



TP2

Algoritmos de Clasificación Supervisada

Grupo 3:

**Tomas Marengo
Santiago Rivas
Franco De Simone
Gastón Francois**

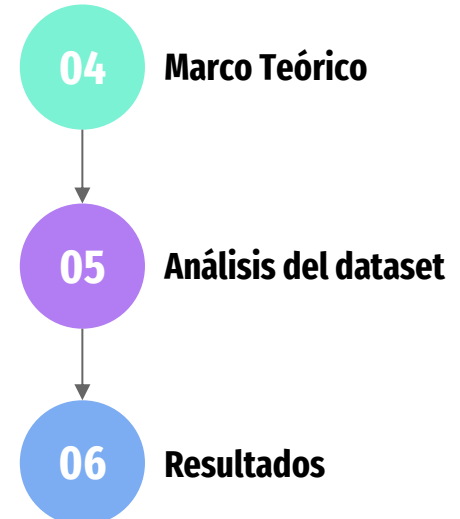


Índice

ID3 & Random Forest



KNN



Ejercicio 1: Devolución de Crédito

ID3 y Random Forest

Marco Teórico

Métricas y Matriz de Confusión

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precisión = \frac{TP}{TP + FP}$$

$$F1 - Score = \frac{2 \cdot Precisión \cdot Recall}{Precisión + Recall}$$

$$Tasa\ de\ Falsos\ Positivos = \frac{FP}{FP + TN}$$

$$Recall\ (Tasa\ de\ Verdaderos\ Positivos) = \frac{TP}{TP + FN}$$

Matriz de Confusión:

	Predicción Positiva	Predicción Negativa
Real Positivo	TP	FN
Real Negativo	FP	TN

Marco teórico

ID3 & Entropía de Shannon

Función información de ganancia: Para cada atributo se calcula la función ganancia y el de máximo valor es el nodo que sigue en el árbol. A **mayor ganancia, mejor** es el atributo para **separar** clases.

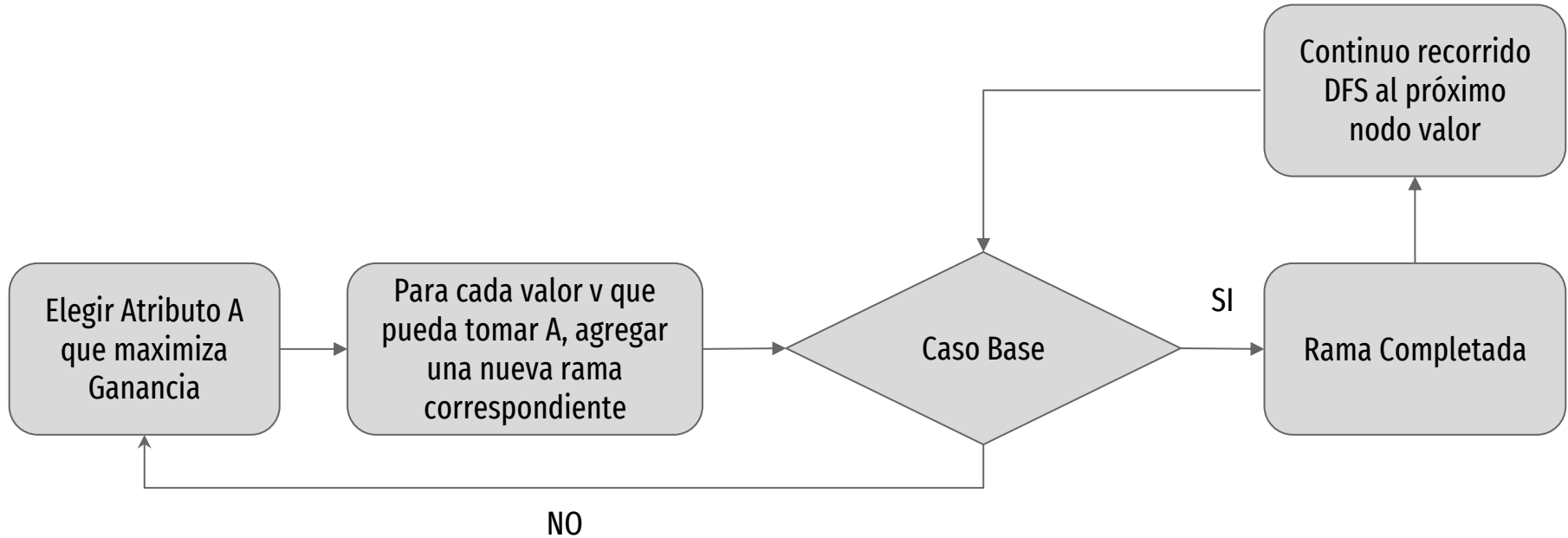
La **Entropía de Shannon** es una función típica usada en ganancia. Esta es utilizada para medir el grado de (des)organización de una muestra.

$$H(X) = - \sum_{i=1}^n P(x_i) \log_b(P(x_i))$$

Con $b=2$ típicamente (b usado).

Marco teórico

ID3 & Entropía de Shannon



Marco teórico

ID3 & Entropía de Shannon

ID3 Casos Base:

- Caso Base 1:
 - Si todos los ejemplos pertenecen a la misma clase, devolver un árbol de un nodo raíz, con rótulo del valor de la clase.
- Case Base 2:
 - Si los atributos están vacíos, devolver un árbol de un único nodo raíz, con rótulo el valor más frecuente del Atributo Objetivo en los ejemplos.
- Caso Base 3:
 - No hay más atributos para clasificar.

Devolución de Crédito

Problemática

Determinar si una persona devolverá el crédito o no, utilizando todas las variables del dataset.

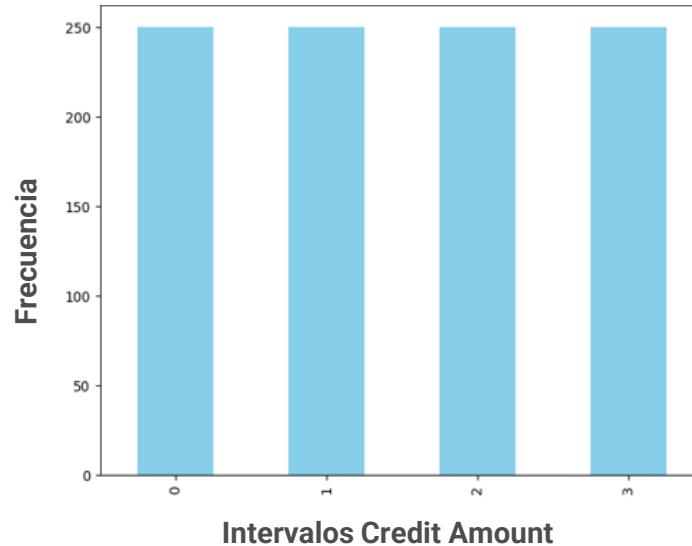
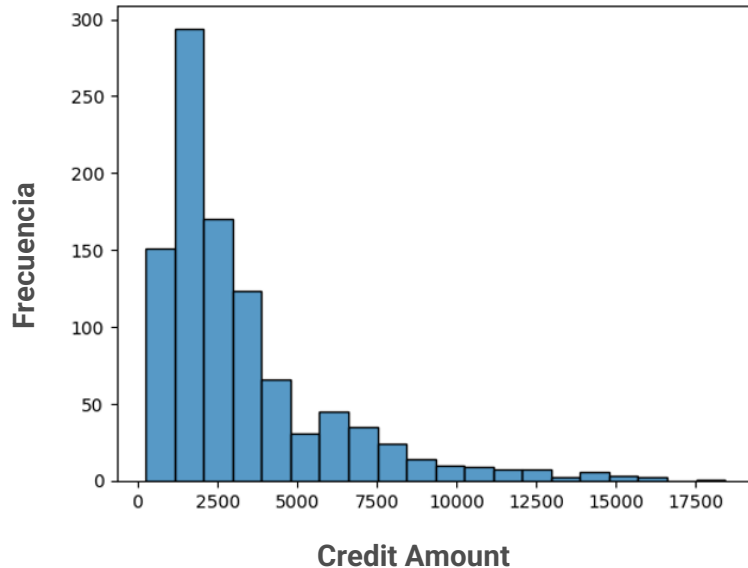
- **Dataset total:** 1000 muestras
 - Creditability: si devolvió el crédito (1) o no (0). **Variable objetivo.**
 - Diferentes **características** como Account Balance, Purpose, Credit Amount, etc.
- **División del dataset:** 80% Entrenamiento - 20% Testeo
 - Bootstrapping del conjunto de entrenamiento para Random Forest.

Devolución de Crédito

Análisis del Dataset

Se buscó discretizar las variables continuas, y reducir la cantidad de categorías para aquellas variables discretas que tuvieran muchos valores.

Credit Amount:



0 - (249.999, 1365.6]

1 - (1365.5, 2319.5]

2 - (2319.5, 3972.25]

3 - (3972.25, 18424.0]

Devolución de Crédito

Análisis del Dataset

Duration of Credit (month):

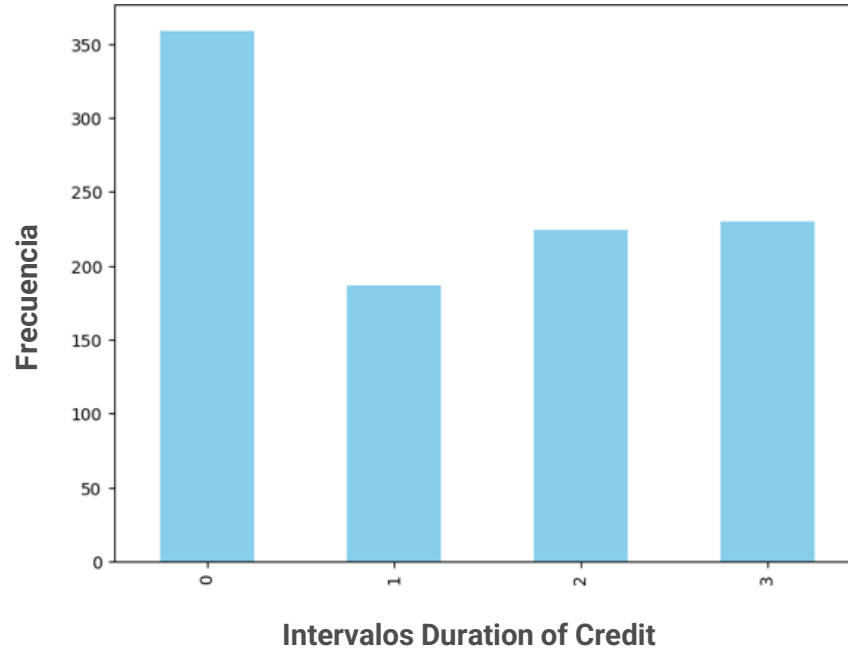
Separamos en cuatro intervalos:

0 - (3.999, 12]

1 - (12.0, 18.0]

2 - (18.0, 24.0]

3- (24.0, 72.0]



Devolución de Crédito

Análisis del Dataset

Age (years):

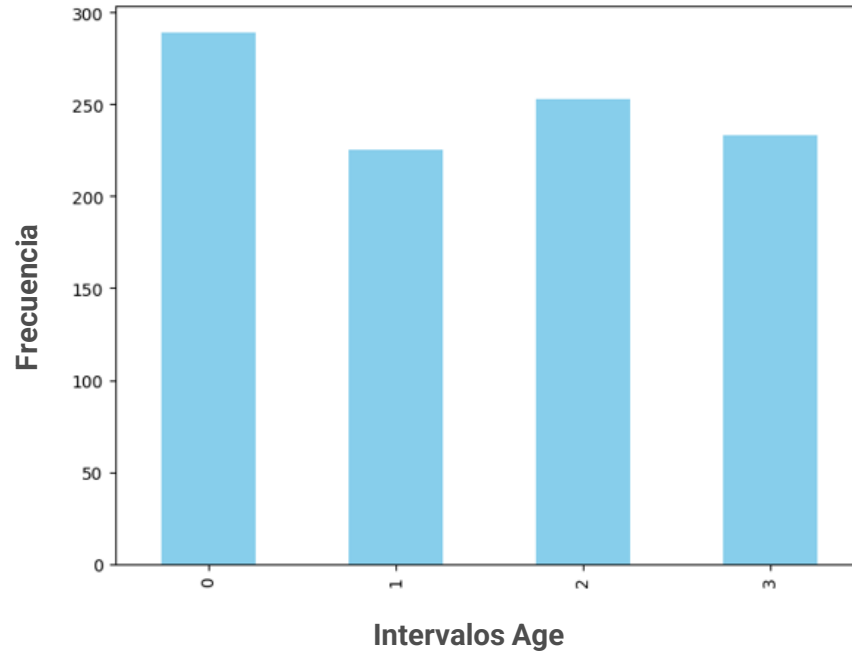
Separamos en cuatro intervalos:

0 - (18.999, 27.0]

1 - (27.0, 33.0]

2 - (33.0, 42.0]

3- (42.0, 75.0]

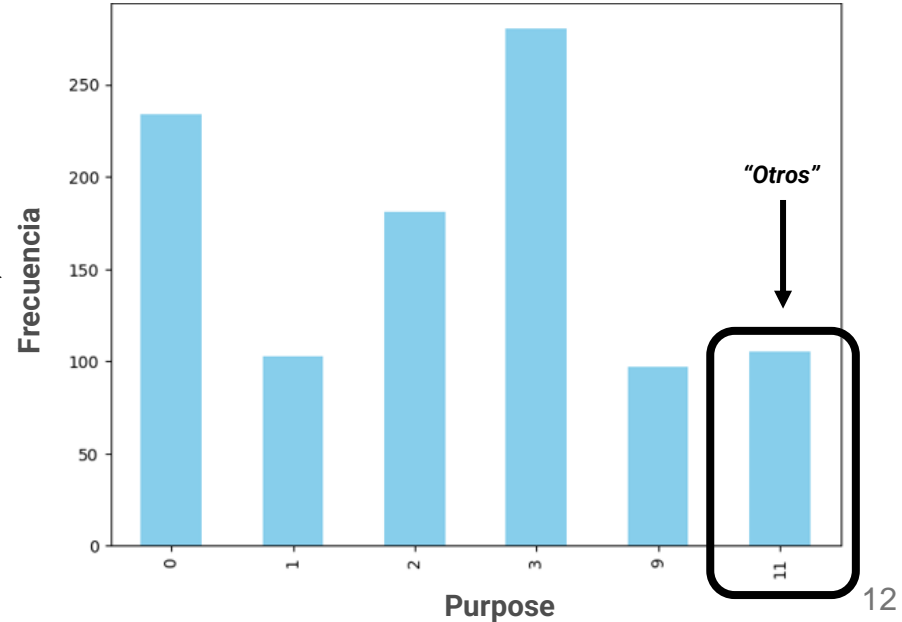
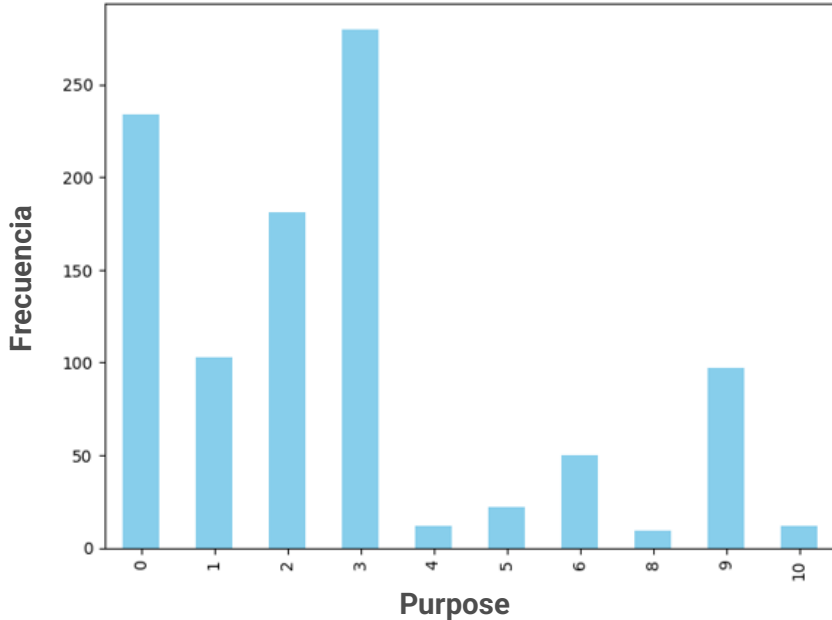


Devolución de Crédito

Análisis del Dataset

Purpose:

Agrupamos las categorías que tuvieran un conteo menor al 5% del total de registro. Las categorías **4, 5, 6, 7, 8 y 10** pasaron a ser agrupadas en **11 ("Otros")**.



Devolución de Crédito

ID3 Resultados: Matriz de Confusión

		Predicción	
Real		Positiva	Negativa
	Positivo	53,5% (107)	16,5% (33)
	Negativo	17,5% (35)	12,5% (25)

Características:

- Account Balance
- Duration of Credit Month
- Purpose
- Credit Amount
- Age
- Etc.

Accuracy: 0,660

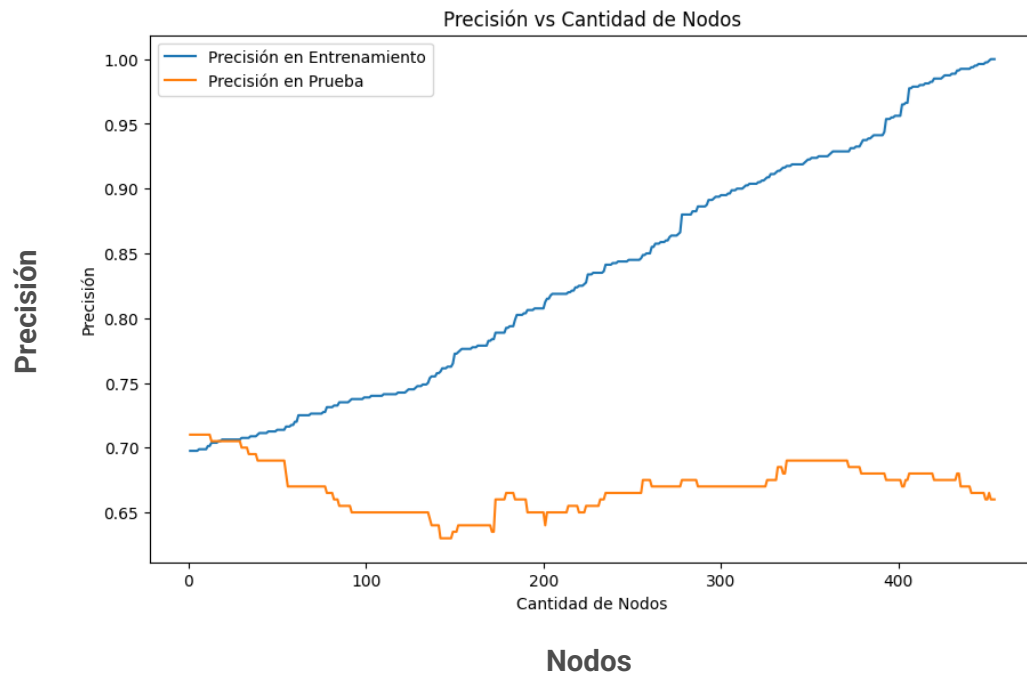
Recall: 0,754

Precision: 0,764

F1-Score: 0,759

Devolución de Crédito

ID3: Precisión vs Nodos



Marco teórico (IV)

Bootstrapping & Random Forest

Método Bootstrapping:

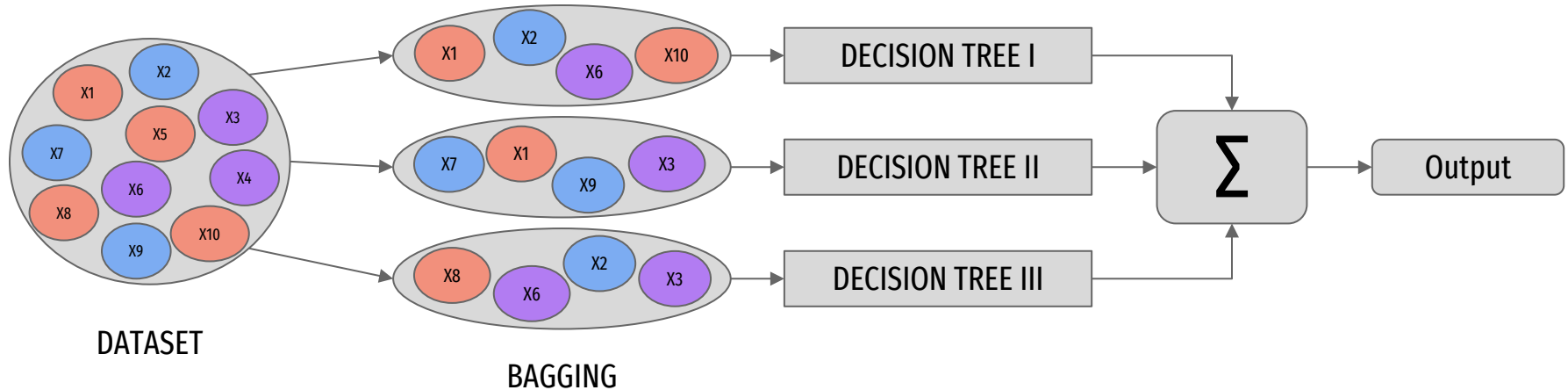
1. Extraer una muestra del tamaño del conjunto de datos original con reemplazo (Esto es 1 bag).
2. Repetir el paso 1 S veces, para que tengamos S bags.
3. Calcular nuestro valor en cada uno de los bags, de modo que tengamos S estimaciones.
4. Utilizar la distribución de estimaciones para realizar inferencias.

Marco teórico (IV)

Bootstrapping & Random Forest

Idea en Random Forest:

- Crear S modelos de "ID3" usando S muestras.
- Combinar resultados tomando la clasificación con más votos.



Devolución de Crédito

Random Forest: Matriz de Confusión

		Predicción	
		Positiva	Negativa
Real	Positivo	55% (110)	16% (32)
	Negativo	16,5% (33)	12,5% (25)

Parámetros usados:

- Cantidad de árboles: 10
- Min Features: 4
- Max Features: 15
- Bootstrap: false

Accuracy: 0,675

Recall: 0,775

Precision: 0,769

F1-Score: 0,772

Devolución de Crédito

Random Forest: Matriz de Confusión

		Predicción	
Real		Positiva	Negativa
	Positivo	54% (108)	17% (34)
	Negativo	12,5% (25)	16,5% (33)

Parámetros usados:

- Cantidad de árboles: 10
- Min Features: 4
- Max Features: 15
- Bootstrap: true

Accuracy: 0,705

Recall: 0,761

Precision: 0,812

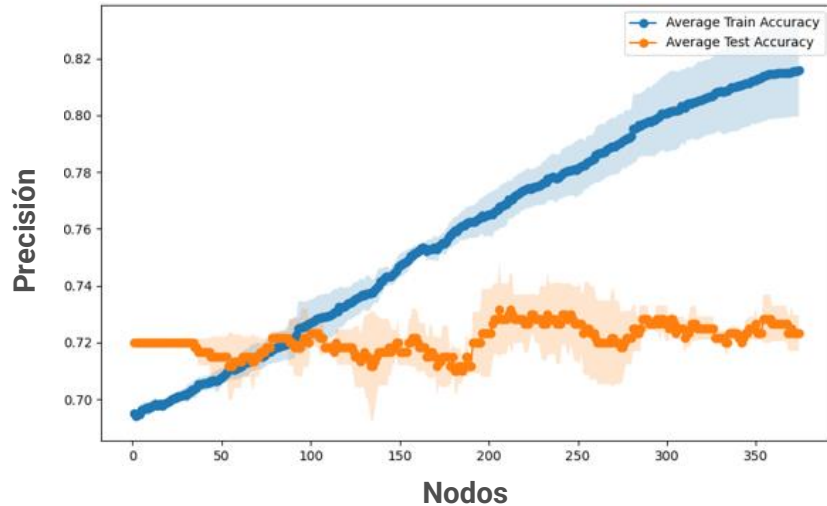
F1-Score: 0,786

Devolución de Crédito

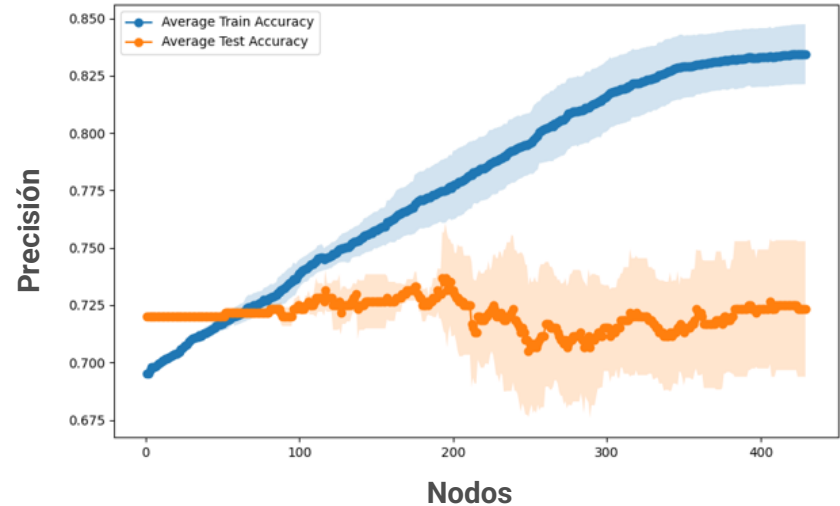
Random Forest: Comparación por parámetros - Cantidad de árboles

Para todos los casos se tomó `min_features = 4`.

5 Árboles



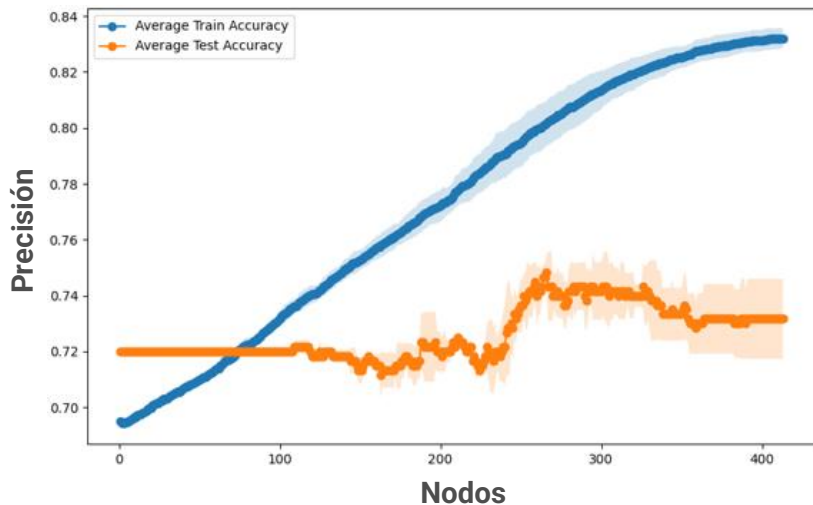
10 Árboles



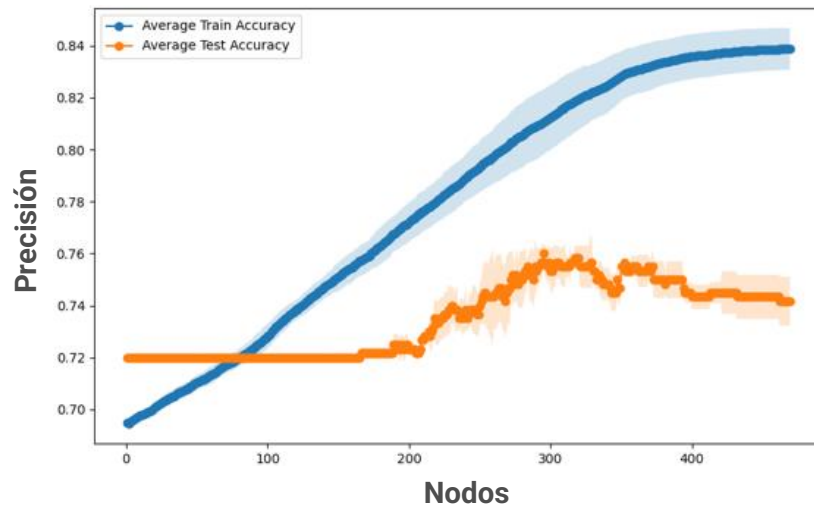
Devolución de Crédito

Random Forest: Comparación por parámetros - cantidad de árboles

20 Árboles

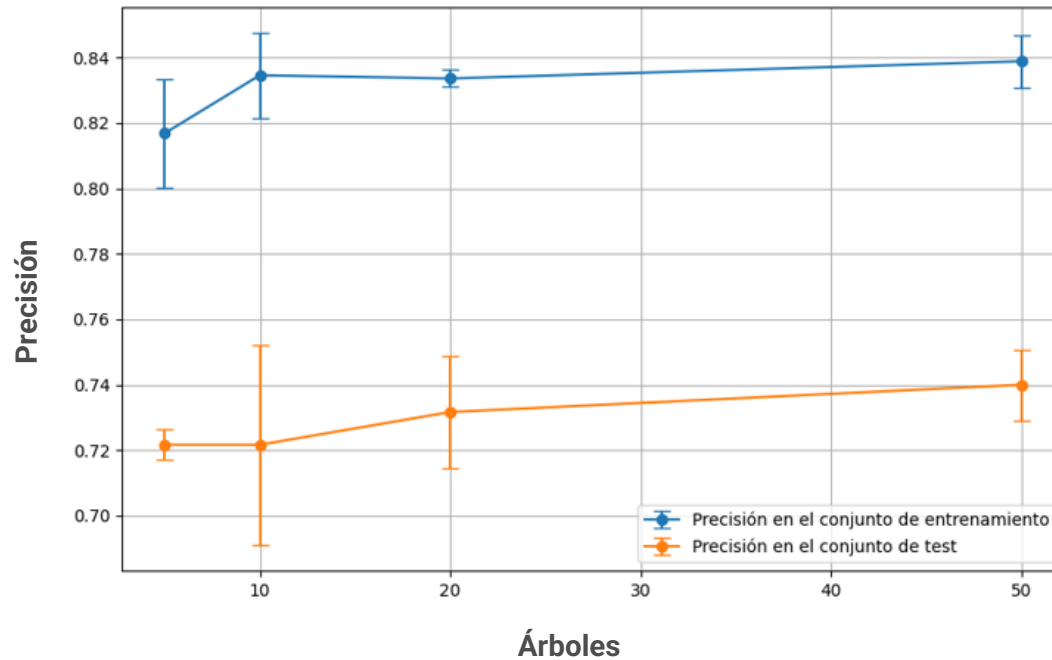


50 Árboles



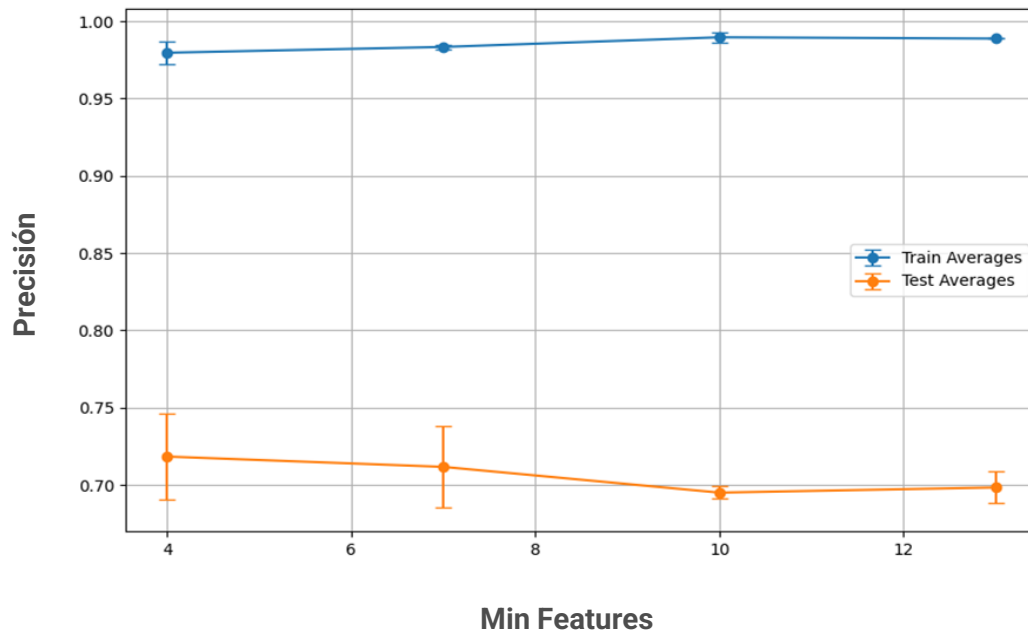
Devolución de Crédito

Random Forest: Comparación por parámetros



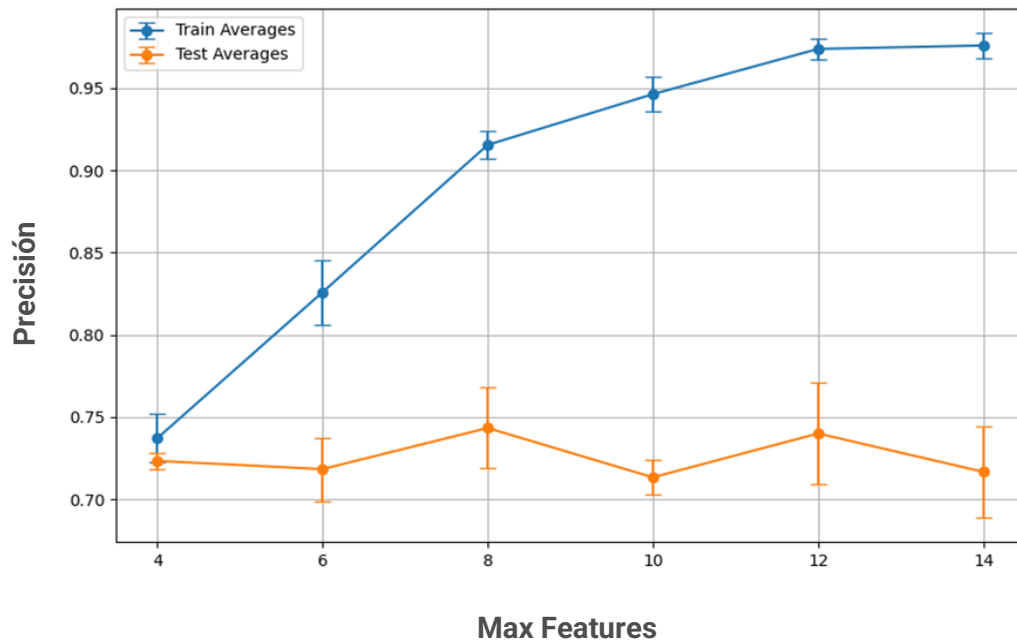
Devolución de Crédito

Random Forest: Comparación por parámetros



Devolución de Crédito

Random Forest: Comparación por parámetros



Ejercicio 2: Sentimiento de Opiniones

KNN

Marco Teórico

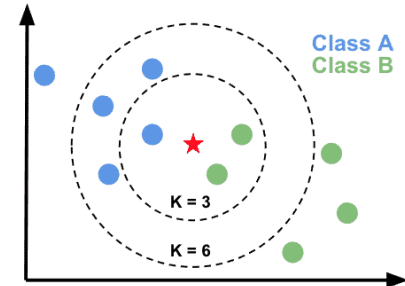
KNN

El método KNN consiste en clasificar una muestra determinando un número de vecinos a partir de alguna métrica de distancia y tomando la clase con más vecinos. Matemáticamente:

$$\hat{f}(x_q) = \underset{v \in V}{\operatorname{arg\,m\acute{a}x}} \sum_{i=1}^k 1_{\{v=f(x_i)\}}$$

Es decir, para cada clase V se suma 1 por cada vecino perteneciente a la clase y luego se elige a la clase V que mayor sumatoria posea.

Ejemplo: tomando distancia euclídea entre puntos, si se utiliza K=3 el punto sería **clase B** (2 vecinos B y 1 Vecino A) y si se utiliza K=6 el punto sería **clase A** (4 vecinos A y 2 vecinos B).



Marco Teórico

Weighted KNN

Uno de los problemas notables de KNN es que no tiene en cuenta la distancia entre los puntos. Se puede considerar estas distancias y “mejorar” el método al agregar a la sumatoria el inverso al cuadrado de la distancia.

$$\hat{f}(x_q) = \arg \underbrace{\max}_{v \in V} \sum_{i=1}^k \frac{1}{d(x_q, x_i)^2}_{\{v=f(x_i)\}}$$

Como caso particular, si con otra muestra $d = 0$, entonces se directamente se replica la clase.

En cuanto a la función de distancia, se suele usar la distancia euclídea, aunque esto depende del problema.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^m (a_r(x_i) - a_r(x_j))^2}$$

Sentimiento de Opiniones

Problemática

El objetivo del ejercicio recae en ***clasificar opiniones utilizando como variable objetivo Star Rating y como variables explicativas Word Count, Title Sentiment y Sentiment Value mediante Método KNN y Método Weighted KNN.***

- **Dataset total:** 257 comentarios sobre el uso de una aplicación
- **Características:** Review Title, Review Text, Word Count, Title Sentiment, Text Sentiment, Star Rating y Sentiment Value.
- **Variación de parámetros:**
 - Diferentes K.
 - Diferente vector de características.
 - Diferentes tratamientos sobre el dataset.

Análisis de dataset

Review general

El dataset cuenta con 257 datos de reseñas estructuradas en 7 características.

N°	Review Title <i>Título de la reseña</i>	Review Text <i>Comentario de la reseña</i>	Word Count <i>Número de palabras del comentario</i>	Title Sentiment <i>valoración del título (positivo - negativo)</i>	Text Sentiment <i>valoración del comentario (positivo – negativo)</i>	Star Rating <i>Estrellas de la reseña (1 a 5)</i>	Sentiment Value <i>Indicador de positividad (-4 a 4)</i>
1	Sin conexión	Hola desde hace algo más...	23	negative	negative	1	-0.486389
2	Es muy buena lo recomiendo	Andres e puto amoooo	4	NaN	negative	1	-0.602240
...
258	Esta bien	Sin ser la biblia....	6	negative	negative	1	-0.651784

Análisis de dataset

Reemplazo datos faltantes

Un análisis sobre datos faltantes arroja 26 datos blancos (~10% del dataset) en Title Sentiment, por lo que surgen 3 posibles reemplazos

1. Eliminar los valores.

- Problema: eliminar un número considerable dentro del dataset.

2. Reemplazar con la moda.

- Problema: Etiqueta dependiente de las características del dataset y no de la reseña.

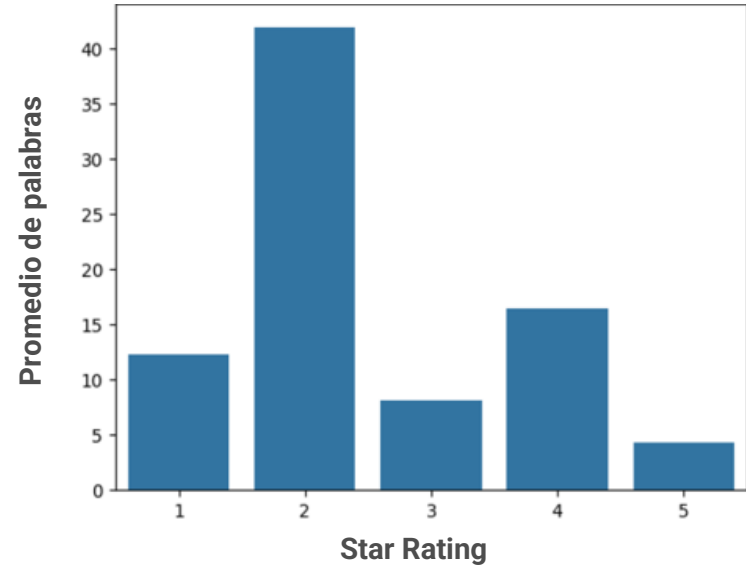
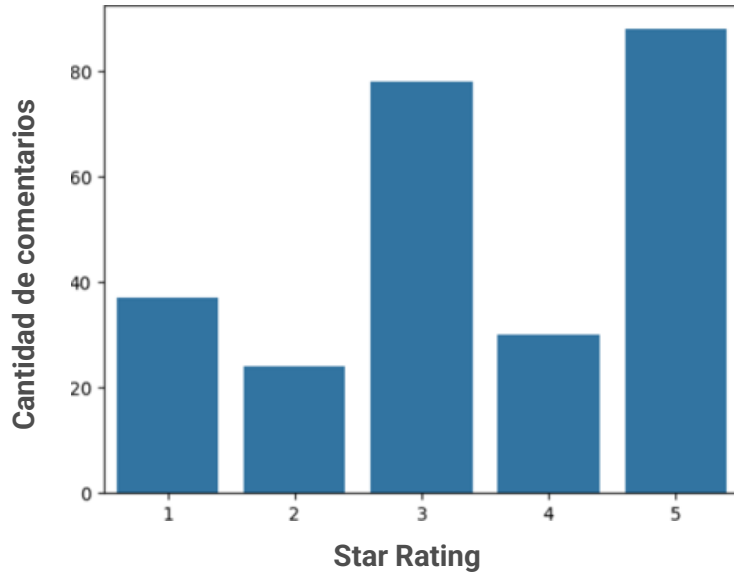
3. Reemplazar con el valor de Text Sentiment.

- En general el título y el comentario deben tener un hilo conductor y tener el mismo sentimiento. Pueden ser diferentes, pero 85% de las muestras poseen el mismo valor en Text Sentiment y Title Sentiment.

Análisis de dataset

Análisis de Star Rating

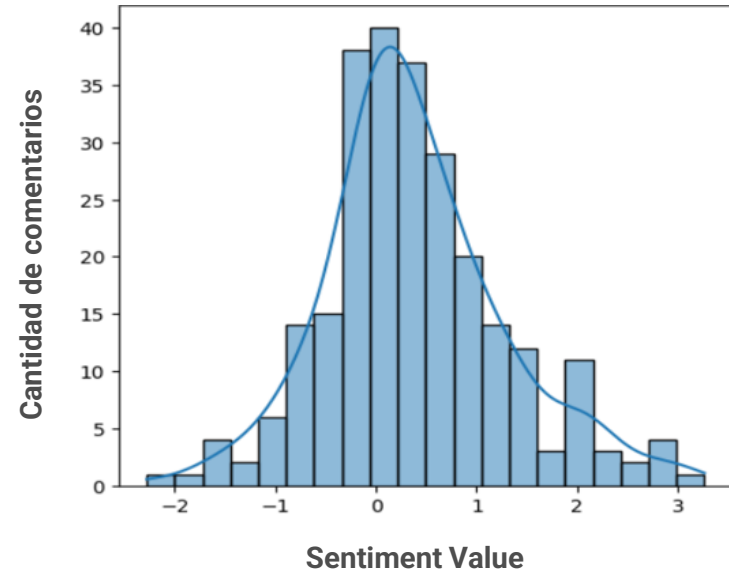
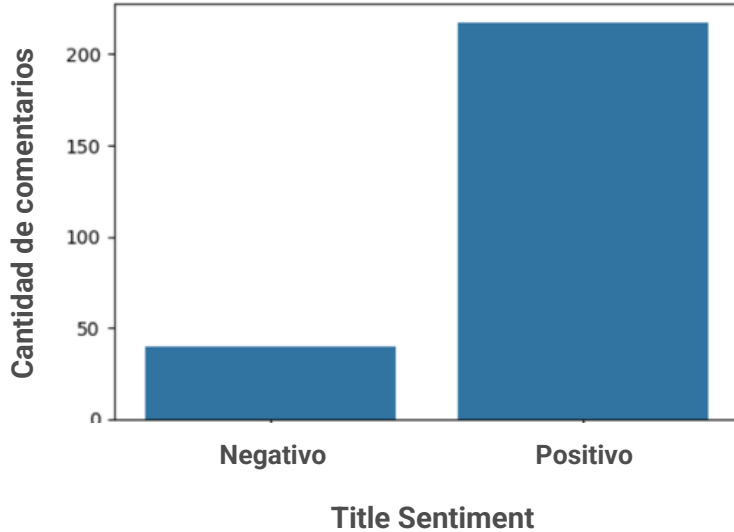
Analizamos la variable objetivo Star Rating con respecto a otras características.



Análisis de dataset

Análisis de características Sentiment

Notamos 84% reseñas positivas pero una distribución más uniforme en cuanto al valor general.



Método KNN

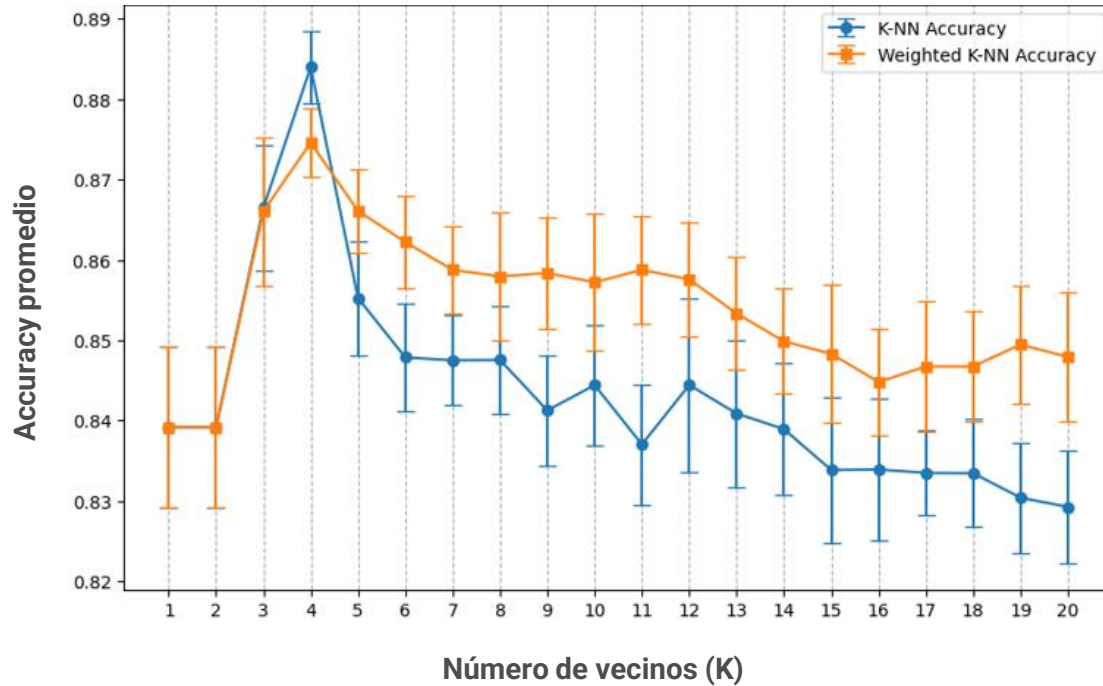
Tratamiento de datos

Para el correcto manejo en el método KNN se consideró:

- Por tener que trabajar con valores numéricos, se **mapea** “positive” a 1 y “negative” a 0.
- Por utilizar métodos donde la distancia entre features resulta importante, en un primer análisis se **estandarizan** los valores.
- Para la división en conjuntos de Train y Test y el testeo correspondiente, se utiliza **Cross Validation** de 10 conjuntos.
- Se realiza el método para diferentes valores de K, en busca del mejor.

Método KNN

Resultados



Características:

- Word Count
- Title Sentiment
- Sentiment Value

Mejores resultados:

KNN

K = 4

Accuracy = 0.884

Weighted KNN

K = 4

Accuracy = 0.875











Método KNN

Best K: Matrices de confusión

		Predicción				
		1 	2 	3 	4 	5 
Real	1 	0.89	0.027	0.08	0	0
	2 	0.12	0.79	0.083	0	0
	3 	0.064	0.013	0.86	0.013	0.051
	4 	0	0.067	0.067	0.73	0.13
	5 	0	0	0.045	0.035	0.92

KNN (Best K)

- Precision: 0.8616
- Recall: 0.8599
- F1 Score: 0.8598

		Predicción				
		1 	2 	3 	4 	5 
Real	1 	0.89	0.027	0.081	0	0
	2 	0.12	0.75	0.12	0	0
	3 	0.051	0.013	0.86	0.013	0.064
	4 	0	0.033	0.067	0.60	0.30
	5 	0	0	0.057	0.034	0.91

Weighted KNN (Best K)

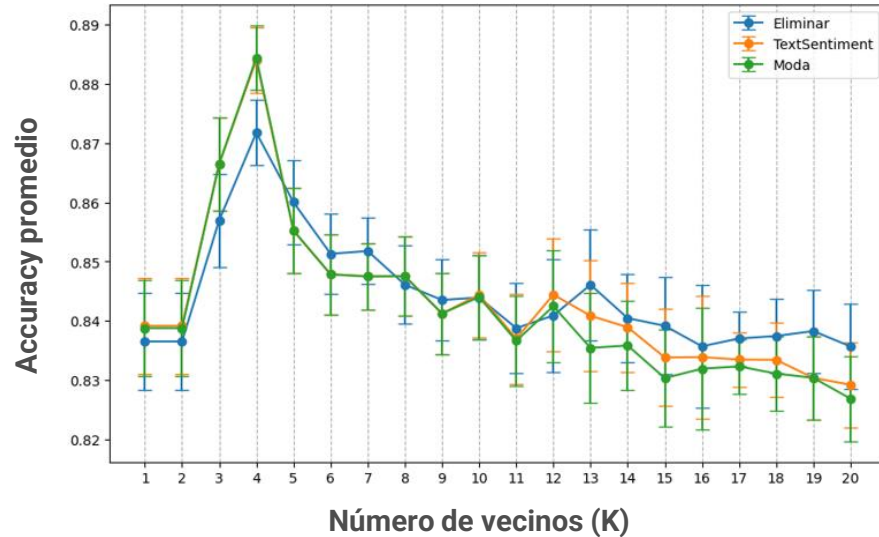
- Precisión: 0.8575
- Recall: 0.8560
- F1 Score: 0.8561

Variantes al método

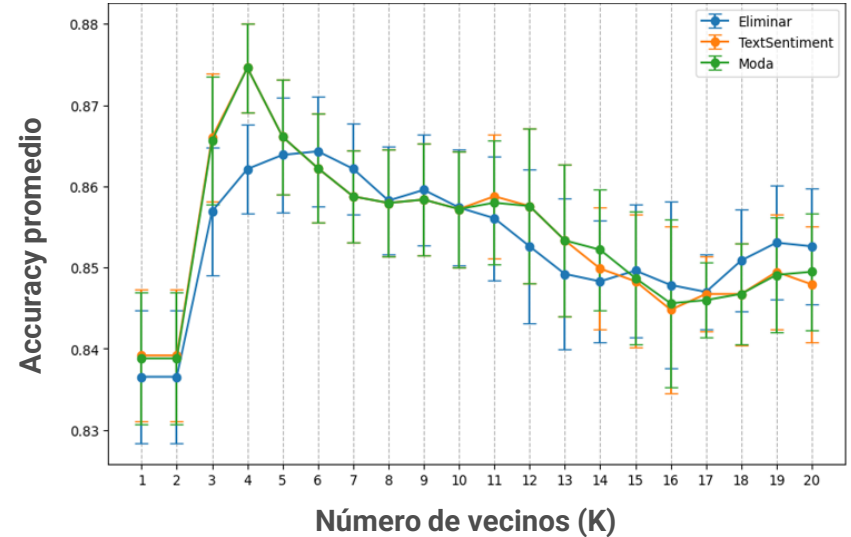
Diferentes tratamientos de datos faltantes

Utilizar la moda o reemplazar por Text Sentiment producen resultados parecidos en el mejor de los casos.

KNN



Weighted KNN

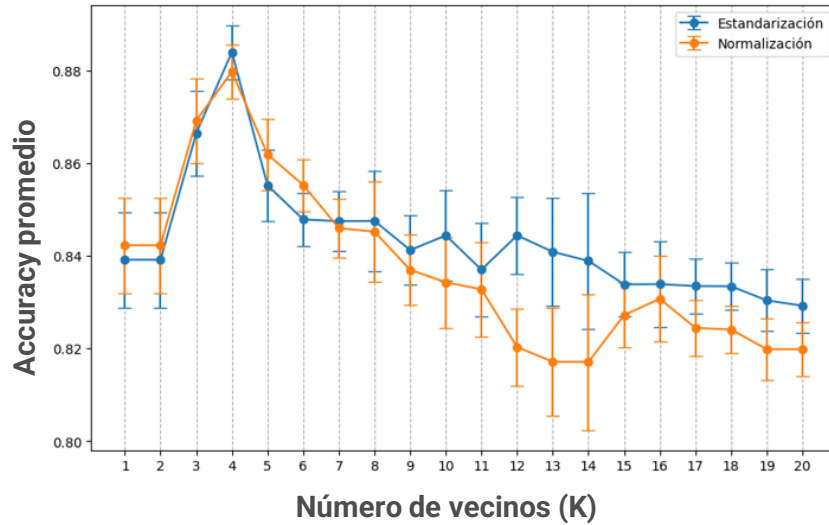


Variantes al método

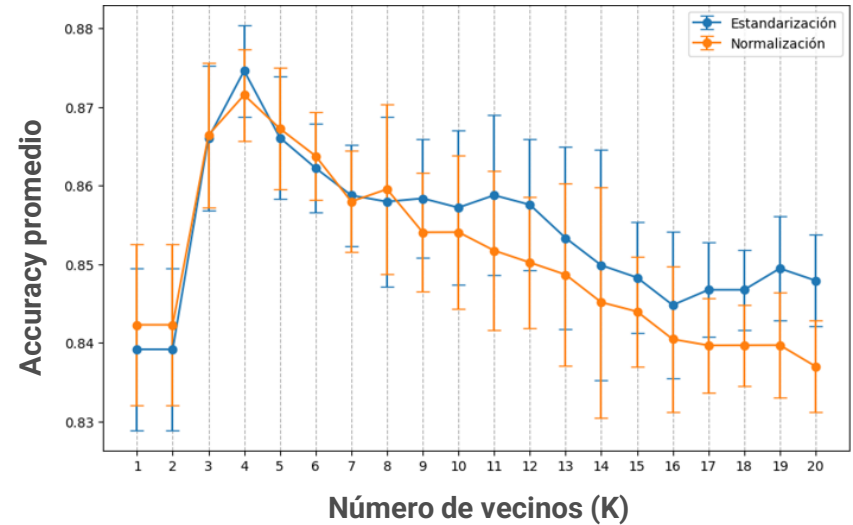
Estandarización vs Normalización

Tomando reemplazo Text Sentiment, al comparar resulta mejor utilizar estandarización en los datos.

KNN



Weighted KNN



Variantes al método

Variables explicativas

Hasta ahora utilizamos las características de la consigna (Word Count, Title Sentiment, Sentiment Value).

Pero lo correcto es realizar un **análisis más detallado** de qué **variables** utilizar (aún más en un dataset con más cantidad de características).

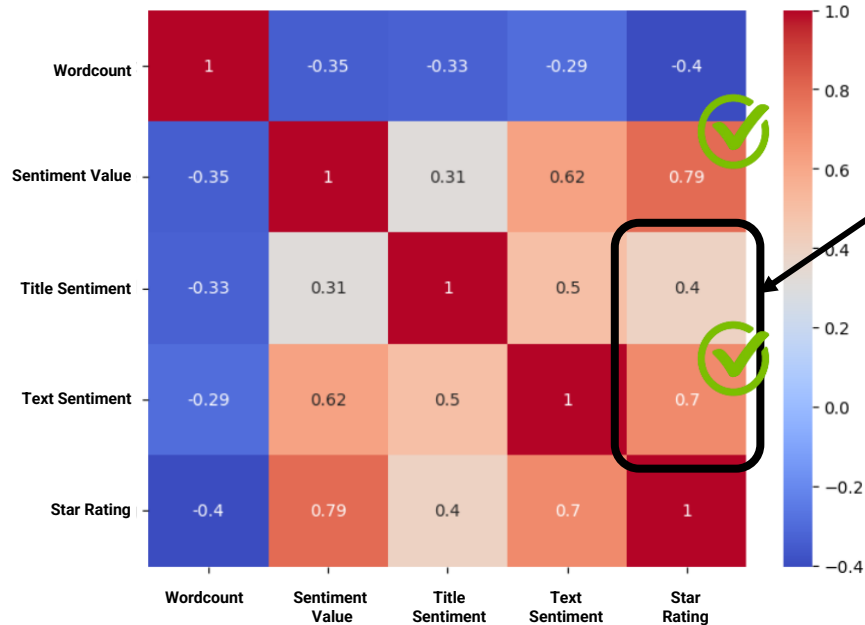
Una variante interesante a la hora de realizar el categorizador es cambiar las variables explicativas del método en busca de las mejores variables:

- Variables con más **correlación** (Coeficiente de Pearson) con Star rating.
- **Reducción** a 2 variables mediante análisis **PCA**.
- Variables con que **aportan más información** sobre Star Rating, mediante la **Entropía de Shannon**.

Variantes al método

Variable explicativa: Correlación con Star rating

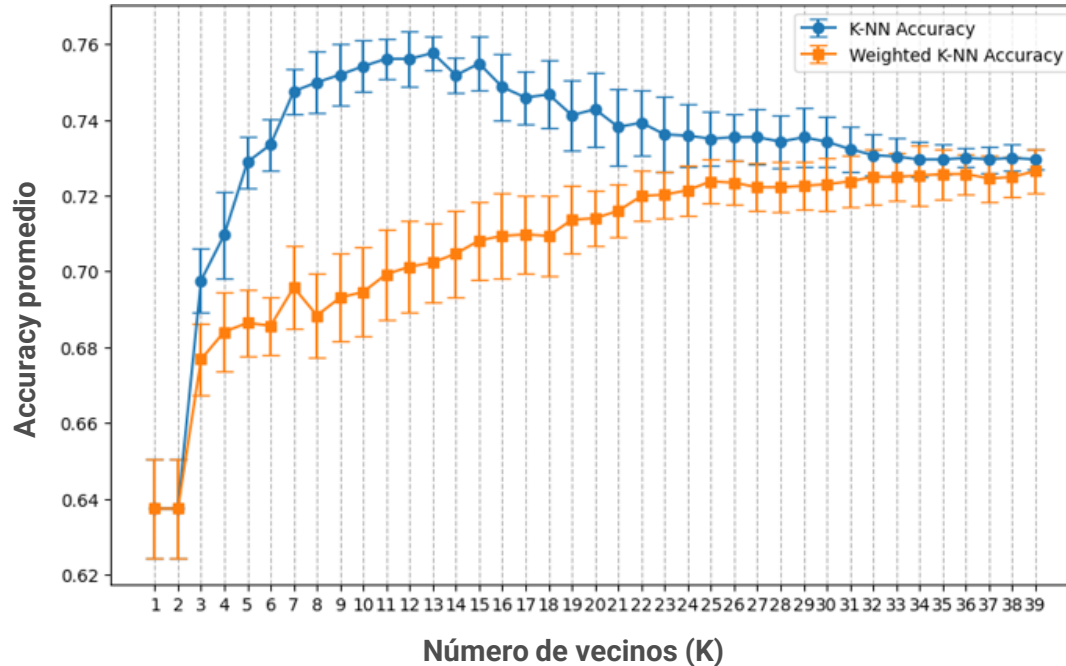
El objetivo de las variables explicativas es tener alguna incidencia en la variable objetivo, en este sentido analizar si existe alguna correlación resulta importante



Diferencia entre correlación del sentimiento del texto con sentimiento del título ya que el título cuenta con menor número de palabras lo que repercute en una menor precisión para predecir la puntuación.

Variantes al método

Variable explicativa: Correlación con Star rating



Características:

- Text Sentiment
- Sentiment Value

Mejores resultados:

KNN

K = 13

Accuracy = 0.758

Weighted KNN

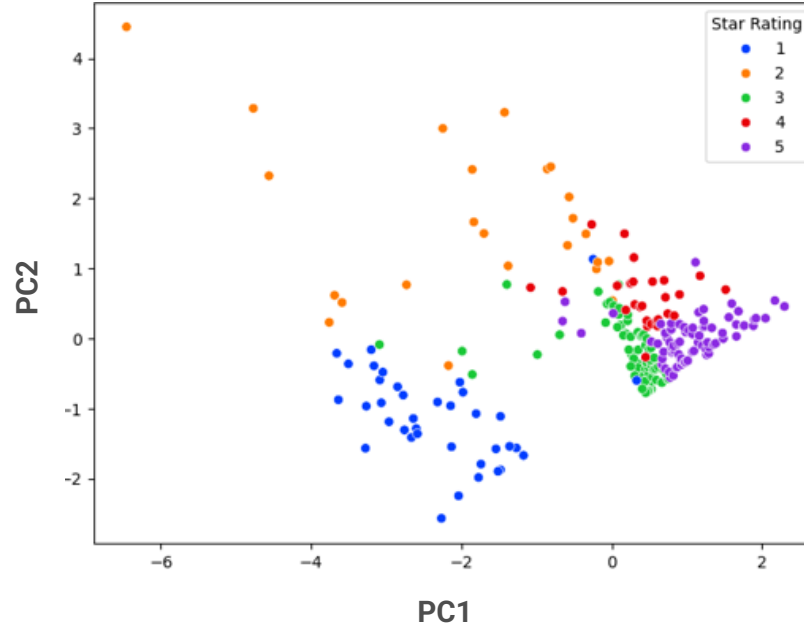
K = 34

Accuracy = 0.726

Variantes al método

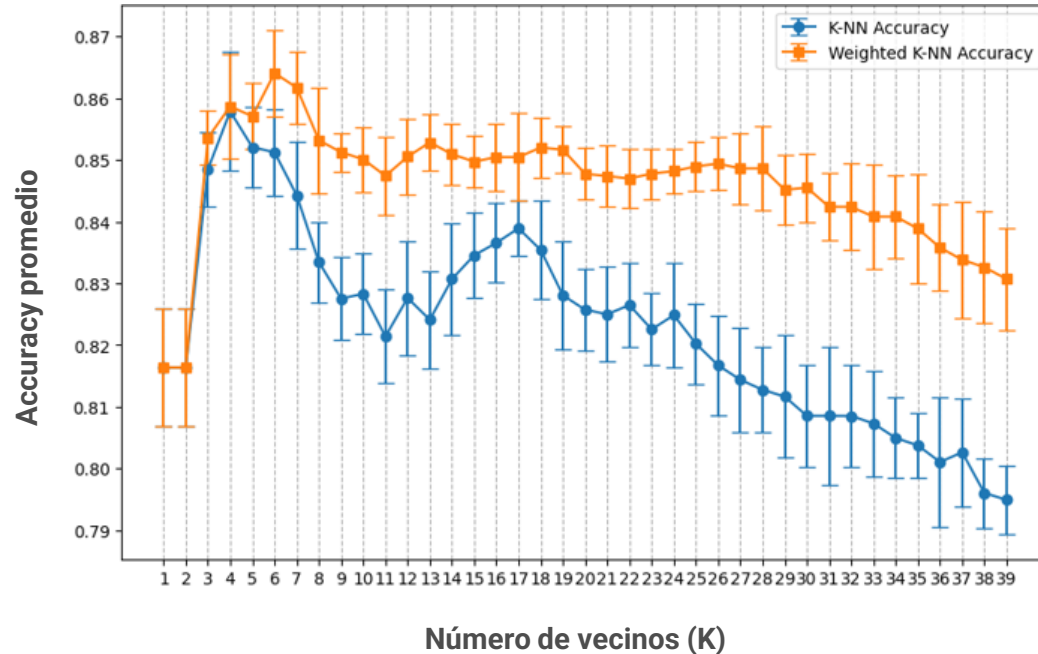
Variable explicativa: PCA

PCA consiste en un método de reducción dimensional, donde como entrada recibe un conjunto de features y devuelve un conjunto de componentes principales. En este caso se redujo de las 4 variables categóricas a 2 características.



Variantes al método

Variable explicativa: PCA



Características:

- PC1
- PC2

Mejores resultados:

KNN

K = 4

Accuracy = 0.858

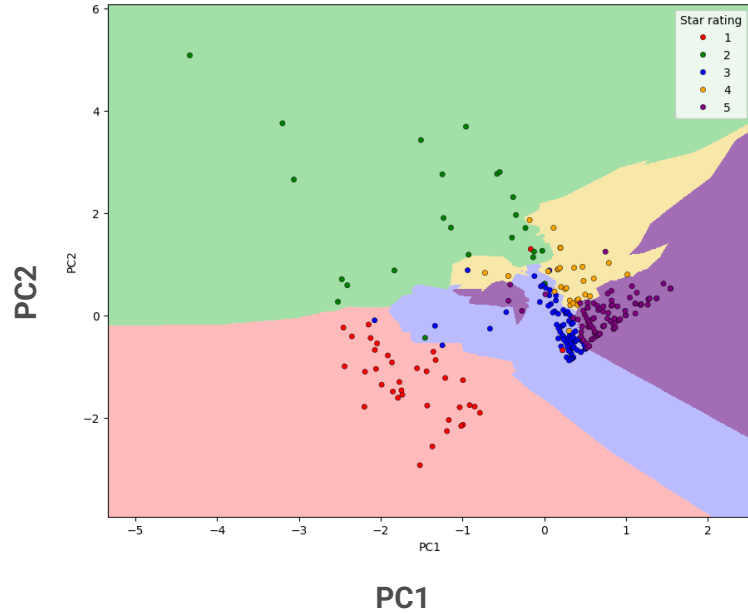
Weighted KNN

K = 6

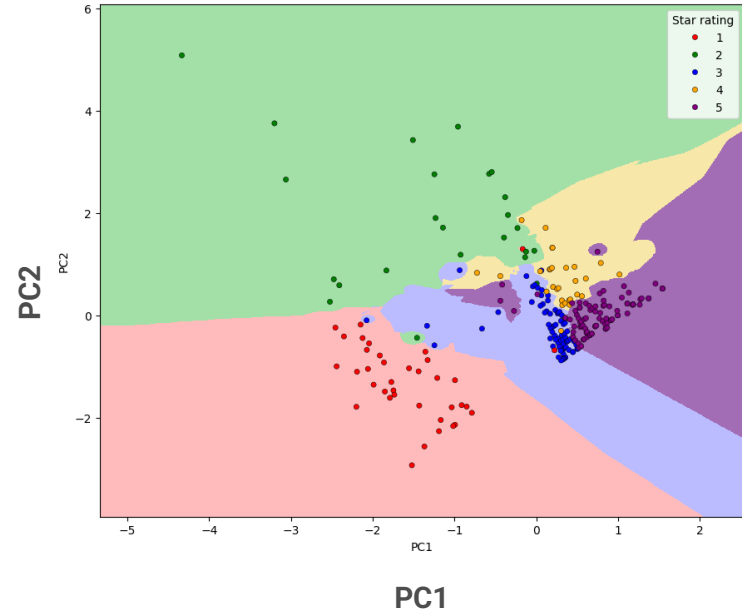
Accuracy = 0.864

Método KNN

PCA: Decision Boundaries



KNN (Best K)

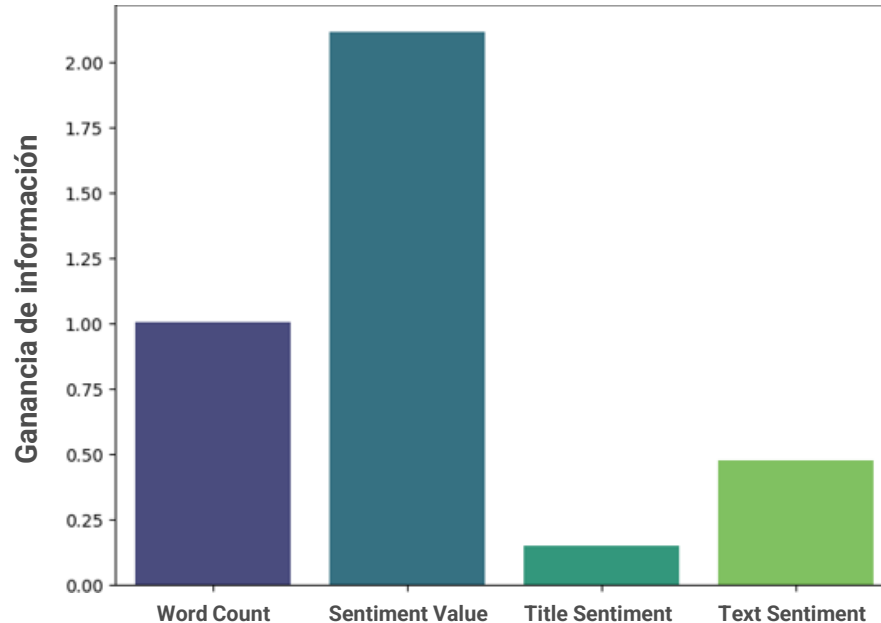


Weighted KNN (Best K)

Variantes al método

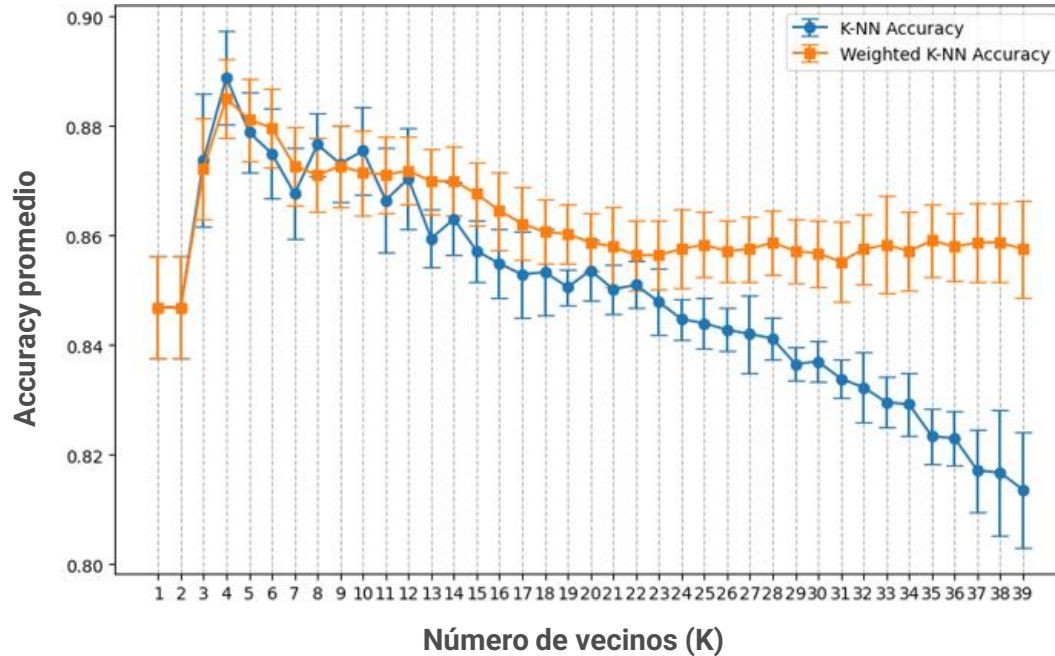
Variable explicativa: Shannon Entropy

La Entropía de Shannon es un método para determinar qué variable aporta más información al dataset. En este sentido, utilizamos como variables explicativas las 3 de mayor entropía.



Variantes al método

Variable explicativa: Shannon Entropy



Características:

- Sentiment Value
- Word Count
- Text Sentiment

Mejores resultados:

KNN

K = 4

Accuracy = 0.889

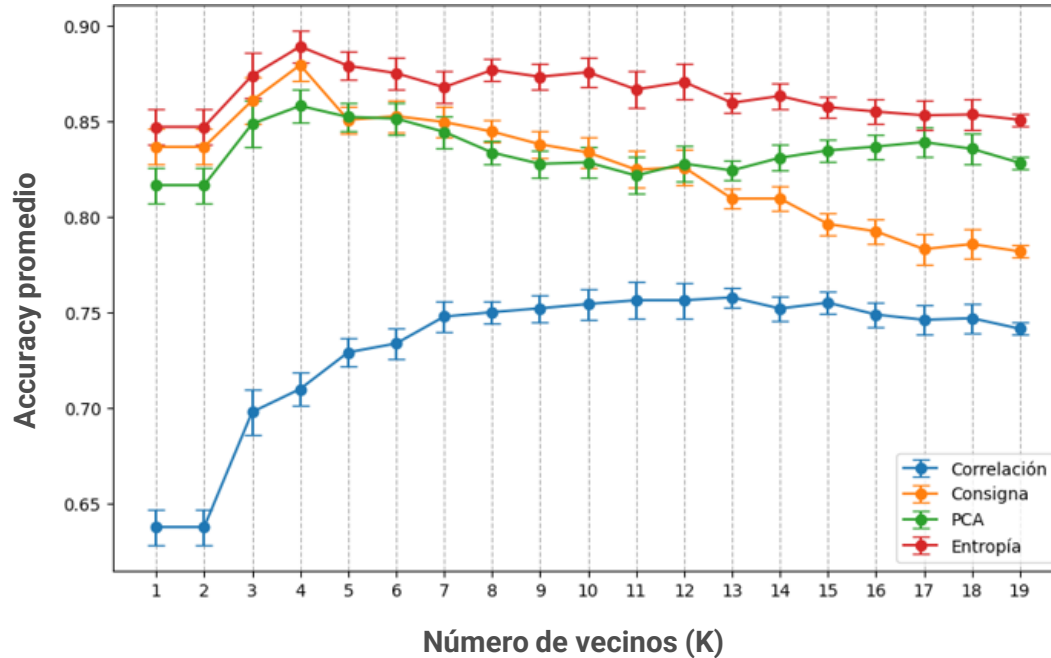
Weighed KNN

K = 4

Accuracy = 0.885

Variantes al método

Comparativa de variables explicativas (KNN)



Mejores resultados:

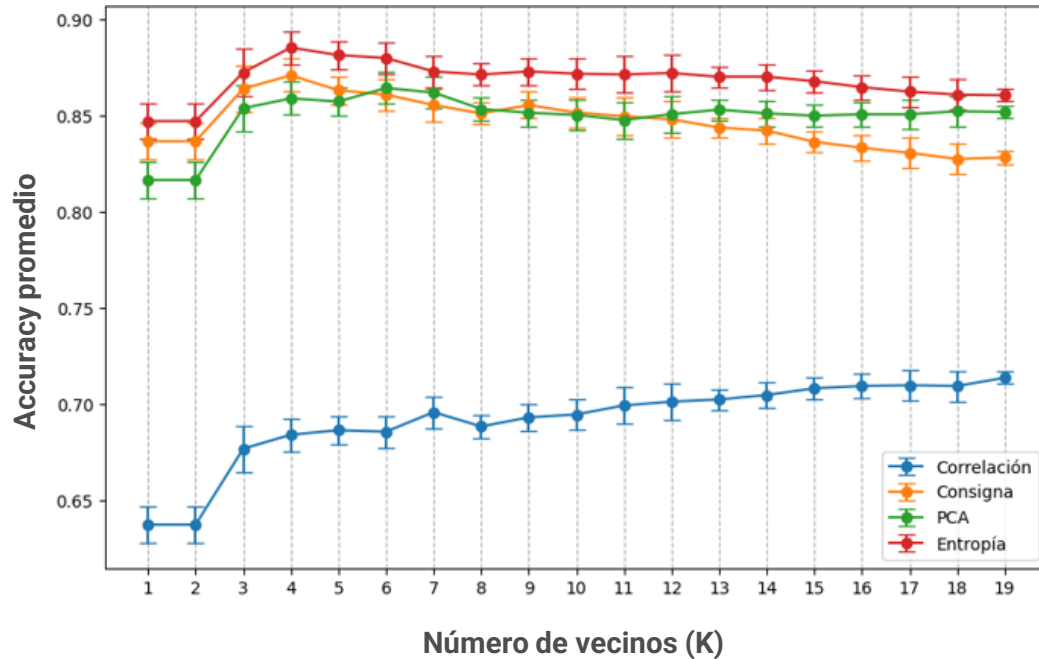
Entropía de Shannon

K = 4

Accuracy = 0.889

Variantes al método

Comparativa de variables explicativas (Weighted KNN)



Mejores resultados:

Entropía de Shannon

K = 4

Accuracy = 0.885

Conclusiones

Conclusiones (I)

Ejercicio 1 – ID3 & Random Forest

- **Conclusión 1:** Bootstrapping es un buen método para mejorar el sobreajuste que presenta el modelo ID3.
- **Conclusión 2:** La precisión para el conjunto de prueba no tiene cambios significativos según el número de nodos. A veces mejora a veces no (necesitarían poda).
- **Conclusión 3:** Agregar árboles a un bosque Random Forest mejora las métricas de precisión y ayuda al sobreajuste.
- **Conclusión 4:** Variar la cantidad de atributos seleccionables para los árboles no tiene un efecto significativo en la performance del modelo.

Conclusiones (II)

Ejercicio 2 – KNN

- **Conclusión 1:** La mejor manera de definir las variables explicativas es mediante el método de la Entropía de Shannon.
- **Conclusión 2:** Resulta mejor estandarizar los datos antes que normalizarlos para este problema.
- **Conclusión 3:** Se aprecian diferencias ligeras entre eliminar datos en blanco y reemplazarlos con el valor de Text Sentiment.
- **Conclusión 4:** Resulta posible construir un clasificador mediante el método KNN y Weighted KNN que pueda predecir Star Rating con una precisión mayor al 88%.
- **Conclusión 5:** Se obtienen mejores resultados con K chicos.

¡Gracias!