0,34**
     - Gune soziolinguistikoa: **0,34**
     - Gipuzkoakoa izan: **0,33**
2. Aldagai azaltzaileek *Random Forest* modeloan duten **garrantzia** aztertu da.
3. Korrelazioaren modulua <0,1 eta garrantzia <0,01 duten aldagaiak ezabatu dira → 11 aldagai.
4. Garrantzia <0,01 duten aldagaiak ezabatu dira → 28 aldagai.

25 aldagai azaltzaile

SORTUTAKO MODELOAK

1. Modeloak 25 aldagai azaltzailerekin:

	Algoritmoa	Balioa
Sentikortasun altuena baliozkotze-multzoan	*Random Forest*	0,743
Sentikortasun altuena testatze-multzoan	*Decision Tree*	0,745

Bi modeloetan garrantzia <0,01 duten aldagaiak ezabatu → 8 aldagai

2. Modeloak 17 aldagai azaltzailerekin:

	Algoritmoa	Balioa
Sentikortasun altuena baliozkotze-multzoan	*Decision Tree*	0,743
Sentikortasun altuena testatze-multzoan	*Decision Tree*	0,755

Modeloan garrantzia <0,01 duten aldagaiak ezabatu → 10 aldagai

3. Modeloak 7 aldagai azaltzailerekin:

	Algoritmoa	Balioa
Sentikortasun altuena baliozkotze-multzoan	*Decision Tree, Random Forest*	0,757
Sentikortasun altuena testatze-multzoan	*Random Forest*	0,781

SORTUTAKO MODELOAK

- Algoritmo bakoitzarekin lortutako modelorik onena:

Algoritmoa	Aldagai azaltzaile kopurua	Sentikortasuna (testatze-multzoan)
Random Forest	**7**	**0,781**
Decision Tree	7	0,759
SVM	7	0,763
KNN	25	0,675
Logistic Regression	25	0,661

AUKERATUTAKO MODELOA

- Aukeratutako modeloaren **algoritmoa**: *Random Forest*
  - *Decision Tree*-a erregistroei buruzko galderak eginez eta hauen erantzunaren arabera sailkatuz eraikitzen da.
    - Lehen erabaki-nodoa: Erregistroak banatzeko egokiena den ezaugarria hautatzen da → Bi erabaki-nodo berri sortzen dira → Erabaki-nodoen kalitatea kalkulatzen da → Kalitatea ez bada nahikoa, erragistroak kalitate handiagoko beste bi nodotan banatzeko ezaugarri berria hautatzen da.
    - Prozesuak jarraitu egiten du geldialdi-irizpide bat betetzen den arte.
    - Amaierako orri-nodoak klase bakarra errepresentatzen duten erregistro-multzoak dira.
  - *Decision Tree* ugari konbinatzen dira: Zuhaitz bakoitza erregistroen eta ezaugarrien zorizko azpimultzo batekin entrenatzen da eta amaieran aurresandako kategoria, zuhaitz guztien artean gehien errepikatzen den aurresandako kategoria da.

- Aukeratutako modeloaren **hiper parametroak**:
  - *n_estimators* (zuhaitz-kopurua) = 32.
  - *min_samples_split* (erabaki-nodo bat zatitzeko behar den gutxieneko erregistro-kopurua) = 100.
  - *min_samples_leaf* (orri-nodo batean egon behar den gutxieneko erregistro-kopurua) = 70.
  - *max_depth* (zuhaitzaren gehieneko sakonera) = 8.
  - *max_features* (nodoen zatiketarik onena bilatzean kontuan hartutako ezaugarri kopurua) = *None* → Ez dago gehienezkorik, guztiak hartzen dira kontuan.

AUKERATUTAKO MODELOA

- Aukeratutako modeloan dauden **aldagai azaltzaileak** eta haien **garrantzia**:

Aldagaia	Modeloan duen garrantzia
Lehen hizkuntza euskara bai/ez	**%63,0**
Gipuzkoakoa bai/ez	%10,9
Euskaraz irakurtzeko gaitasuna	%7,1
Adin-taldea	%7,0
Gune soziolinguistikoa	%6,2
Eskura dituen aparailu kopurua	%3,4
Lehen hizkuntza gaztelania bai/ez	%2,4

ONDORIOAK

- Aldagai azaltzaileek menpeko aldagaiarekin ez dute erlaziorik edo erlazioa ahula da.
  - **Euskara lehen hizkuntza** izatea da euskarazko hedabide tradizionalak kontsumitzea aldagaiarekin erlazio sendoena duen aldagaia.
  - Ondoren, euskara ezagutza mailarekin lotutako aldagaiek dituzte menpeko aldagaiarekiko erlazio sendoenak.

- Euskarazko kontsumoa hoberen aurresaten duen modeloan (sentikortasuna= 0.781) **zazpi aldagai azaltzaile** daude.
  - Euskarazko kontsumoaren **%63**a **euskara lehen hizkuntza** izatea aldagaiak azaltzen du.
  - Modelo hau momentura arte eskura dauden aldagai azaltzaileekin eraiki da → Aldagai azaltzaile berriak sartuko dira (CIES + Inkesta Soziolinguistikoa).

Matematikari Euskaldunen
VI. Topaketak

behategia
EUSKAL HEDABIDEEN BEHATOKIA

EUSKARAZKO HEDABIDEEN KONTSUMOAN ERAGITEN DUTEN ERABILTZAILE ALDAGAIAK

Naroa Burreso Pardo
naroa.burreso@ehu.eus

2024ko ekainak 17