eman ta zabal zazu
UPV EHU
ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA
50 URTE AÑOS 1968 - 2018
Biba Zientzia! Ciencia Viva

Machine Learning bidezko mezu toxikoen detekzioa

Ane Acha Gonzalez

Gradu Amaierako Lana
Matematikako Gradua

Mikel Peñagarikano
irakasleak zuzendutako lana

Leioan, 2024ko uztailaren 11a

Sarrera

Helburua

Sare sozial zehatz bateko mezu toxikoak automatikoki iragaztea da, Machine Learning-ean oinarritutako klasifikazio ereduak erabiliz.

AURKIBIDEA

1. Problemaren eta datu basearen azalpena.
2. Naive Bayes.
3. Bektoreen bidezko eredua.
4. Erregresio Logistikoa.
5. Support Vector Machine (Euskarri bektoreko makinak).
6. Decision Tree (Erabaki zuhaitzak) eta Random Forest (Ausazko basoak).
7. Ondorioak.

1. Problemaren eta datu basearen azalpena.

- Kaggle → *Quora Insincere Questions Classification*

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0
...	...	...	...
1306117	ffffcc4e2331aaf1e41e	What other technical skills do you need as a c...	0
1306118	ffffd431801e5a2f4861	Does MS in ECE have good job prospects in USA ...	0
1306119	ffffd48fb36b63db010c	Is foam insulation toxic?	0
1306120	ffffec519fa37cf60c78	How can one start a research project based on ...	0
1306121	ffffed09fedb5088744a	Who wins in a battle between a Wolverine and a...	0

1. Problemaren eta datu basearen azalpena.

- Datuen analisia:
  - Aurreprozesatze teknikak
  - Erantzun aldagaia etiketak: 0 klasea (ez - toxikoa) $\rightarrow$ %94
    1 klasea (toxikoa) $\rightarrow$ %6

- Datuen banaketa:
  Entrenamendua (% 80), balidazioa (% 10) eta testa (% 10)

- Ebaluazioa:
  - Zehaztasuna (Accuracy)
  - Prezisioa (Precision)
  - Sentikortasuna (Recall)
  - Espezifikotasuna (Specificity)
  - ROC AUC
  - **Precision-Recall AUC**

2. Naive Bayes.

2. 1. Oinarri teorikoa

Helburua: $P(Y = y_j \mid \mathbf{X} = \mathbf{x}_i)$

$$P(y_j \mid \mathbf{x}_i) = \frac{P(y_j) \prod_{k=1}^{n} P(x_k \mid y_j)}{P(\mathbf{x}_i)} \tag{1}$$

$$\hat{y} = \underset{y_j \in Y}{\operatorname{argmax}} P(y_j \mid \mathbf{x}_i) = \underset{y_j \in Y}{\operatorname{argmax}} P(y_j) \prod_{k=1}^{n} P(x_k \mid y_j) \tag{2}$$

$$\hat{y} = \underset{y_j \in Y}{\operatorname{argmax}} \log P(y_j \mid \mathbf{x}_i) = \underset{y_j \in Y}{\operatorname{argmax}} \log P(y_j) + \sum_{k=1}^{n} \log P(x_k \mid y_j) \tag{3}$$

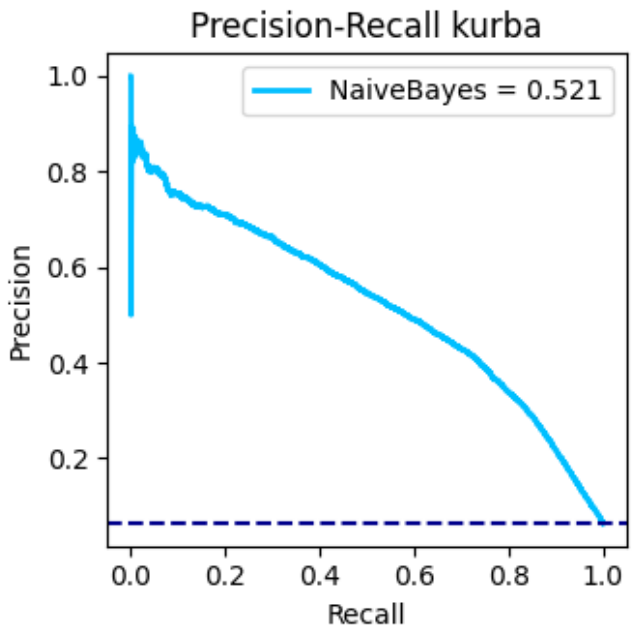
2. Naive Bayes.

2. 2. Entrenamendua

- $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ mezuen multzoa eta $\mathbf{x}_i = (x_1, \ldots, x_n)$
- $Y = \{0, 1\}$ klaseen multzo bitarra.
- $\hat{y} = \underset{y_j \in \{0,1\}}{\operatorname{argmax}} \log P(y_j \mid \mathbf{x}_i) = \underset{y_j \in \{0,1\}}{\operatorname{argmax}} \log P(y_j) + \sum_{k=1}^{n} \log P(x_k \mid y_j)$

2.3. Ebaluazioa

ALGORITMOA	NAIVE BAYES
PR_AUC	**0.52**
ROC_AUC	0.92
Zehaztasuna	0.92
Prezisioa	0.41
Sentikortasuna	0.73
Espezifikotasuna	0.93



3. Bektoreen bidezko eredua.

3. 1. Bektorizazioa

$$M = \begin{pmatrix} m_{11} & \dots & m_{1q} \\ \vdots & \ddots & \vdots \\ m_{p1} & \dots & m_{pq} \end{pmatrix} \in \mathbb{N}^{p \times q} \tag{4}$$

3. 2. Eredu bektorialaren oinarriak

$$M \cdot V + J_{p\times 1} \cdot L = \begin{pmatrix} \log(P(\mathbf{x}_1 \mid 0)P(0)) & \log P((\mathbf{x}_1 \mid 1)P(1)) \\ \log(P(\mathbf{x}_2 \mid 0)P(0)) & \log P((\mathbf{x}_2 \mid 1)P(1)) \\ \vdots & \vdots \\ \log P((\mathbf{x}_p \mid 0)P(0)) & \log P((\mathbf{x}_p \mid 1)P(1)) \end{pmatrix} \tag{5}$$

non

$$V = \begin{pmatrix} \log P(x_1 \mid 0) & \log P(x_1 \mid 1) \\ \log P(x_2 \mid 0) & \log P(x_2 \mid 1) \\ \vdots & \vdots \\ \log P(x_q \mid 0) & \log P(x_q \mid 1) \end{pmatrix}, \quad J_{p\times 1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \; eta \quad L = \begin{pmatrix} \log P(0) & \log P(1) \end{pmatrix}$$

4. Erregresio Logistikoa

4. 1. Oinarri teorikoa.

Izan bitez Y banaketa binomialari darraion erantzun aldagaia eta $\mathbf{X} = (X_1, X_2, ..., X_p)^T$ non $X_1, X_2, ..., X_p$ p aldagai aske diren.

$$P(Y = 1 \mid \mathbf{X}) = p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \in (0, 1) \qquad (6)$$

Logit transformazioa:

$$g(p(\mathbf{X})) = logit(p(\mathbf{X})) = \ln\left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right) = \beta_0+\beta_1 X_1+\ldots+\beta_p X_p \qquad (7)$$

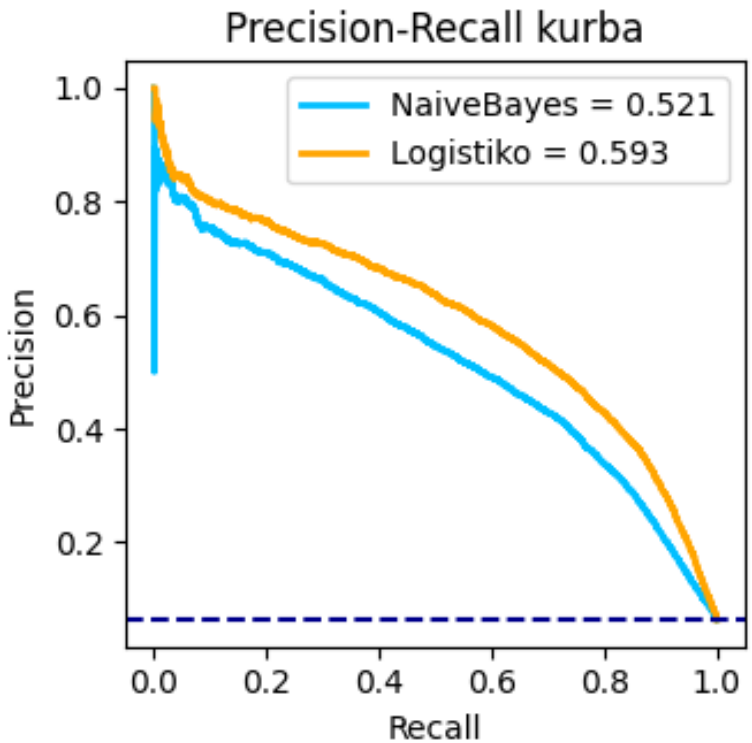
Parametroen estimazioa eta interpretazioa

4. Erregresio Logistikoa

4. 2. Entrenamendua eta ebaluazioa.

- *LogisticRegression*

ALGORITMOA	NAIVE BAYES	LOGISTIKOA
PR_AUC	0.52	**0.59**
ROC_AUC	0.92	0.94
Zehaztasuna	0.92	0.95
Prezisioa	0.41	0.68
Sentikortasuna	0.73	0.40
Espezifikotasuna	0.93	0.99



5. Support Vector Machine.

5. 1. Oinarri teorikoa.

Linealki banandutako kasuak

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{8}$$

Hiperplano optimoaren bilaketa **w** eta b balioak aurkitzeko optimizazio problema gisa adieraz daiteke:

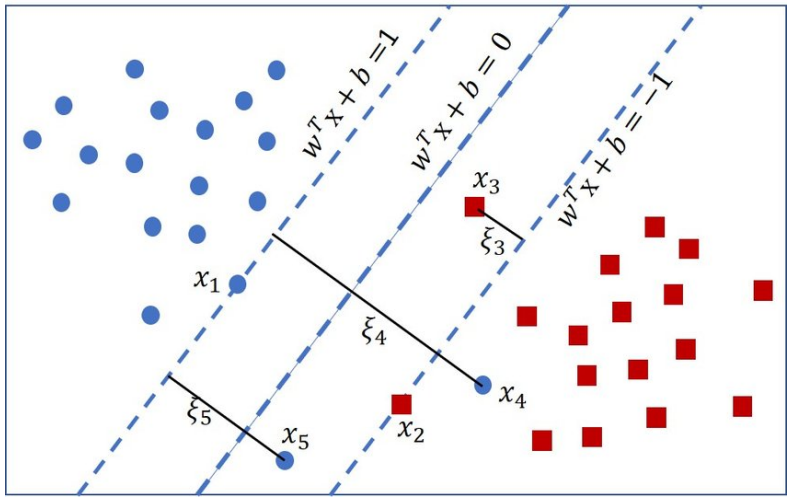
$$\begin{aligned} &\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad i = 1, \ldots, n \end{aligned} \tag{9}$$

Problema duala:

$$\begin{aligned} &\max_{\alpha} \mathcal{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ &\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{eta} \quad \alpha_i \geq 0 \quad i = 1, \ldots, n \end{aligned} \tag{10}$$

5. Support Vector Machine.

Linealki kuasi-bananduak



Optimizazio problema:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{11}$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) + \xi_i - 1 \geq 0, \quad \xi_i \geq 0, \quad i = 1,\ldots,n$$

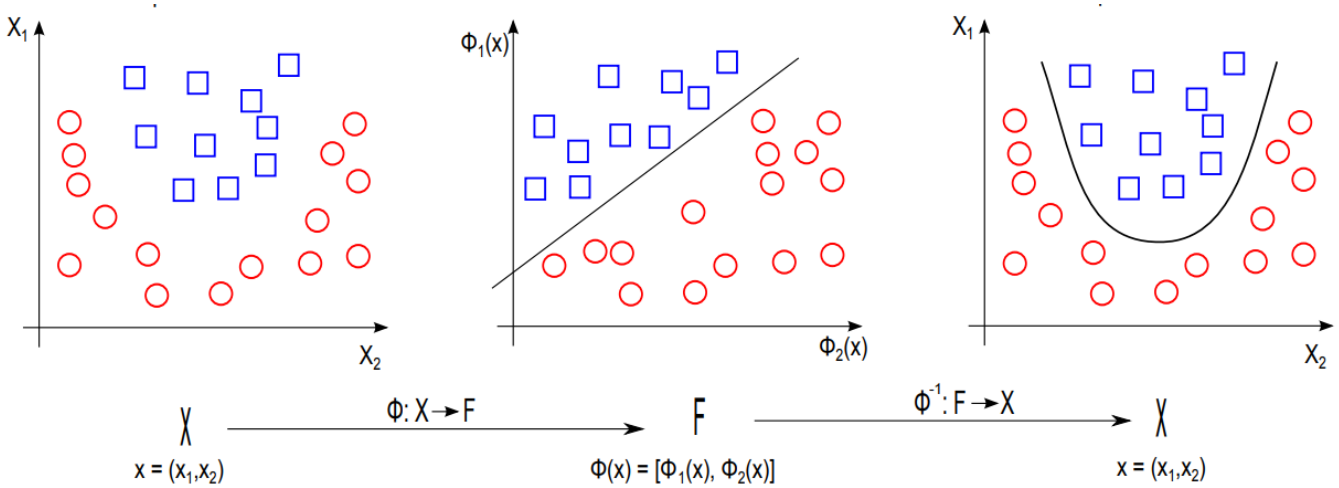
Lortzen den problema duala:

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \tag{12}$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1,\ldots,n$$

5. Support Vector Machine.

Linealki banangarriak ez diren kasuak



Problema duala:

$$\max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \tag{13}$$

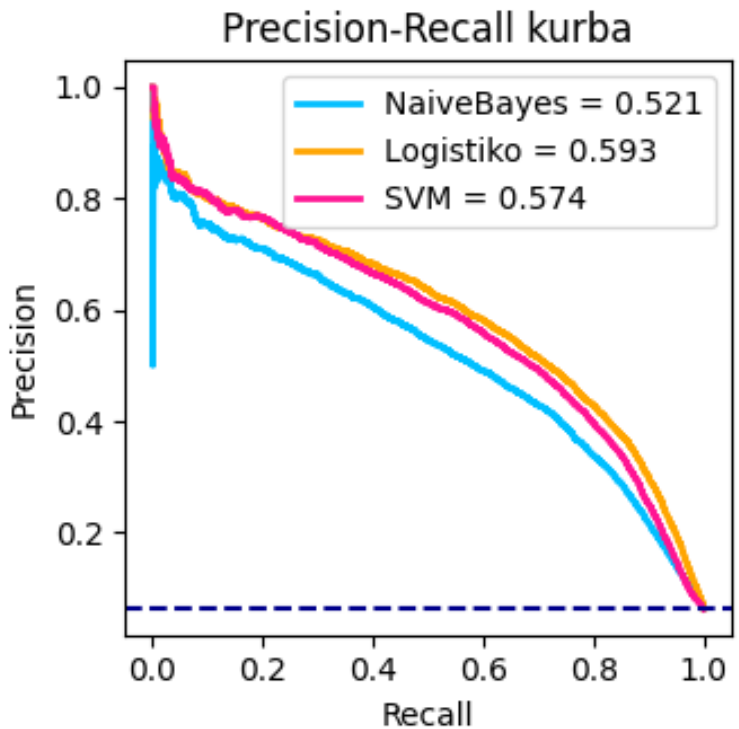
non $K(\mathbf{x}_i, \mathbf{x}_j)$ Kernelen funtzioa den.

5. Support Vector Machine.

5. 2. Entrenamendua eta ebaluazioa

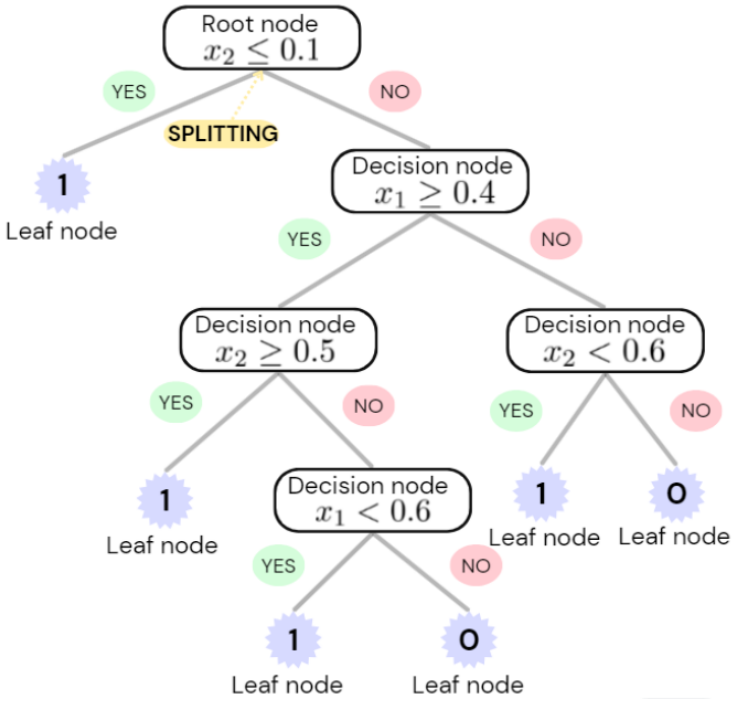
- *LinearSVC*

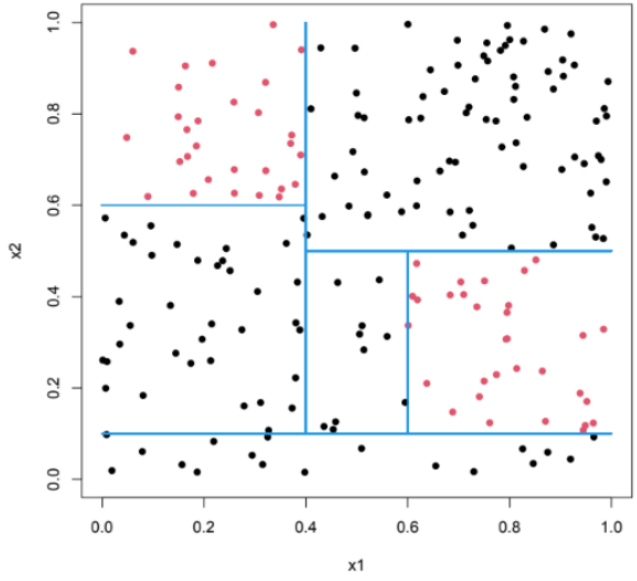
ALGORITMOA	NAIVE BAYES	LOGISTIKOA	SMV
PR_AUC	0.52	**0.59**	0.57
ROC_AUC	0.92	0.94	0.93
Zehaztasuna	0.92	0.95	0.95
Prezisioa	0.41	0.68	0.67
Sentikortasuna	0.73	0.40	0.39
Espezifikotasuna	0.93	0.99	0.99



6. Decision Tree eta Random Forest.

6. 1. Oinarri teorikoa





6. Decision Tree eta Random Forest.

Zatiketa bitarraren ezpurutasunaren murrizketa:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{14}$$

Zatiketa optimoa:

$$s^* = \arg\max_{s \in Q} \Delta i(s,t) \tag{15}$$

Shannon-en entropia eta *Gini indizea*, hurrenez hurren:

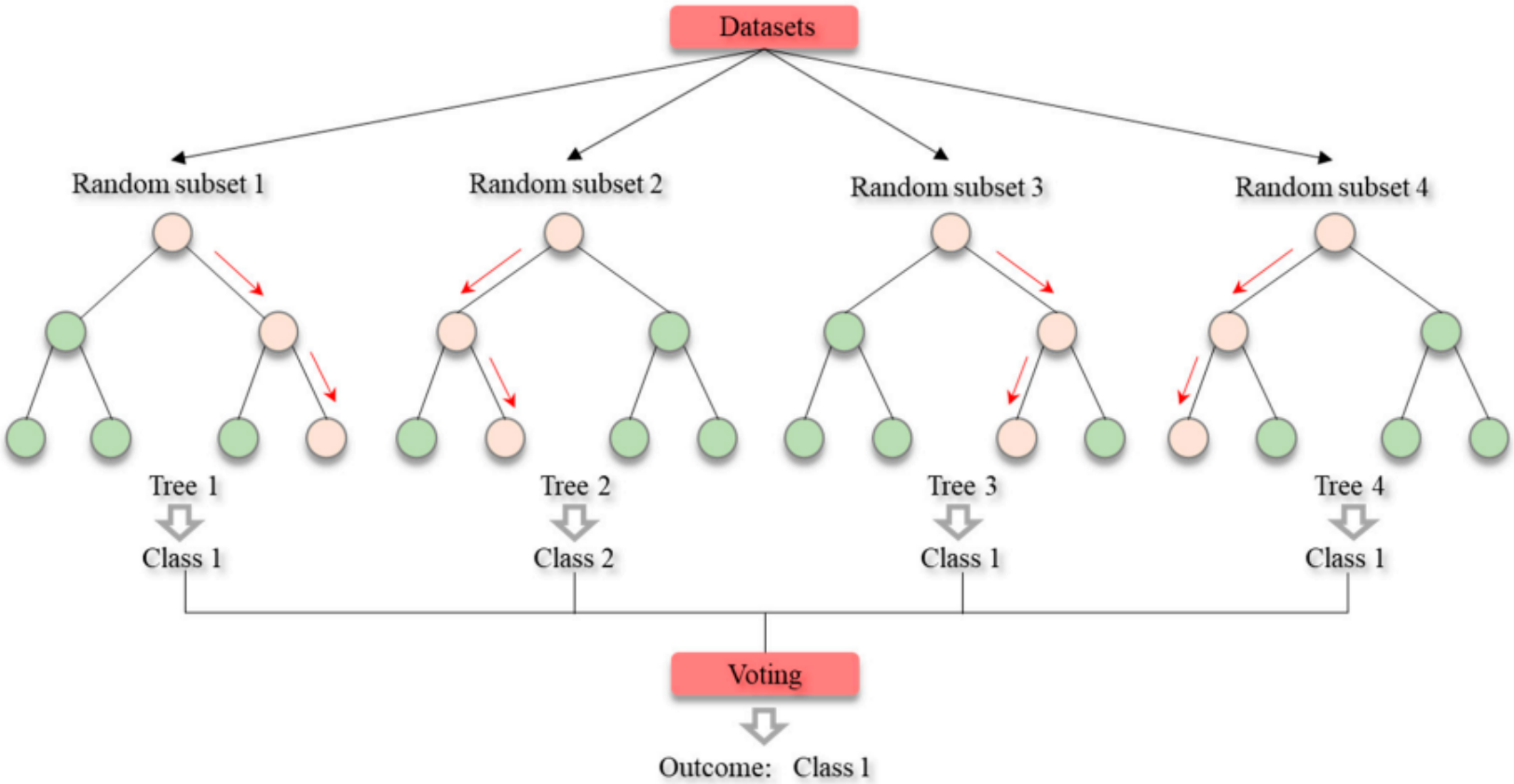
$$i_H(t) = -\sum_{k=1}^{J} p(c_k|t) \log_2(p(c_k|t)) \tag{16}$$

$$i_G(t) = \sum_{k=1}^{J} p(c_k|t) \left(1 - p(c_k|t)\right) \tag{17}$$

non $p(c_k|t)$ t nodoa c_k klasekoa izateko probabilitatea den.

6. Decision Tree eta Random Forest.

Gehiegizko doikuntza (overfitting) ekiditeko: Random Forest.



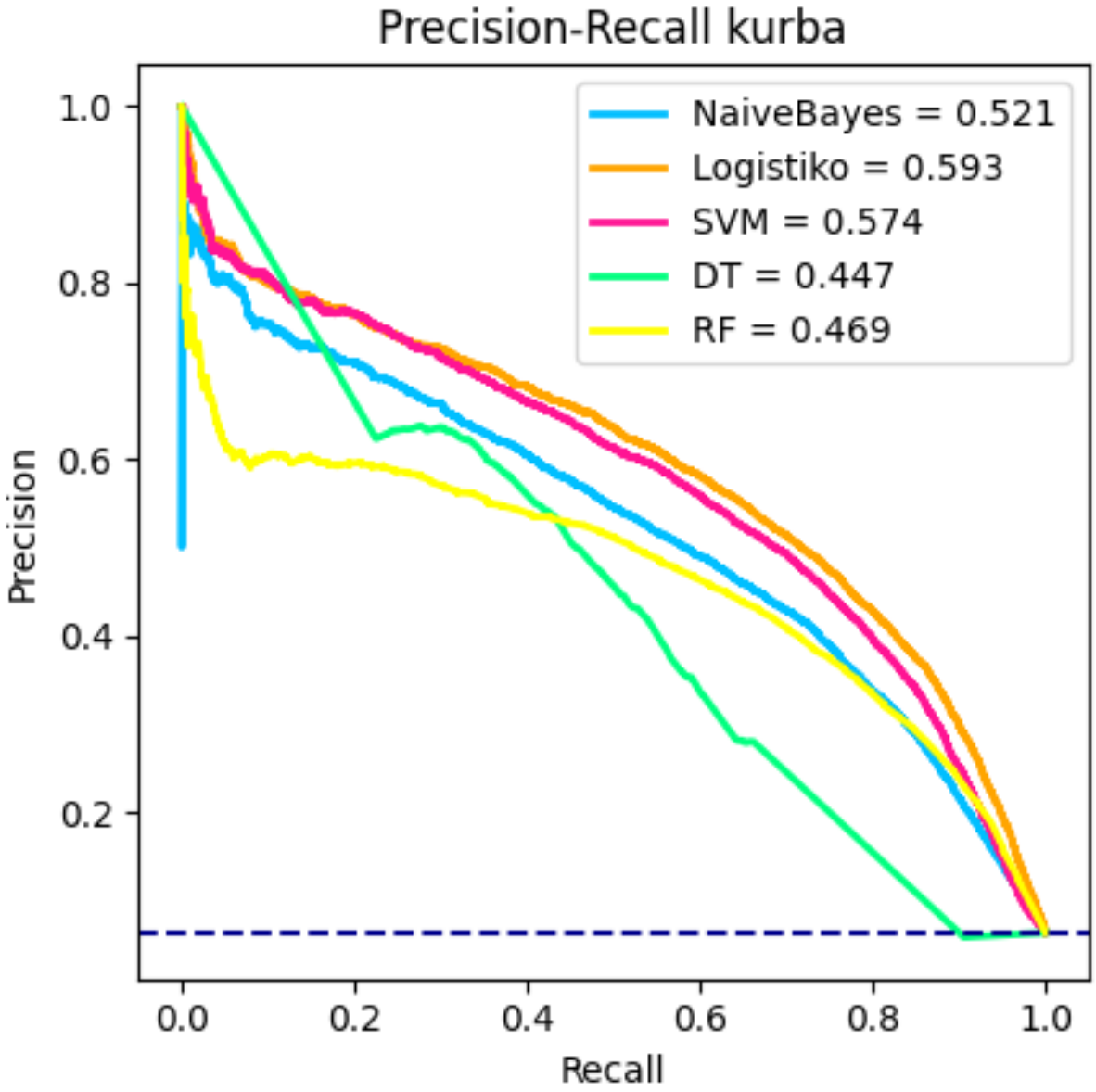
6. Decision Tree eta Random Forest.

6. 2. Entrenamendua eta ebaluazioa

- *DecisionTreeClassifier* eta *RandomForestClassifier*

ALGORITMOA	NAIVE BAYES	LOGISTIKOA	SMV	DECISION TREE	RANDOM FOREST
PR_AUC	0.52	**0.59**	0.57	0.45	0.47
ROC_AUC	0.92	0.94	0.93	0.75	0.92
Zehaztasuna	0.92	0.95	0.95	0.94	0.94
Prezisioa	0.41	0.68	0.67	0.62	0.60
Sentikortasuna	0.73	0.40	0.39	0.33	0.15
Espezifikotasuna	0.93	0.99	0.99	0.99	0.99

6. Decision Tree eta Random Forest.



7. Ondorioak.

- Algoritmo bakoitzak bere abantailak eta desabantailak.
- Hainbat arazori egin behar diete aurre.
- Erregresio logistikoa eta SVM eredurik eraginkorrenak.
- Ezinbestekoa da hauen mugak ezagutzea, matematikoki ulertzea eta egoki ebaluatzea.

Eskerrik asko !