Autor: Ivonne Campoverde Fecha:23-feb-24

INFORME MODELOS DE ENTRENAMIENTO

Supuestos:

Supongamos que usted trabaja en el servicio de salud y recibe muestras que provienen de mujeres con cáncer de mama.

Los médicos han extraído características y las han anotado, su trabajo es crear un modelo que sea capaz de identificar si un paciente tiene o no cáncer.

Recordemos que un falso positivo no es tan preocupante como un falso negativo, ya que en el futuro se les hacen más pruebas a las pacientes y hay oportunidades de descubrir que estábamos en un error.

Sin embargo, un falso negativo puede llevar a que el cáncer se desarrolle sin supervisión durante más tiempo del necesario y podría llevar a daños más graves o incluso la muerte de la paciente.

Teniendo esto en cuenta, desarrolla un modelo que funcione lo mejor posible y explica qué decisiones has tomado en su elaboración y por qué.

Problemática:

Desarrollar un modelo que prediga el cáncer basada en las diferentes características que puede presentar un tumor, tomando en cuenta que los tumores pueden se benignos o malignos

Información relevante:

- **Tumor Benigno:**
  Se refiere a una afección, tumor o crecimiento que no es canceroso. Esto significa que no se propaga a otras partes del cuerpo ni invade el tejido adyacente. Algunas veces, una afección se denomina benigna para sugerir que no es peligrosa o grave, generalmente crece en forma lenta y no es dañino.
  Un tumor benigno puede crecer mucho o encontrarse cerca de vasos sanguíneos, el cerebro, nervios u órganos. Como resultado de esto, puede causar problemas localmente sin diseminarse a otra parte del cuerpo. Algunas veces, estos problemas pueden ser serios.
- **Tumor Maligno:**
  El término "malignidad" se refiere a la presencia de células cancerosas que tienen la capacidad de diseminarse a otros sitios en el cuerpo (hacer metástasis) o invadir y destruir tejidos cercanos (localmente). Estas células malignas tienden a tener un crecimiento rápido e incontrolable y no mueren de la manera normal debido a cambios en su estructura genética.

Características:

- Radio: El radio de un tumor es una medida importante que se utiliza para evaluar y caracterizar el tumor, se refiere a la distancia desde el centro del tumor hasta su borde exterior. Esta medida, junto con otras características, proporciona información valiosa sobre la naturaleza del tumor y su posible comportamiento.
- Simetría: la simetría de un tumor es otra característica importante que se evalúa en el contexto médico para determinar la naturaleza del tumor y tomar decisiones sobre diagnóstico y tratamiento.
- Tamaño del Tumor: El radio del tumor es una medida directa del tamaño del tumor. Tumores más grandes pueden indicar un mayor riesgo de malignidad.
- Seguimiento del Crecimiento: El monitoreo del radio del tumor a lo largo del tiempo puede proporcionar información sobre la velocidad de crecimiento del tumor. Un crecimiento rápido podría ser una señal de malignidad.
- Evaluación de la Invasión: El radio también puede ayudar a determinar si el tumor ha invadido tejidos circundantes. Tumores con radios mayores pueden indicar una mayor invasión.
- Caracterización del Tumor: La combinación de medidas como el radio con otras características, como la forma y el tipo celular, ayuda a los profesionales médicos a caracterizar el tumor y clasificarlo como benigno o maligno.

Modelo de Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Modelos utilizados

En los tres modelos se utilizaron como características importantes los parámetros de simetría y radio del tumor.

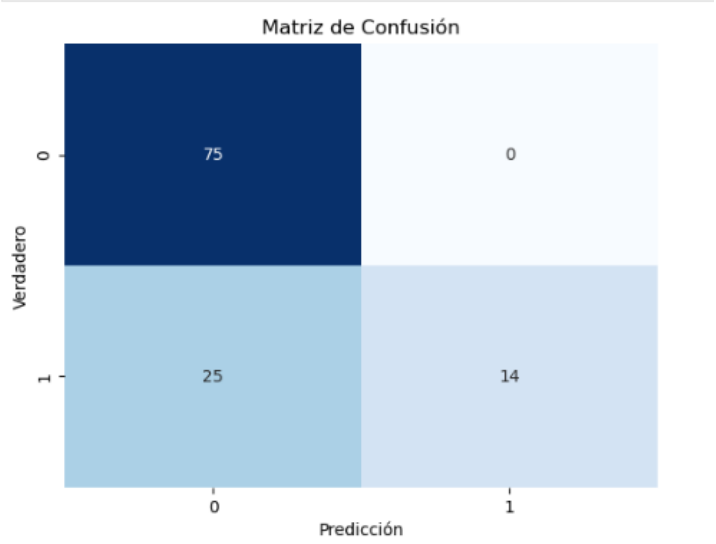
En cada modelo implementado se obtuvieron 2 resultados, ya que se ejecutaron técnicas de reducción de dimensionalidad, en el caso de regresión lineal se utilizó la técnica PCA (Principal Component Analysis) y en el caso de los otros modelos TSNE (Distributed Stochastic Neighbor Embedding)

1. **Regresión Lineal**

Resultados sin PCA

Verdaderos Positivos: 75 Falsos Negativos: 0

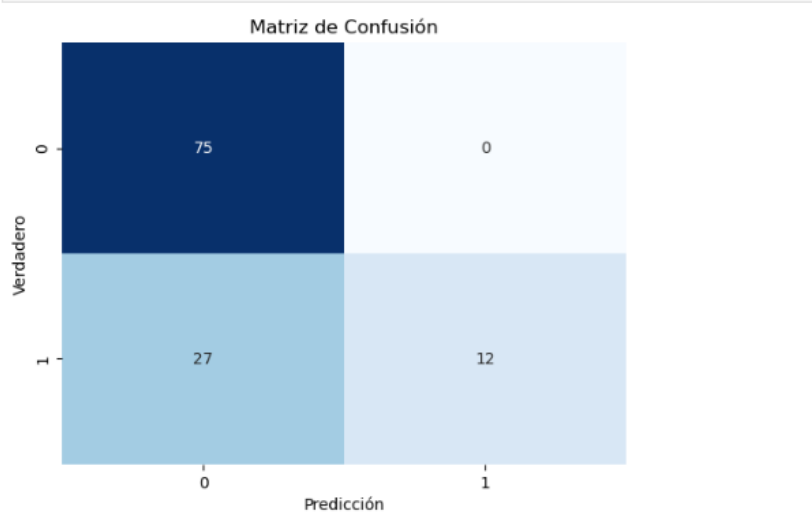
Falsos Positivos: 25 Verdaderos Negativos: 14



El modelo funciona relativamente bien, predice en un nivel más alto el número de Verdaderos positivos y en 0 los Falsos negativos, es decir que ninguno de los casos que son positivos se definieron como negativos.

Resultados con PCA

En el caso del ejercicio realizado implementado técnicas de reducción de dimensionalidad, en el caso de Verdaderos positivos y falsos negativos tenemos los mismos valores, pero aumentan los casos en ellos que se predice que se presenta un tumor maligno, cuando realmente noes el caso.



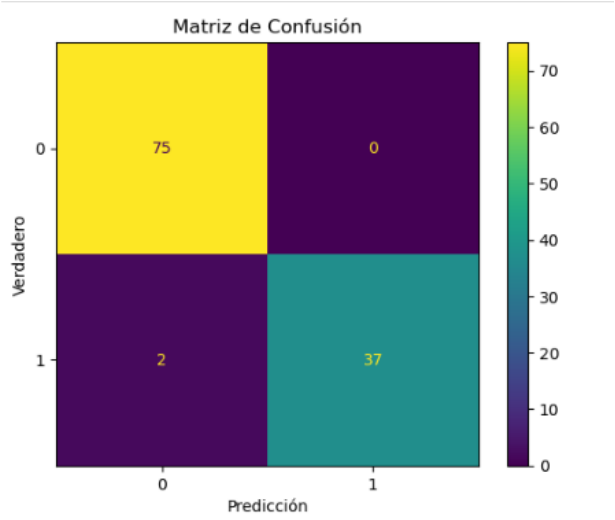
2. Random Forest

Resultados sin TSNE

Verdaderos Positivos: 75 Falsos Negativos: 0

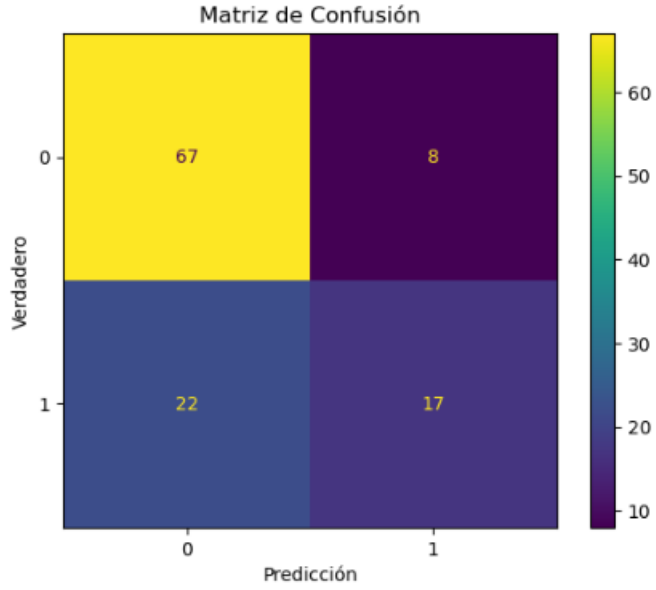
Falsos Positivos: 2 Verdaderos Negativos: 37

El modelo presenta un alto nivel de éxito tomando en cuenta que no presenta falsos negativos y el número de falsos positivos es menor.



Resultados con TSNE

Una vez aplicada la reducción de dimensionalidad el modelo presenta una disminución en el nivel de predicción.



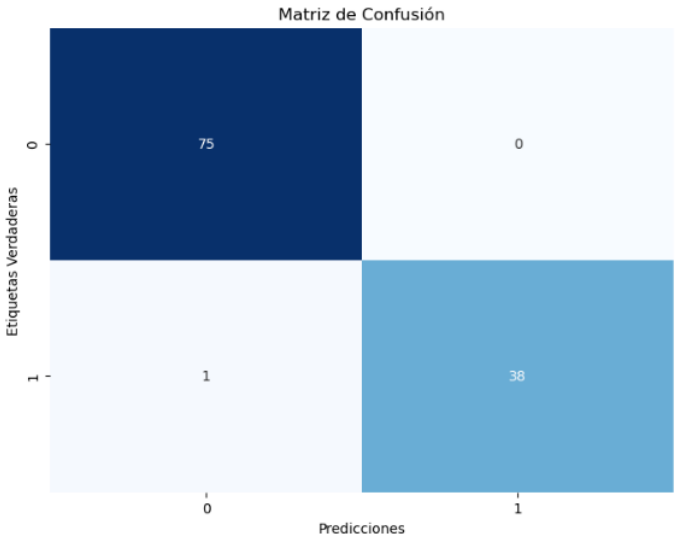
3. **XGBClassifier**

Resultados sin TSNE

Verdaderos Positivos: 75 Falsos Negativos: 0

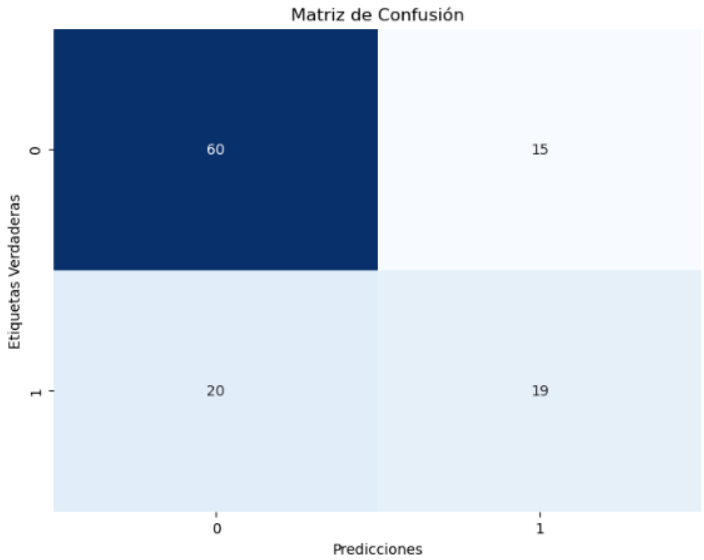
Falsos Positivos: 1 Verdaderos Negativos: 38

El modelo presenta un alto nivel de éxito tomando en cuenta que no presenta falsos negativos, el número de falsos positivos es aún menor que usando Random Forest.



Resultados con TSNE

Una vez aplicada la reducción de dimensionalidad el modelo presenta una disminución en el nivel de predicción con respecto al modelo aplicado sin reducción de dimensionalidad, pero presentando más errores tanto en falsos positivos como negativos.



Conclusiones:

- La reducción de dimensionalidad es importante en la aplicación de modelos, sin embargo, esto depende de lo que s quiera obtener como resultado, en el caso de enfermedades como el cáncer, no basta aplicar una, dos o un número limitado de dimensiones, pues al ser células que van mutando constantemente en base al sistemas en el que se encuentran, en cada cuerpo el comportamiento es diferente, es por ello que hay personas que no pueden darse cuenta de que sufren de esta enfermedad hasta que llega a etapa terminal.
- El modelo implementado con mayor porcentaje de acierto fue XGBClassifier ya que al estar basado en el modelo de árboles supera el resultado del modelo Random Forest, aunque los valores realmente no presentan un nivel de diferencia tan alto.
- Las características o parámetros de cada modelo deben adaptarse al objetivo al que queremos visualizar, tomando en cuenta que en base a estos parámetros podemos llegar a entrenar de manera errónea a los modelos haciendo que no sean viables.