Trabajo Práctico Nro. 2: Algoritmos de Clasificación Supervisada

Ezequiel Agustin Perez, Gonzalo Nahuel Baliarda, Lucas Agustin Vittor

Instituto Tecnológico de Buenos Aires

{*eperez, gbaliarda, lvittor*}*@itba.edu.ar*

20 de septiembre de 2023

Ejercicio 1

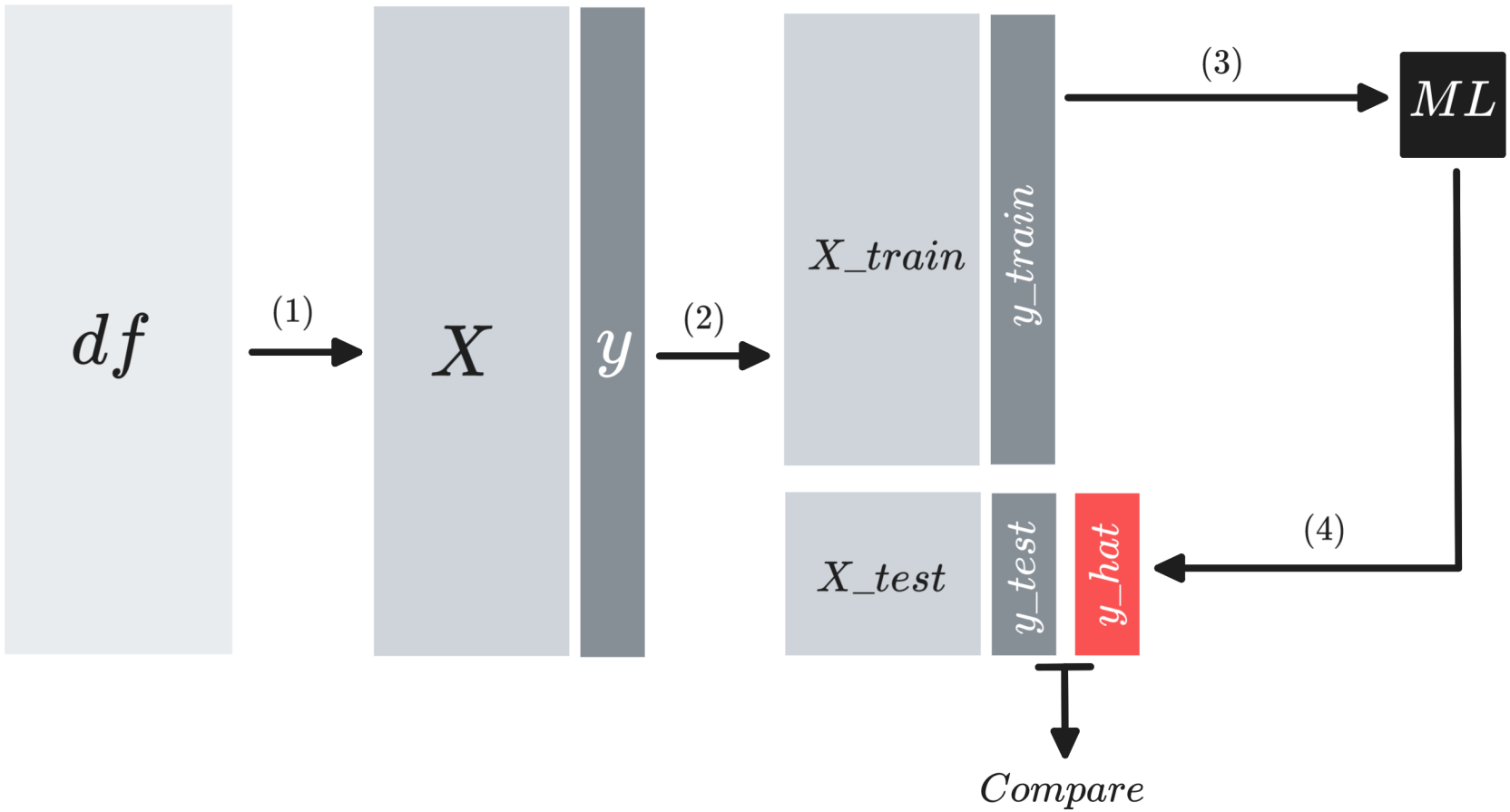
Enunciado

Dado *german_credit.csv*:

1. Dividir datos en entrenamiento y prueba.
2. Implementar algoritmo ID3 con entropía de Shannon. Clasificar.
3. Clasificar usando Random Forest.
4. Construir matriz de confusión y comparar resultados.
5. Graficar curvas de precisión vs. nodos para cada método.

Dataset

Randomizado y dividido en 60 % train, 40 % test.



Árbol de decisión: modelo

Entropia de Shannon

$$H(X) = -\sum_{i=1}^{n} P(x_i)\ log_2 P(x_i) \quad \text{con } H(X) \in \{0, 1\}$$

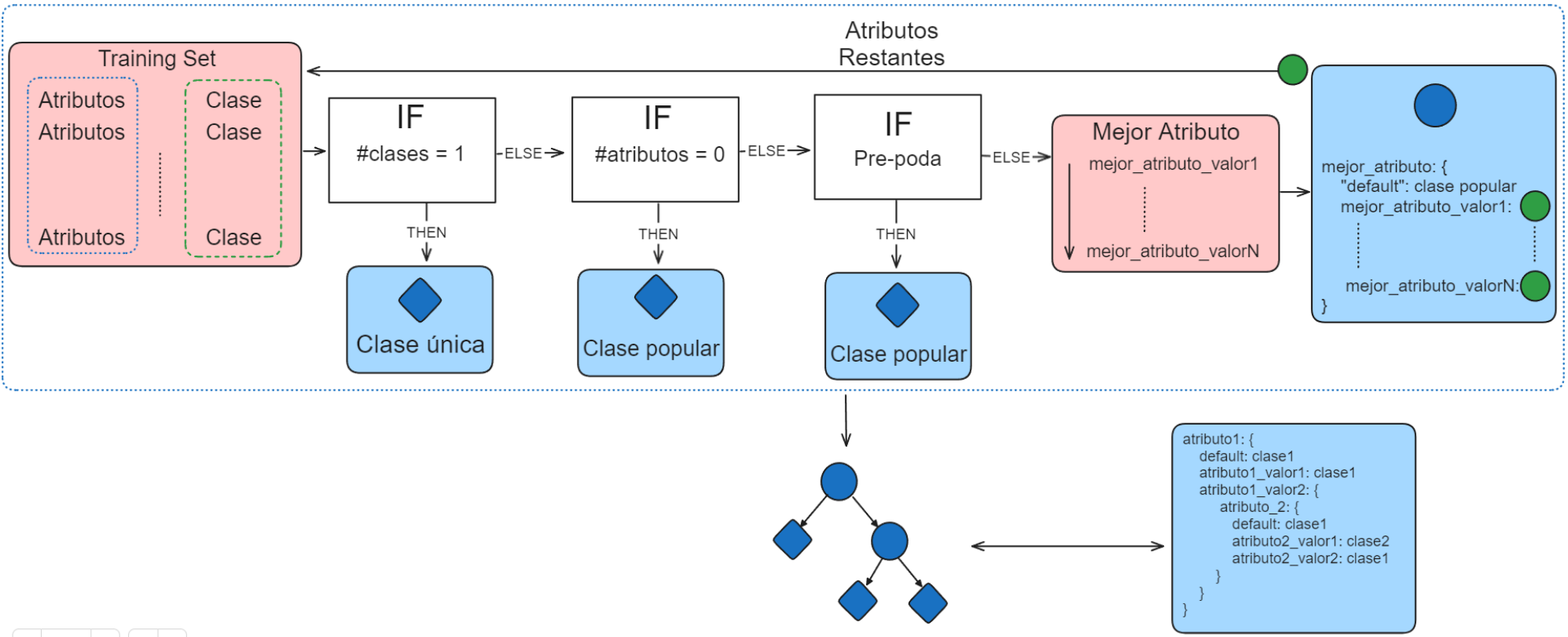
Informacion de Ganancia

$$Gain(S, A) = H(S) - \sum_{v \in Valores(A)} \left|\frac{S_v}{S}\right| H(S_v)$$

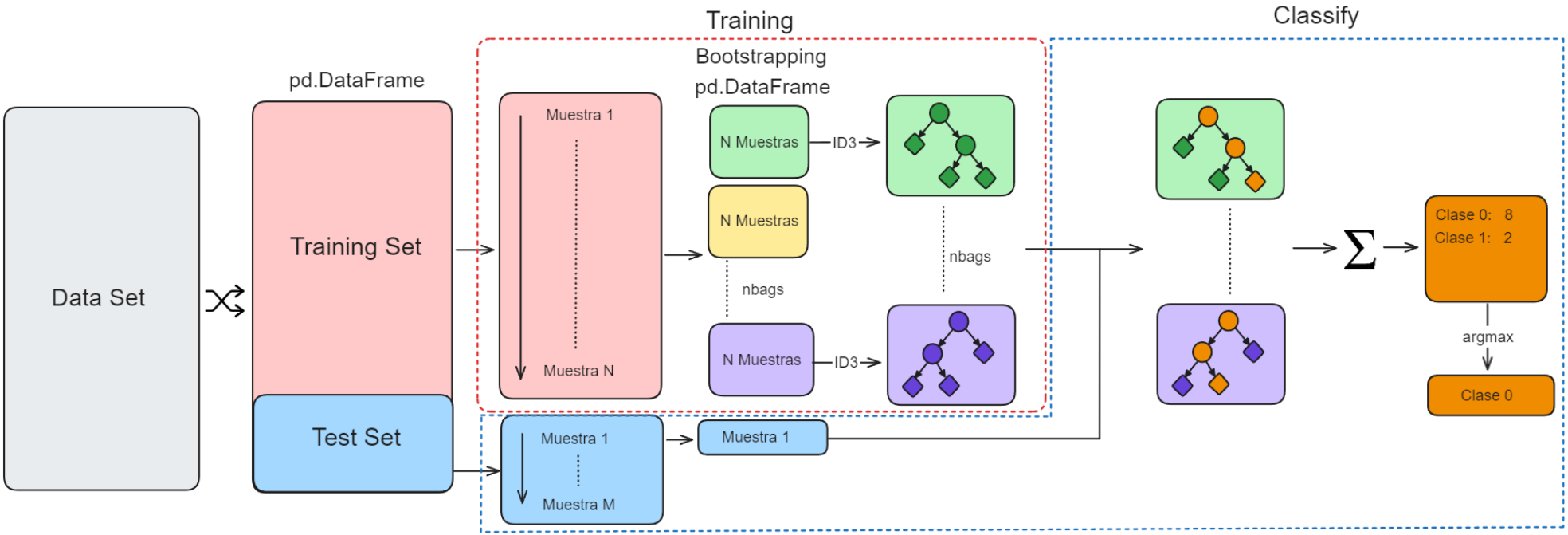
Mejor atributo

$$A^* = \underset{\{A \in Atributos\}}{\operatorname{argmax}} \ Gain(S, A)$$

Árbol de decisión: implementación



Random Forest: implementación



Resultados: matriz de confusión

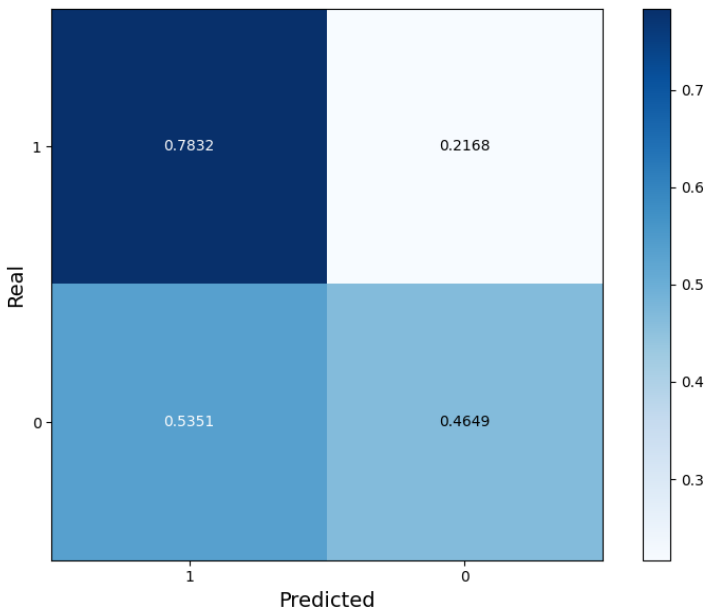


Figura: Decision Tree

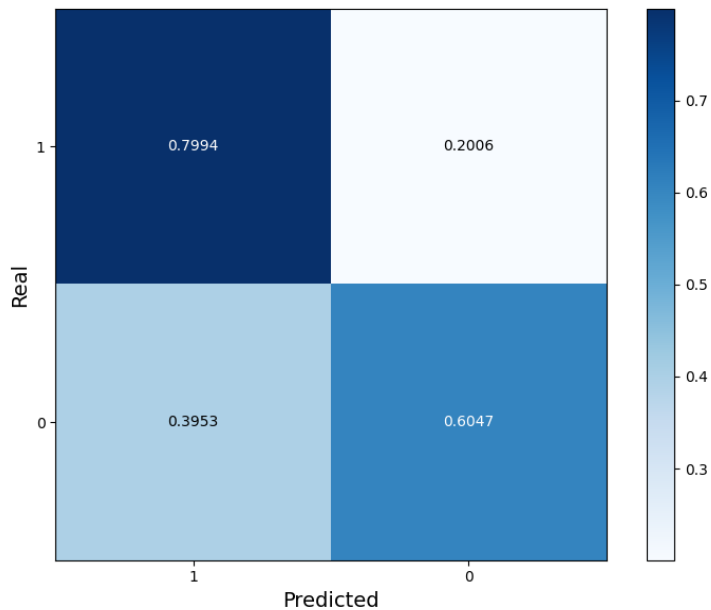


Figura: Random Forest

Resultados: Precision

- 10 experimentos ; error = std.
- Se ajustan la máxima altura del árbol y la mínima cantidad de muestras para ramificar (pre-poda).

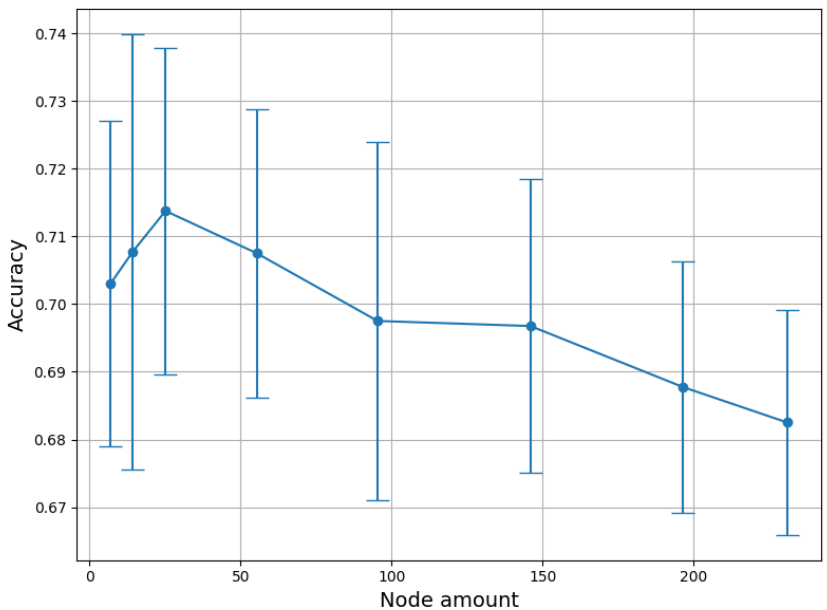


Figura: Decision Tree

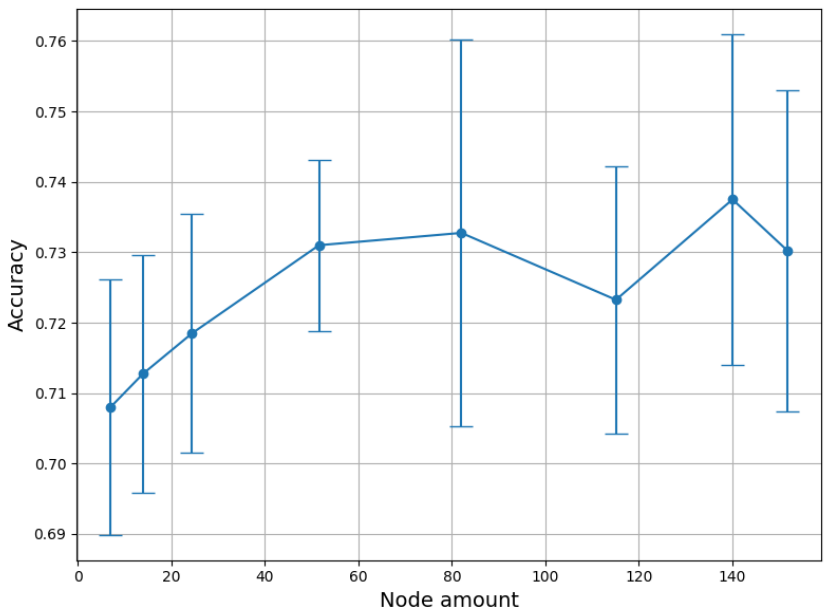


Figura: Random Forest

Conclusiones

- Random Forest mejora notablemente la precisión sobre una de las clases, pero muy levemente la precisión sobre la otra.
- Ninguno de los modelos predice bien la no devolución del crédito.
- El Árbol de Decisión funciona mejor con una cantidad de nodos baja, mientras que Random Forest funciona mejor con una cantidad promedio de nodos más alta.

Ejercicio 2

Enunciado

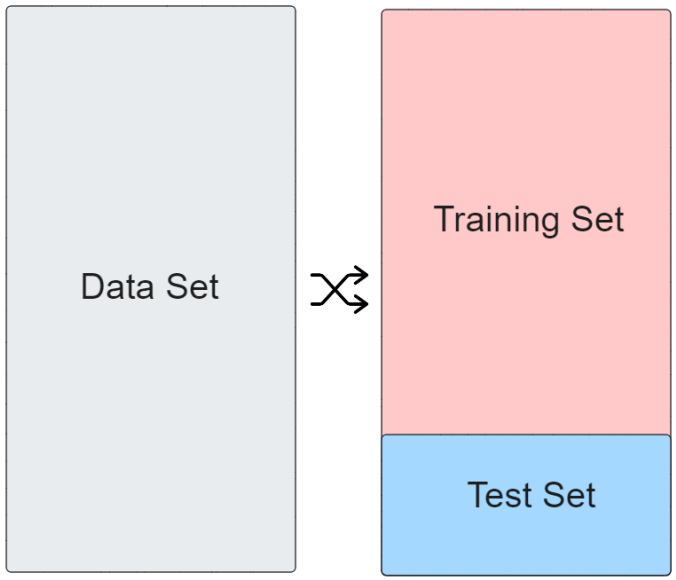
A partir de un dataset de opiniones, clasificarlas.

1. Calcular la cantidad promedio de palabras para comentarios con 1 estrella.
2. Dividir dataset en training y testing set.
3. Aplicar K-NN y K-NN con distancias pesadas para clasificar opiniones.
4. Calcular la precisión y la matriz de confusión.

Dataset

Reviews Sentiment

- Multiclase segun el Star Rating variable de 1 a 5
- Variables explicativas: wordcount, titleSentiment y sentimentValue
- Variables estandarizadas
- 70-30 división de entrenamiento y testeo
- Promedio de palabras para comentarios con 1 estrella: 12.216



Modelo

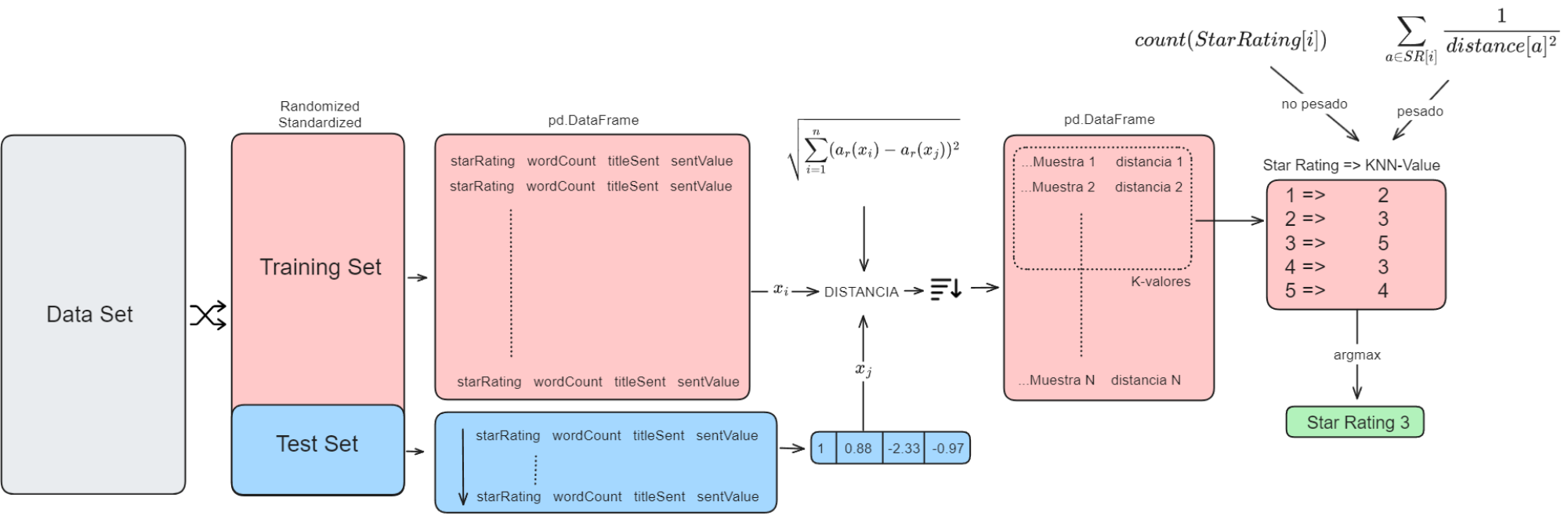
Distancia Euclidea:

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

Algoritmo K-NN

$$\hat{f}(x_q) = \underset{\{v \in V\}}{\operatorname{argmax}} \sum_{i=1}^{k} w_i \cdot 1_{\{v = f(x_i)\}} \qquad \text{con } w_i = \begin{cases} \frac{1}{d(x_q, x_i)^2} & \text{pesado} \\ 1 & \text{sino} \end{cases}$$

Implementación



Resultados: matriz de confusión

k = 2 ; split 70/30

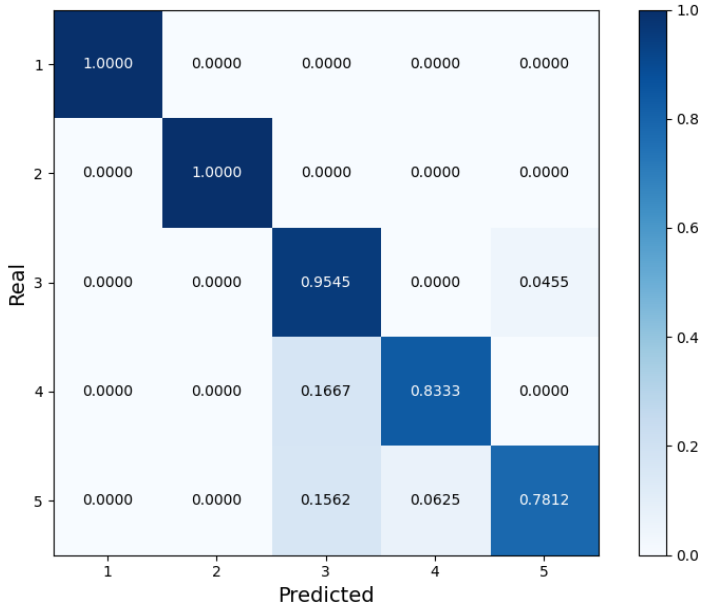


Figura: unweighted

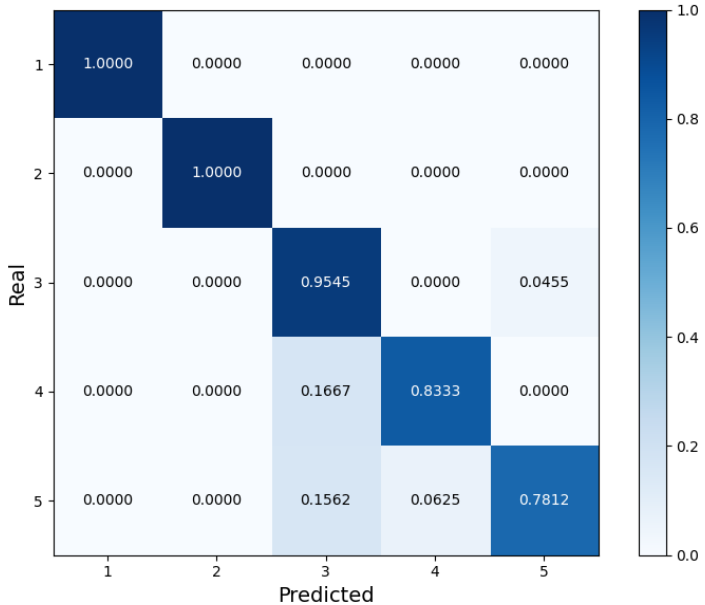


Figura: weighted

Resultados: matriz de confusión

k = 3 ; split 70/30

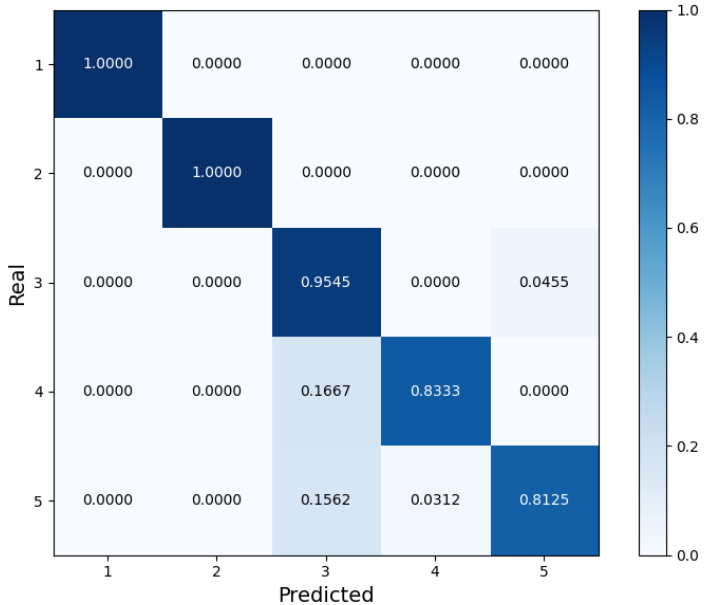


Figura: unweighted

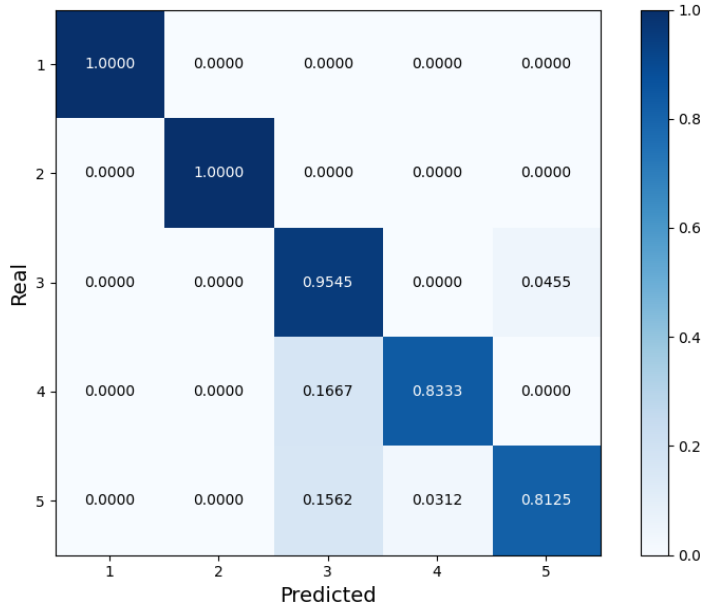


Figura: weighted

Resultados: matriz de confusión

k = 5 ; split 70/30

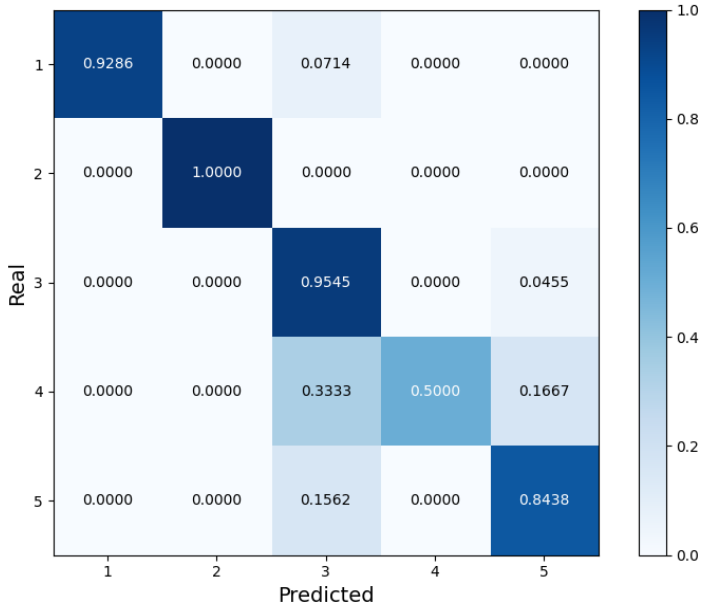


Figura: unweighted

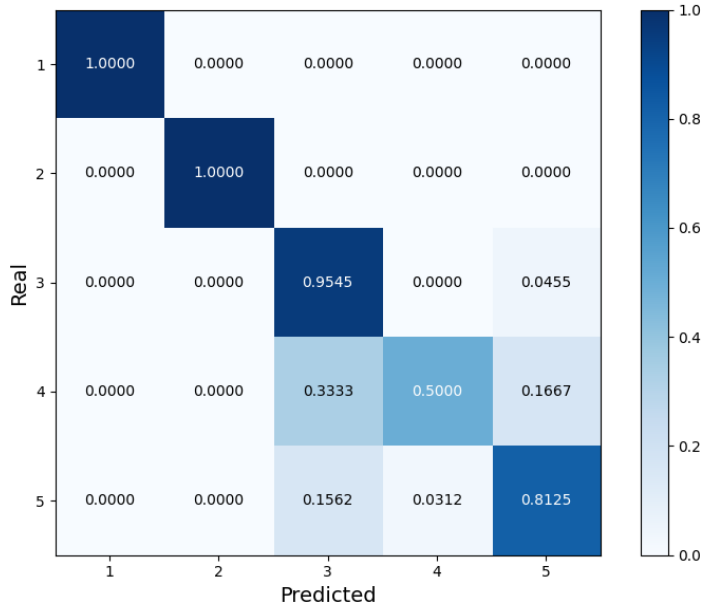


Figura: weighted

Resultados: matriz de confusión

k = 10 ; split 70/30

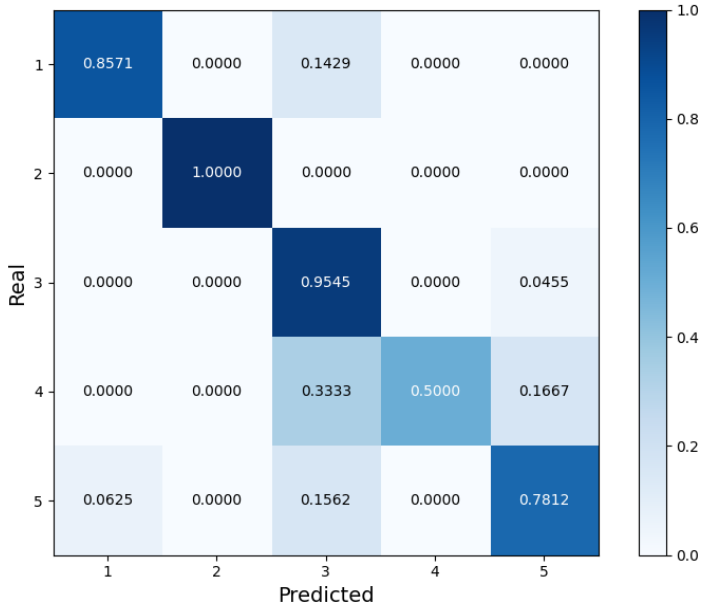


Figura: unweighted

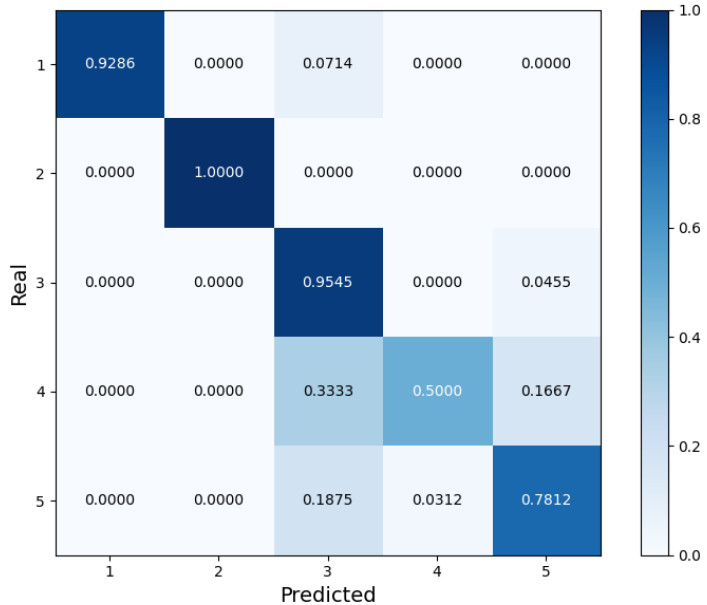


Figura: weighted

Resultados: precisión

- Se analiza un KNN con distancias pesadas.
- 5 experimentos ; error = std.

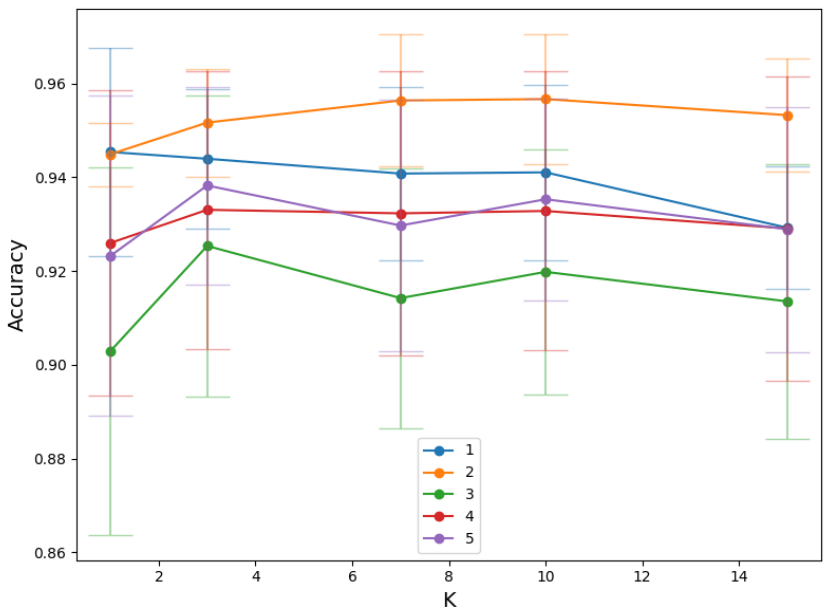


Figura: split 70/30

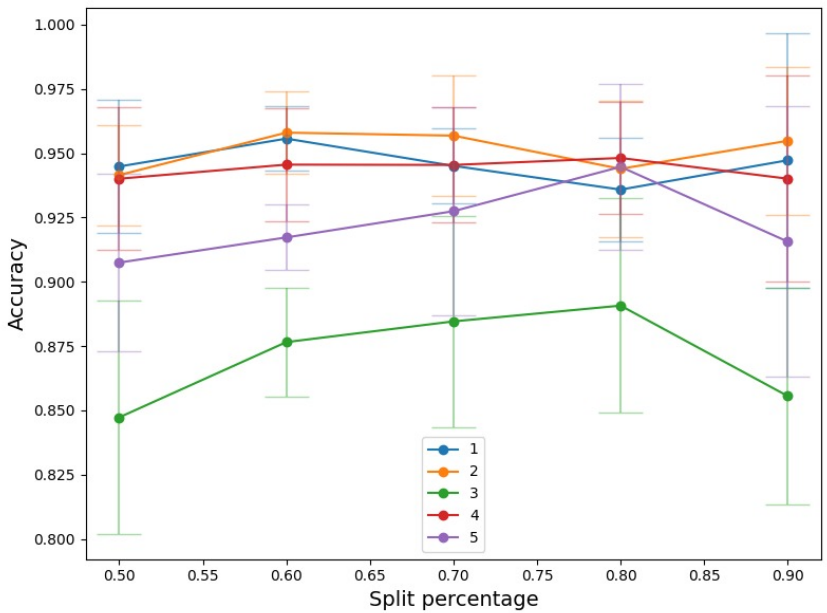


Figura: k=3

Conclusiones

- Los valores de k que mejor funcionaron fueron $3 \leq k \leq 10$, siendo k=3 el de mejor precisión general.
- La precisión general es mayor cuando se usa entre 70 % y 80 % del dataset para entrenamiento.
- Para $k \leq 3$, no se observaron diferencias entre KNN y KNN con distancias pesadas.
- KNN con distancias pesadas tiene una mayor precisión general para los k > 3 analizados.

Muchas gracias