

Curso: Data Science
Comisión: 39960
Fecha de entrega: 17 de julio de 2023
Nombre: Catalina Miranda y Basilia Alvarado

Desafío N°15: Ingeniería de atributos y selección de variables

ABSTRACT

Debido a la imposibilidad de aplicar un modelo de Machine Learning al conjunto de datos que veníamos tratando (Hot 100 del Billboard chart), decidimos cambiar el tema del proyecto final. De ahora en adelante, trataremos con un conjunto de datos de tumores mamarios, con el objetivo de clasificarlos en "benignos" o "malignos". De esta manera, el análisis a realizar será predictivo.

Los campos incluidos en el set de datos derivan de imágenes obtenidas a través de una técnica de aspiración con aguja fina. Cada uno de los atributos describe una característica del núcleo de las células obtenidas luego de la realización del estudio.

Los resultados del proyecto permitirán detectar tumores malignos que requieran de tratamientos posteriores de forma rápida y efectiva sin necesidad de realizar intervenciones quirúrgicas complejas.

PROBLEMA Y CONTEXTO COMERCIAL

Según la OMS, el cáncer de mama es el tipo de cáncer más común, siendo la principal causa de mortalidad femenina. Una de cada 12 mujeres tiene una alta probabilidad de sufrirlo en algún momento de su vida. Por tales motivos, la detección temprana de la enfermedad es el objetivo de todas las instituciones médicas del mundo. La aplicación de modelos predictivos para identificar y clasificar tumores permite aumentar la efectividad del diagnóstico, a la vez que minimiza la complejidad de los estudios necesarios para realizarlo.

Los resultados del proyecto serán de interés para hospitales y clínicas de todo el mundo que quieran mejorar la calidad de sus diagnósticos.

ANALISIS EXPLORATORIO DE DATOS (EDA)

En esta sección se realizará el análisis exploratorio de datos, conocido por sus siglas en inglés como EDA. Este proceso tiene como objetivo examinar la base de datos seleccionada para encontrar patrones, descubrir anomalías y probar hipótesis usando medidas estadísticas.

A la hora de llevar a cabo el EDA, se siguieron los siguientes pasos recomendados por el libro *"Hands-on exploratory data analysis with Python"* de Suresh Kumar Mukhiya y Usman Ahmed.

1. Definición del problema;
2. Recolección de datos;
3. Carga de datos en el formato deseado;
4. Procesamiento de datos;
5. Limpieza de datos;
6. Análisis de datos. Es el EDA propiamente dicho, que incluye la aplicación de conceptos de estadística descriptiva para conocer más a fondo los datos.

A continuación, se detallarán cada una de las etapas mencionadas previamente.

1. DEFINICIÓN DEL PROBLEMA

A partir de los atributos con los cuales se cuenta, será necesario identificar aquellos de mayor importancia en la identificación de tumores malignos. Se usará un modelo supervisado de clasificación.

2. RECOLECCIÓN DE DATOS

El conjunto de datos está formado por un archivo llamado "datos" que contiene 12 columnas, de las cuales 2 son variables de tipo categórico. A continuación, se enumeran los campos del archivo:

1. ID: identificador de la muestra extraída al paciente
2. Diagnóstico: resultado del análisis. Solo toma dos valores: maligno o benigno.
3. Radio: de la muestra extraída
4. Textura: de la muestra extraída
5. Perímetro: de la muestra extraída
6. Área: de la muestra extraída
7. Suavidad: de la muestra extraída
8. Compactibilidad: de la muestra extraída
9. Concavidad: de la muestra extraída
10. Puntos cóncavos: de la muestra extraída
11. Simetría: de la muestra extraída
12. Dimensión fractal: de la muestra extraída

Los atributos 3 – 12 son los resultantes de analizar las imágenes obtenidas en el estudio.

Link de descarga del conjunto de datos seleccionado:

<https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

3. OBTENCIÓN DE DATOS EN PYTHON

El archivo fue cargado en Google Colab como un Pandas Dataframe, a partir de una carpeta de Google Drive.

4. PROCESAMIENTO DE DATOS

Incluye las etapas de estructuración y ordenamiento de los datos para que su posterior análisis sea más sencillo. Dado que se cuenta únicamente con un DataFrame, no fue necesario realizar ninguna reestructuración de los datos. Las variables cuantitativas están todas en milímetros.

Tabla 1. DataFrame

| Columna | Tipo de variable | Tipo de dato | Descripción |
|-------------------|------------------|--------------|---|
| ID | Cualitativa | int64 | Es la identificación de cada muestra obtenida. |
| Diagnóstico | Cualitativa | string | Es la clasificación del tumor (benigno o maligno). |
| Radio | Cuantitativa | float64 | Radio de la muestra, calculado como la media de las distancias del centro del tumor a los puntos del perímetro. |
| Textura | Cuantitativa | float64 | Textura de la muestra, considerada como la desviación estándar del color con respecto a una escala de grises. |
| Perímetro | Cuantitativa | float64 | Perímetro de la muestra. |
| Área | Cuantitativa | float64 | Área de la muestra. |
| Suavidad | Cuantitativa | float64 | Se considera como la variación local del radio. |
| Compactibilidad | Cuantitativa | float64 | Es el perímetro al cuadrado, dividido el área menos 1. |
| Concavidad | Cuantitativa | float64 | Es la severidad de las porciones cóncavas del contorno de la muestra. |
| Puntos cóncavos | Cuantitativa | float64 | Es la cantidad de porciones cóncavas del contorno de la muestra. |
| Simetría | Cuantitativa | float64 | Simetría de la muestra |
| Dimensión fractal | Cuantitativa | float64 | Exponente que da cuenta de cuán completamente parece llenar un fractal el espacio conforme se amplía el primero hacia escalas más y más finas |

Para cada uno de los atributos de la tabla se tomar el valor medio, la desviación estándar, identificada como se, y el peor valor registrado.

5. LIMPIEZA DE DATOS

Consiste en la eliminación de *outliers* y en el tratamiento de datos ausentes.

DATOS AUSENTES

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   diagnóstico                               569 non-null    int64
1   radio medio                               569 non-null    float64
2   textura media                             569 non-null    float64
3   perímetro medio                           569 non-null    float64
4   área media                                569 non-null    float64
5   suavidad media                            569 non-null    float64
6   compactibilidad media                     569 non-null    float64
7   concavidad media                          569 non-null    float64
8   puntos_cóncavos medios                    569 non-null    float64
9   simetría media                            569 non-null    float64
10  dimensión_fractal media                    569 non-null    float64
11  radio se                                   569 non-null    float64
12  textura se                                 569 non-null    float64
13  perímetro se                               569 non-null    float64
14  área se                                    569 non-null    float64
15  suavidad se                               569 non-null    float64
16  compactibilidad se                         569 non-null    float64
17  concavidad se                             569 non-null    float64
18  puntos_cóncavos se                        569 non-null    float64
19  simetría se                               569 non-null    float64
20  dimensión_fractal se                      569 non-null    float64
21  peor radio                                569 non-null    float64
22  peor textura                              569 non-null    float64
23  peor perímetro                            569 non-null    float64
24  peor área                                 569 non-null    float64
25  peor suavidad                             569 non-null    float64
26  peor compactibilidad                      569 non-null    float64
27  peor concavidad                          569 non-null    float64
28  peores puntos_cóncavos                    569 non-null    float64
29  peor simetría                             569 non-null    float64
30  peor dimensión_fractal                    569 non-null    float64
31  Unnamed: 32                               0 non-null     float64

```

Tal como se observa en la figura 1, ninguna de las variables tiene valores ausentes.

OUTLIERS

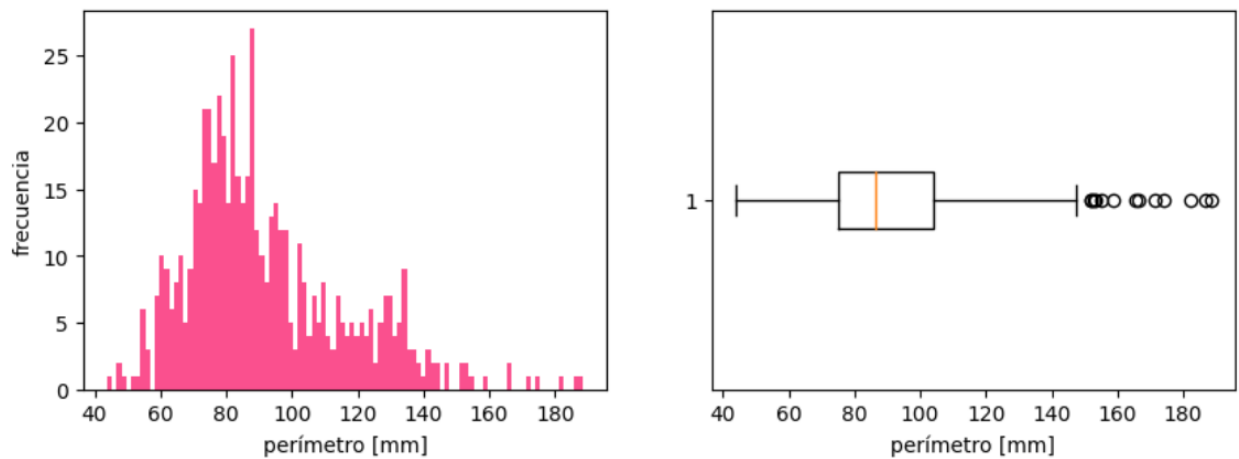
Las variables presentan todas distribuciones similares. No podrían considerarse distribuciones normales, ya que se observa un ligero sesgo hacia la izquierda, y en algunos casos un leve pico secundario. Se comprueba que la moda es menor que la mediana y que la media.

Los valores que se alejan del *inlier* izquierdo no se considerarán outliers y se mantendrán dentro del conjunto de datos. En los histogramas no hay grandes valores atípicos notables que justifiquen una mayor limpieza.

| | diagnóstico | radio | textura | perímetro | área | suavidad | compactibilidad | concavidad | puntos_cóncavos | simetría | dimensión_fractal |
|-------|-------------|------------|------------|------------|-------------|------------|-----------------|------------|-----------------|------------|-------------------|
| count | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 |
| mean | 0.372583 | 14.127292 | 19.289649 | 91.969033 | 654.889104 | 0.096360 | 0.104341 | 0.088799 | 0.048919 | 0.181162 | 0.062798 |
| std | 0.483918 | 3.524049 | 4.301036 | 24.298981 | 351.914129 | 0.014064 | 0.052813 | 0.079720 | 0.038803 | 0.027414 | 0.007060 |
| min | 0.000000 | 6.981000 | 9.710000 | 43.790000 | 143.500000 | 0.052630 | 0.019380 | 0.000000 | 0.000000 | 0.106000 | 0.049960 |
| 25% | 0.000000 | 11.700000 | 16.170000 | 75.170000 | 420.300000 | 0.086370 | 0.064920 | 0.029560 | 0.020310 | 0.161900 | 0.057700 |
| 50% | 0.000000 | 13.370000 | 18.840000 | 86.240000 | 551.100000 | 0.095870 | 0.092630 | 0.061540 | 0.033500 | 0.179200 | 0.061540 |
| 75% | 1.000000 | 15.780000 | 21.800000 | 104.100000 | 782.700000 | 0.105300 | 0.130400 | 0.130700 | 0.074000 | 0.195700 | 0.066120 |
| max | 1.000000 | 28.110000 | 39.280000 | 188.500000 | 2501.000000 | 0.163400 | 0.345400 | 0.426800 | 0.201200 | 0.304000 | 0.097440 |

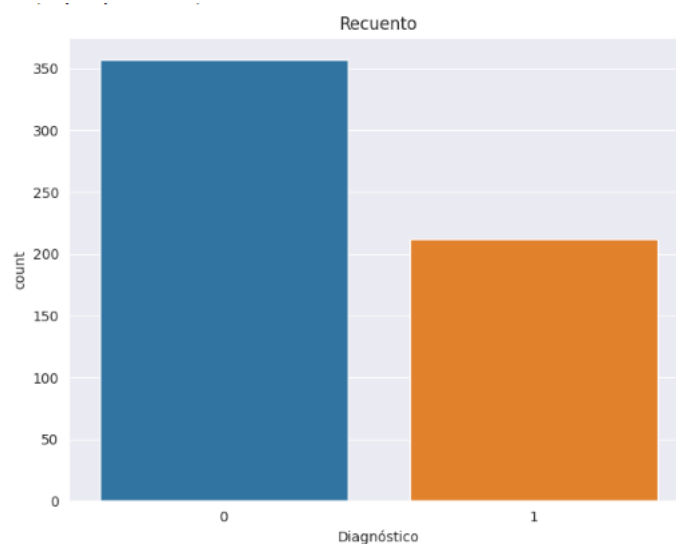
Figura 1. Resumen de las características estadísticas de los datos.

Las variables compactibilidad, concavidad, puntos_cóncavos y simetría se toman en una escala que varía entre 0 y 1, siendo cero nada y 1 todo. Por ejemplo, 0 simetría significa que es totalmente asimétrico, mientras que 1 de simetría indica que es totalmente simétrico.



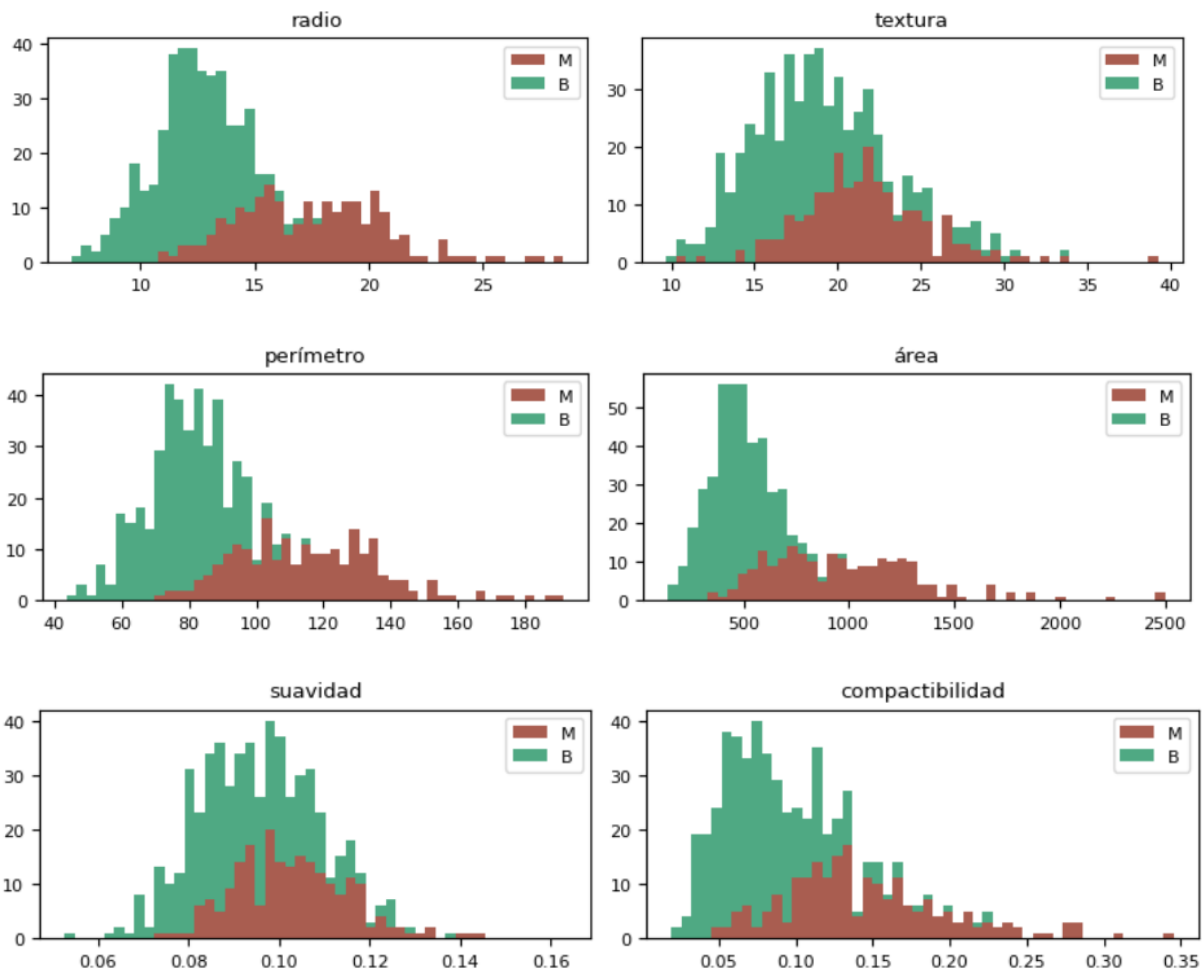
6. ANÁLISIS DE DATOS

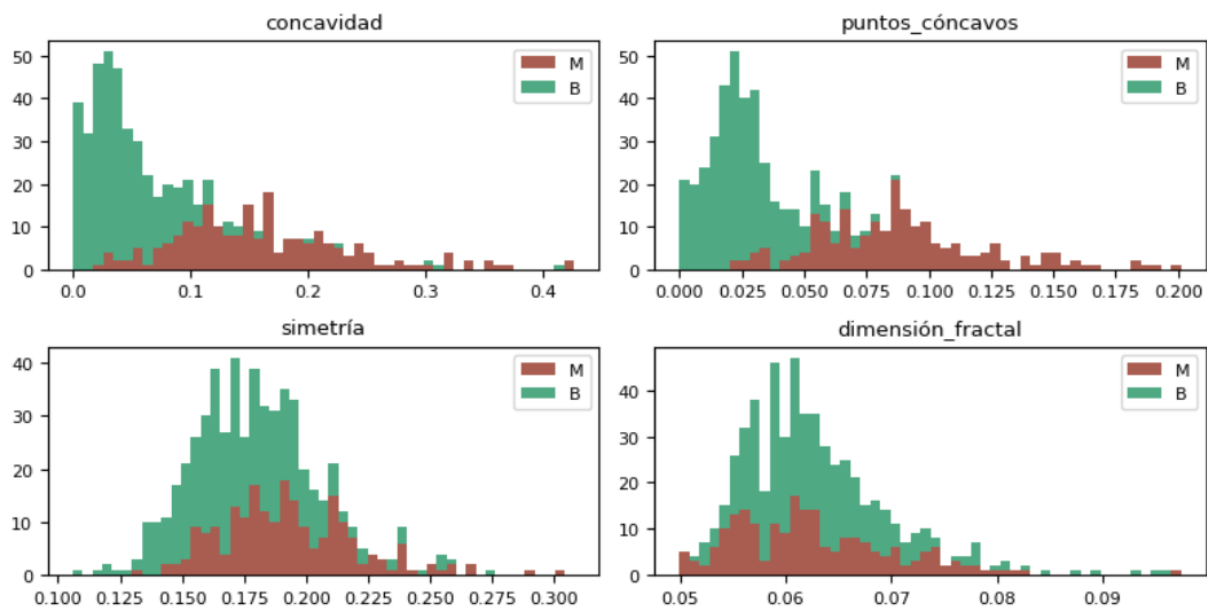
En primer lugar, se realizó un recuento de los tumores analizados



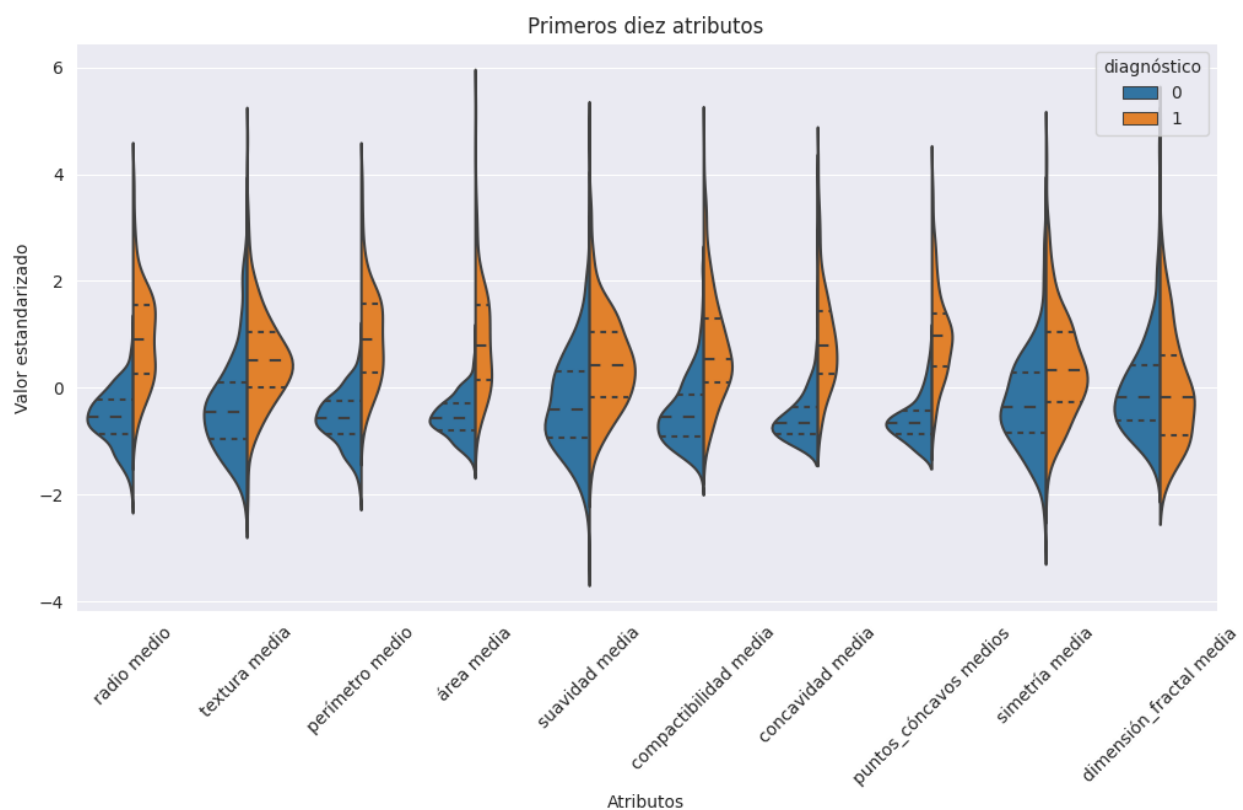
Para conocer la relación entre las variables cuantitativas y el diagnóstico del tumor se realizaron histogramas en función de si el tumor es benigno o maligno. El análisis de la distribución tomando como tercera variable al diagnóstico muestra que fue una decisión acertada no eliminar los datos que parecían outliers del análisis univariado, ya que cada pico observado corresponde a un tipo de tumor diferente.

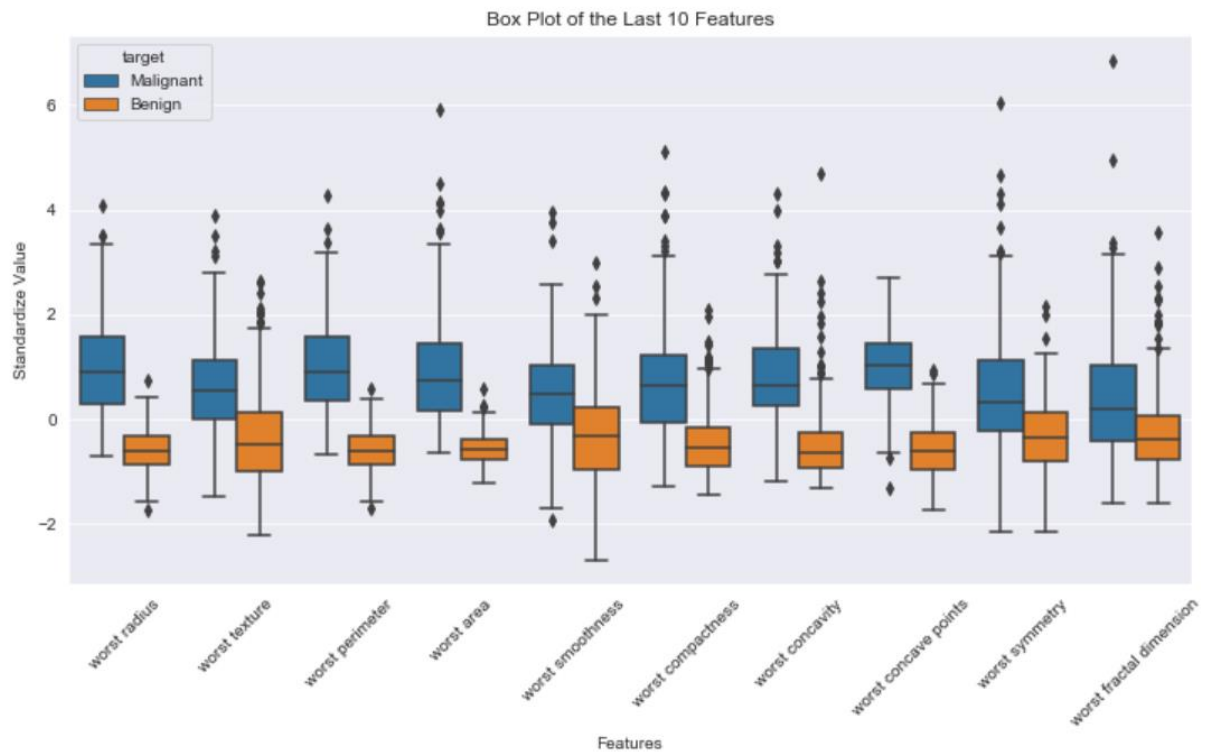
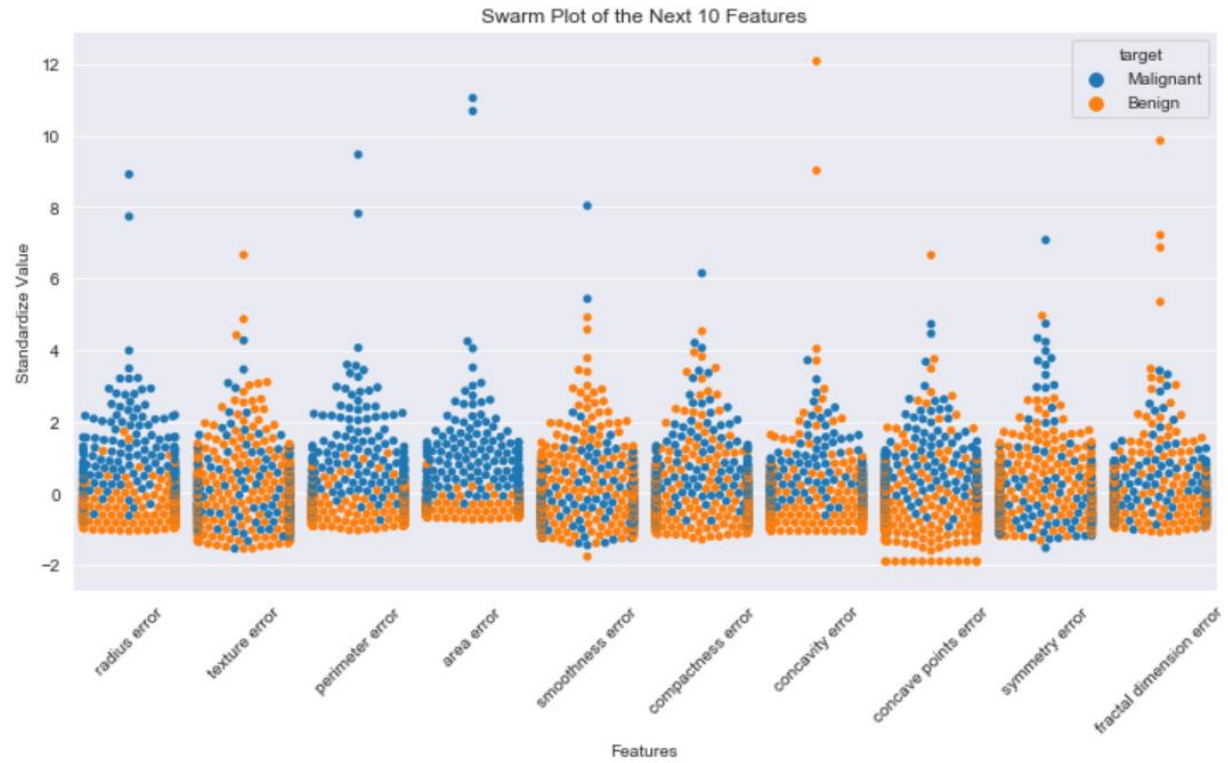
Los valores medios de radio celular, perímetro, área, concavidad y puntos cóncavos pueden usarse en la clasificación del cáncer, ya que difieren en función del tipo de tumos. Los valores más grandes de estos parámetros tienden a mostrar una correlación con los tumores malignos. Por otro lado, los valores medios de textura, suavidad, simetría y dimensión fraccional no muestran una preferencia particular de un diagnóstico sobre el otro.





Por otro lado, se hicieron gráficos de violín, de enjambre y de caja para analizar la correlación entre los datos.

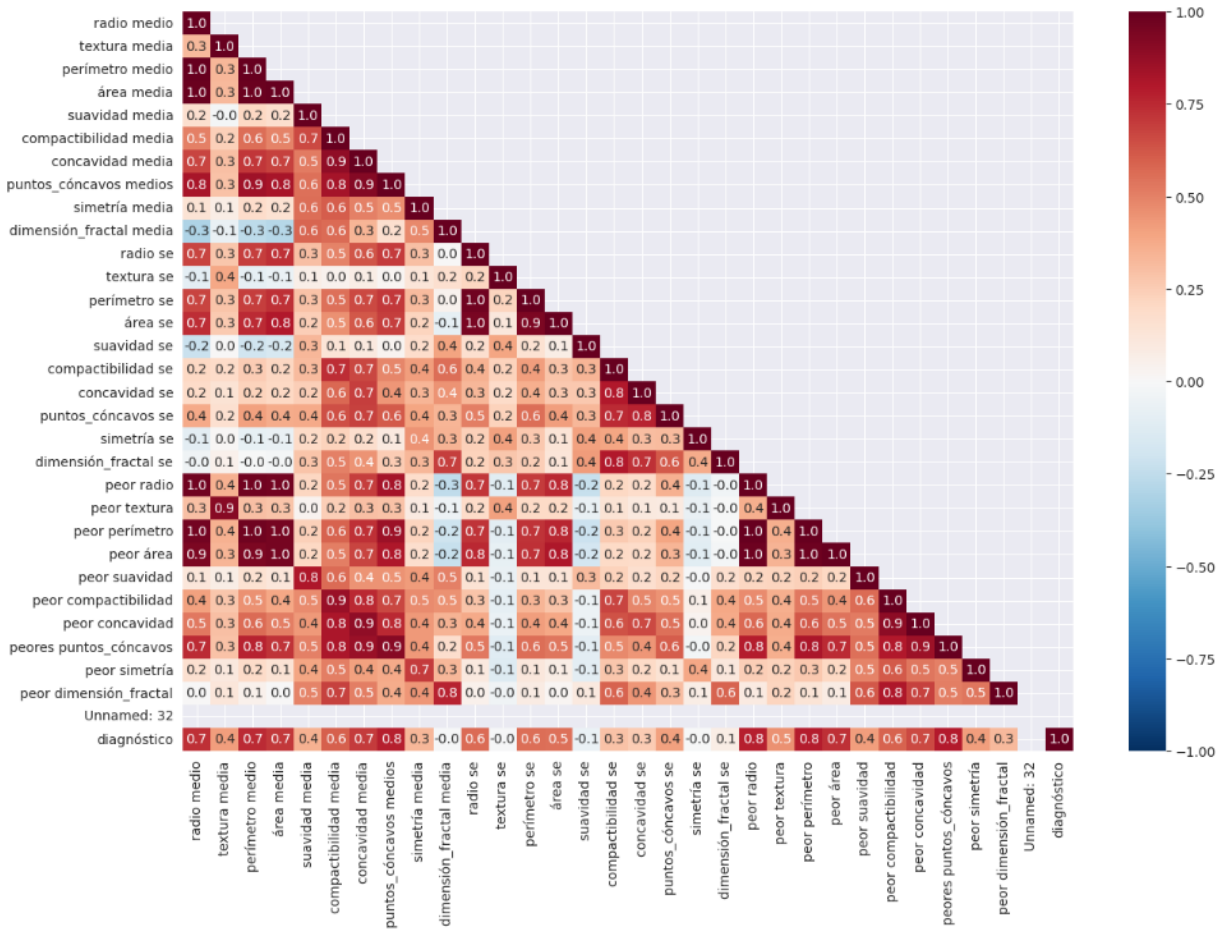


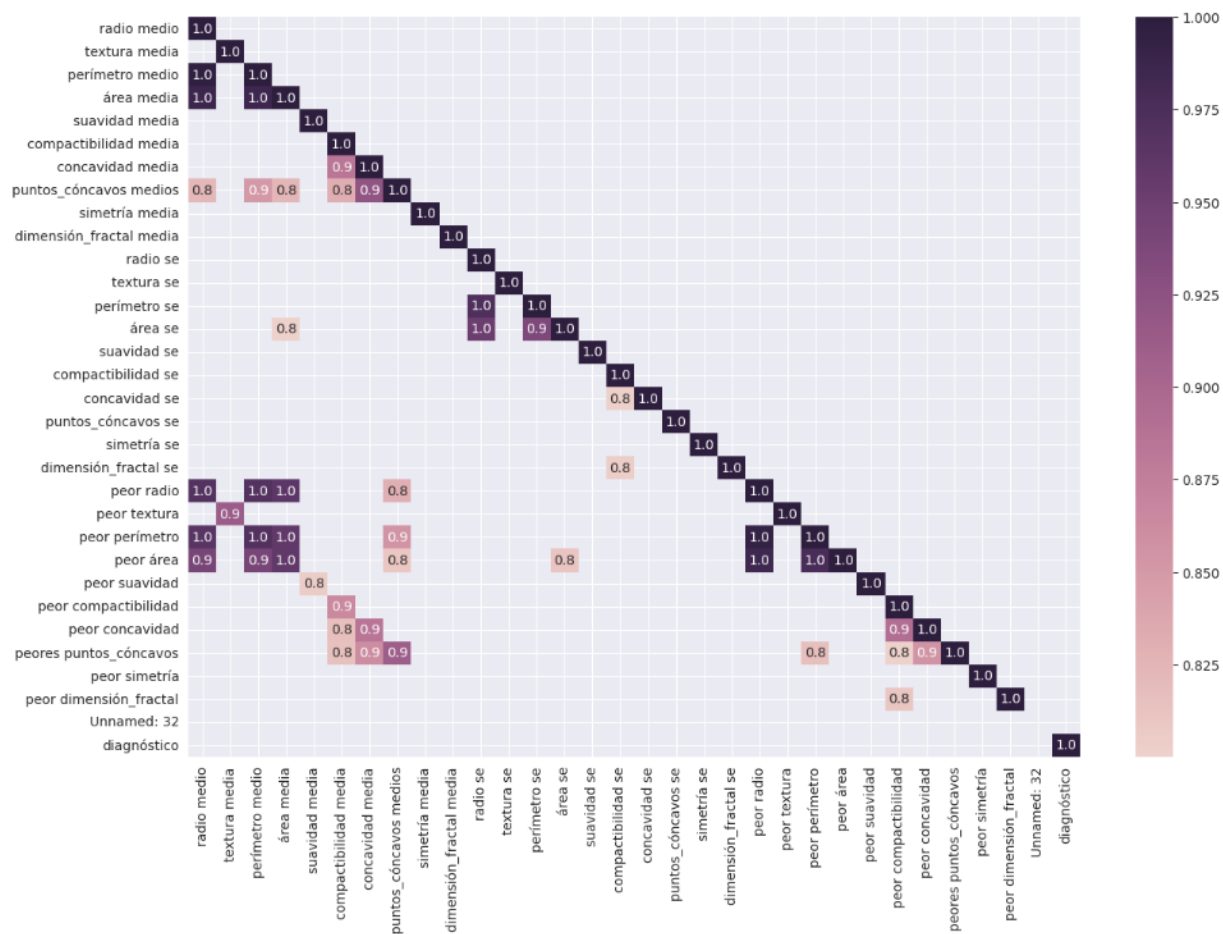


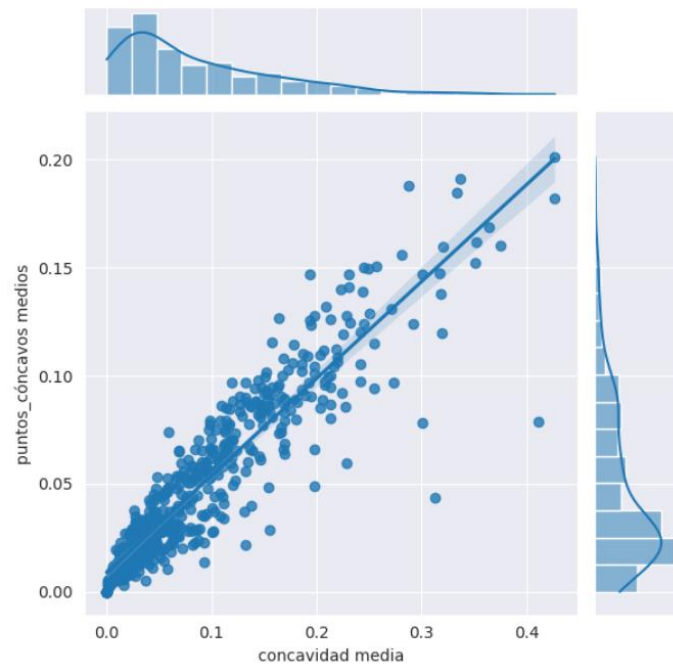
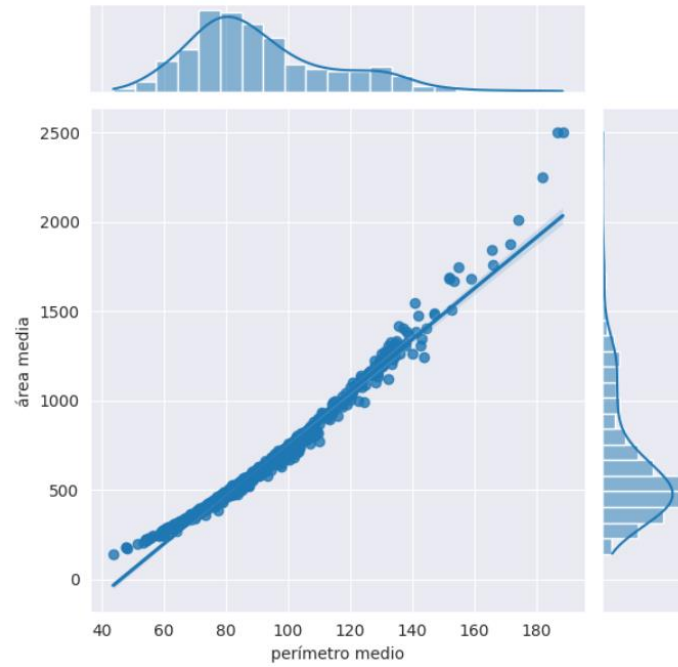
El diagrama de violín es muy eficiente para comparar distribuciones de diferentes variables, ya que en él la clasificación se vuelve clara Finalmente, los diagramas de caja son útiles para comparar la mediana y detectar valores atípicos.

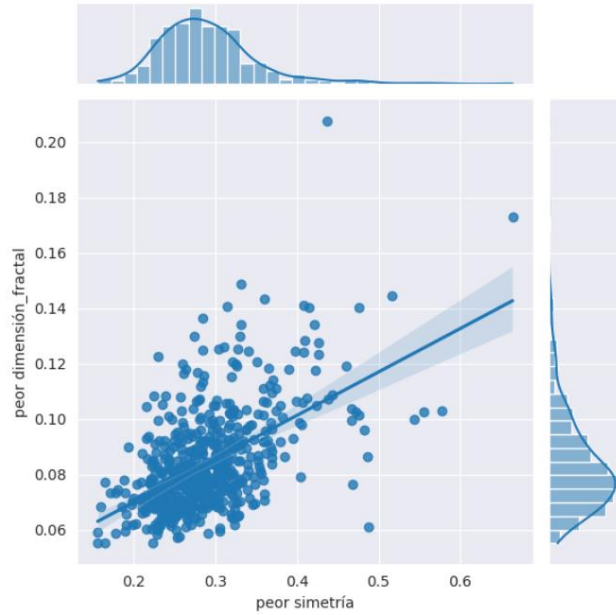
De los gráficos anteriores se pueden extraer algunas conclusiones:

- Las medianas de algunas características son muy diferentes entre "malignas" y "benignas". Esta separación se puede ver claramente en los diagramas de caja. Las siguientes pueden ser muy buenas características para el modelo de clasificación: radio medio, área media, puntos cóncavos medios, peor radio, peor perímetro, peor área, peores puntos cóncavos.
- Hay distribuciones que parecen similares entre "malignas" y "benignas". Por ejemplo: suavidad media, simetría media, dimensión fractal media, error de suavidad. Estas características son débiles en la clasificación de datos.
- Algunas características tienen distribuciones similares, por lo que pueden estar altamente correlacionadas entre sí. Por ejemplo: perímetro medio vs. área media, concavidad media vs. puntos cóncavos medios y peor simetría vs. peor dimensión fractal. A continuación se analizan las correlaciones.









INGENIERÍA DE ATRIBUTOS

La ingeniería de atributos se realizó usando el análisis de componentes principales (PCA). Se obtuvo la siguiente información:

The five selected features are: `['perímetro medio', 'área media', 'área se', 'peor perímetro', 'peor área']`



IMPLEMENTACIÓN DEL MODELO

Se aplicó un modelo de SVM con los siguientes resultados:

```
Matriz de confusión:
[[116  5]
 [  5 62]]
```

Reporte:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.96 | 0.96 | 0.96 | 121 |
| 1 | 0.93 | 0.93 | 0.93 | 67 |
| accuracy | | | 0.95 | 188 |
| macro avg | 0.94 | 0.94 | 0.94 | 188 |
| weighted avg | 0.95 | 0.95 | 0.95 | 188 |

CONCLUSIÓN

En la primera parte de este proyecto se realizó un análisis exploratorio de datos para comprender cada una de las 30 características originales y cómo podrían estar asociadas con el cáncer.

A continuación, se seleccionaron las 5 mejores características para el modelo usando la selección de características univariante. Se utilizó PCA para encontrar los dos componentes principales y crear una visualización basada en estas dos variables. La visualización muestra que con solo dos variables, se pueden separar claramente los datos entre cáncer y no cáncer. Finalmente, se aplicó nuevamente un modelo Support Vector Machines para predecir el cáncer basado en PCA. La tasa de precisión de este modelo fue del 95%.