



Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Escuela de Ciencias e Ingeniería

Inteligencia artificial avanzada para la ciencia de datos I

TC3006C

Grupo 101

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Fecha de entrega:

09 de Septiembre del 2023

Profesores:

Jorge Adolfo Ramirez Uresti

Alumno:

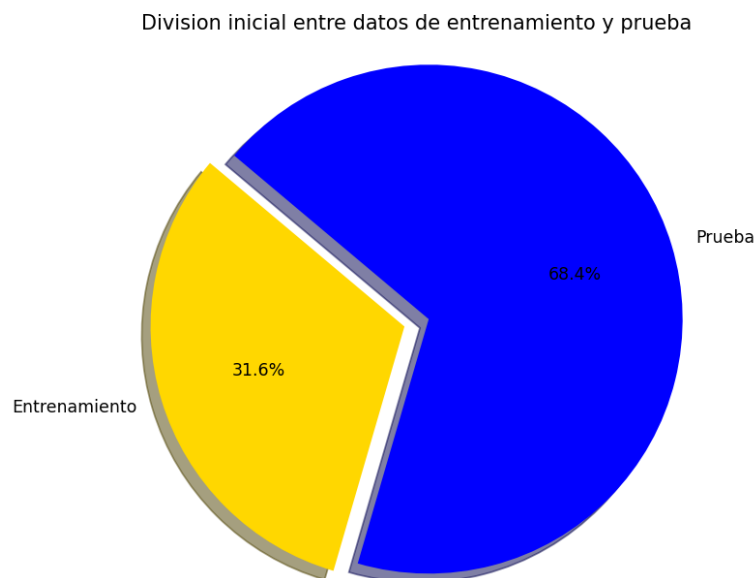
José Miguel García Gurtubay Moreno -A01373750

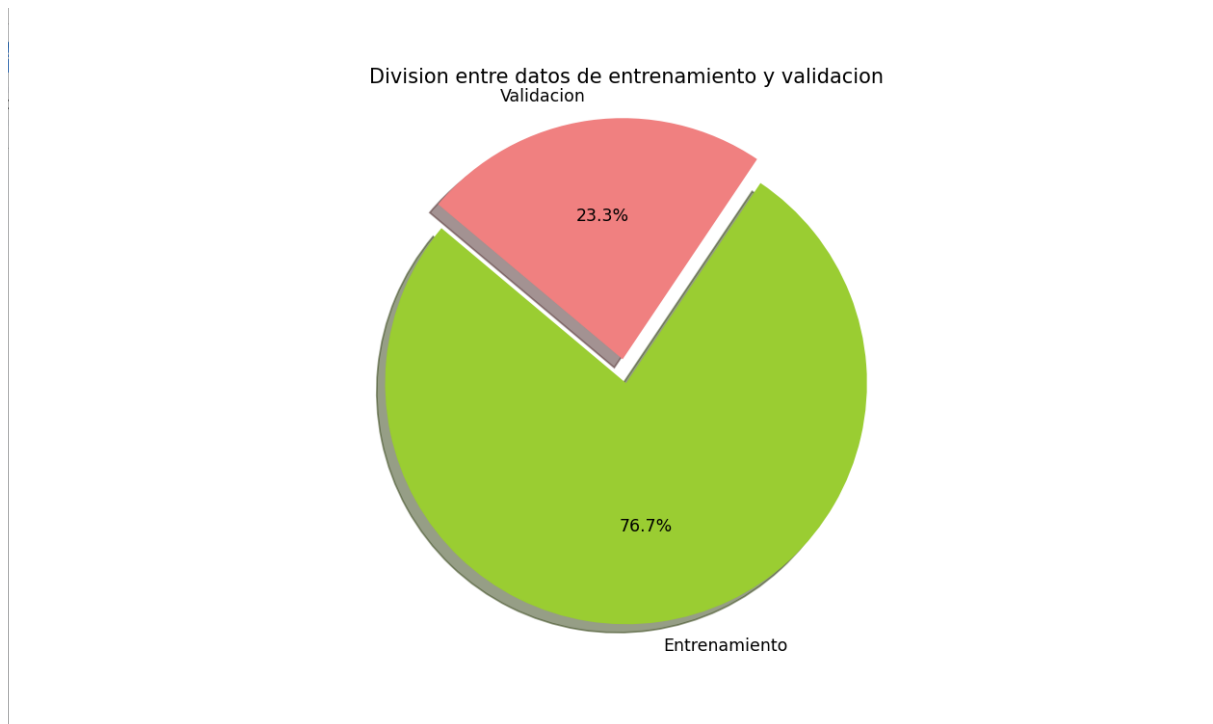
Introducción

El diagnóstico temprano de enfermedades, en particular del cáncer, es esencial para aumentar las tasas de supervivencia y mejorar la calidad de vida de los pacientes. En el ámbito de la medicina, los modelos de machine learning, como el K-Nearest Neighbors (KNN), juegan un papel crucial en la detección y clasificación de enfermedades basadas en características medidas.

Desarrollo

Es muy claro que en cuanto hay de variables que puedan afectar al modelo, la división de datos cuentan como parte de estas variables que tanto influyen al modelo, en una primera instancia se dividieron los datos de esta forma.

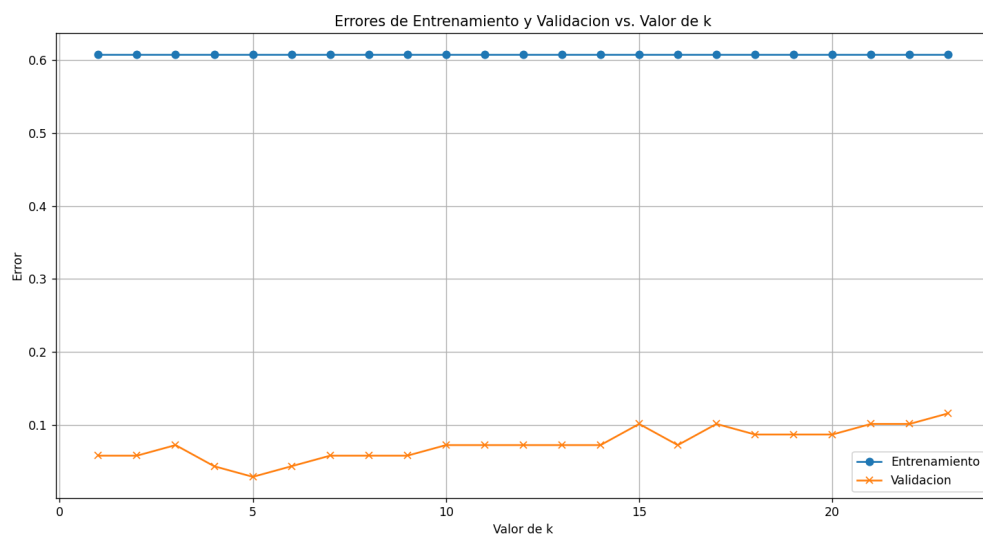




Tras dividir y entrenar los datos tenemos la primera gráfica que nos dará información sobre el modelo, en particular los errores de entrenamiento y validación en contraste con los valores de k . En ella podremos ver ambas líneas de división, las cuales nos dan información sobre el sesgo y la varianza, en este caso vemos que tanto el sesgo como la varianza son bajos.

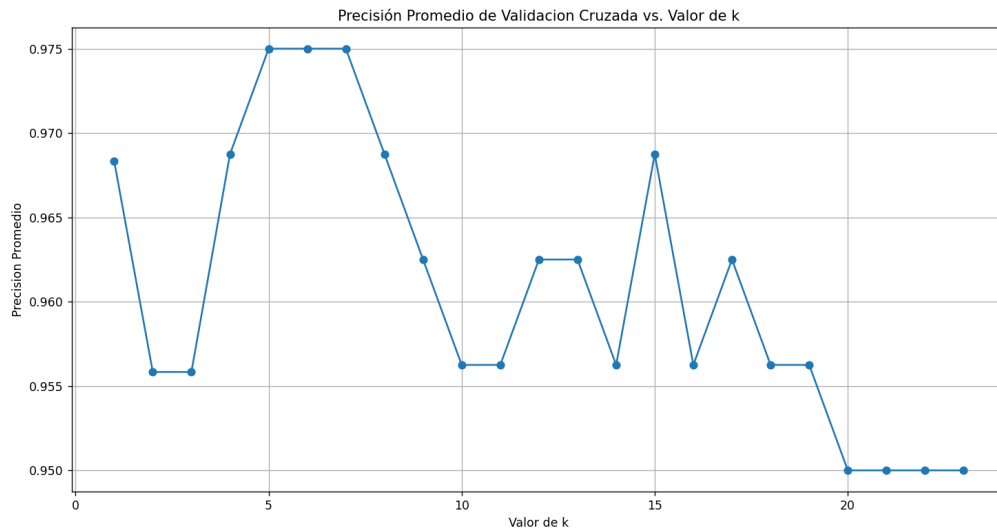
Por parte de la varianza ya que la diferencia entre ambos errores, entre la distancia de ambas líneas.

También tenemos al sesgo bajo, el cual se puede distinguir con la pendiente que hay entre punto y punto en contraste, si ambas líneas se comportan de manera similar tienen menos sesgo, como es el caso de este modelo



En la siguiente gráfica podemos apreciar la precisión cruzada en contraste con los valores de k , obteniendo las precisiones de las diferentes particiones para así determinar cuál sería en teoría el mejor valor de k posible para el modelo.

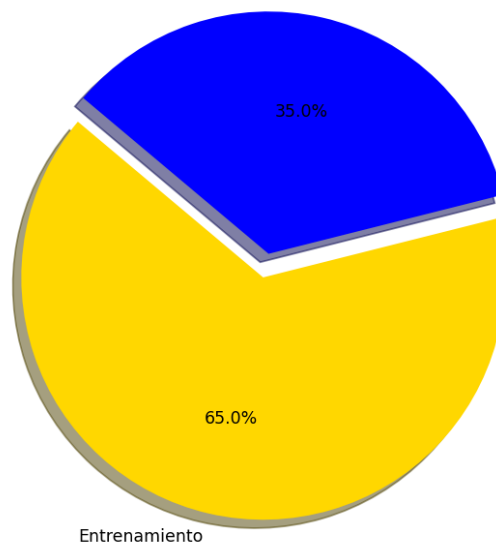
En este caso podemos ver que el 5,6 y 7 tienen prácticamente la misma precisión.



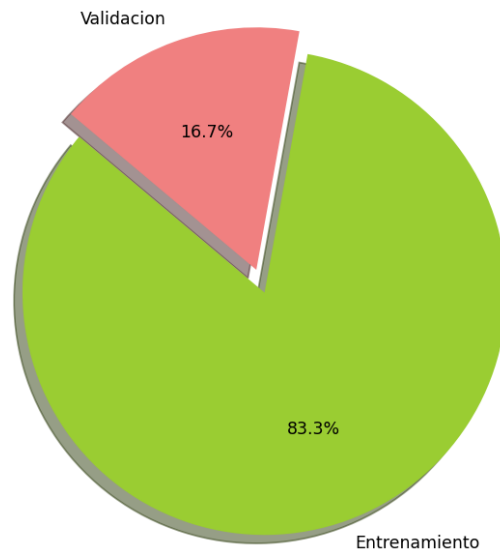
Propuestas para mejoras

Con la intención de ver si se puede mejorar el promedio, se obito por dividir los datos de manera distinta, en este caso la división de datos para el entrenamiento, `prueba y validación fue el siguiente:

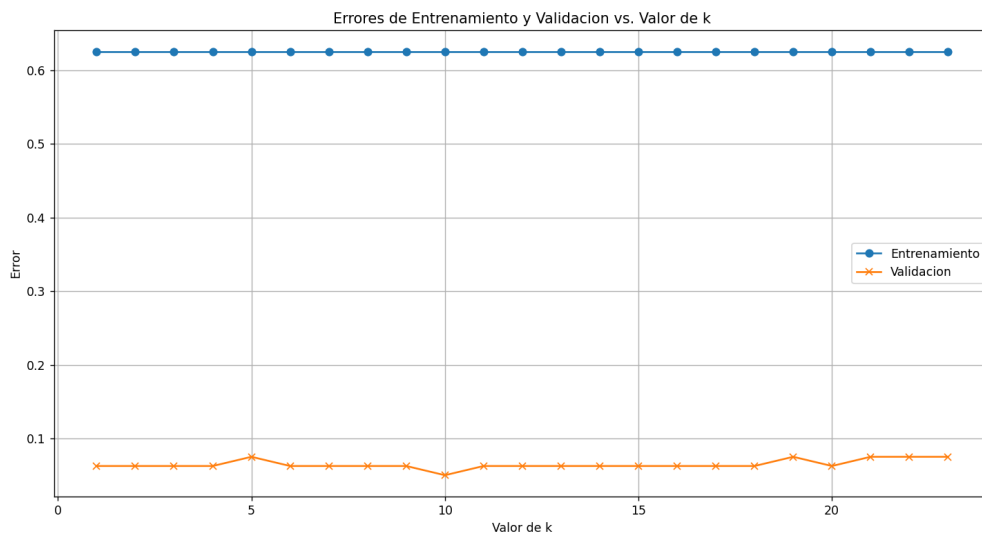
Division inicial entre datos de entrenamiento y prueba



Division entre datos de entrenamiento y validacion



En el caso de la gráfica de los errores, podemos ver que el sesgo mejorar y disminuye algo ya que se comportan todavía más similar que en el caso pasado, asimismo para mejorar aún más la percusión se puede cambiar el mejor valor de k, y en vez de usar el recomendado, usar 5.



Conclusión

El modelo KNN ha demostrado ser eficaz en la clasificación de tumores como benignos o malignos con una alta precisión. Aunque se exploraron técnicas adicionales para mejorar el rendimiento, el modelo ya está adecuadamente ajustado y presenta un rendimiento robusto.

Para futuros trabajos, se podría considerar la exploración de otros algoritmos, la selección de características o la ingeniería de características para ver si pueden ofrecer mejoras en el rendimiento.