CLASIFICACIÓN Y REGRESIÓN EN DATASET DE PELÍCULAS

Marco
Pérez González

marco.perez
@alumnos.upm.es

David
Cano Rosillo

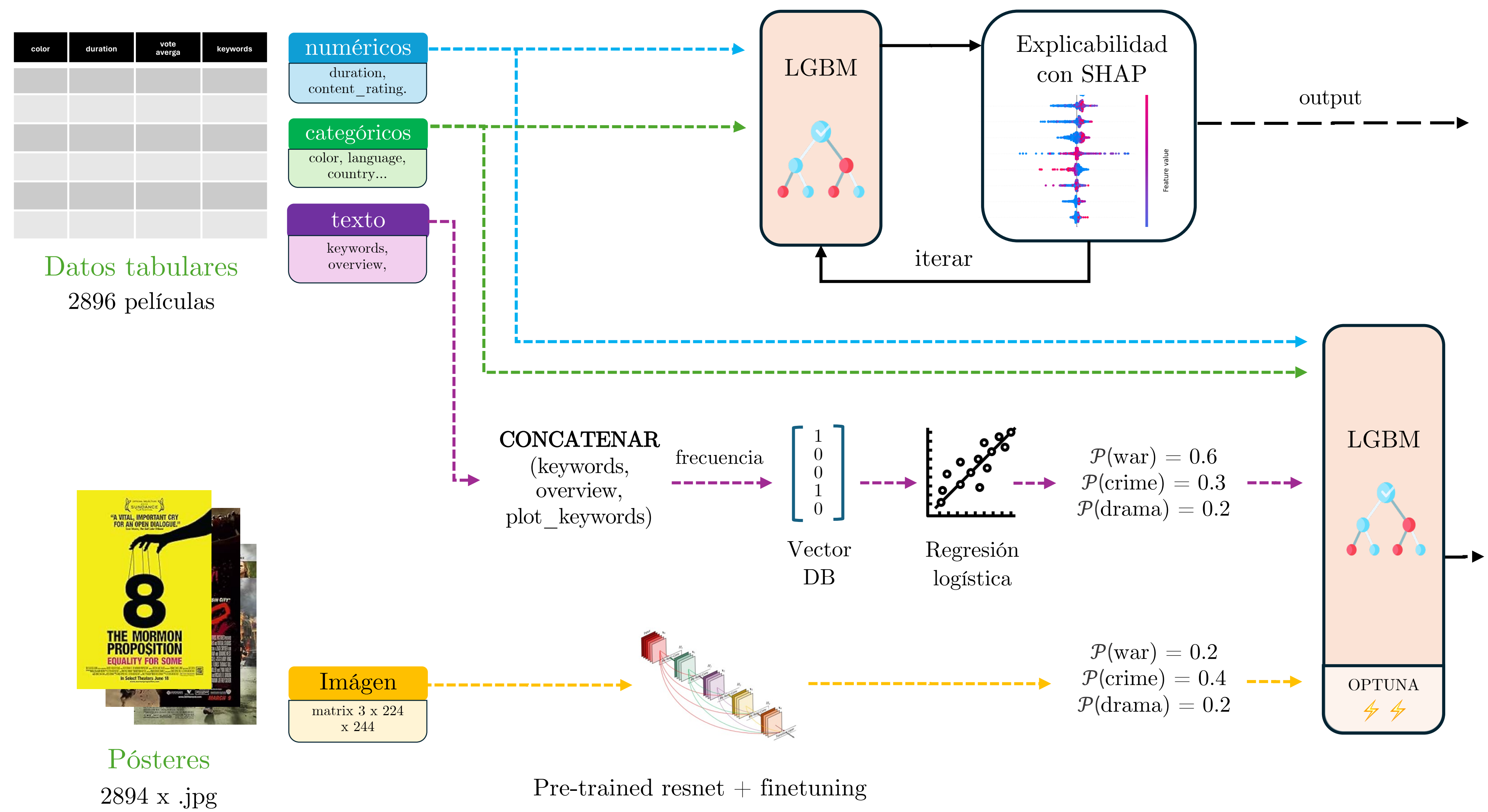
david.cano.rosillo
@alumnos.upm.es

José Javier
Román Camacho

josejavier.roman
@alumnos.upm.es

CLASIFICACIÓN Y REGRESIÓN EN DATASET DE PELÍCULAS

Datos | Procesado | Métricas



Datos tabulares

2896 películas

numéricos

duration, content_rating.

categóricos

color, language, country...

texto

keywords, overview,

LGBM

Explicabilidad con SHAP

iterar

output

CONCATENAR (keywords, overview, plot_keywords)

frecuencia

Vector DB

Regresión logística

$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

LGBM

OPTUNA

Pósteres

2894 x .jpg

Imágen

matrix 3 x 224 x 244

Pre-trained resnet + finetuning

$\mathcal{P}(\text{war}) = 0.2$
$\mathcal{P}(\text{crime}) = 0.4$
$\mathcal{P}(\text{drama}) = 0.2$

Predicción de score

	CV	TEST
MAE	**0.24**	**0.22**
MSE	**0.13**	**0.09**

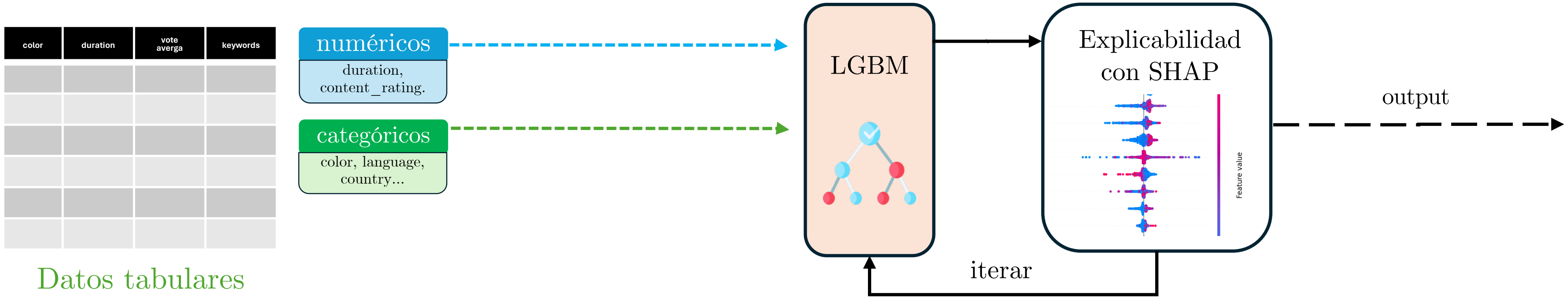
Predicción de género

	CV	TEST
ACC	0.92	0.27

Datos

Procesado

Métricas



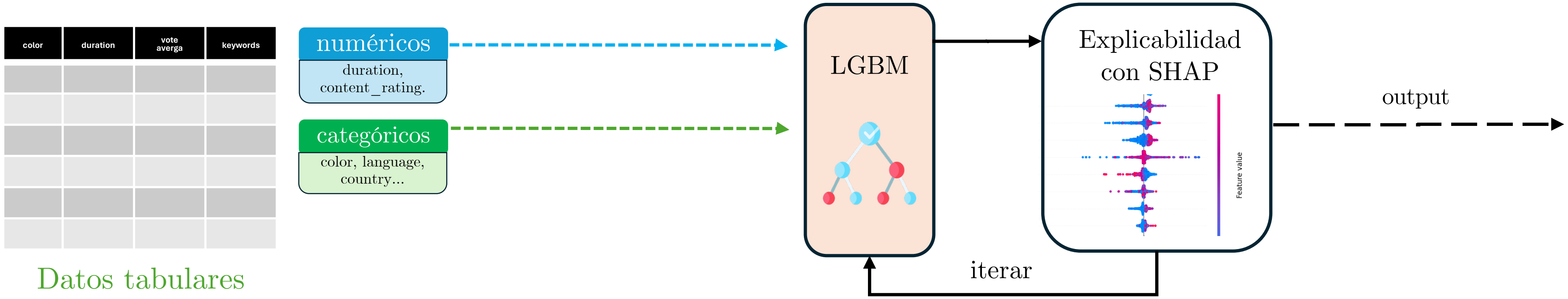
Datos tabulares

2896 películas

Predicción de score

	CV	TEST
MAE	**0.24**	**0.22**
MSE	**0.13**	**0.09**

Datos | Procesado | Métricas



Datos tabulares

2896 películas

numéricos

duration, content_rating.

categóricos

color, language, country…

LGBM

Explicabilidad con SHAP

output

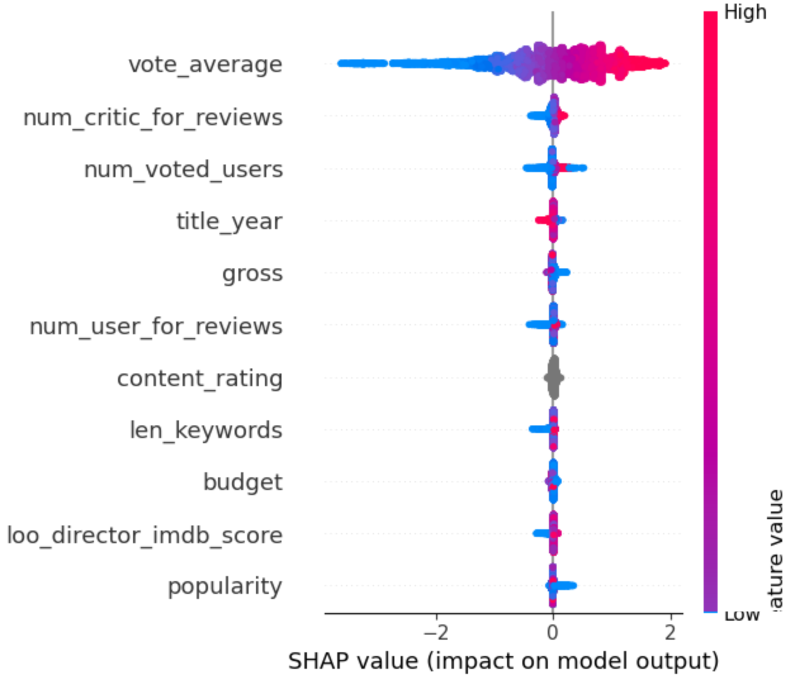
iterar

Predicción de score

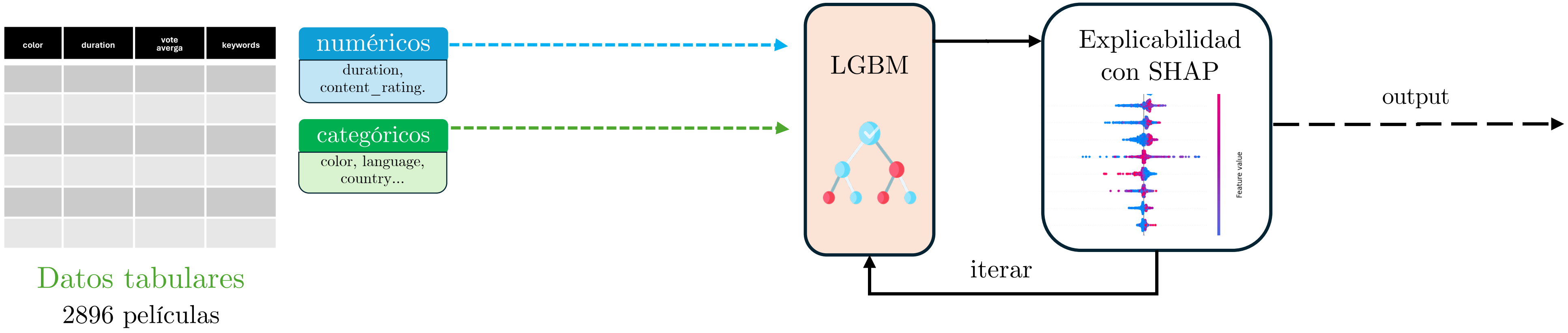
	CV	TEST
MAE	0.24	0.22
MSE	0.13	0.09

Ingeniería de características

- Variables asociadas a los directores (LOO target encoding)
- Variables asociadas a longitudes (del reparto, descripciones, etc)
- Proporciones de likes en facebook (película, actores, etc)



Datos | Procesado | Métricas



Datos tabulares

2896 películas

numéricos

duration, content_rating.

categóricos

color, language, country...

LGBM

Explicabilidad con SHAP

output

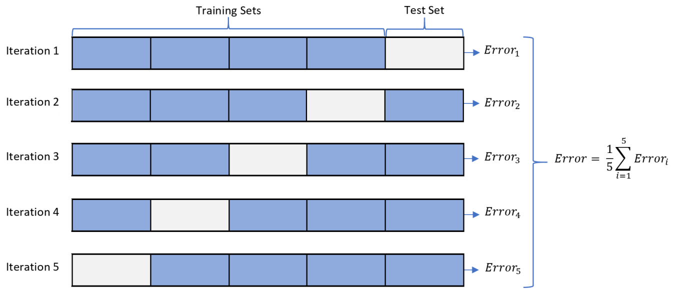
iterar

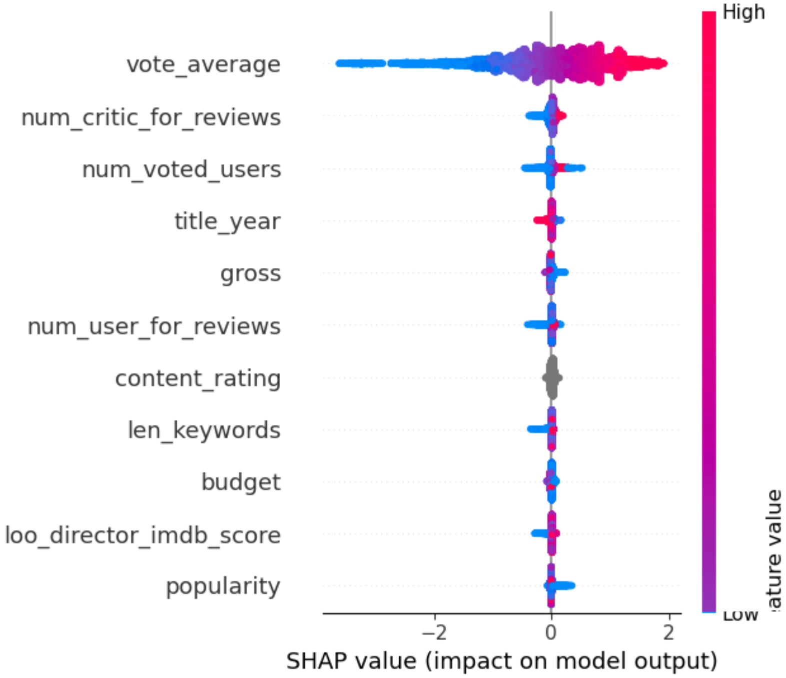
Predicción de score

	CV	TEST
MAE	0.24	0.22
MSE	0.13	0.09

Ingeniería de características

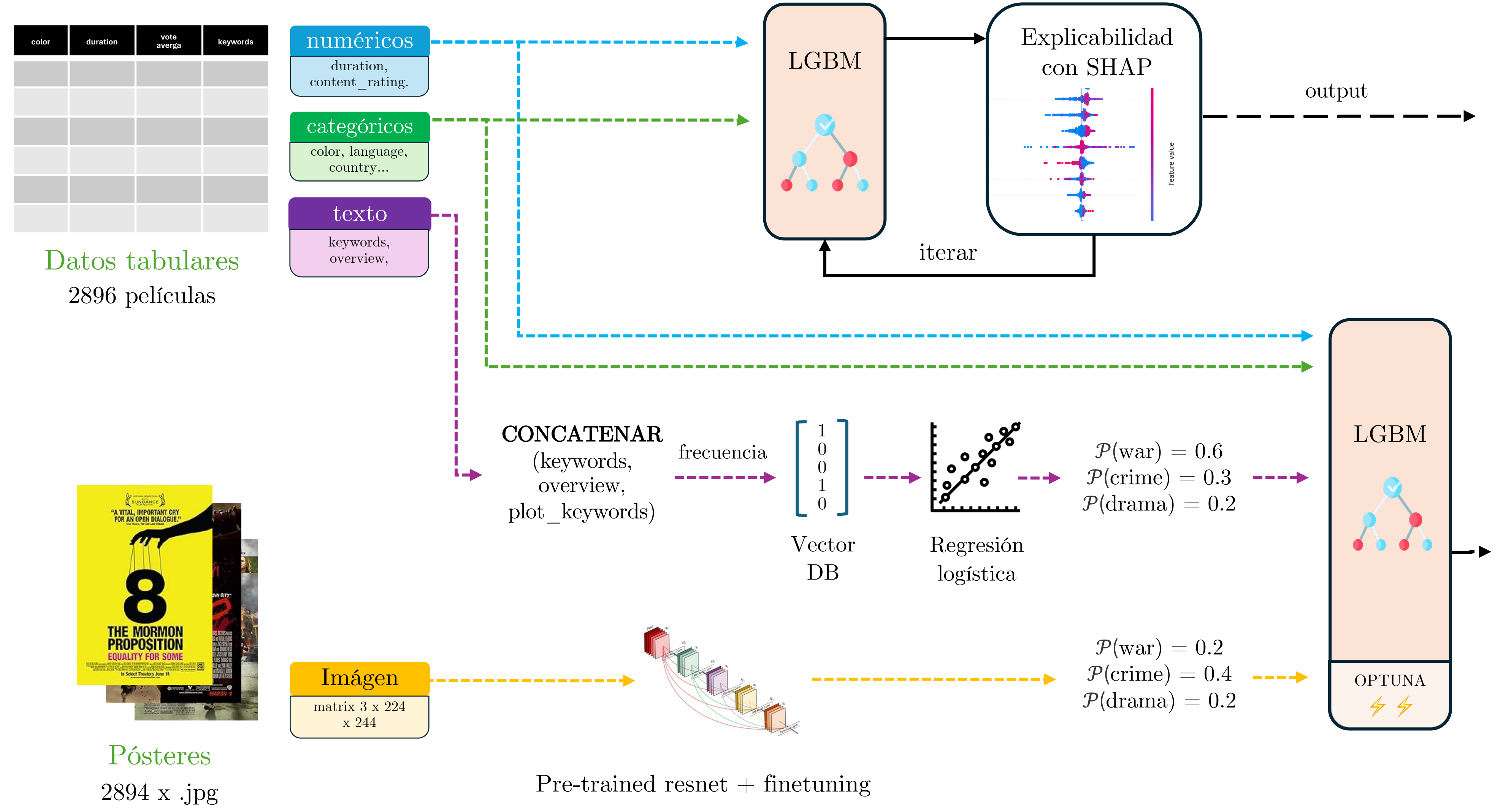
- Variables asociadas a los directores (LOO target encoding)
- Variables asociadas a longitudes (del reparto, descripciones, etc)
- Proporciones de likes en facebook (película, actores, etc)





CLASIFICACIÓN Y REGRESIÓN EN DATASET DE PELÍCULAS

Datos | Procesado | Métricas



Datos tabulares

2896 películas

numéricos

duration, content_rating.

categóricos

color, language, country...

texto

keywords, overview,

LGBM

Explicabilidad con SHAP

iterar

output

CONCATENAR (keywords, overview, plot_keywords)

frecuencia

$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$

Vector DB

Regresión logística

$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

LGBM

OPTUNA

Pósteres

2894 x .jpg

Imágen

matrix 3 x 224 x 244

Pre-trained resnet + finetuning

$\mathcal{P}(\text{war}) = 0.2$
$\mathcal{P}(\text{crime}) = 0.4$
$\mathcal{P}(\text{drama}) = 0.2$

Predicción de score

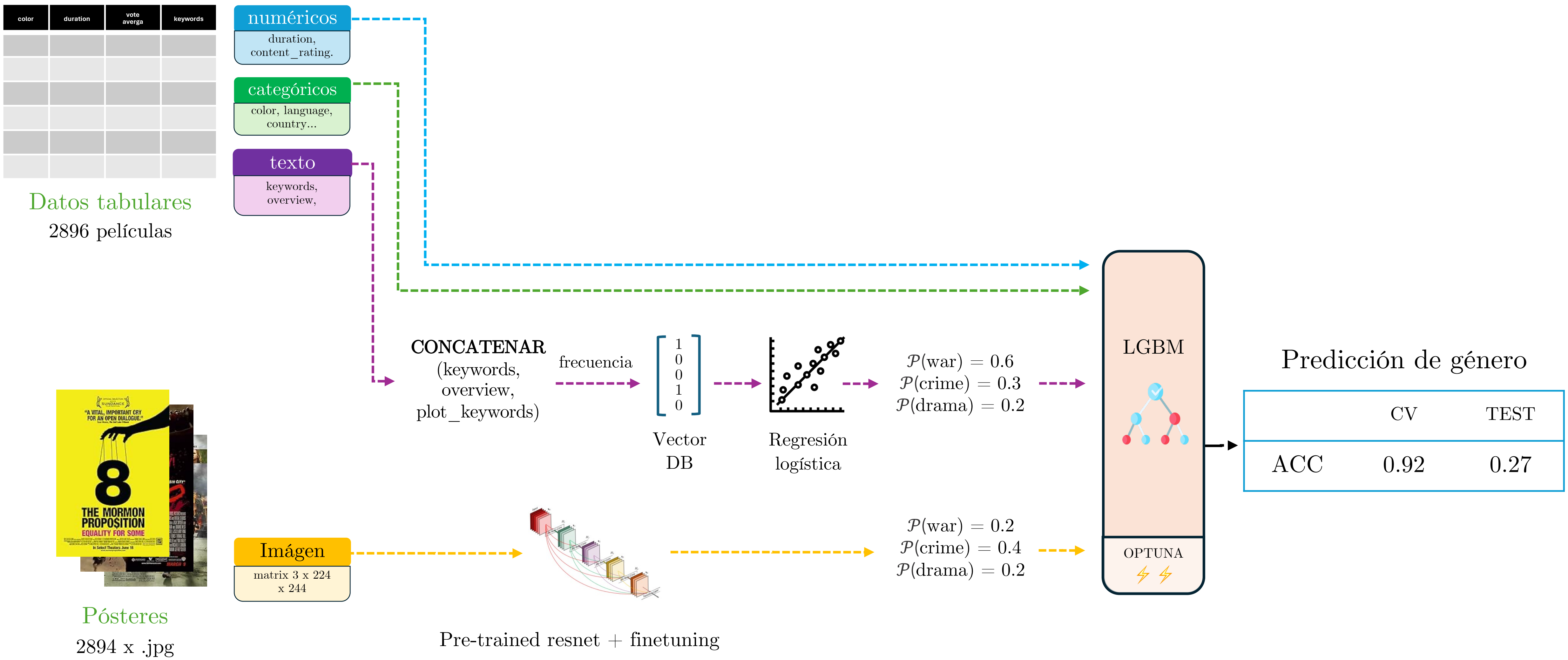
	CV	TEST
MAE	**0.24**	**0.22**
MSE	**0.13**	**0.09**

Predicción de género

	CV	TEST
ACC	0.92	0.27

CLASIFICACIÓN Y REGRESIÓN EN DATASET DE PELÍCULAS

Datos | Procesado | Métricas



Datos tabulares

2896 películas

numéricos

duration, content_rating.

categóricos

color, language, country...

texto

keywords, overview,

CONCATENAR
(keywords, overview, plot_keywords)

frecuencia

$\begin{bmatrix}1\\0\\0\\1\\0\end{bmatrix}$

Vector DB

Regresión logística

$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

Pósteres

2894 x .jpg

Imágen

matrix 3 x 224 x 244

Pre-trained resnet + finetuning

$\mathcal{P}(\text{war}) = 0.2$
$\mathcal{P}(\text{crime}) = 0.4$
$\mathcal{P}(\text{drama}) = 0.2$

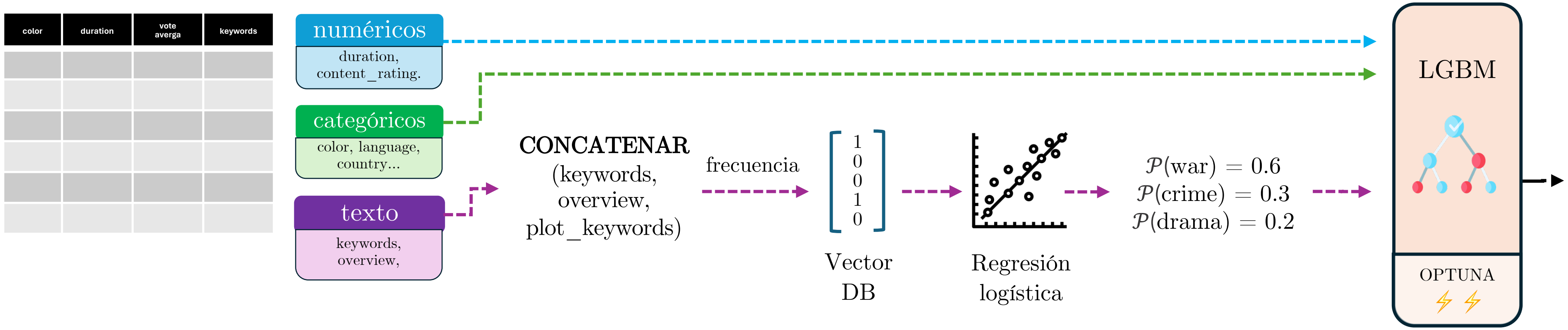
LGBM

OPTUNA

Predicción de género

	CV	TEST
ACC	0.92	0.27

Datos | Procesado | Métricas



numéricos
duration, content_rating.

categóricos
color, language, country...

texto
keywords, overview,

CONCATENAR
(keywords, overview, plot_keywords)

frecuencia

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Vector DB

Regresión logística

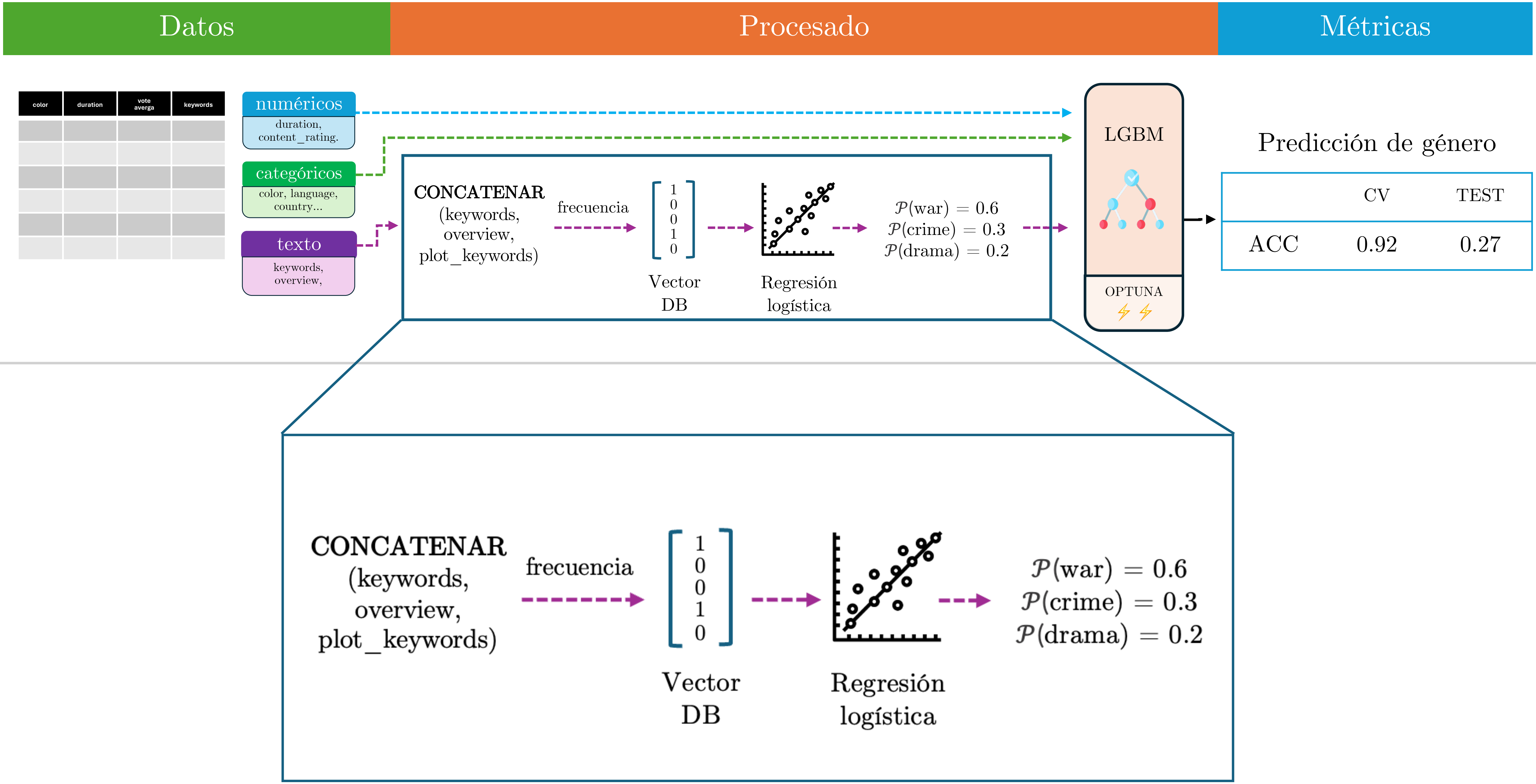
$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

LGBM

OPTUNA ⚡⚡

Predicción de género

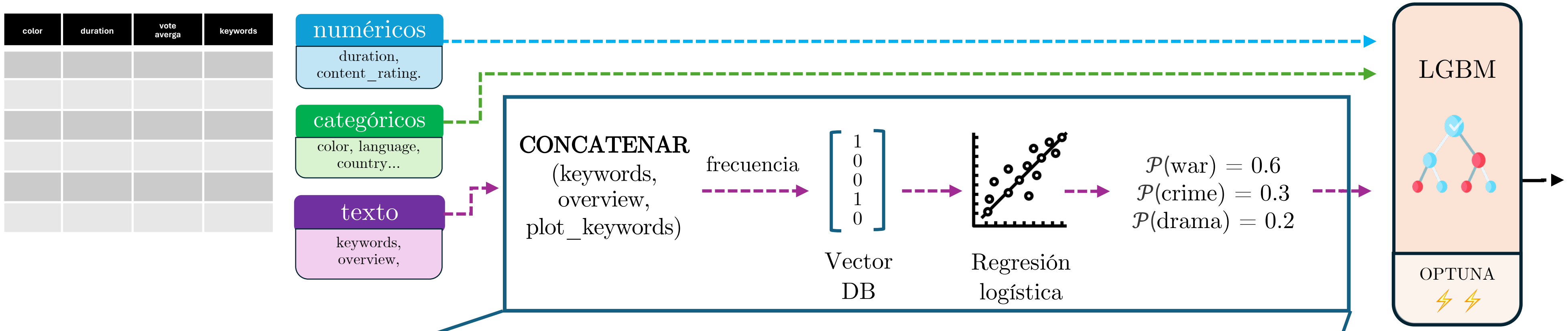
	CV	TEST
ACC	0.92	0.27


Datos
Procesado
Métricas
color
duration
vote averga
keywords
numéricos
duration, content_rating.
categóricos
color, language, country...
texto
keywords, overview,
LGBM
OPTUNA
CONCATENAR (keywords, overview, plot_keywords)
frecuencia
1 0 0 1 0
Vector DB
Regresión logística
$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$
CONCATENAR (keywords, overview, plot_keywords)
frecuencia
1 0 0 1 0
Vector DB
Regresión logística
$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

Predicción de género

	CV	TEST
ACC	0.92	0.27

Datos | Procesado | Métricas



numéricos

duration, content_rating.

categóricos

color, language, country...

texto

keywords, overview,

CONCATENAR (keywords, overview, plot_keywords) → frecuencia → Vector DB → Regresión logística →

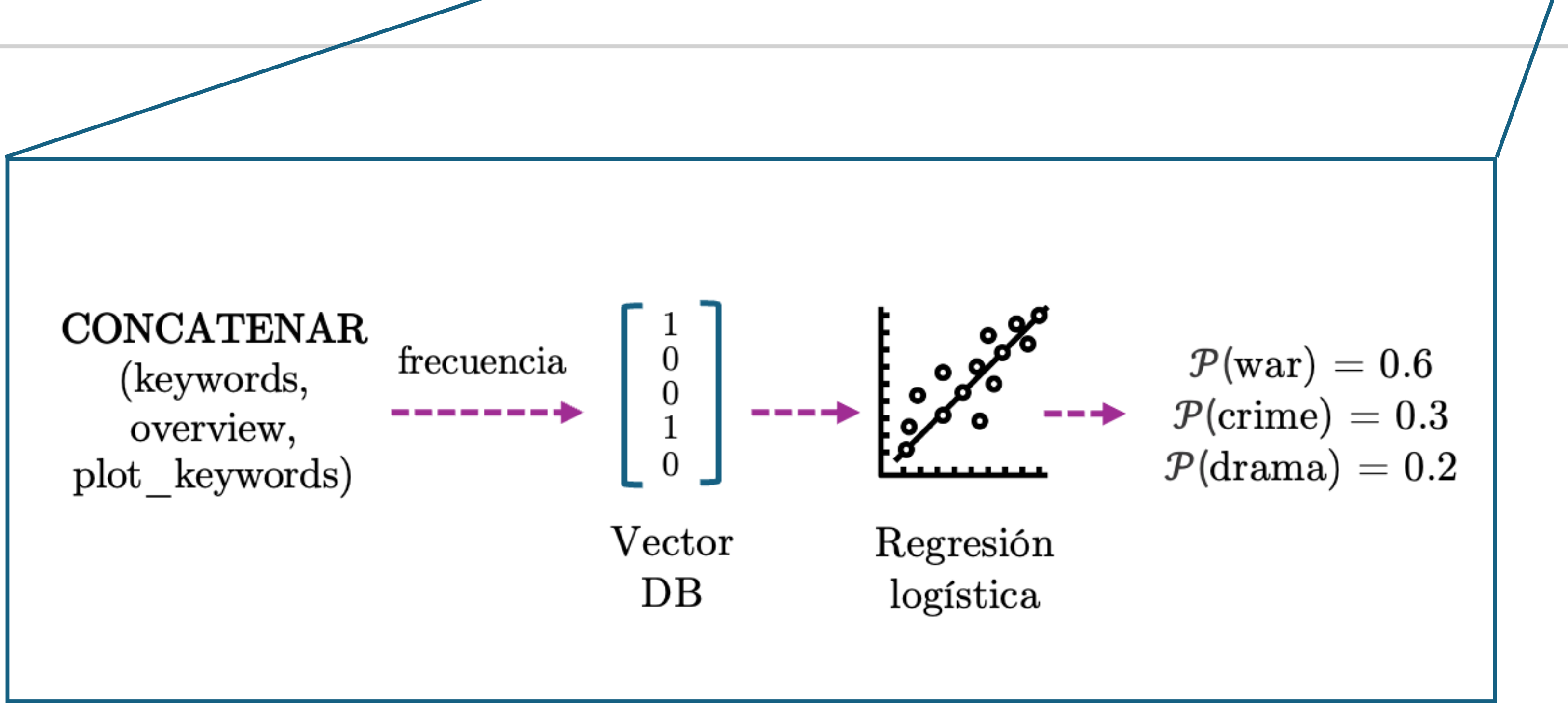
$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

LGBM

OPTUNA ⚡ ⚡

Predicción de género

	CV	TEST
ACC	0.92	0.27



CONCATENAR (keywords, overview, plot_keywords) → frecuencia → Vector DB → Regresión logística →

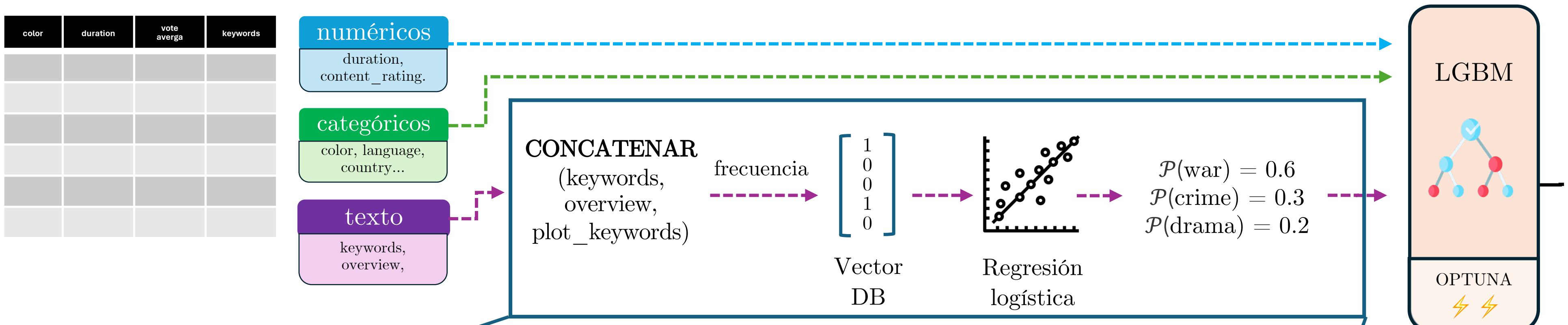
$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

Cosas que no funcionaron

- BERT
- CountVectorizer

Datos | Procesado | Métricas



color | duration | vote averga | keywords

numéricos
duration, content_rating.

categóricos
color, language, country...

texto
keywords, overview,

CONCATENAR (keywords, overview, plot_keywords)

frecuencia

$\begin{bmatrix}1\\0\\0\\1\\0\end{bmatrix}$

Vector DB

Regresión logística

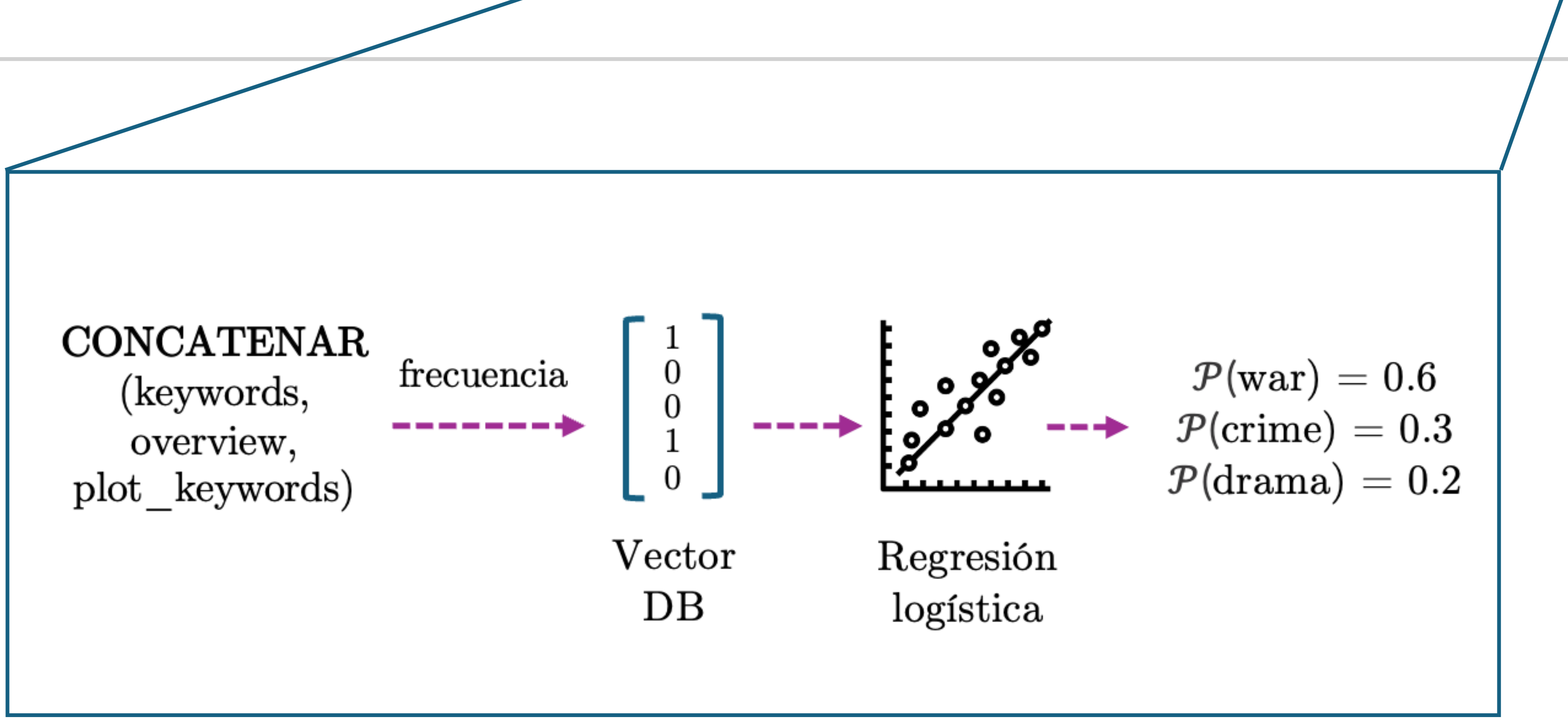
$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

LGBM

OPTUNA

Predicción de género

	CV	TEST
ACC	0.92	0.27



CONCATENAR (keywords, overview, plot_keywords)

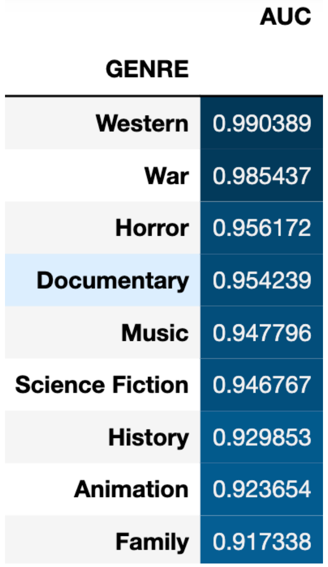
frecuencia

$\begin{bmatrix}1\\0\\0\\1\\0\end{bmatrix}$

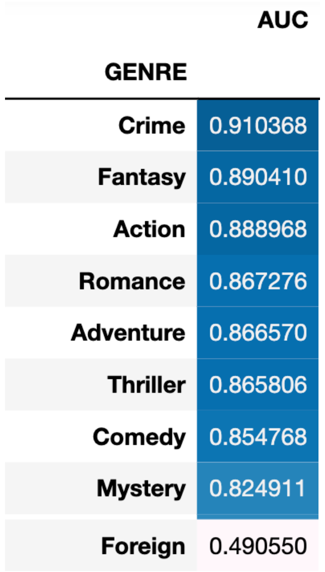
Vector DB

Regresión logística

$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

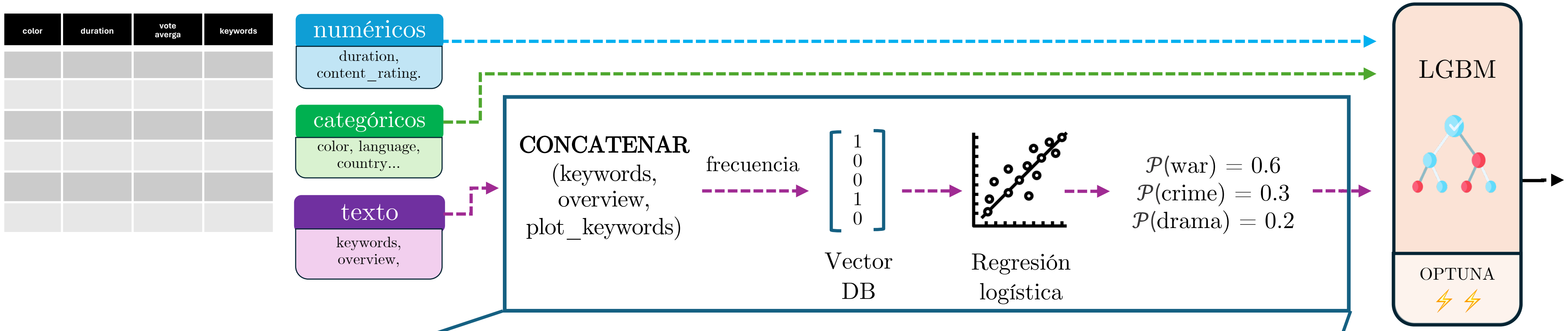


GENRE	AUC
Western	0.990389
War	0.985437
Horror	0.956172
Documentary	0.954239
Music	0.947796
Science Fiction	0.946767
History	0.929853
Animation	0.923654
Family	0.917338



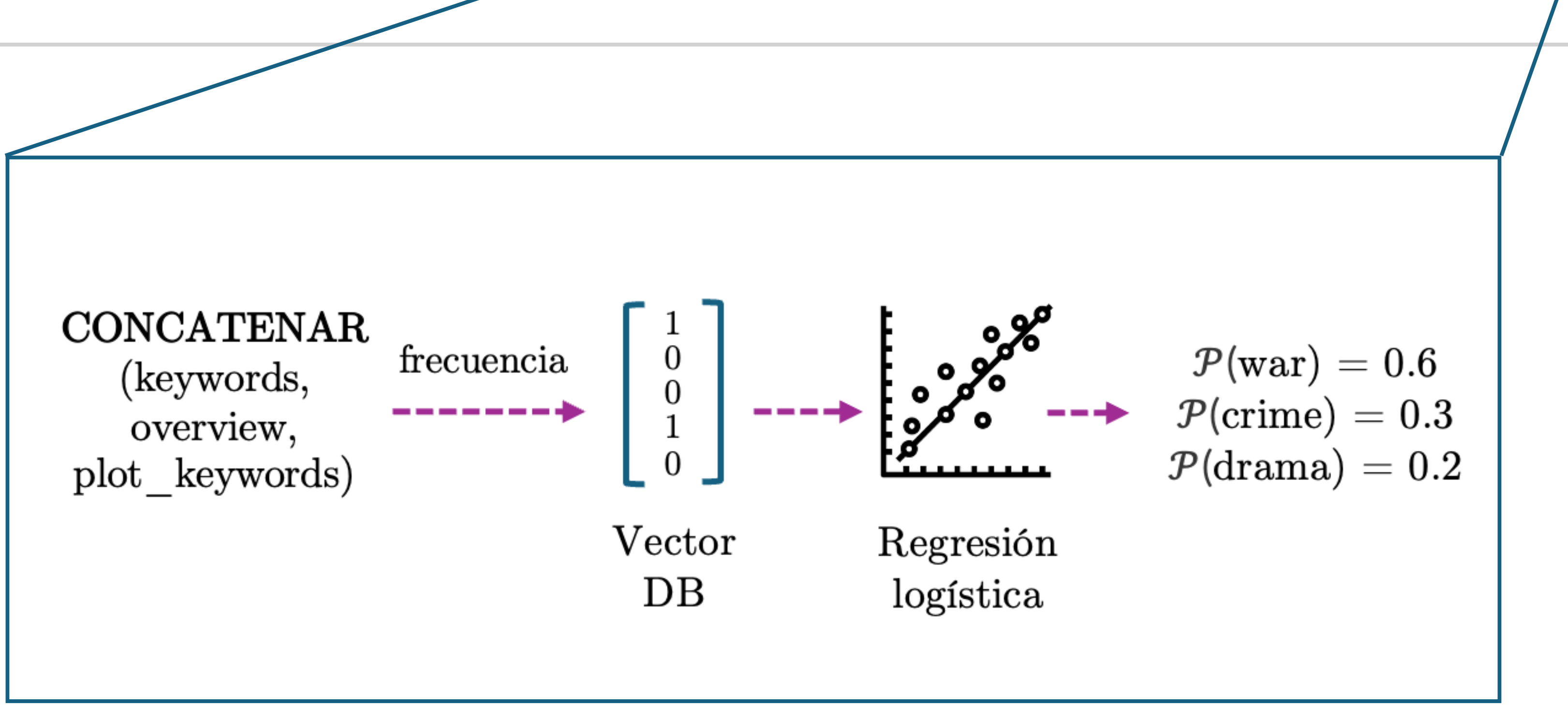
GENRE	AUC
Crime	0.910368
Fantasy	0.890410
Action	0.888968
Romance	0.867276
Adventure	0.866570
Thriller	0.865806
Comedy	0.854768
Mystery	0.824911
Foreign	0.490550

Datos | Procesado | Métricas



Predicción de género

	CV	TEST
ACC	0.92	0.27

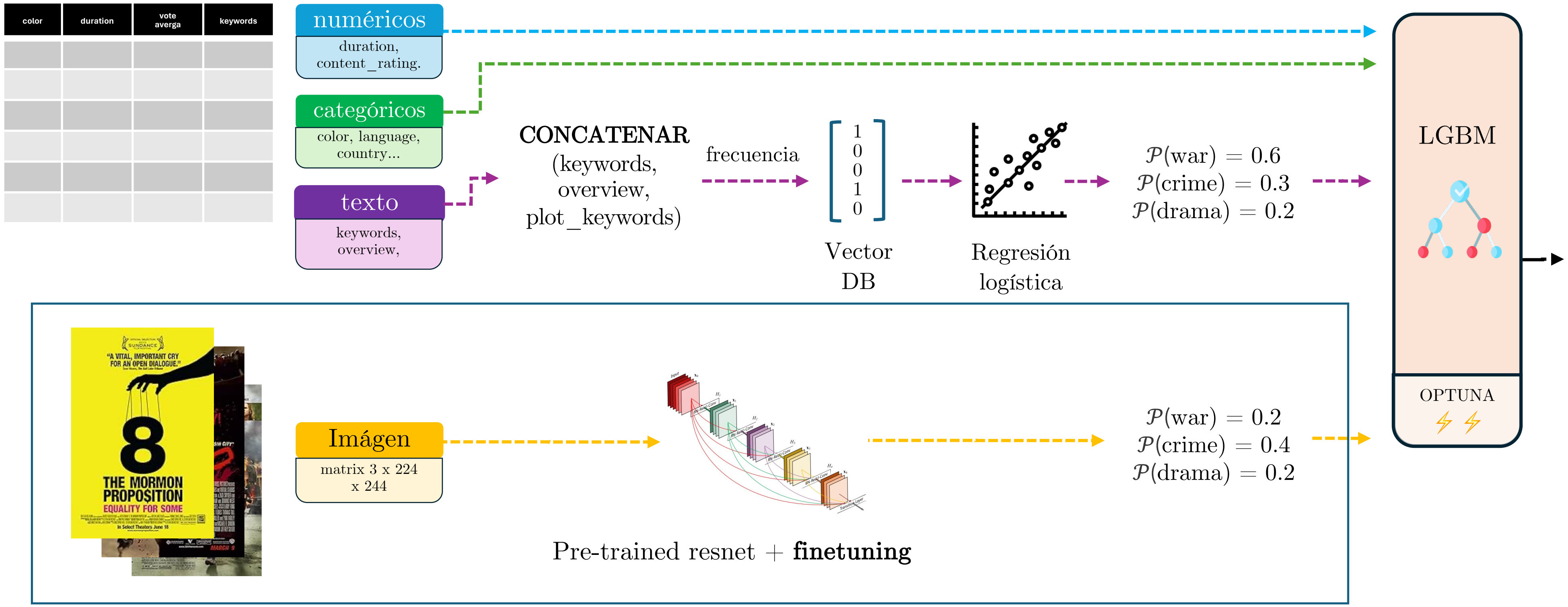

CONCATENAR (keywords, overview, plot_keywords)
frecuencia
1 0 0 1 0
Vector DB
Regresión logística
$\mathcal{P}$(war) = 0.6
$\mathcal{P}$(crime) = 0.3
$\mathcal{P}$(drama) = 0.2

“ Nuestro objetivo es generar las mejores *features* para LGBM ”

Datos

Procesado

Métricas



numéricos

duration, content_rating.

categóricos

color, language, country...

texto

keywords, overview,

CONCATENAR (keywords, overview, plot_keywords)

frecuencia

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Vector DB

Regresión logística

$$\mathcal{P}(\text{war}) = 0.6$$
$$\mathcal{P}(\text{crime}) = 0.3$$
$$\mathcal{P}(\text{drama}) = 0.2$$

Imágen

matrix 3 x 224 x 244

Pre-trained resnet + **finetuning**

$$\mathcal{P}(\text{war}) = 0.2$$
$$\mathcal{P}(\text{crime}) = 0.4$$
$$\mathcal{P}(\text{drama}) = 0.2$$

LGBM

OPTUNA

Predicción de género

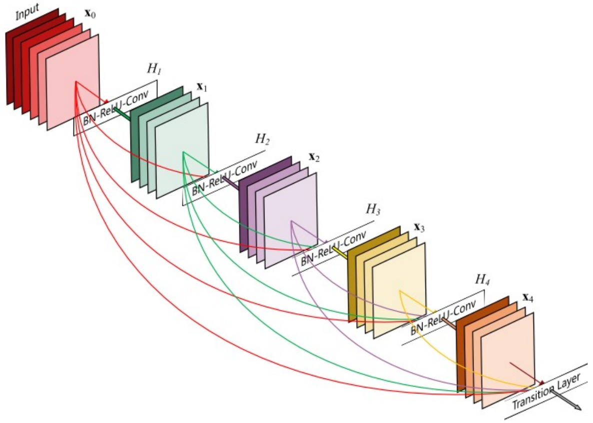
	CV	TEST
ACC	0.92	0.27

Pre-trained resnet + **finetuning**

SUNDANCE
"A VITAL, IMPORTANT CRY FOR AN OPEN DIALOGUE."
8
THE MORMON PROPOSITION
EQUALITY FOR SOME
In Select Theaters June 18
SIN CITY
MARCH 9

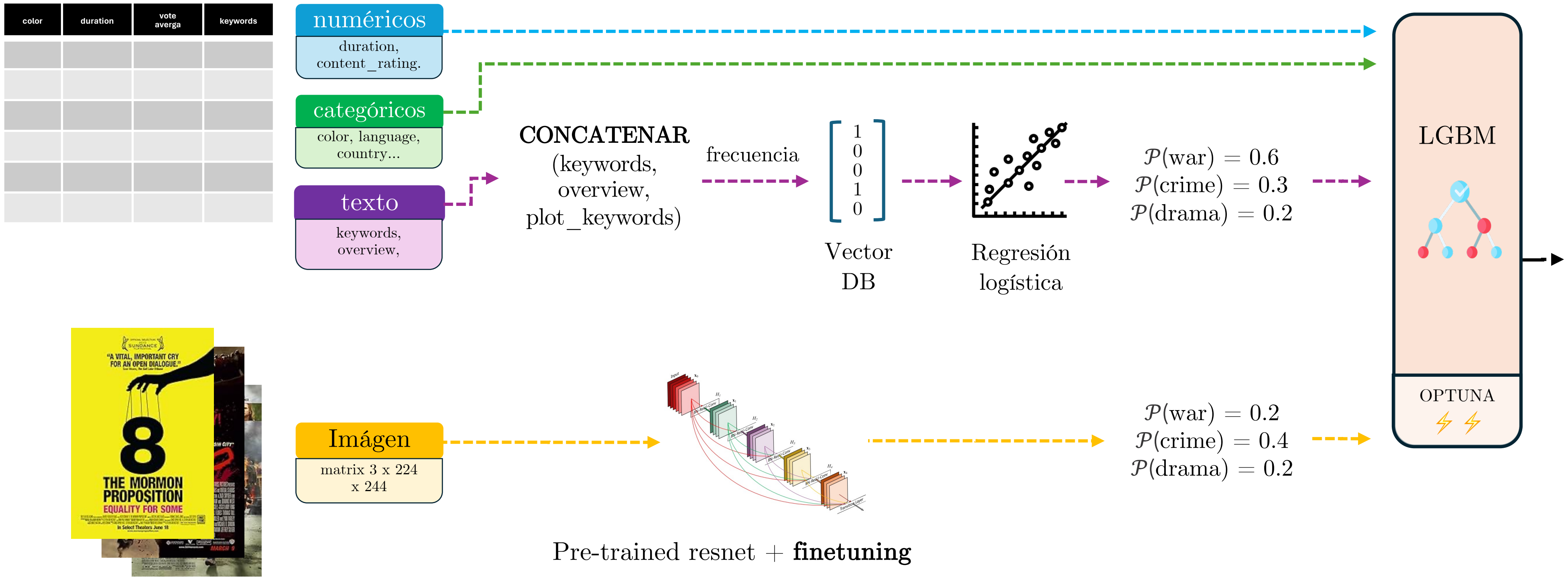
Imágen

matrix 3 x 224 x 244



$$\mathcal{P}(\text{war}) = 0.2$$
$$\mathcal{P}(\text{crime}) = 0.4$$
$$\mathcal{P}(\text{drama}) = 0.2$$

Datos | Procesado | Métricas



numéricos
duration, content_rating.

categóricos
color, language, country...

texto
keywords, overview,

CONCATENAR
(keywords, overview, plot_keywords)

frecuencia

$$\begin{bmatrix}1\\0\\0\\1\\0\end{bmatrix}$$

Vector DB

Regresión logística

$$\mathcal{P}(\text{war}) = 0.6$$
$$\mathcal{P}(\text{crime}) = 0.3$$
$$\mathcal{P}(\text{drama}) = 0.2$$

Imágen
matrix 3 x 224 x 244

Pre-trained resnet + **finetuning**

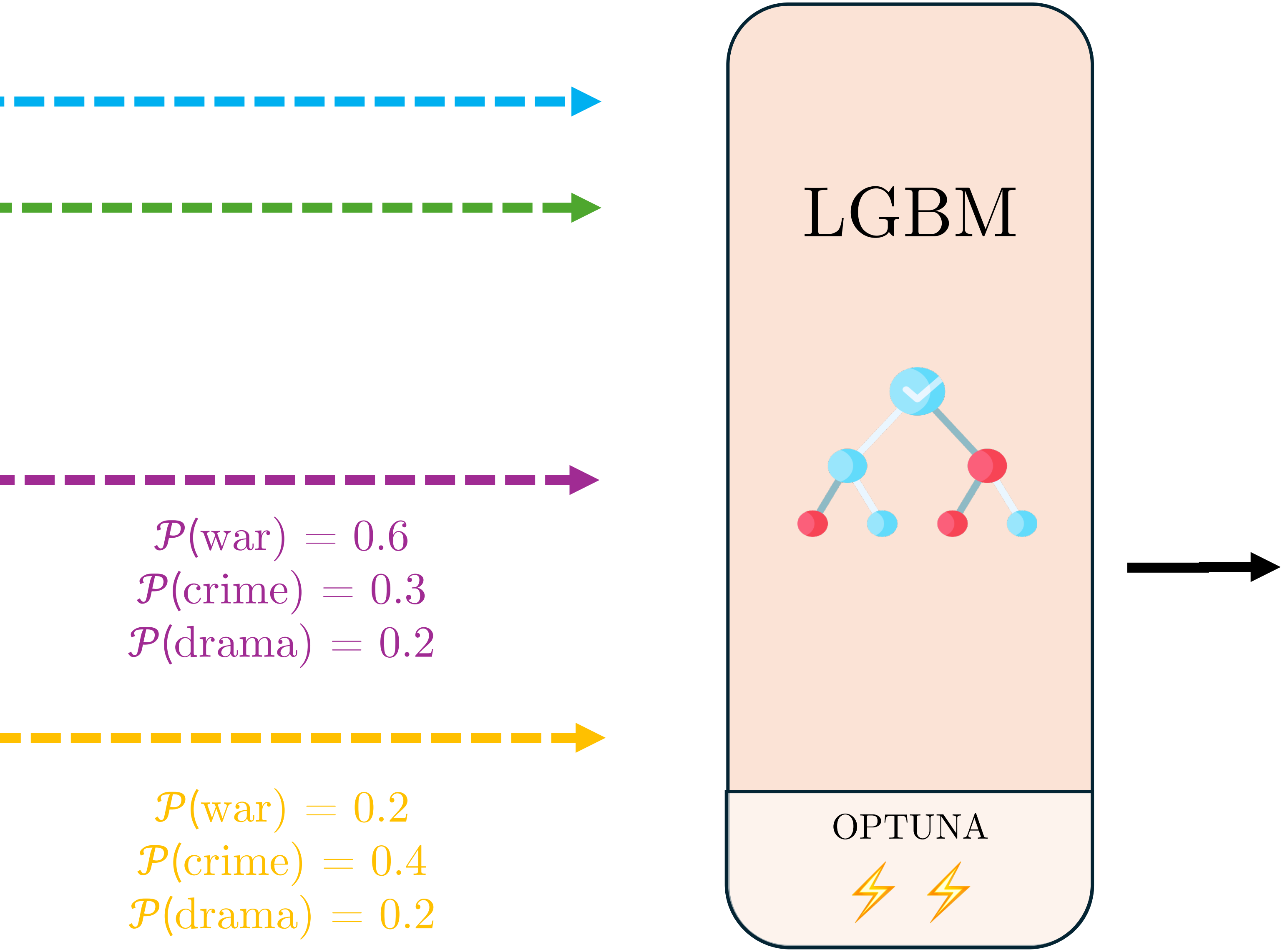
$$\mathcal{P}(\text{war}) = 0.2$$
$$\mathcal{P}(\text{crime}) = 0.4$$
$$\mathcal{P}(\text{drama}) = 0.2$$

LGBM

OPTUNA

Predicción de género

	CV	TEST
ACC	0.92	0.27



LGBM

$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$

$\mathcal{P}(\text{war}) = 0.2$
$\mathcal{P}(\text{crime}) = 0.4$
$\mathcal{P}(\text{drama}) = 0.2$

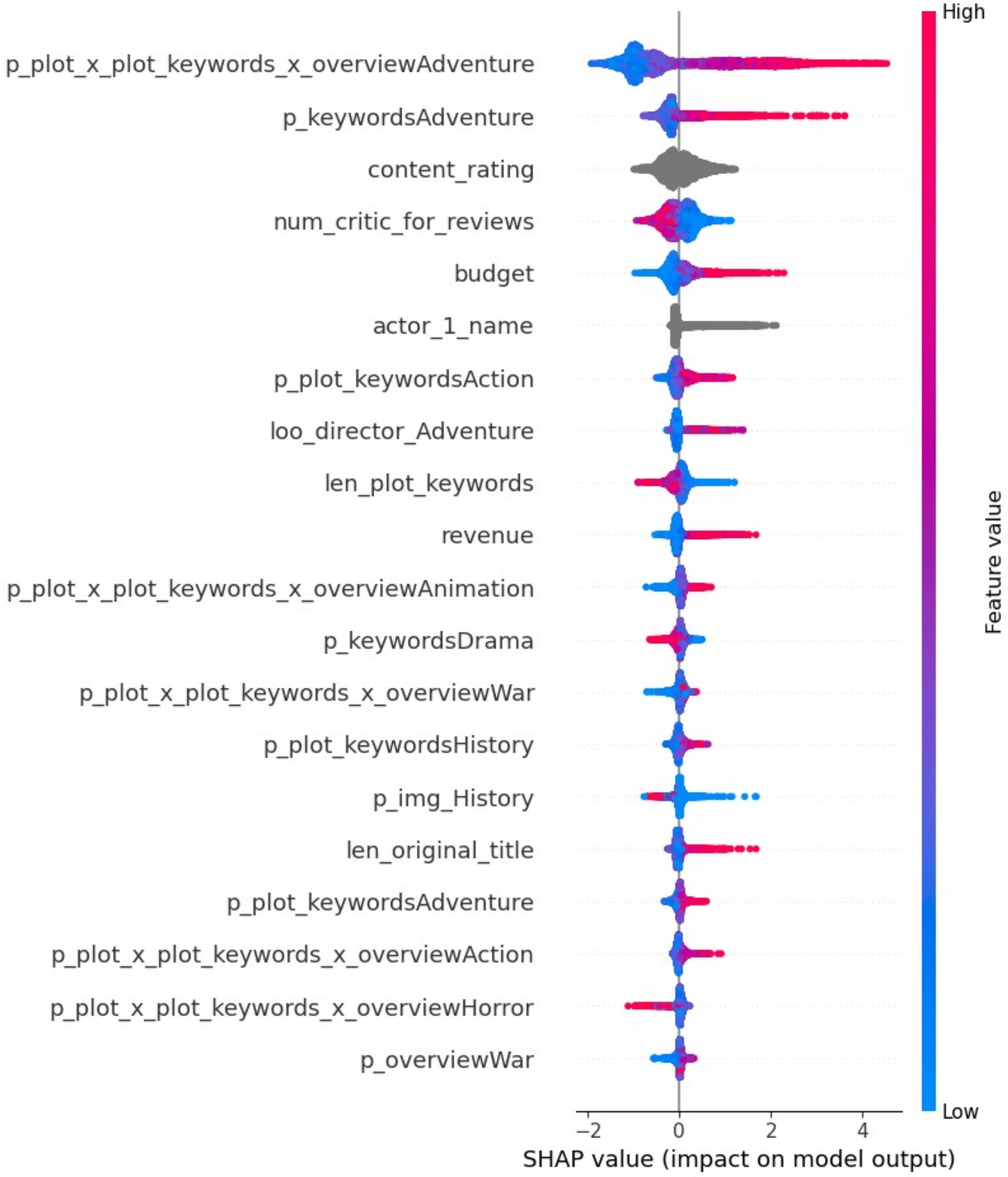
OPTUNA

Predicción de género

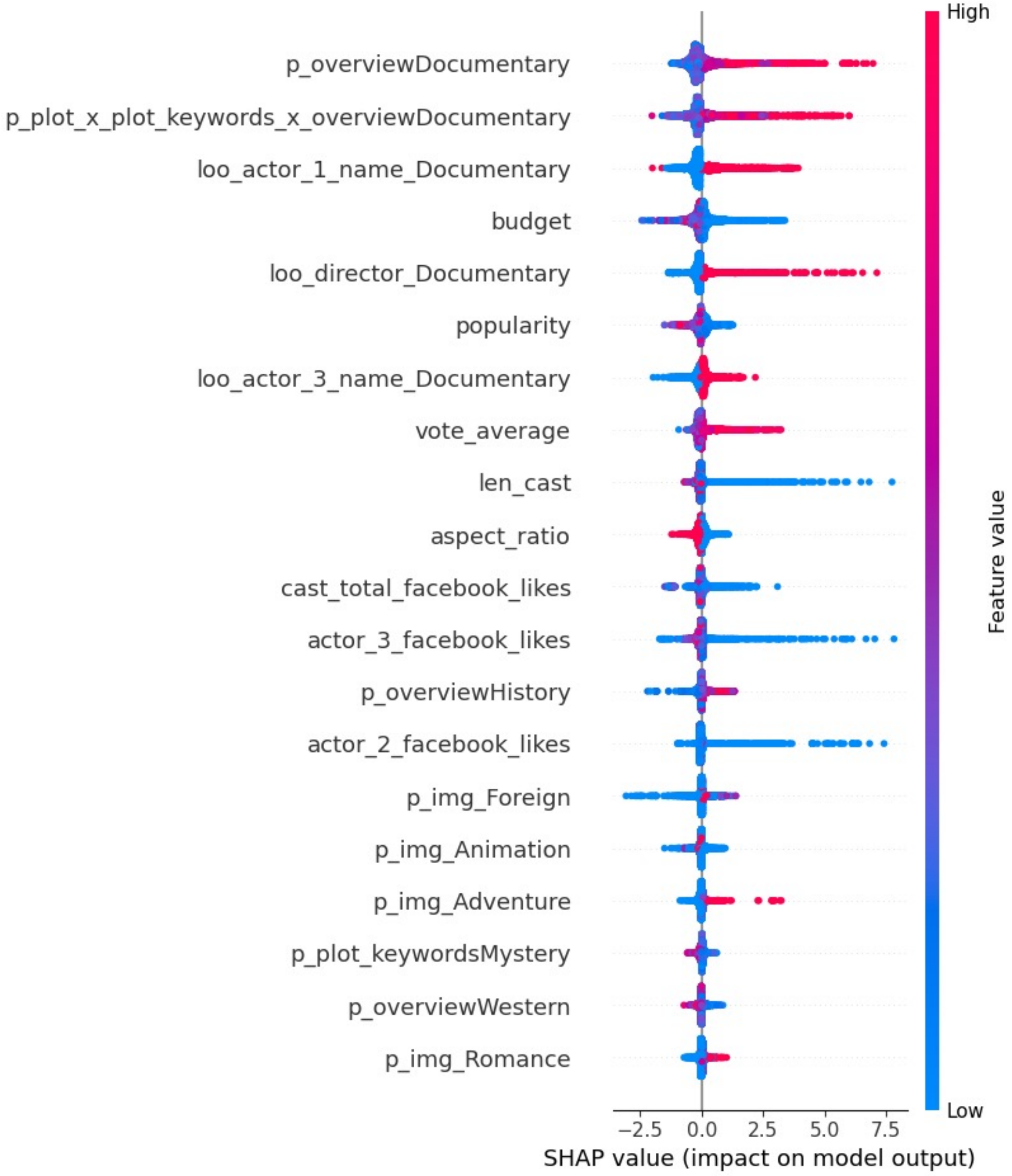
	CV	TEST
ACC	0.92	0.27

EL MODELO FINAL

Adventure

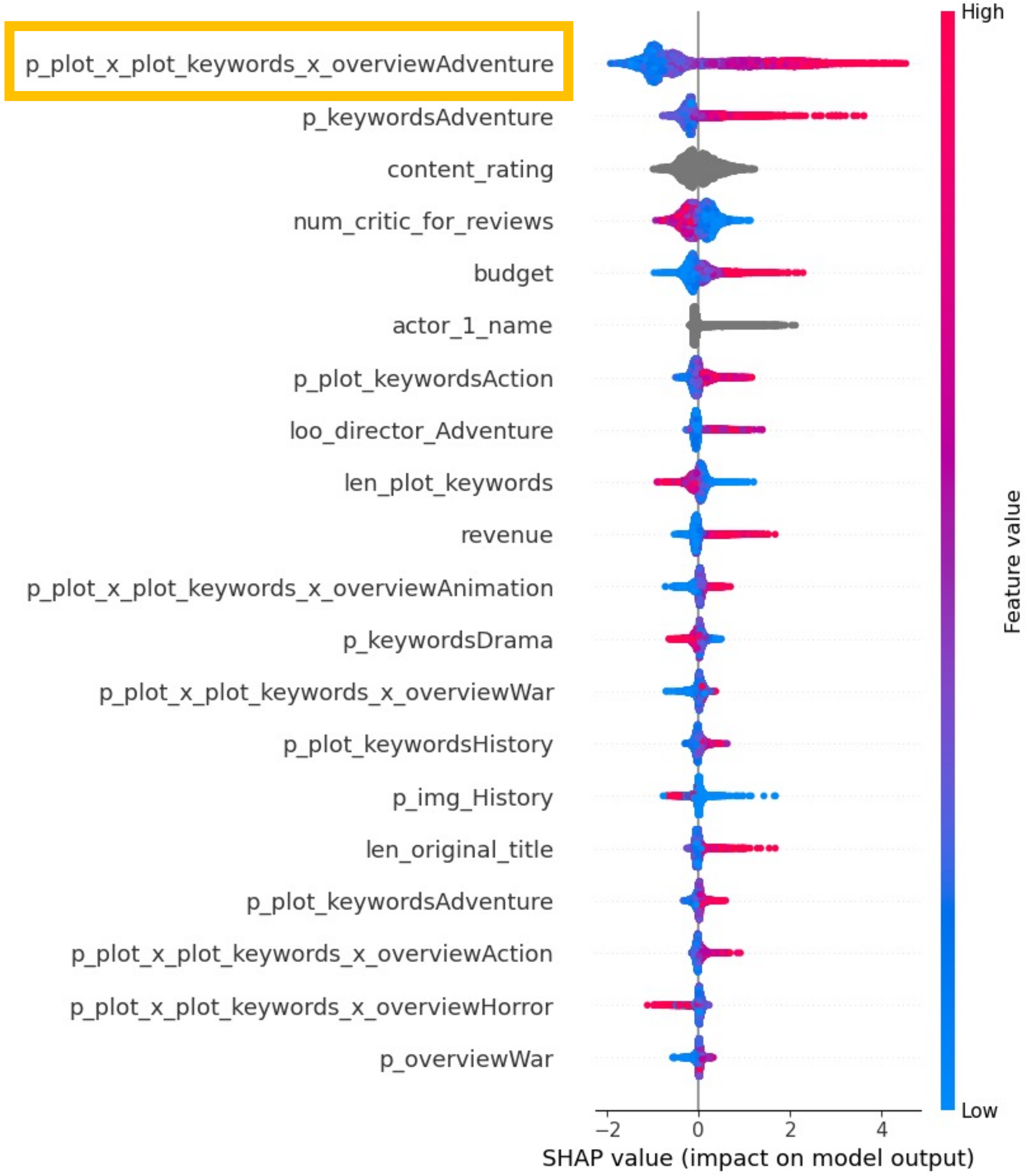


Documentary

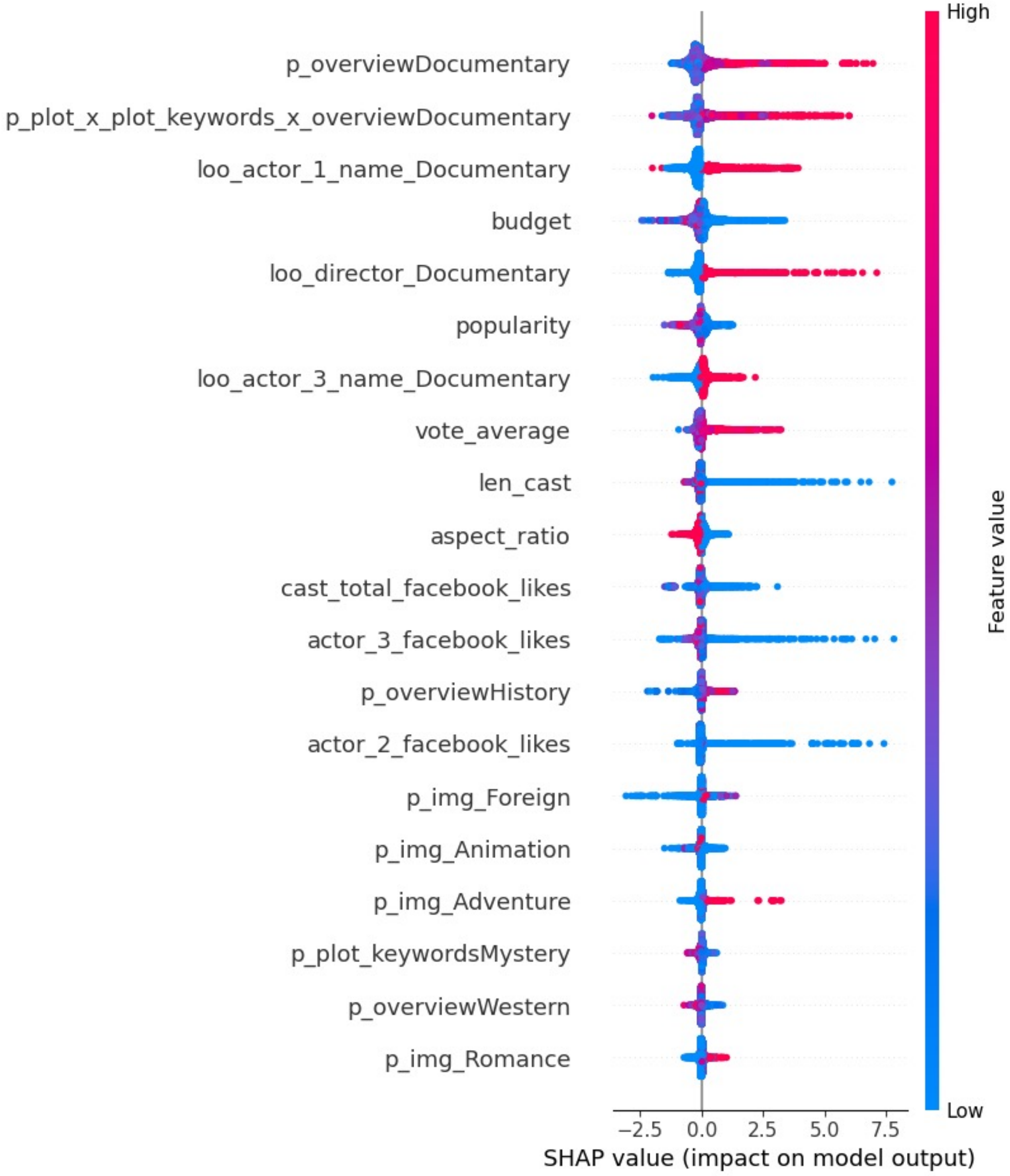


EL MODELO FINAL

Adventure

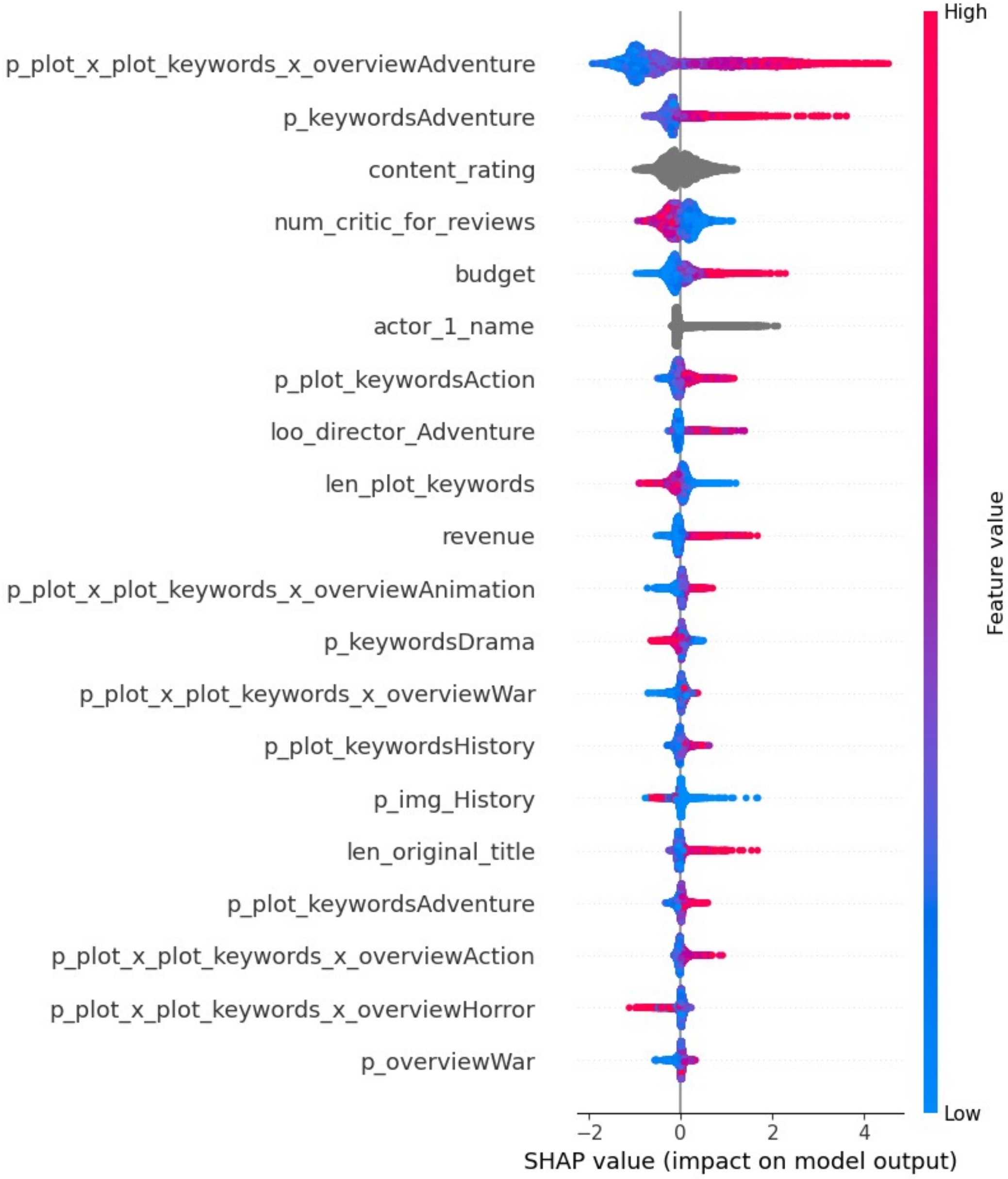


Documentary

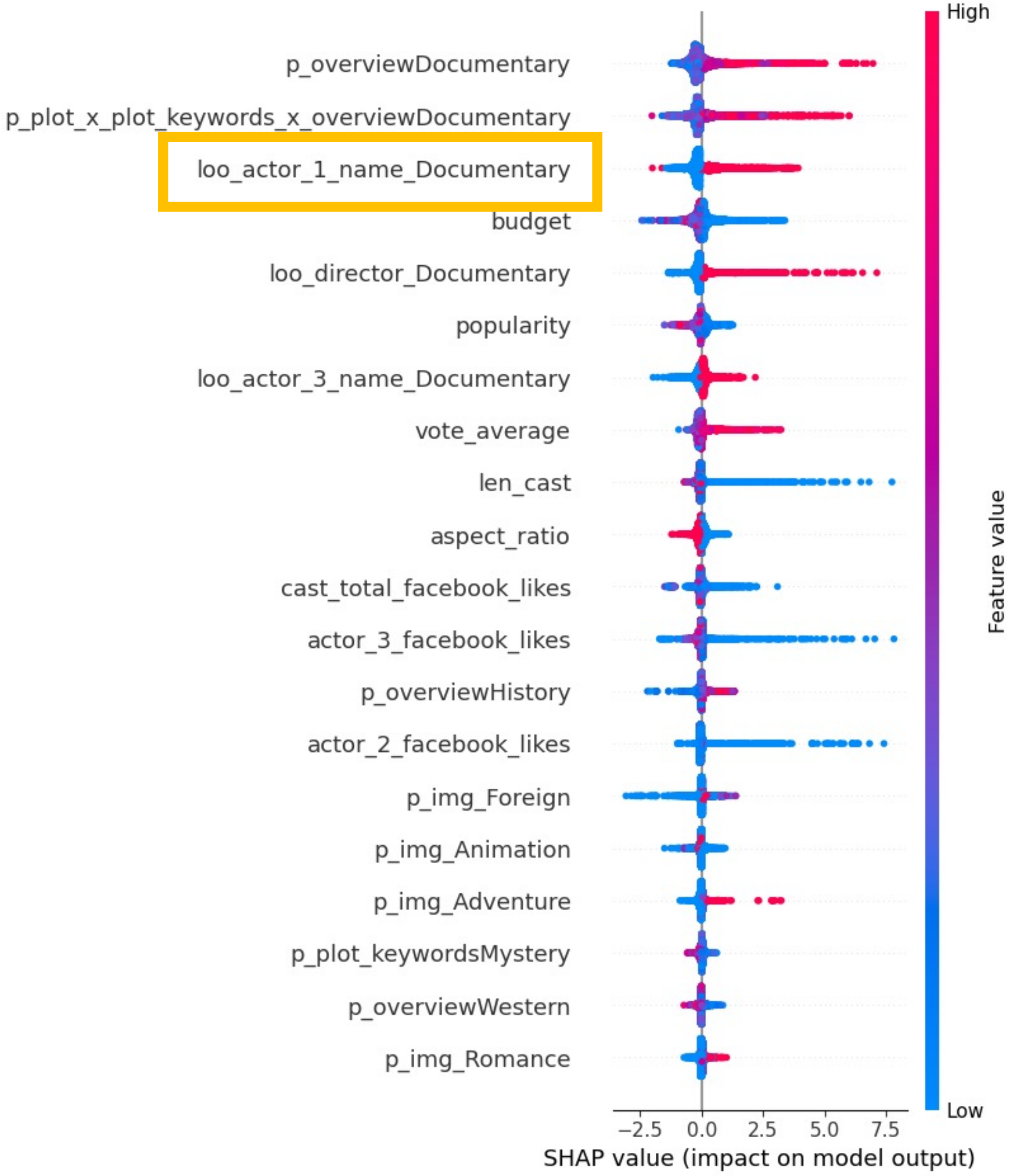


EL MODELO FINAL

Adventure

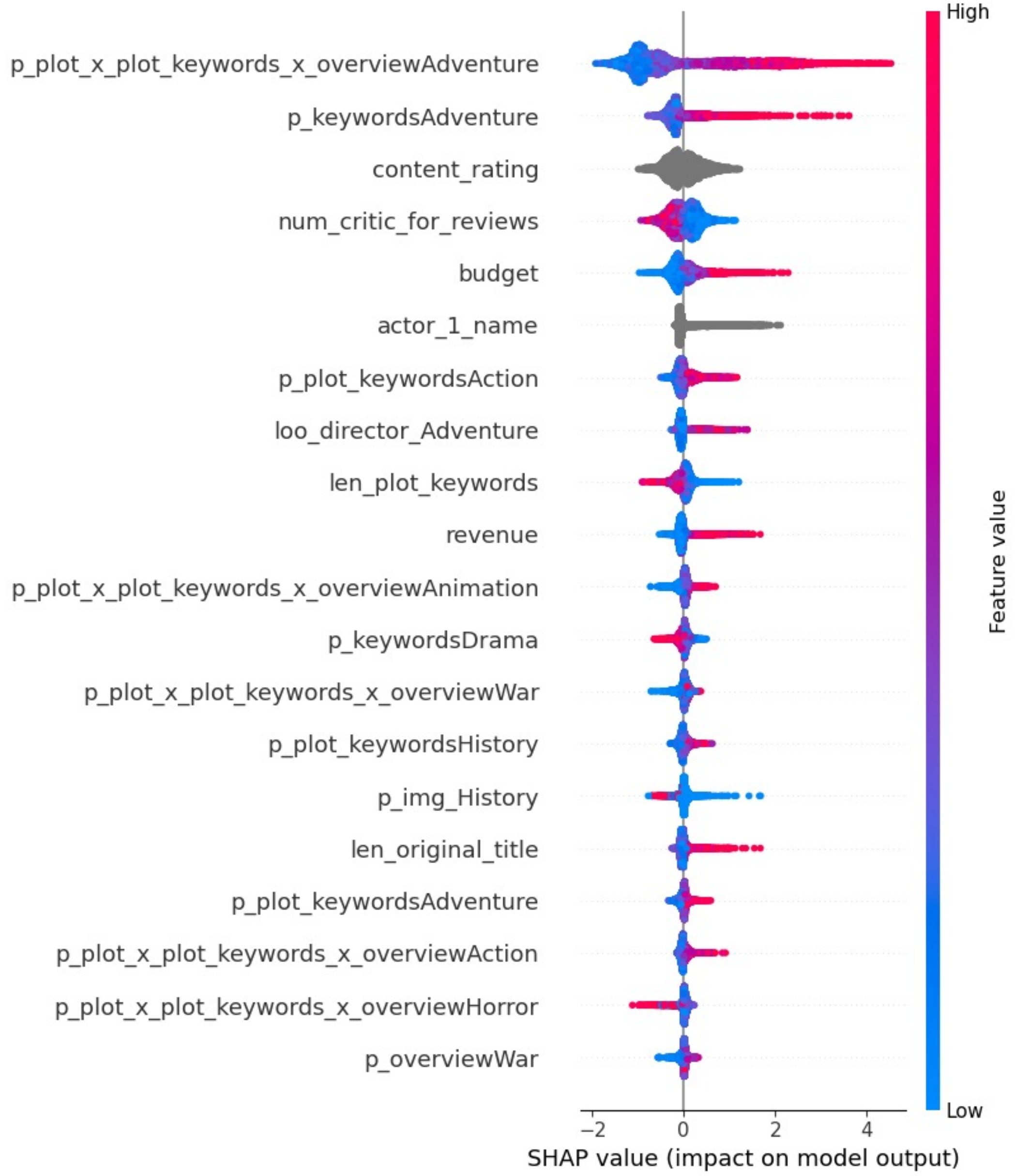


Documentary

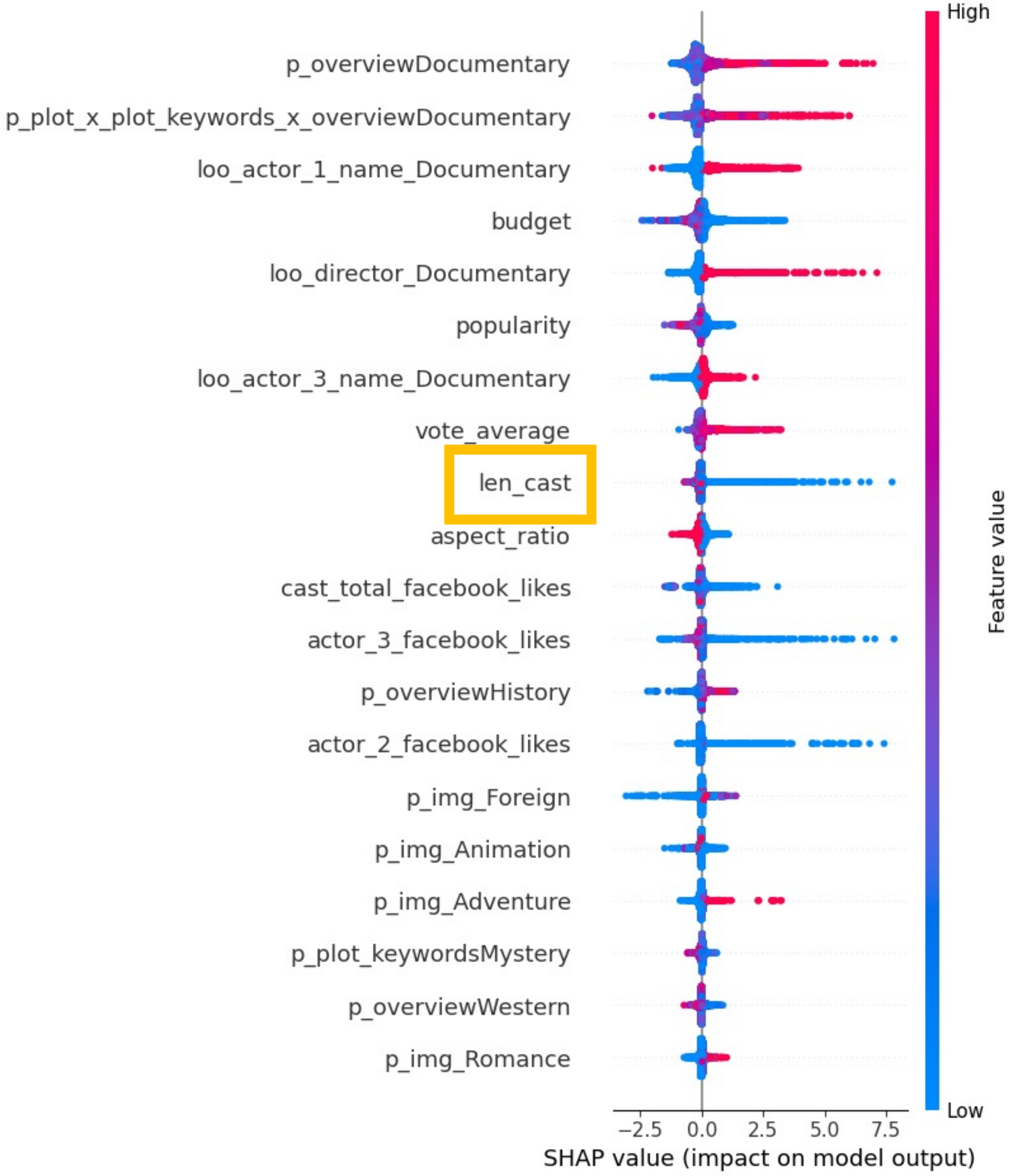


EL MODELO FINAL

Adventure

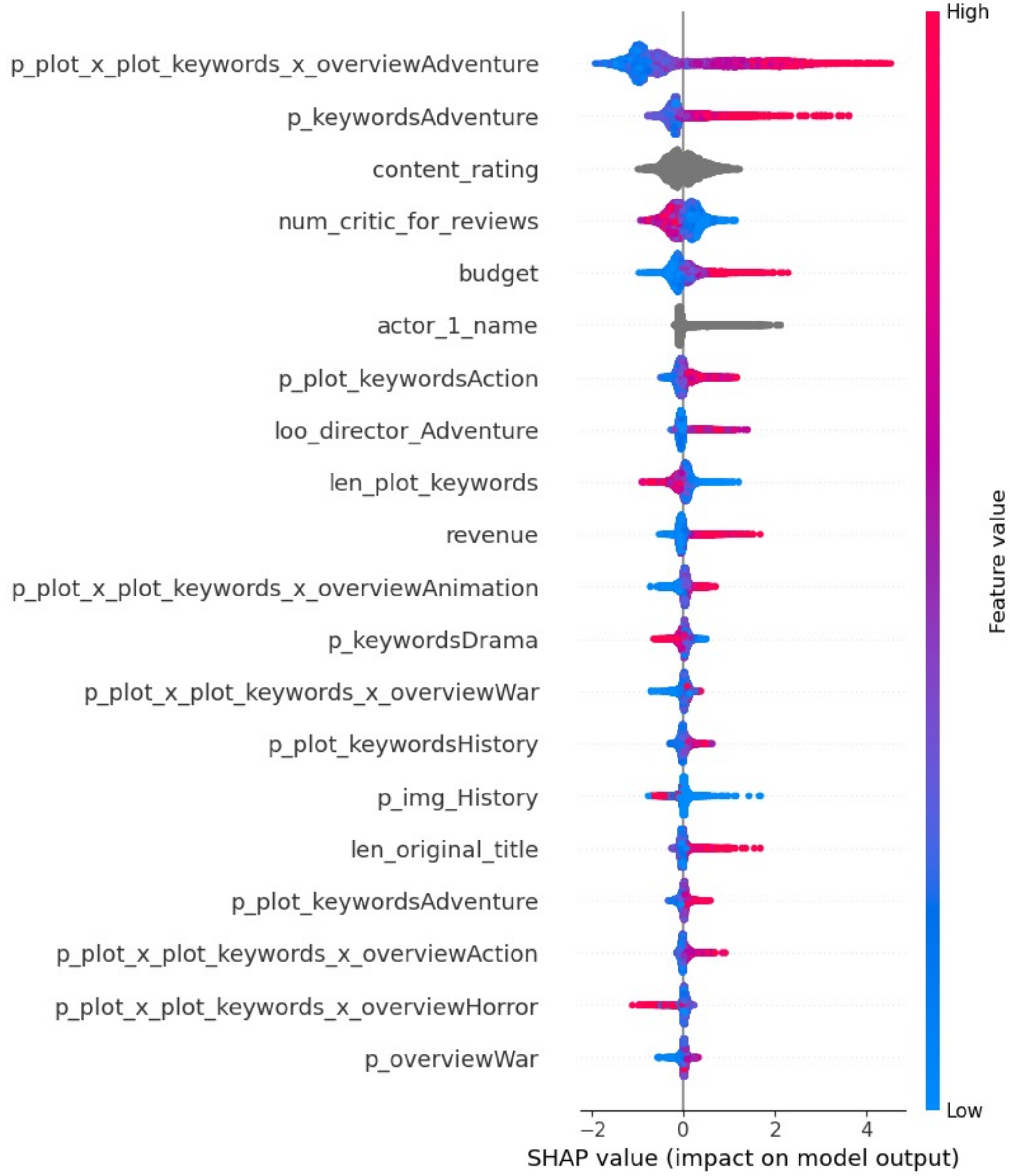


Documentary

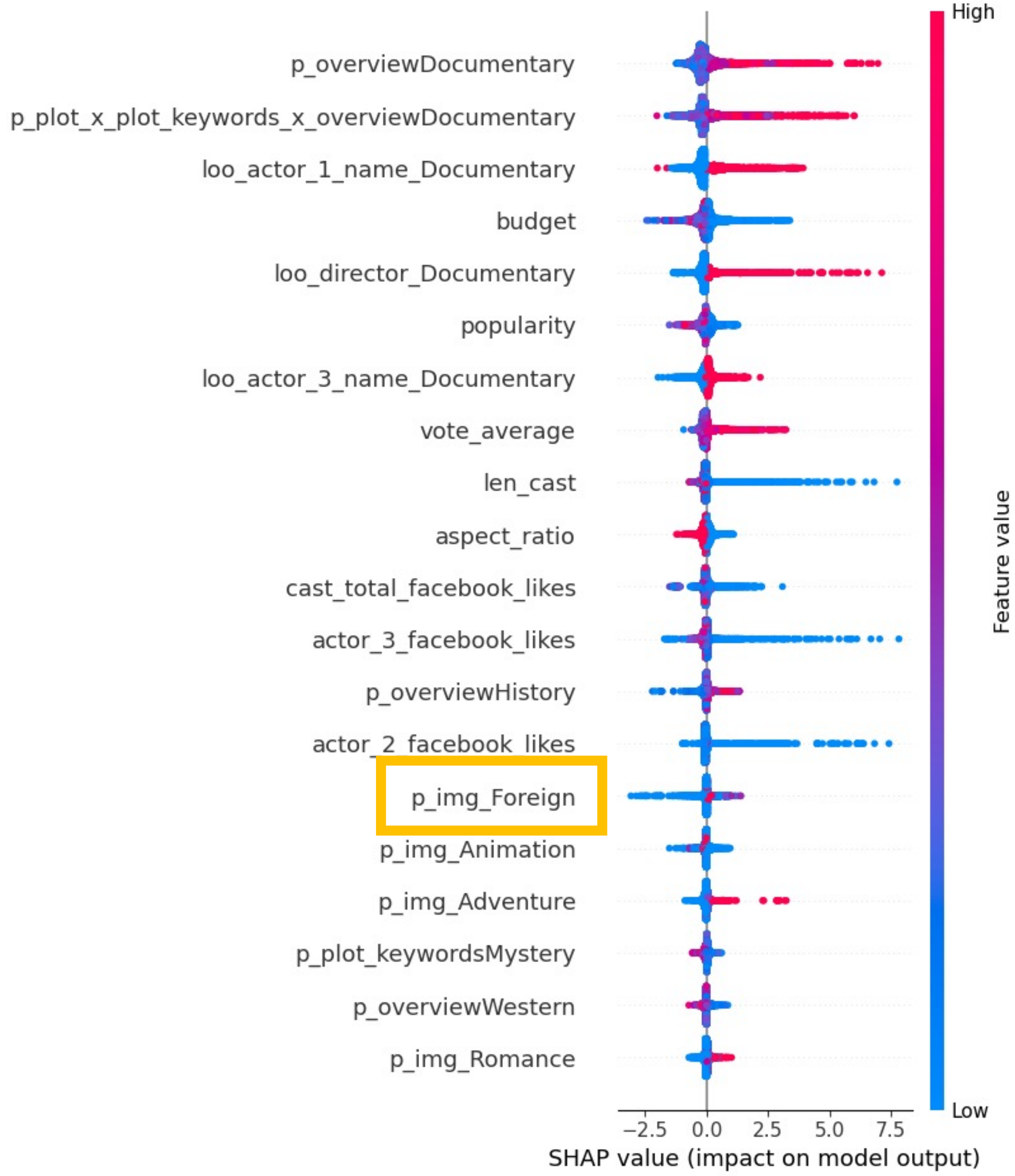


EL MODELO FINAL

Adventure

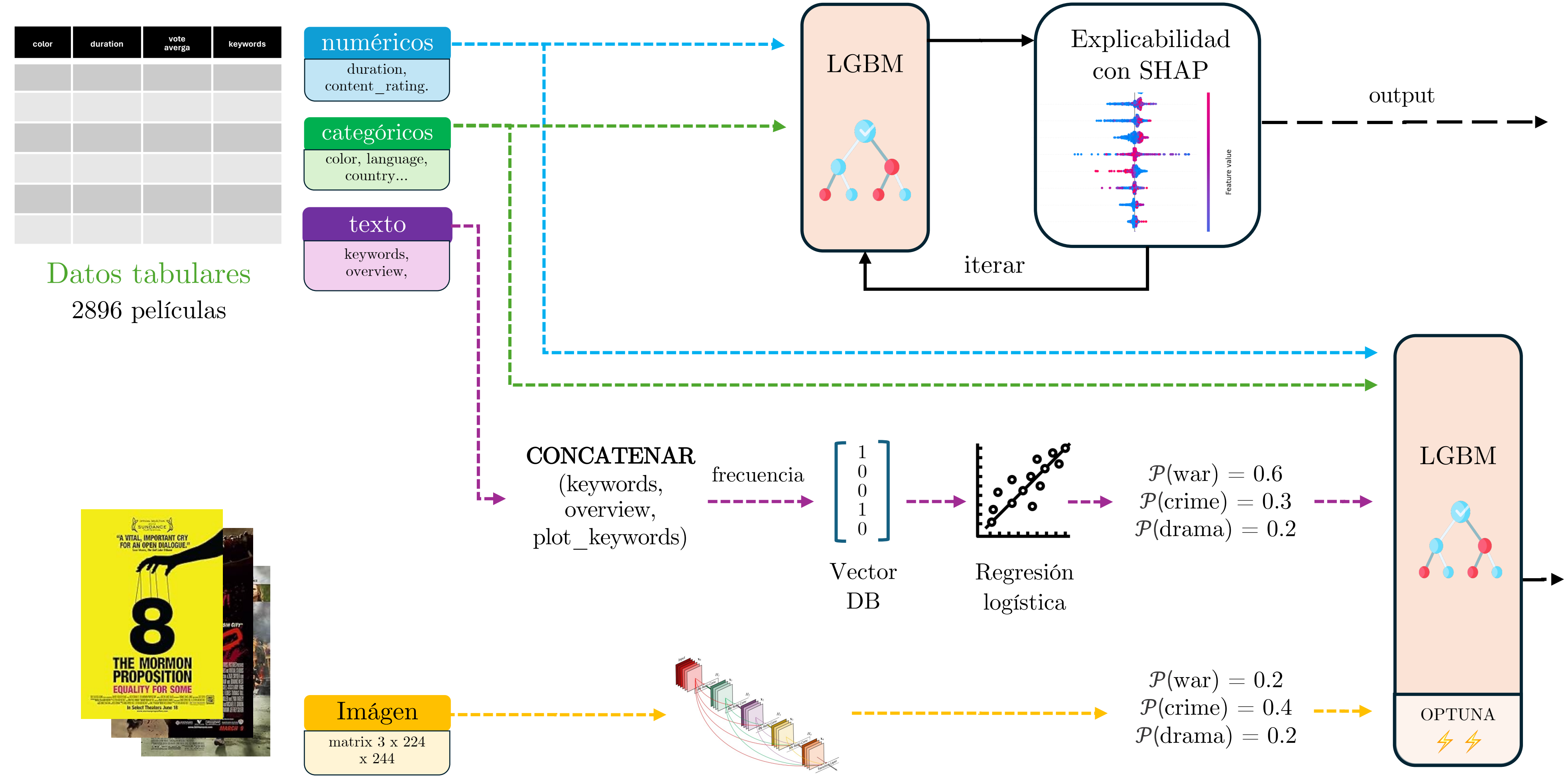


Documentary



CLASIFICACIÓN Y REGRESIÓN EN DATASET DE PELÍCULAS

Datos | Procesado | Métricas



color
duration
vote averga
keywords
numéricos
duration, content_rating.
categóricos
color, language, country...
texto
keywords, overview,
LGBM
Explicabilidad con SHAP
Feature value
iterar
output
CONCATENAR (keywords, overview, plot_keywords)
frecuencia
1 0 0 1 0
Vector DB
Regresión logística
$\mathcal{P}(\text{war}) = 0.6$
$\mathcal{P}(\text{crime}) = 0.3$
$\mathcal{P}(\text{drama}) = 0.2$
LGBM
OPTUNA
Imágen
matrix 3 x 224 x 244
$\mathcal{P}(\text{war}) = 0.2$
$\mathcal{P}(\text{crime}) = 0.4$
$\mathcal{P}(\text{drama}) = 0.2$
"A VITAL, IMPORTANT CRY FOR AN OPEN DIALOGUE."
8
THE MORMON PROPOSITION
EQUALITY FOR SOME

Datos tabulares
2896 películas

Pósteres
2894 x .jpg

Pre-trained resnet + finetuning

Predicción de score

	CV	TEST
MAE	**0.24**	**0.22**
MSE	**0.13**	**0.09**

Predicción de género

	CV	TEST
ACC	0.92	0.27