Comparación de clasificadores

Atzi Merino

En este proyecto, se comparan las predicciones de 5 algoritmos de clasificación distintos sobre el diagnóstico de un tumor.

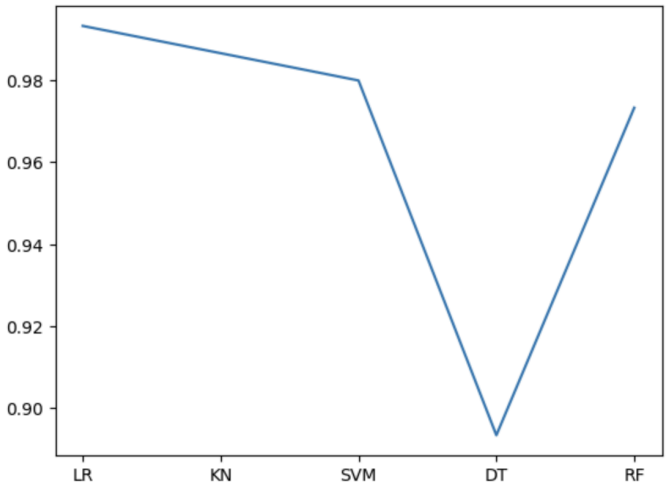
- Logistic Regression
- K-Nearest Neighbors
- Máquinas de soporte vectorial (SVM)
- Árboles de decisión
- Random Forest

La predicción cae en un caso binario, pues el tumor puede ser maligno o benigno; si no es uno de las dos, entonces es la otra.

Se hizo un programa que procesa los datos de un archivo csv y aplica los algoritmos mencionados para generar las predicciones correspondientes, luego, usando diferentes técnicas de valoración, genera resultados valiosos para concluir cuál es el mejor algoritmo.

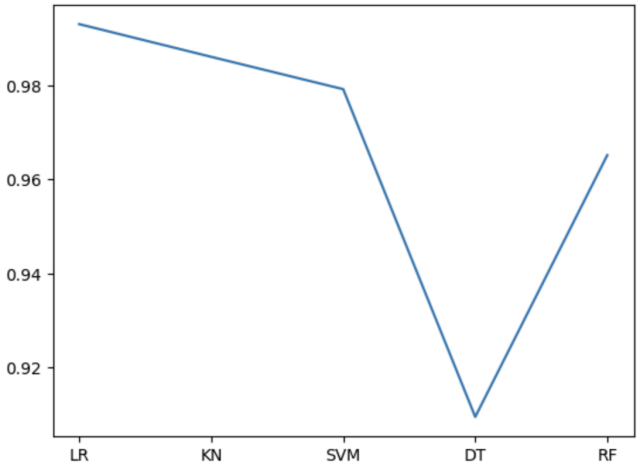
La primera técnica que usa el programa es la evaluación cruzada, donde dato por dato checa si la predicción del algoritmo es correcta, y en base a eso arroja una calificación del 0 al 1. Para la base de datos de tumores, el programa contestó que el mejor algoritmo con esta medición era Logistic Regression, con una valoración de 0.9933, es decir, que tuvo 99.333% correcto.

En el diagrama siguiente se ve la exactitud de Logistic Regression comparada al resto de algoritmos:



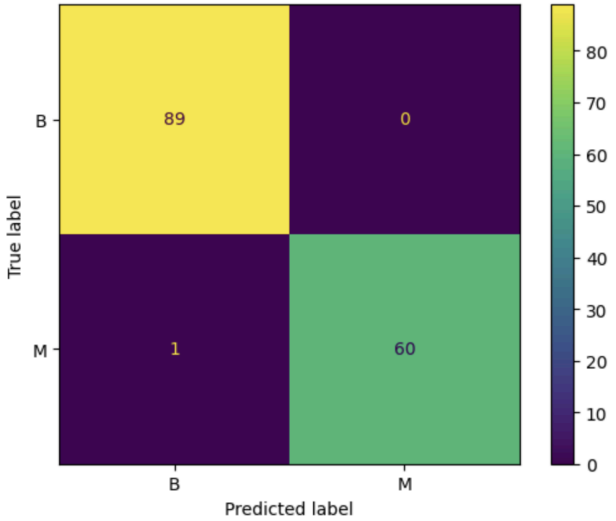
La segunda técnica usada en el programa es la valoración de F1. Para ella, se importó de la paquetería de sklearn la función ‘f1_score’. Al usarla, se debe elegir un tipo de promedio, pues esta función en particular recibe la entrada ‘average’, que se debe igualar a un promedio ‘micro’, ‘macro’ o ‘weighted’. La selección del tipo de promedio dependerá de la naturaleza de la base de datos proporcionada. Una de las cuestiones importantes para elegir el tipo de promedio es el balance que hay en los datos. Se considera desbalanceado cuando hay el doble o más de datos de un tipo que de otro. Se suele seleccionar ‘micro’ cuando hay un desbalance pequeño en los datos (existe una razón mayor que dos pero menor que 3), ‘weighted’ cuando hay un desbalance más grande, y ‘macro’ cuando no hay desbalance (todo esto lo investigué por mi cuenta). En este caso, al ser un caso binario, se consideró la razón que hay entre la cantidad de tumores malignos y la cantidad de benignos. Haciendo un apartado sencillo en el programa, se vio que la base de datos estaba bastante balanceada, pues había aproximadamente 190 de un tipo y 300 del otro. Por lo cuál, cayó en el caso de ‘average = macro’. Para la base de datos de tumores, el programa contestó que el mejor algoritmo con esta medición era Logistic Regression.

En el diagrama siguiente se ve la exactitud de Logistic Regression comparada al resto de algoritmos:

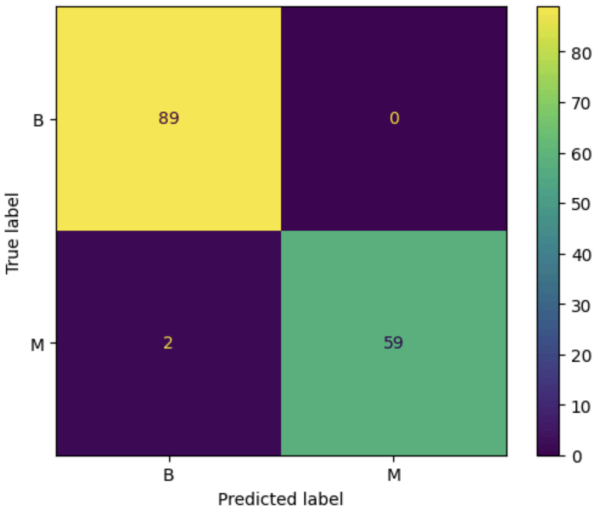


Finalmente, para visualizar mejor las predicciones de cada algoritmo, se hizo uso de la matriz de confusión. Esta contrasta las predicciones del programa, con los datos reales. Se muestran a continuación las matrices de confusión por cada algoritmo:

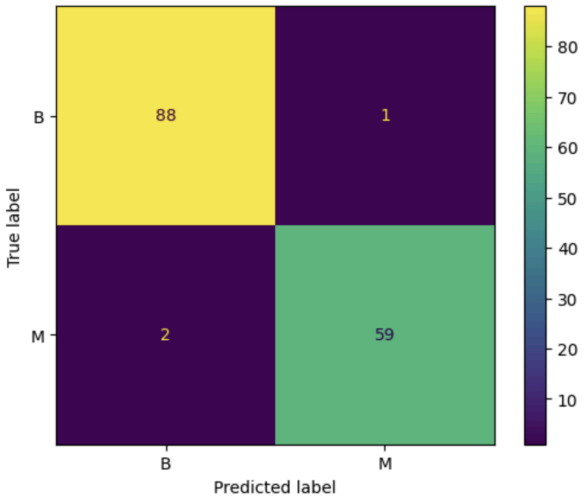
Logistic Regression



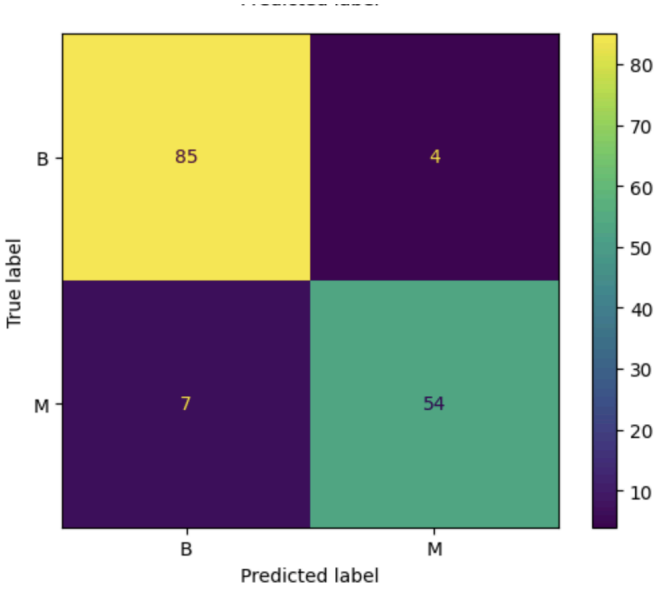
K-Nearest Neighbors



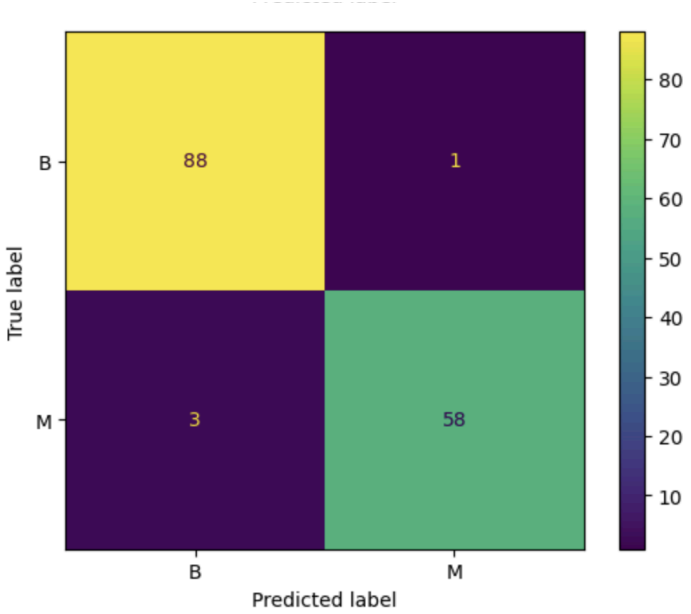
Máquinas de soporte vectorial (SVM)



Árboles de decisión



Random Forest



De las matrices de confusión, queremos fijarnos en los recuadros superior derecho e inferior izquierdo, pues son en los que se muestran los fallos del algoritmo. Nuevamente, podemos observar que Logistic Regression fue el algoritmo con menos fallos, seguido por K-Neighbors que solo tuvo una falla más.

Tomando en cuenta todas las medidas anteriores, podemos concluir que **Logistic Regression es el mejor algoritmo para predecir si un tumor es maligno o benigno** (dada nuestra base de datos). Pues en todos los casos, se demostró la exactitud de sus predicciones.

Agregado a lo anterior, adjunté un extra del algoritmo de K-Means (fue mi investigación del punto extra), que, aunque no es como tal un clasificador, se puede aplicar como uno al tratar los agrupamientos que arroja como clasificaciones de los datos. En este caso se tomó k=2 para ver qué datos incluía en el mismo grupo, y viendo si se asemejaba a los diagnósticos de los tumores. Tuvo una exactitud menor que la de los clasificadores, pero aún así logró un 84.769% de sus predicciones correctamente