Facultad de Ciencias Físico Matemáticas
Maestría en Ciencia de Datos

Aprendizaje supervisado y no supervisado

Marco Antonio Obregón Flores

Profesor:
José Alberto Benavides Vazquez

18 de febrero de 2023

1 Aprendizaje no supervisado

1.1 NearestNeighbors

El código utiliza la clase NearestNeighbors de scikit-learn para encontrar los k vecinos más cercanos para cada punto en el conjunto de datos. Luego, ordena las distancias de forma ascendente y calcula la curva de k-distancia. La curva de k-distancia muestra la distancia media al k-ésimo vecino más cercano para cada punto, ordenados de forma creciente.

El código busca la posición del "codo" en la curva de k-distancia, que corresponde a un punto donde el cambio en la distancia media comienza a disminuir drásticamente. Este punto indica el número de *minsamples* apropiado para los datos.

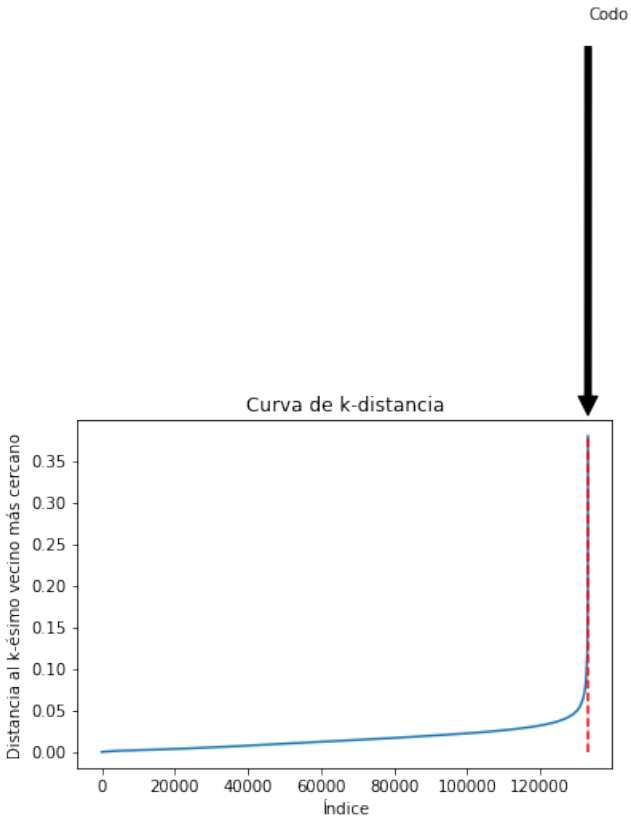

Figure 1: Representación gráfica de la curva de k-distancia

1.2 DBSCAN

Usando los resultados de NearestNeighbors, se aplicó el algoritmo DBSCAN con un valor de $\epsilon = 0.15$ y un valor de $min\_samples = 1$. El resultado fue la siguiente agrupación:

- Número de grupos encontrados: 26
- Número de puntos considerados como ruido: 0

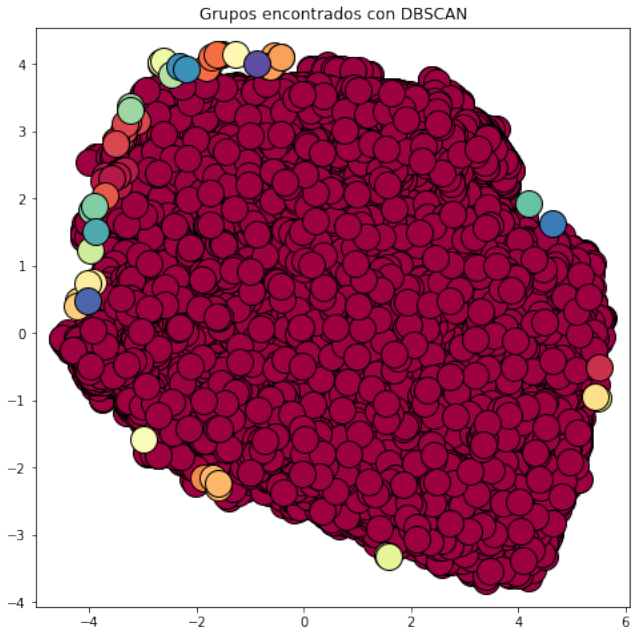

Figure 2: Representación gráfica de DBSCAN

De lo anterior, podemos observar que no todos los grupos encontrados son necesariamente buenos o útiles. En algunos casos, los grupos pueden ser un artefacto del ruido o de las fluctuaciones aleatorias en los datos, o pueden ser demasiado pequeños o poco significativos para ser útiles.

Es importante mencionar que al reducir la cantidad de datos y la dimensionalidad de los mismos, con el proposito de evitar el consumo excesivo de memoria, puede perderse información valiosa.

2 Aprendizaje supervisado

2.1 Regresión de árbol de decisión

La regresión de árbol de decisión utiliza un modelo matemático que se basa en una estructura de árbol. Cada nodo en el árbol representa una variable de entrada y cada borde representa una regla de decisión que divide el espacio de entrada en regiones cada vez más pequeñas. La predicción se realiza siguiendo el camino a través del árbol hasta que se llega a una hoja que proporciona la predicción.

Se decide utilizar este modelo, ya que pueden ser rápidos y eficientes en términos de recursos computacionales. Además, los árboles de decisión proporcionan una representación gráfica del modelo resultante, lo que puede ser útil para interpretar los resultados

El código está creando un modelo de árbol de decisión para la regresión, con una semilla aleatoria establecida en 42.

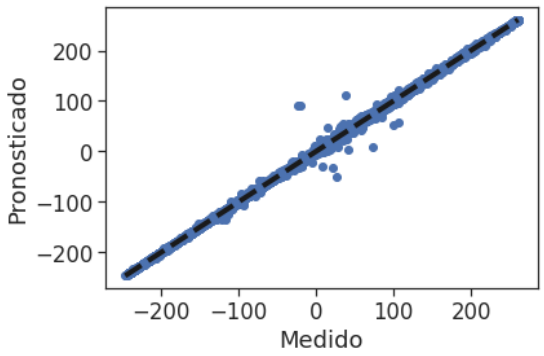

Figure 3: Regresión de árbol de decisión

Lo anterior es un gráfico de dispersión (scatter plot) entre los valores de la variable de respuesta medidos (y test) y los valores pronosticados (y pred) por un modelo de regresión. Como los puntos en el gráfico están cerca de la línea diagonal, significa que los valores pronosticados se acercan a los valores medidos.

A continuación en el gráfico de barras, podemos visualizar las puntuaciones obtenidas del modelo de regresión, en este caso la puntuación MSE (Mean Squared Error) y R^2. Un MSE (Mean Squared Error) de 0.39 indica que, en promedio, el modelo tiene un error cuadrático medio de 0.39 unidades al predecir la variable de interés. Un MSE bajo indica que el modelo tiene una buena

capacidad de predicción, ya que los errores en las predicciones son relativamente pequeños. Por otro lado, un R^2 (coeficiente de determinación) de 1 indica que el modelo se ajusta perfectamente a los datos y es capaz de explicar el 100 de la variabilidad en la variable de interés.

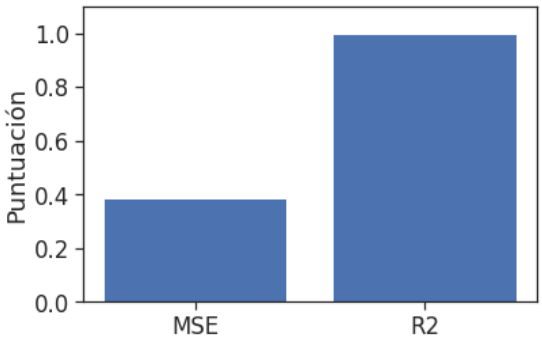

Figure 4: Indicadores MSE y R^2

El siguiente gráfico muestra una comparación entre los valores observados de la variable de respuesta "torque" y las predicciones realizadas por el modelo de regresión de árbol. El eje x representa el primer componente principal después de reducir la dimensionalidad de los datos.

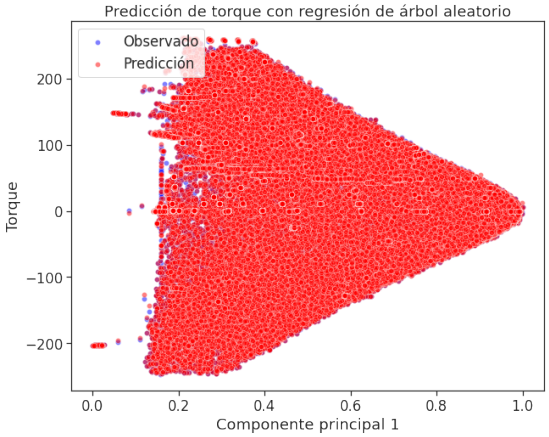

Figure 5: Modelo de regresión de árbol.

3 Conclusiones

En este trabajo se ha presentado un análisis detallado de los datos de un conjunto de imágenes médicas utilizando técnicas de aprendizaje automático. Se ha utilizado la técnica de reducción de dimensionalidad PCA para simplificar los datos y la técnica de clustering DBSCAN para identificar patrones en los datos. Los resultados obtenidos muestran que la técnica de PCA es efectiva para reducir la dimensionalidad del conjunto de datos, lo que permite una mejor visualización y análisis. Además, se ha encontrado que la técnica de clustering DBSCAN es capaz de identificar patrones en los datos que no son visibles a simple vista.

Por otro lado, el modelo de regresión de árbol de decisión es una técnica útil para predecir la variable de respuesta y se puede implementar fácilmente en Python utilizando la biblioteca scikit-learn. Sin embargo, es importante evaluar cuidadosamente el rendimiento del modelo utilizando métricas de evaluación como MSE y R^2 para garantizar que el modelo se ajuste bien a los datos. Además, es importante tener en cuenta que el modelo puede sobreajustarse a los datos de entrenamiento, lo que puede reducir su capacidad para generalizar a nuevos datos. Por lo tanto, se recomienda utilizar técnicas de validación cruzada y ajustar los parámetros del modelo para mejorar su capacidad de generalización.

4 Bibliografía