IES Pere Maria Orts

Sistemas de Aprendizaje Automático

Practicando con el modelo: K vecinos más próximos

Autor:
Kenny Berrones

Profesor:
David Campoy Miñarro

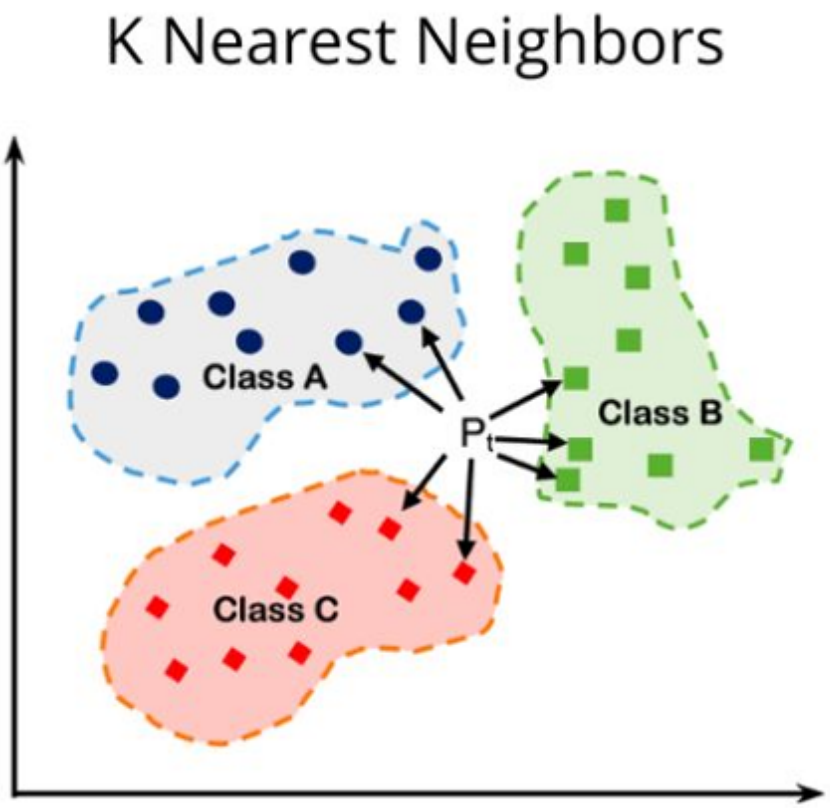
iesperemariaorts

GENERALITAT VALENCIANA

Índice

1. Introducción

El algoritmo kNN se basa en la idea de que los objetos que son similares están cercanos en un espacio n-dimensional. El objetivo del algoritmo kNN es clasificar nuevos puntos de datos basados en los puntos de datos existentes que están más cercanos a ellos en términos de distancia euclidiana.

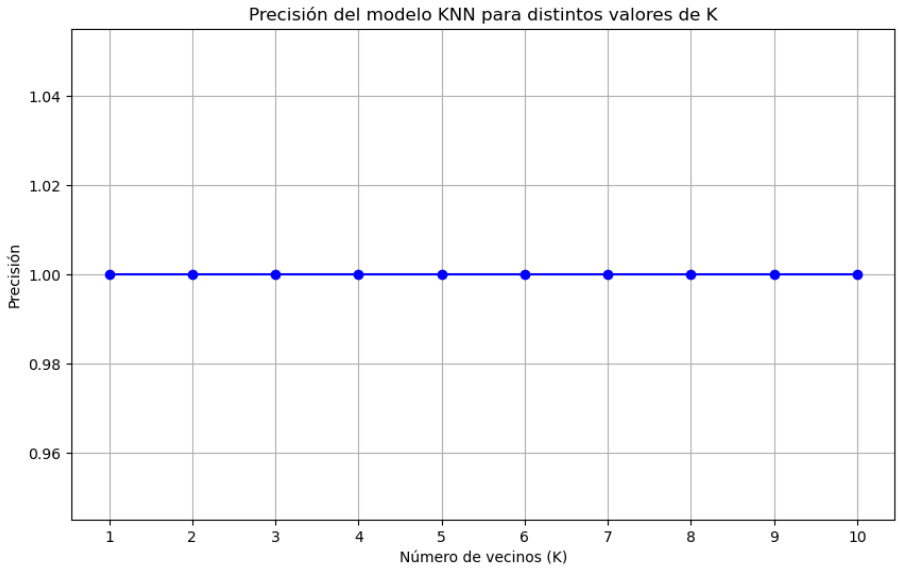


2. Experimentos

Tendremos que realizar diversas actividades para ir entendiendo el funcionamiento de este algoritmo.

2.1. Actividad 1

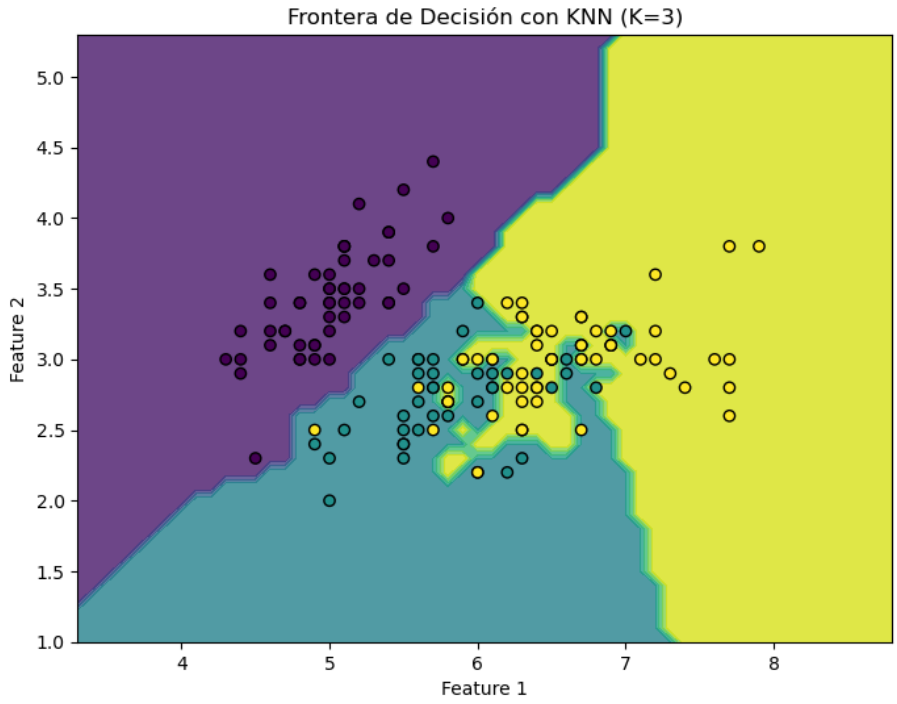
En la primera actividad vamos a usar el dataset del Iris, la idea de esta actividad es ver que tasa de acierto obtenemos al ejecutar este código. Hemos probado con distintos valores de K para que sea una prueba más rigurosa, hemos obtenido los siguientes valores:



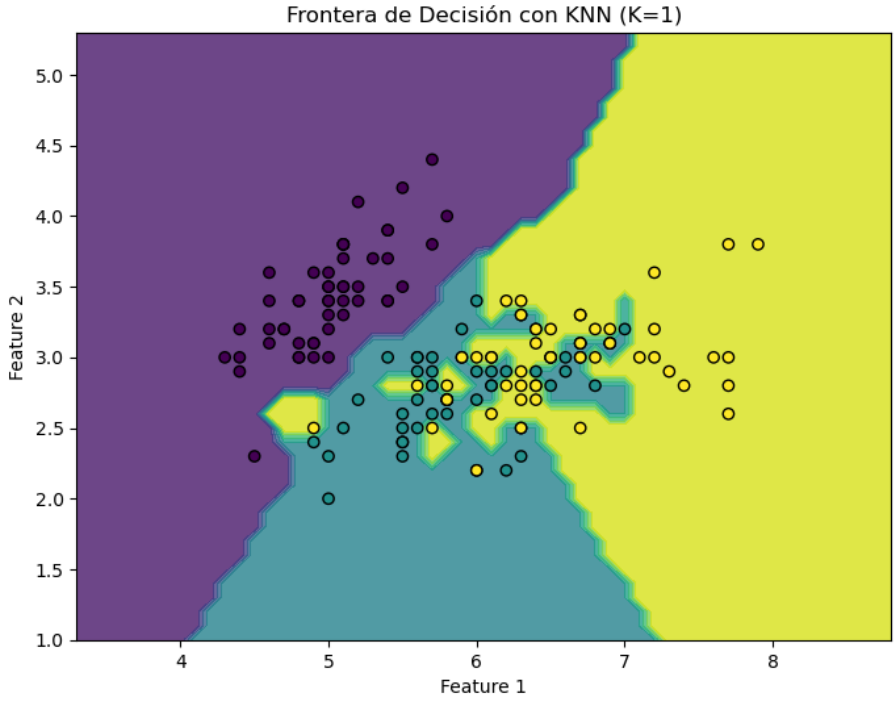
Como se aprecia en la gráfica anterior obtenemos un valor de precisión de 1 para todos los valores de k, entiendo que esto se debe a que es dataset muy sencillo.

2.2. Actividad 2

En la segunda actividad vamos a mostrar una gráfica en la cual podemos apreciar de forma visual el como se separan las distintas clases del dataset anterior, es la siguiente gráfica:

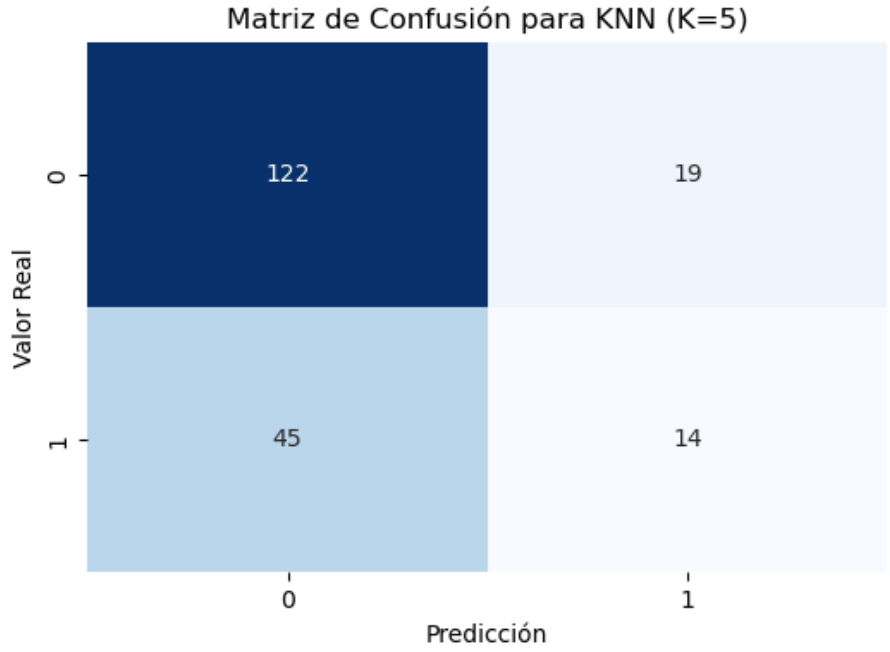


También hemos probado a modificar el valor de k y si que se observa algún cambio en la gráfica, y si que ocurre el cambio, la frontera de las distintas clases varia, en la siguiente imagen vemos un ejemplo de esto:



2.3. Actividad 3

En la tercera actividad vamos a probar un dataset sobre el crédito alemán, es decir, que si un cliente pide un crédito será aprobado o no. Obtenemos la siguiente gráfica:

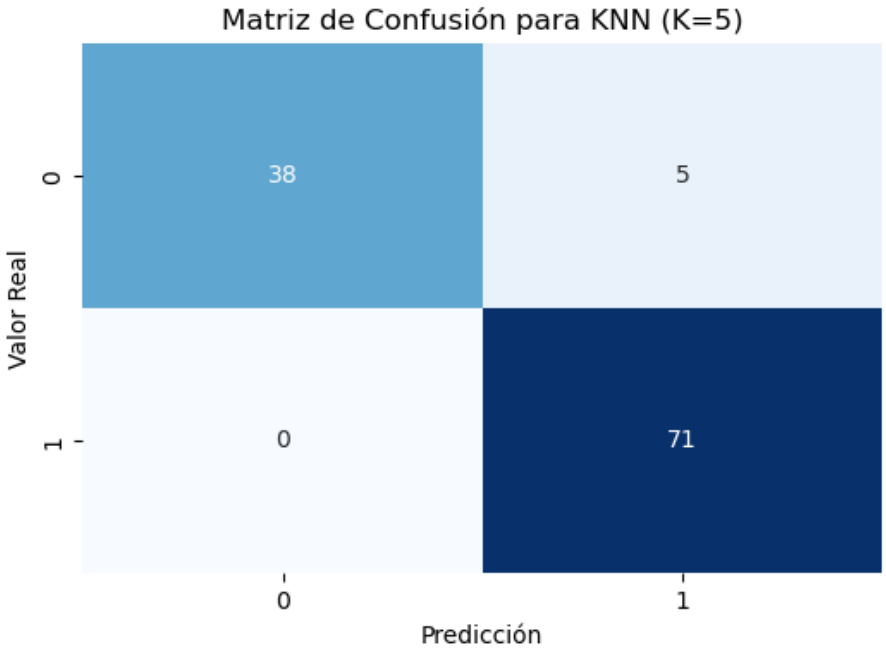


Vemos que tiene un buen acierto para cuando no se ha dado un crédito y se predice que no se ha dado, pero falla mucho cuando se ha dado un crédito y predice el modelo que se ha dado. Además, también da unos falsos negativos, para los casos en el que se ha proporcionado el crédito pero el modelo predice que no se ha dado.

2.4. Actividad 4

Para la cuarta actividad vamos a emplear el modelo KNN para la ámbito de la salud. En este caso vamos a usar el modelo KNN para predecir si un tumor se corresponde con un tumor benigno o maligno.

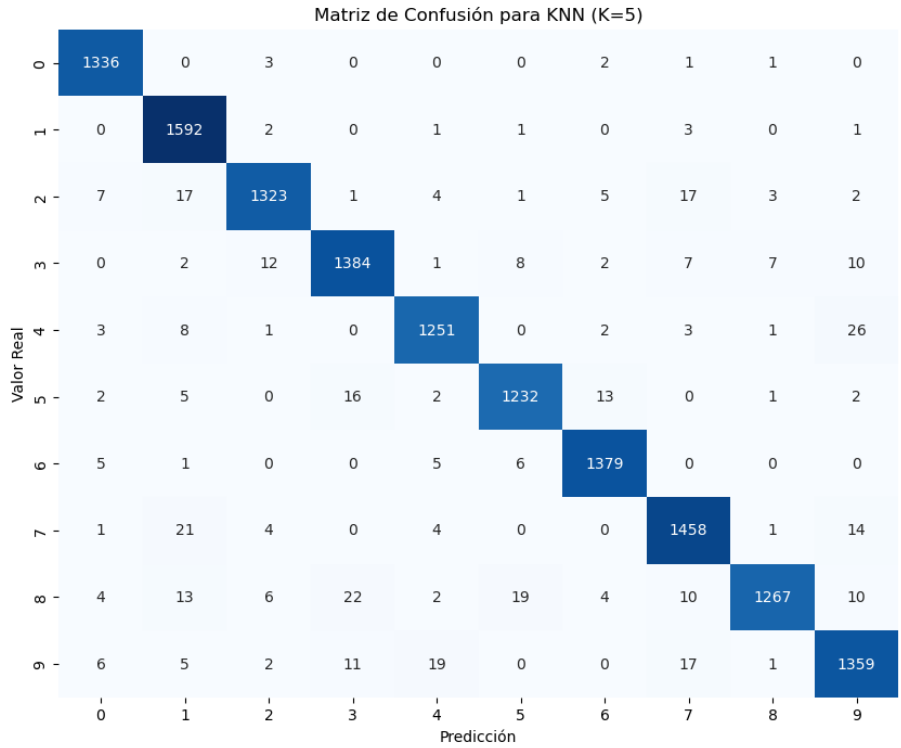
Al ejecutar el código proporciona obtenemos la siguiente matriz de confusión:



Vemos que este modelo funciona bastante bien, ya que obtenemos un 96 % de precisión, además, si observamos la gráfica anterior vemos que solo falla en 5 casos, e incluso en estos casos sería un caso de falso positivo.

2.5. Actividad 5

En la quinta actividad vamos a emplear el modelo KNN para reconocer dígitos manuscritos, vamos a usar el dataset MNIST, al ejecutar el código que se nos proporciona obtenemos el siguiente resultado:



Además, nos da una tasa de acierto del 97 %, y vemos que la mayoría de dígitos los clasifica correctamente.

3. Conclusiones

El algoritmo kNN demuestra ser una herramienta versátil y efectiva para tareas de clasificación en diversos contextos. A través de las actividades realizadas, se evidencia su alta precisión en datasets simples como Iris, su capacidad de adaptación al ajustar k, y su efectividad en problemas más complejos, como la predicción de tumores y el reconocimiento de dígitos manuscritos, alcanzando precisiones superiores al 95 %. Sin embargo, también se observan limitaciones, como en el dataset del crédito alemán, donde surgen problemas con falsos positivos y negativos. Esto resalta la importancia de seleccionar adecuadamente k y evaluar el rendimiento según las características específicas de cada conjunto de datos.