Informe Predicciones Tumores de Mama

Objetivo: A partir del ingreso de características tomadas de imágenes digitalizadas de muestras de tejido predecir con alta exactitud la existencia de tumores malignos.

Método:

MachineLearning. Modelo de Regresión Logística.

La regresión logística es un modelo de aprendizaje automático utilizado, en este caso, para predecir la probabilidad de que un paciente tenga cáncer de mama en función de características específicas.

A diferencia de la regresión lineal, que se utiliza para predecir valores numéricos, la regresión logística se enfoca en problemas de clasificación binaria, como determinar si un tumor es benigno o maligno.

El modelo analiza características clínicas, como el tamaño del tumor y la uniformidad celular, y calcula una probabilidad basada en estos atributos, la que se interpreta como la probabilidad de que el tumor sea maligno.

Al utilizar la regresión logística, los médicos pueden obtener una evaluación cuantitativa del riesgo de cáncer de mama, **lo que les ayuda a tomar decisiones informadas sobre el diagnóstico y tratamiento de los pacientes**.

Entrenamiento del modelo

Para entrenar al modelo en la predicción de imágenes y evaluar su exactitud, utilizamos una base de datos de imágenes de cáncer de mama denominada Breast Cancer Wisconsin.

La base de datos "Breast Cancer Wisconsin" contiene información sobre características de células extraídas de imágenes digitalizadas de muestras de tejido mamario. El objetivo principal es predecir si una muestra es benigna (no cancerosa) o maligna (cancerosa). Cada muestra se describe mediante un conjunto de atributos.

Las características o atributos se calculan a partir de una imagen digitalizada de una aspiración con aguja fina (PAAF) de una masa mamaria. Describen características de los núcleos celulares presentes en la imagen.

Información de atributos

Se calculan diez características de valor real para cada núcleo celular:

a) radius: radio (media de las distancias desde el centro a los puntos del perímetro)

b) texture: textura (desviación estándar de los valores de la escala de grises)

c) perimeter: perímetro

d) área

e) smoothness: suavidad (variación local en las longitudes de los radios)

f) compactness : compacidad (perímetro^2 / área - 1,0)

g) concavity: concavidad (severidad de las porciones cóncavas del contorno)

h) concave points: puntos cóncavos (número de porciones cóncavas del contorno)

o) symmetry: simetría

j) fractal dimensión: dimensión fractal ("coastline approximation" - 1)

De cada característica de la lista anterior, se hicieron varias mediciones calculando los valores estadísticos: Valor medio (mean), Desviación estándar (se), peor valor (worst). Resultando en 30 características ya que por ejemplo para el radio tendríamos:

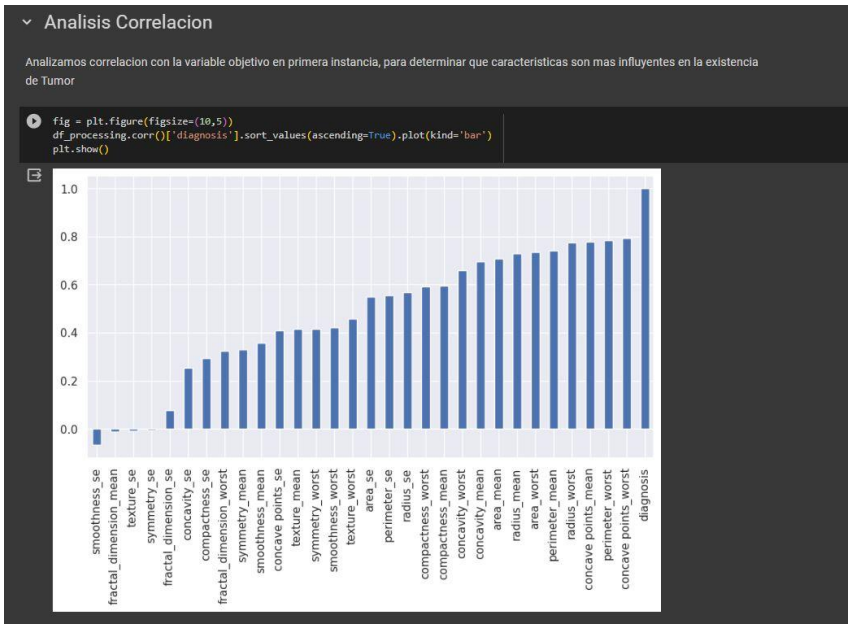
- radius_mean. Valor medio de todas las mediciones del radio
- radius_se. Desviación standard de las mediciones del radio
- radius_worst. Medición más alejada del valor medio

Link a base de datos:

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

DESARROLLO

En nuestro estudio, analizamos en primera instancia, la importancia o peso que tiene cada atributo para el diagnóstico de un tumor maligno.



Los valores “worst” son las mediciones más alejadas del valor medio. Estos valores pueden deberse a errores de medición u otros factores, que pueden perjudicar el entrenamiento de nuestro modelo.

A estos valores en la Ciencia de Datos se los llama “outliers”

Correlación

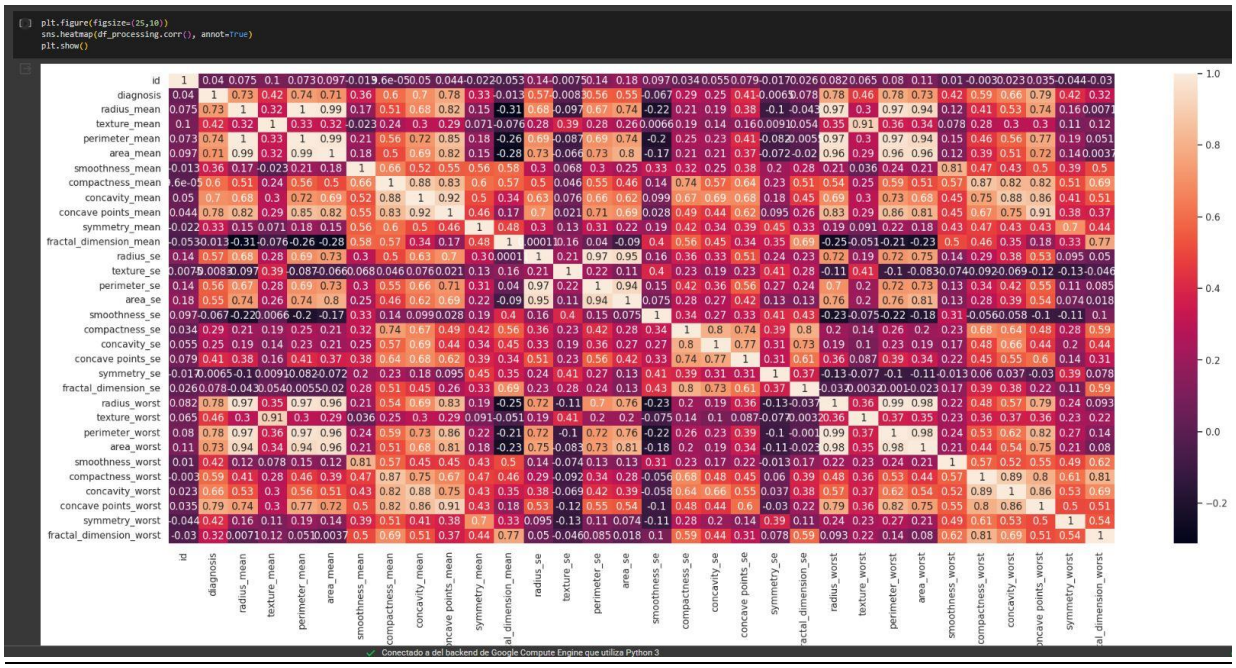
- Variables altamente correlacionadas pueden afectar el desempeño de nuestros modelos

La correlación es un concepto estadístico que describe la relación entre dos variables. En el contexto médico, esto implica cómo dos características o medidas pueden estar relacionadas entre sí. Cuando dos variables están fuertemente correlacionadas, significa que su comportamiento tiende a cambiar de manera similar.

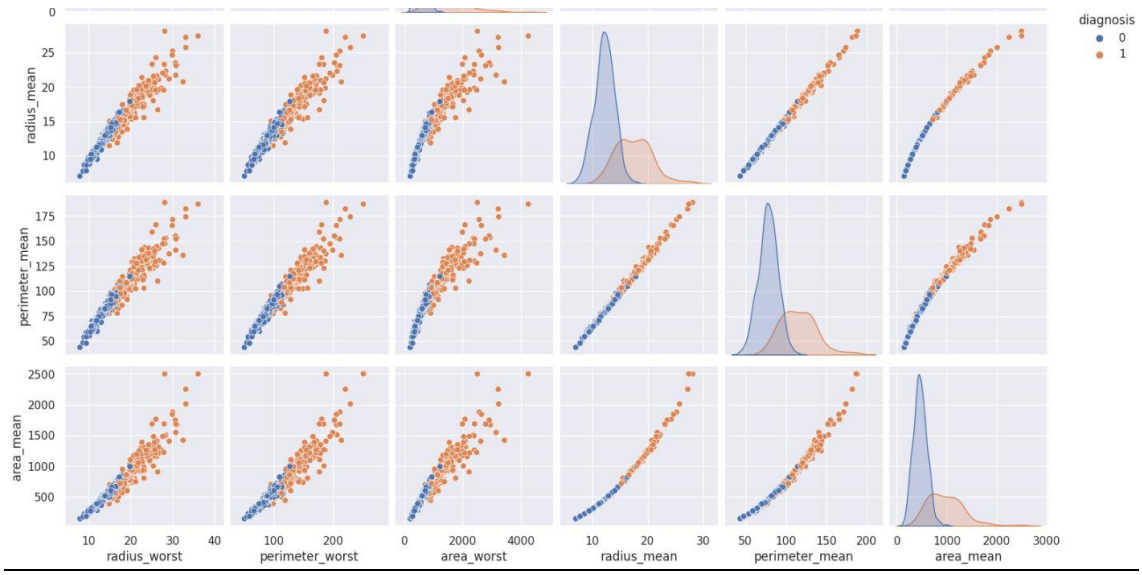
En el modelado de datos, tener variables altamente correlacionadas puede afectar la exactitud del modelo. Esto se debe a que esas variables proporcionan información redundante, lo que puede llevar a una sobrevaloración de su influencia en la predicción.

Por lo tanto*, es importante identificar y eliminar variables correlacionadas antes de construir un modelo para asegurar que se utilicen características informativas y* ***evitar posibles sesgos*** *o errores en las predicciones médicas*.

Correlación con Heat Map (Mapa de Calor)



Analisis correlacion con Grafico de pares



A partir de estos dos ultimos gráficos vemos una correlación fuerte entre los siguientes atributos: radius-mean, area_mean, perimeter_mean, radius_se, área_se, perimeter_se .

Esta correlación Lineal Fuerte entre variables perjudica el entrenamiento del modelo, por lo que hubo que tomar una decisión sobre con que variables nos quedaríamos y cuales podríamos descartar. Para ello buscamos información y asesoramiento de personas con conocimiento en la temática.

Consultando con la doctora Carol Nilda De La Torre, nos dio una primera aproximación: en caso de cáncer de mama las irregularidades de la superficie son un indicio de posible malignidad.

Como son variables geométricas del tumor fuertemente correlacionadas entre sí, decidimos quedarnos con perimeter_mean (acorde con el criterio medico) , eliminando los datos de radius_mean y área_mean.

Con respecto a los datos relacionados con la desviación estándar (radius_se, área_se) decidimos mantenerlos por ser una medida estadística de fuerte impacto.

Previamente ya habíamos decidido eliminar la mediciones worst (outliers).

Datos de entrenamiento y datos de prueba

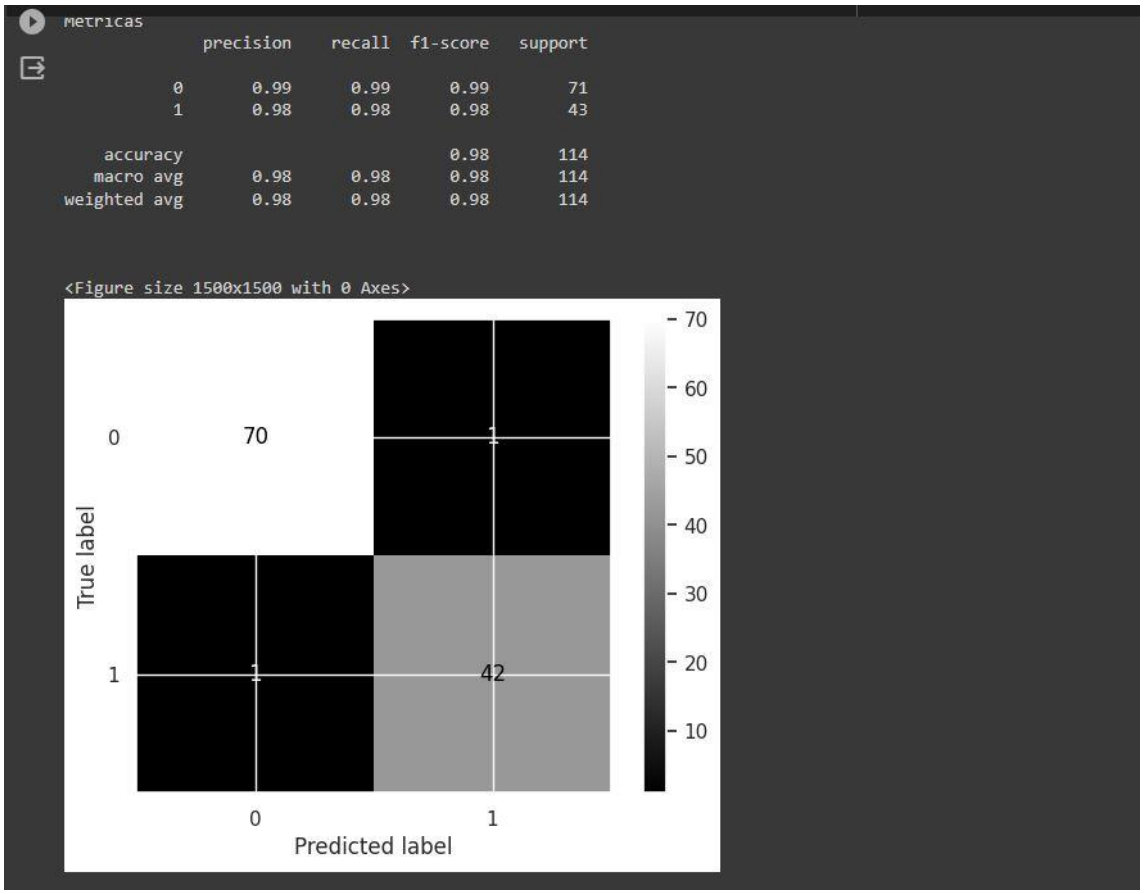
El 80% de los datos del Data set se utilizan para entrenar el modelo y el 20% restante se utilizan para evaluarlo. Es decir este 20% serian datos de entrada que el modelo desconoce y con los cuales generaremos predicciones. Y como de antemano conocemos el resultado exacto, comparamos y establecemos la exactitud del modelo

Aquí la estructura principal del código en Python:

```python
#Elimino ouliers
df.drop(["perimeter_worst","texture_worst","symmetry_worst","concavity_worst","concave points_worst","smoothness_worst"],axis = 1,inplace=True)

# Divide los datos en características (X) y etiquetas (y)
X = df.drop('diagnosis',axis=1)
y = df['diagnosis'].values

#Dividimos el set en datos de entrenmiento y testeo
X_train, X_test, y_train , y_test = train_test_split(X, y, random_state=42, shuffle=True, test_size= .2)

st_x = StandardScaler()
X_train = st_x.fit_transform(X_train)
X_test = st_x.transform(X_test)

model=LogisticRegression(max_iter=10000)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
print(confusion_matrix(y_test,predictions))
print(accuracy_score(y_test,predictions))
print("")
print("")

#Metricas
from sklearn.metrics import classification_report

report = classification_report(y_test,predictions, labels = [0,1])
print("Metricas")
print(report)

print("")
print("")


#matriaz de confucion
```

Evaluación del modelo

Con el 20% de datos que separamos para evaluar al modelo realizamos una predicción y vemos los resultados que se analizan con métricas de precisión y un gráfico denominado” Matriz de confusión”



En las métricas vemos una

- precisión 99% para la clase 0 (TUMOR NO MALIGNO)
- Precisión 98% para la clase 1 (TUMOR MALIGNO)

Matriz de Confusión:

Leyendo la matriz de manera horizontal se ven los casos reales. Entrando por cero, los tumores no malignos, entrando por 1, los tumores malignos.

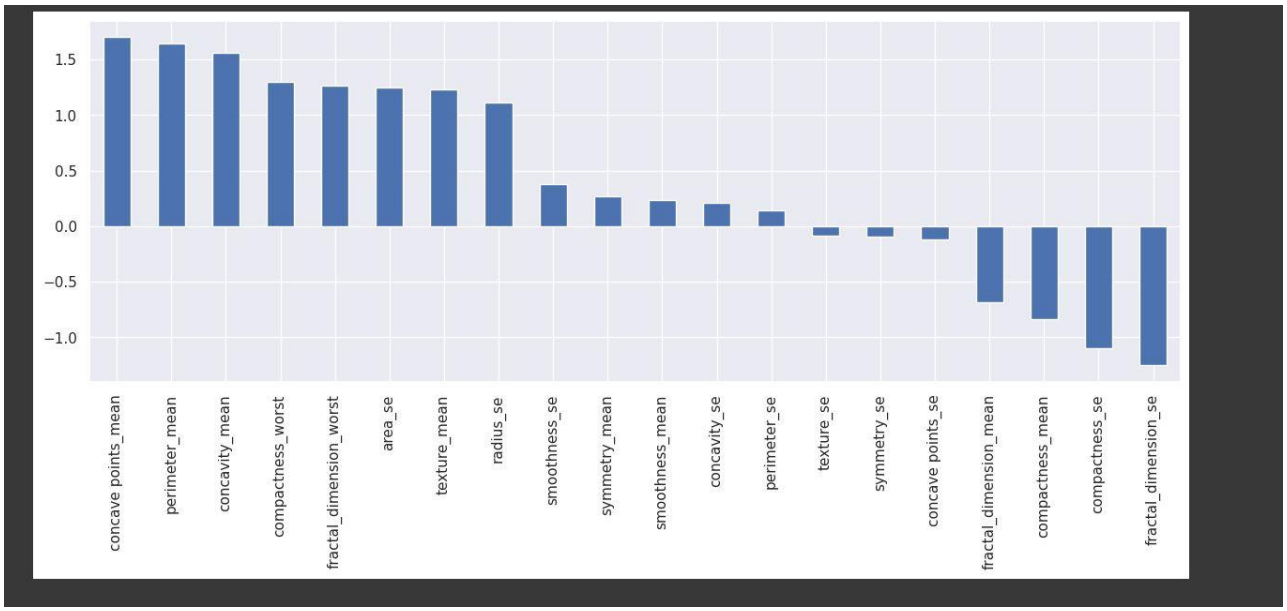
Leyendo la matriz de manera vertical, se ven los diagnósticos que hace el modelo. Entrando por cero, las predicciones de diagnósticos no malignos, entrando por 1, las predicciones de diagnósticos malignos.

Por lo tanto, entrando por cero horizontal, el valor de 70 nos indica que sobre un total de 71 casos reales, el modelo predijo 70 casos de TUMORES NO MALIGNOS en forma correcta. El valor 1 nos indica que predijo 1 caso de TUMOR NO MALIGNO, cuando en realidad era maligno.

Entrando por uno horizontal, el valor 1 nos indica que predijo 1 caso de TUMOR MALIGNO cuando en realidad era NO MALIGNO. El valor 42 nos indica que predijo 42 casos de TUMOR MALIGNO correctamente sobre un total de 43 casos reales.

Por último, en el gráfico siguiente, presentamos la influencia de cada atributo en la determinación de TUMOR MALIGNO.

Los valores positivos indican una mayor influencia en la determinación del diagnóstico maligno y viceversa.



Conclusión

Según el gráfico, los atributos relacionados con el contorno de las imágenes:

- ✓ concavity: concavidad (severidad de las porciones cóncavas del contorno)
- ✓ concave points: puntos cóncavos (número de porciones cóncavas del contorno)
- ✓ perimeter: perímetro

son los que mayor influencia tienen en el diagnóstico del tumor maligno. Esto coincide con el criterio establecido por la médica consultada.