Proyecto final - Data Science comisión 39960

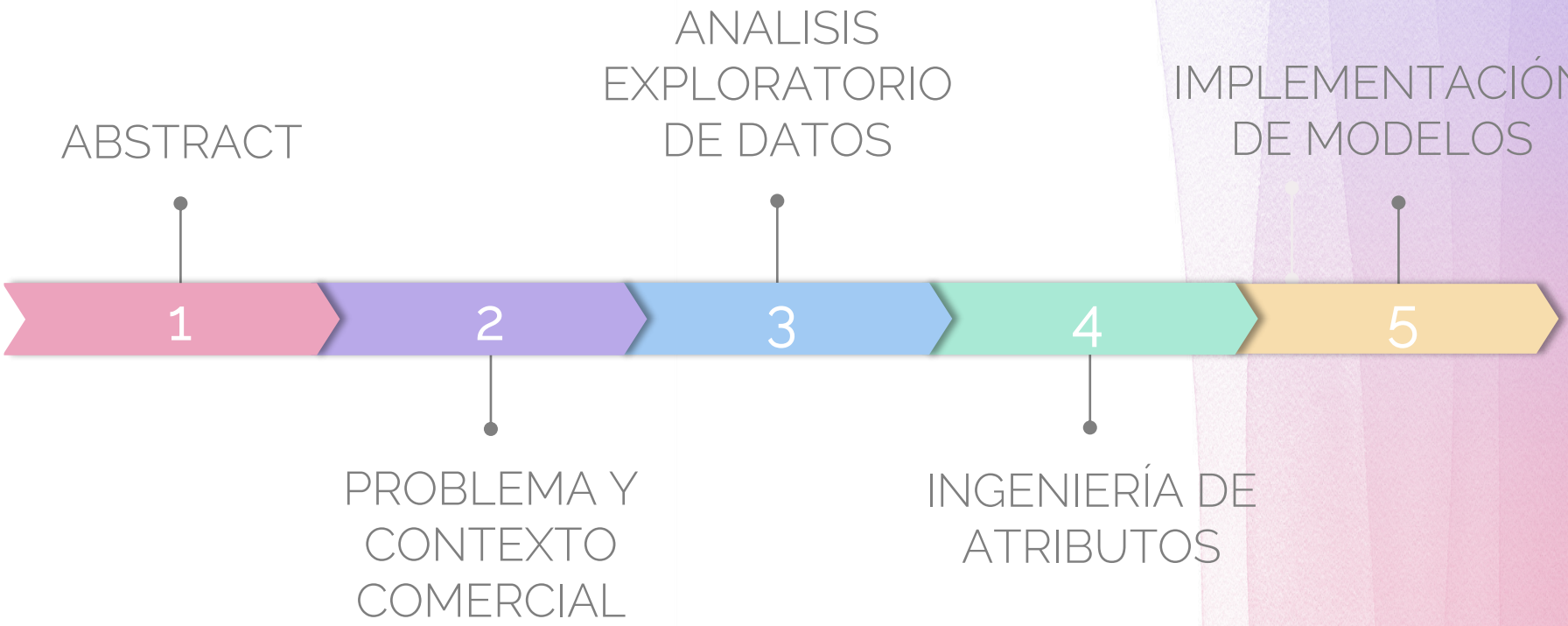
CODERHOUSE

Entrenamiento y optimización de modelos de Machine Learning para la detección precoz de tumores mamarios

26 de agosto de 2023

Catalina Miranda y Basilia Alvarado

GUÍA DE CONTENIDO



ABSTRACT

1

ABSTRACT

El cáncer de mama es el tipo de cáncer más frecuente en las mujeres. Solamente en 2020, en el mundo, se diagnosticaron más de 2,2 millones de casos y alrededor de 685 mil personas fallecieron como consecuencia de dicha enfermedad.

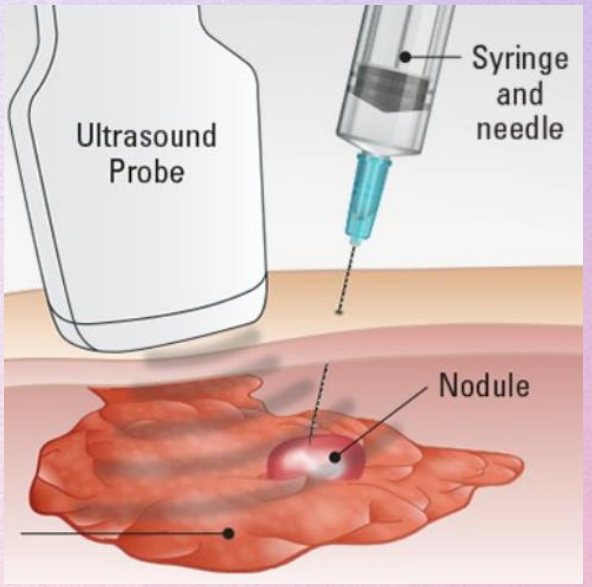
LA DETECCIÓN PRECOZ ES LA MEJOR MANERA DE PREVENIR Y/O TRATAR LA ENFERMEDAD

Entre las técnicas de detección precoz menos invasivas se encuentra la biopsia por aspiración con aguja fina (FNA), que permite obtener una serie de imágenes del tumor para determinar la necesidad de aplicar terapias o descartar la presencia de la enfermedad.

En el presente proyecto se aplicarán modelos de Machine Learning para clasificar en “benignos” o “malignos” a los resultados del estudio de FNA. De esta manera, el análisis a realizar será predictivo.

ABSTRACT

- El modelo se realizará a partir de los atributos de los tumores observados en las imágenes obtenidas a través de FNA.
- Cada uno de los atributos describe una característica del nobtenidas 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.
- La pregunta por responder será: ¿el tumor que posee el paciente es benigno o maligno?



PROBLEMA Y CONTEXTO COMERCIAL

2

ABSTRACT

- La detección temprana del cáncer de mama es uno de los principales objetivos 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 médicos, hospitales y clínicas de todo el mundo que quieran mejorar la calidad de sus diagnósticos.
- Se aplicarán modelos de aprendizaje supervisado, como Random Forest Classifier (RFC) y Support Vector Machine (SVM) y de aprendizaje no supervisado, como principal Component análisis (PCA).

ANÁLISIS EXPLORATORIO DE DATOS

3

ANÁLISIS EXPLORATORIO DE DATOS

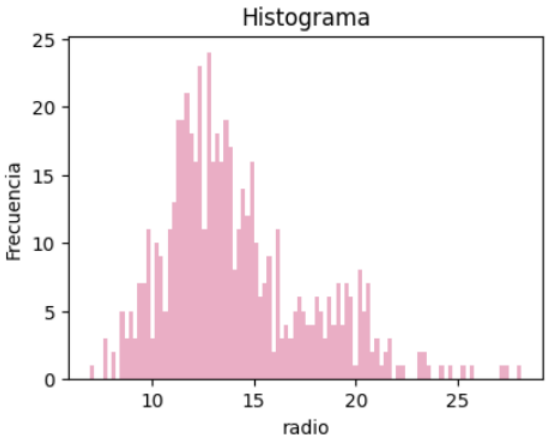
1. ID
2. Diagnóstico
3. Radio
4. Textura
5. Perímetro
6 .Área
7. Suavidad
8. Compactibilidad
9. Concavidad
10. Puntos cóncavos
11. Simetría
12. Dimensión fractal

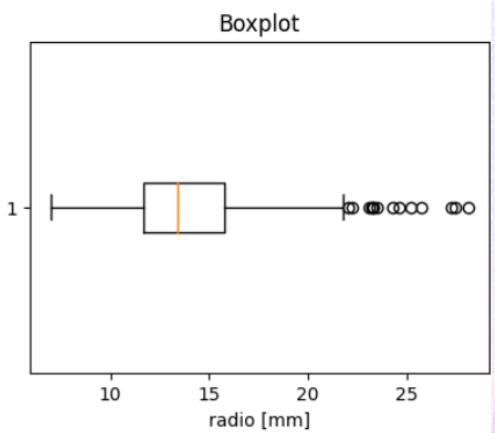
Identificación de cada tumor encontrado.

Clasificación del tumor en benigno o maligno

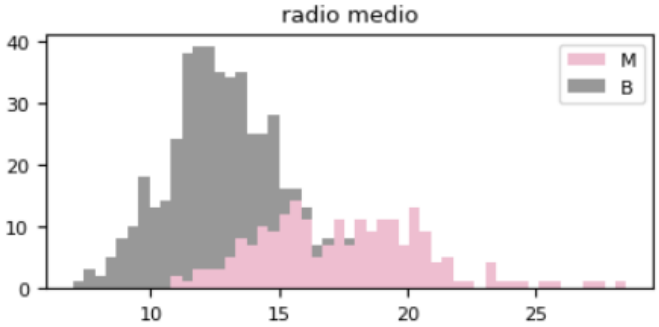
Los atributos 3 – 12 son las variables cuantitativas, resultantes de analizar las imágenes obtenidas en el estudio. Para cada uno de los atributos cuantitativos de la tabla se toma el valor medio, la desviación estándar (identificada como *se*) y el peor valor registrado.

ANÁLISIS EXPLORATORIO DE DATOS



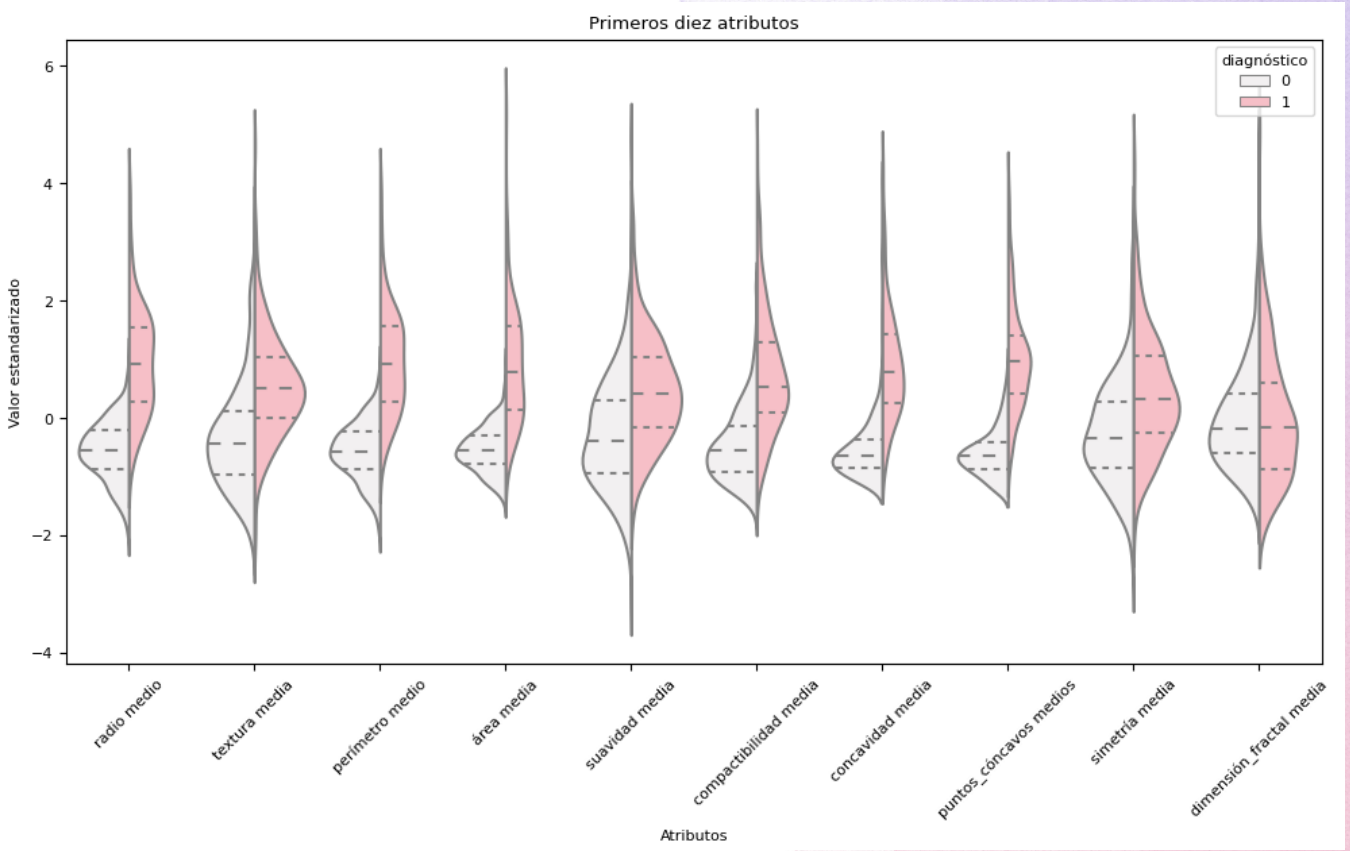


Parece que se trata de outliers....

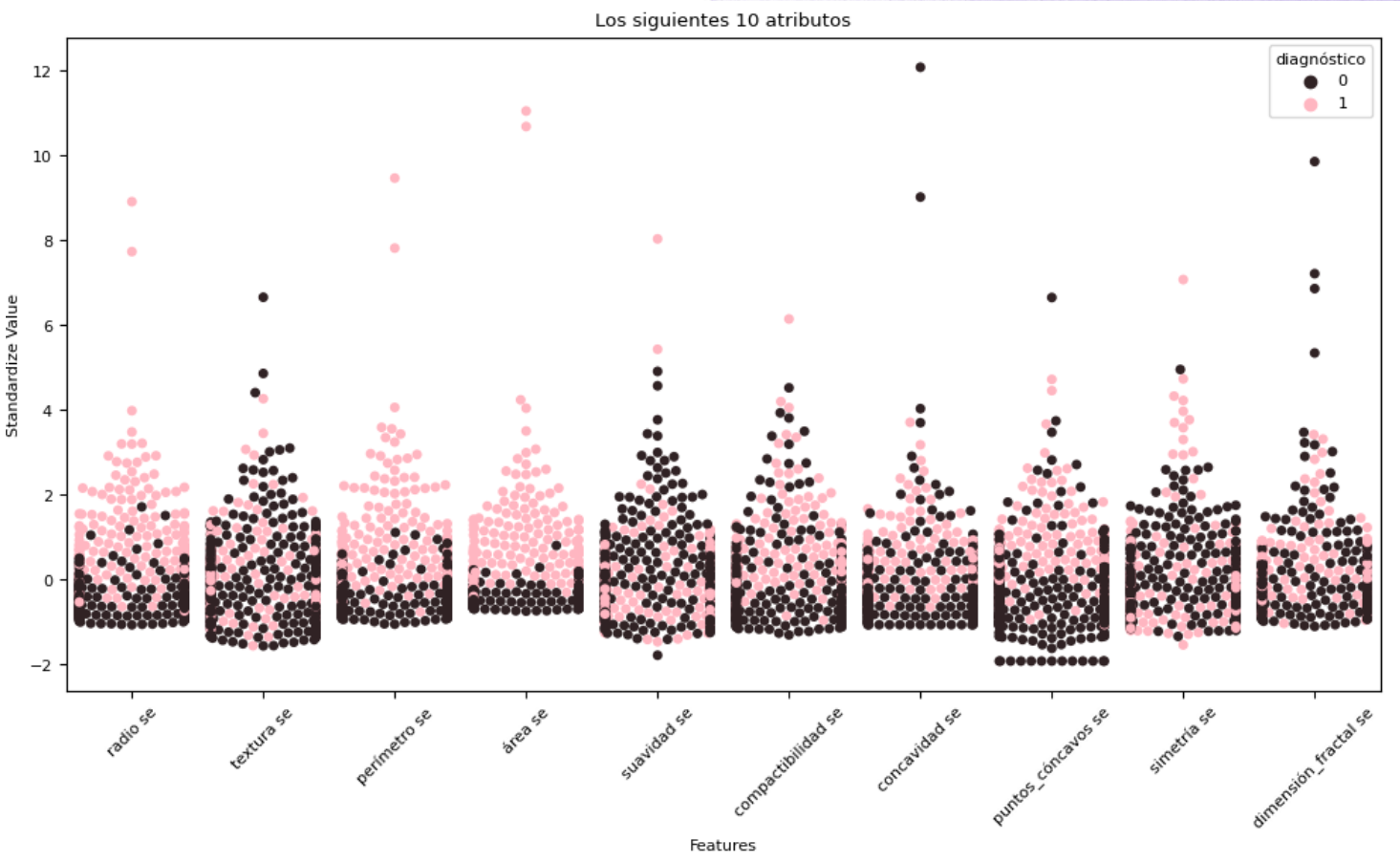


Pero al visualizar los datos con el tipo de tumor como tercera variable se encuentra que cada tumor tiene su propia distribución.

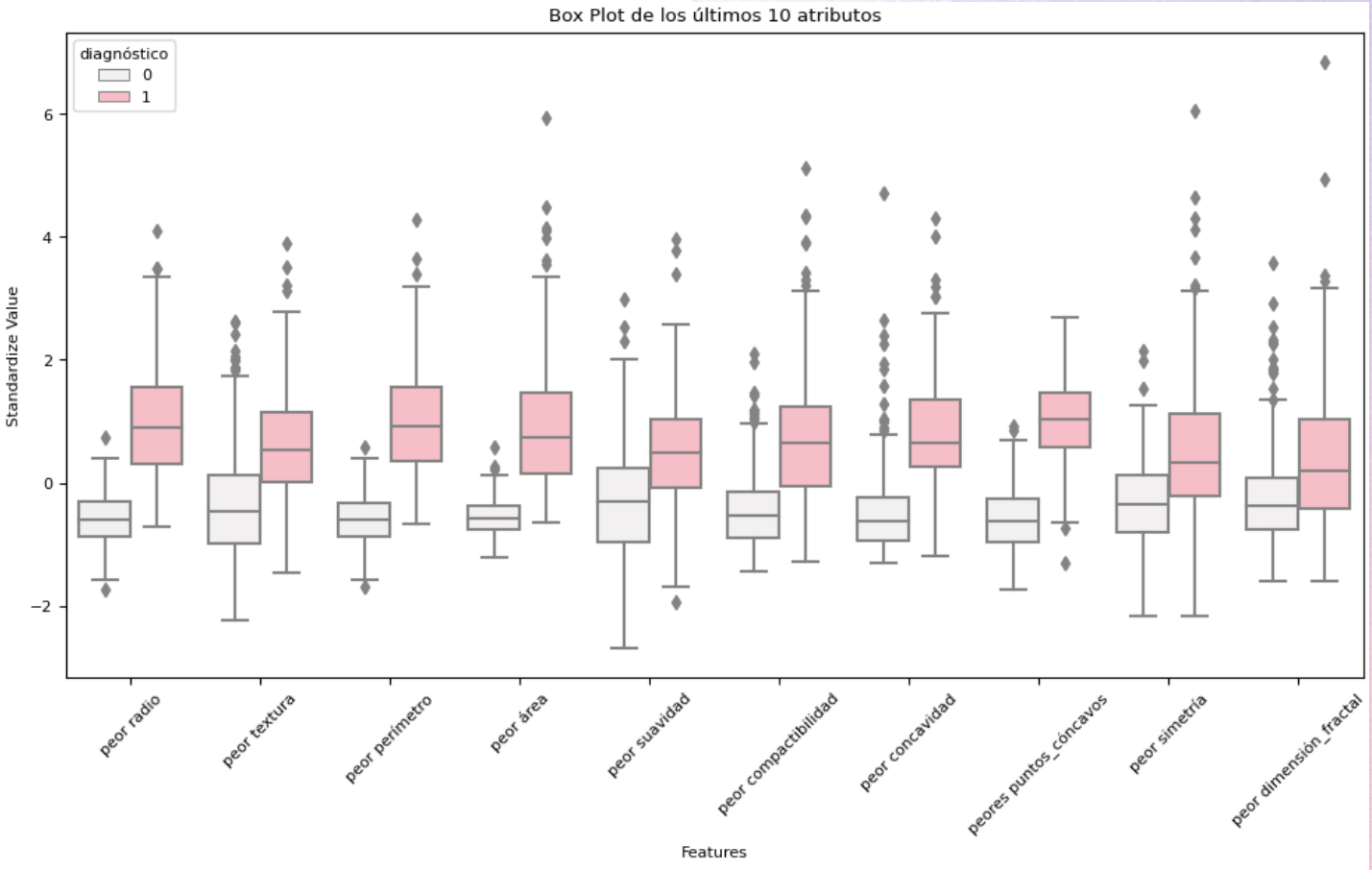
ANÁLISIS MULTIVARIADO



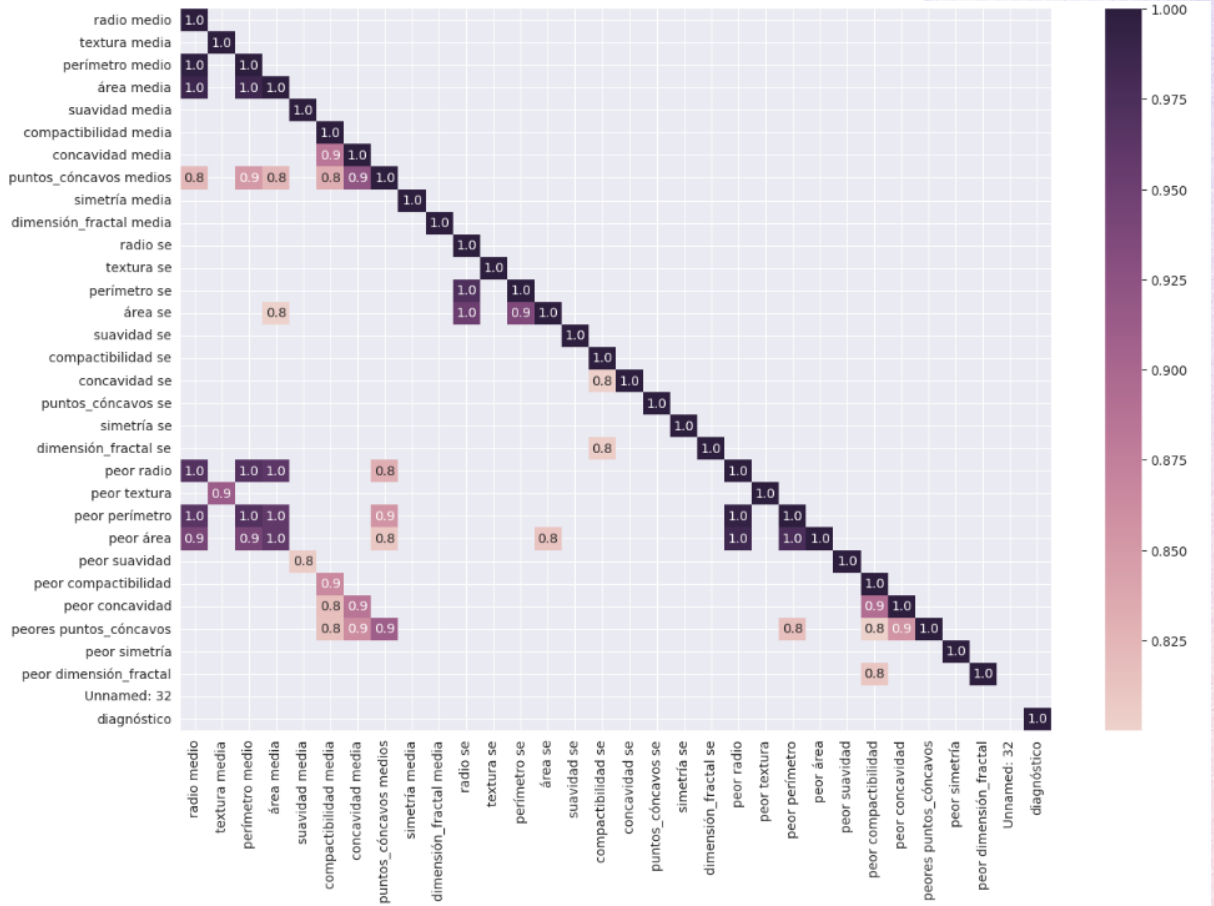
ANÁLISIS MULTIVARIADO



ANÁLISIS MULTIVARIADO



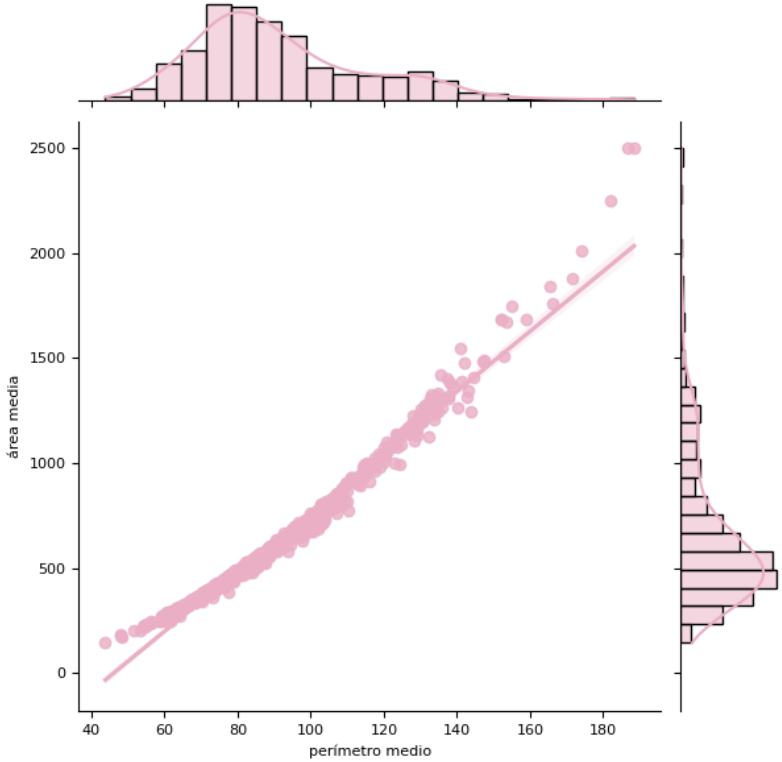
ANÁLISIS MULTIVARIADO



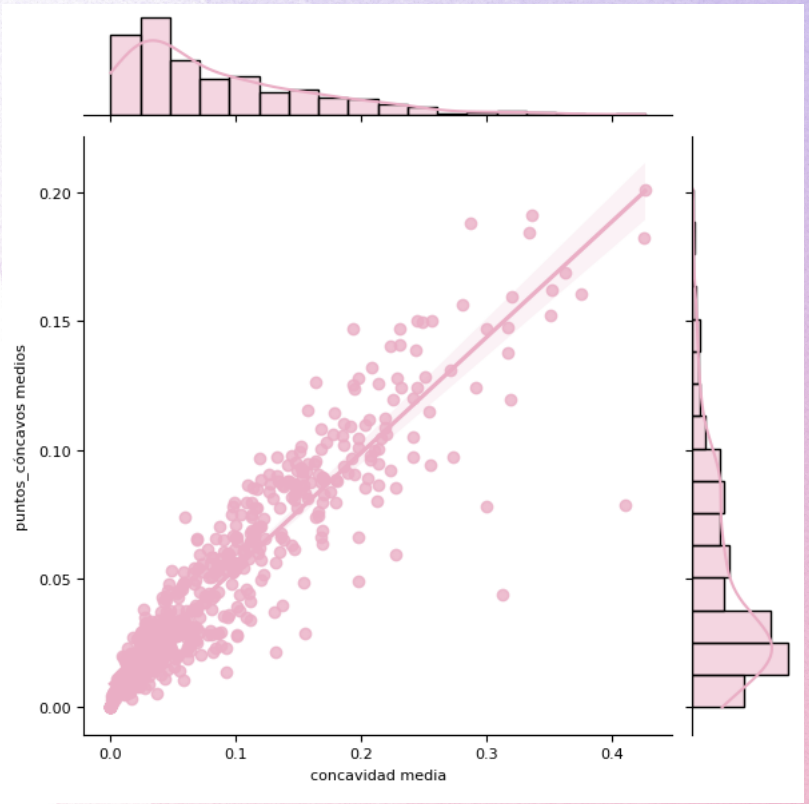
CORRELACIÓN ENTRE TODAS LAS VARIABLES DEL SET DE DATOS

SE DISTINGUEN UNAS POCAS DE RELEVANCIA

ANÁLISIS MULTIVARIADO



Correlación = 0,98



Correlación = 0,92

CONCLUSIONES DEL ANÁLISIS EXPLORATORIO

- 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

4

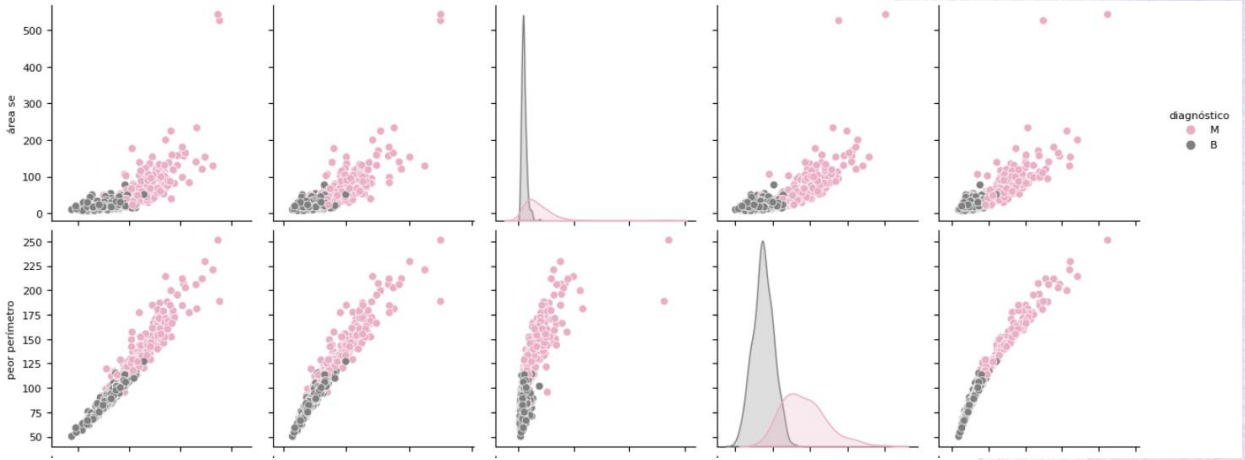
SELECCIÓN DE ATRIBUTOS

```
from sklearn.feature_selection import SelectKBest, chi2
feature_selection = SelectKBest(chi2, k=5)
feature_selection.fit(data, y)
selected_features = data.columns[feature_selection.get_support()]
print("The five selected features are: ", list(selected_features))
```

	perímetro medio	área media	área se	peor perímetro	peor área
0	122.80	1001.0	153.40	184.60	2019.0
1	132.90	1326.0	74.08	158.80	1956.0
2	130.00	1203.0	94.03	152.50	1709.0
3	77.58	386.1	27.23	98.87	567.7
4	135.10	1297.0	94.44	152.20	1575.0

Variables seleccionadas

SELECCIÓN DE ATRIBUTOS



Ejemplo con el área y el perímetro

Se observa que los tumores malignos toman valores mayores de área y de perímetro

Se ve la misma tendencia en las demás variables

IMPLEMENTACIÓN DE MODELOS

5

MODELOS DE MACHINE LEARNING

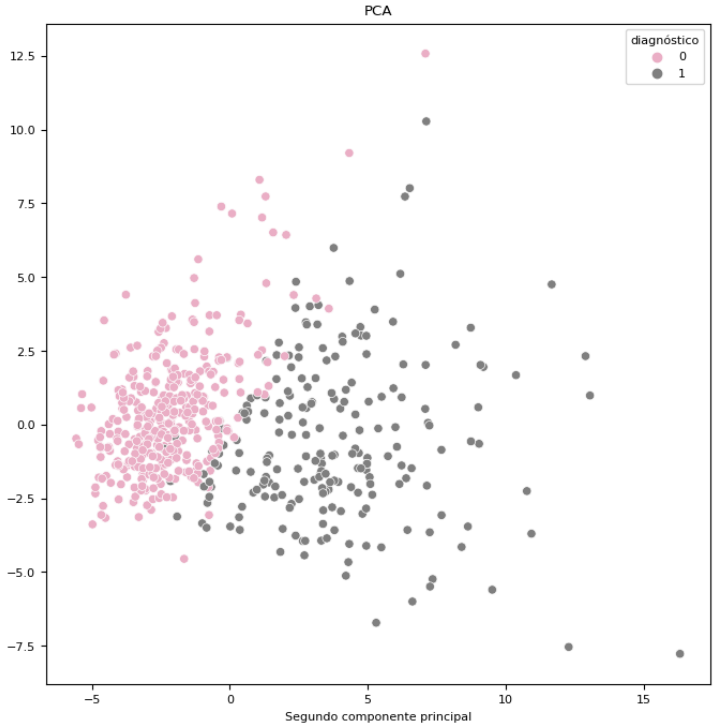
Random forest classifier

```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=200)
rfc.fit(X_train, y_train)
y_pred = rfc.predict(X_test)
```

```
Matriz de confusión:
 [[118   3]
 [  2  65]]


Reporte:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       121
           1       0.96      0.97      0.96        67

    accuracy                           0.97       188
   macro avg       0.97      0.97      0.97       188
weighted avg       0.97      0.97      0.97       188
```

MODELOS DE MACHINE LEARNING

Reducción de la dimensionalidad



Support Vector Machine

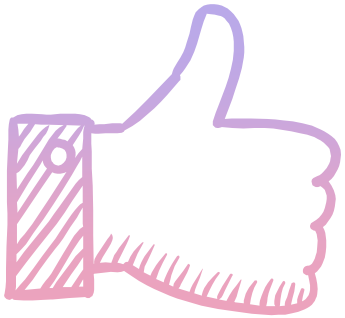
```
Matriz de confusión:
 [[116   5]
 [  5  62]]


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
```

CONCLUSIONES

- Se realizó un análisis exploratorio de datos para identificar cada una de las 30 variables originales y comprender cómo podrían utilizarse para identificar tumores malignos. Se encontró que las medianas de algunas de ellas eran muy diferentes entre tumores malignos y benignos. Además, se encontraron correlaciones entre el perímetro y el área medios, y la concavidad media y los puntos cóncavos medios.

- Utilizando la selección de aributos univariados del módulo sklearn se eligieron las 5 características con las puntuaciones más altas, es decir, con mayor potencial para realizar un análisis predictivo. Los 5 atributos resultantes fueron el perímetro medio, el área media, el área se, el peor perímetro y la peor área.

- Se aplicó el modelo de random forest classfier con una precisión del 97%. Por otro lado, se utilizó PCA para encontrar los dos componentes principales y separar claramente los datos en benignos y malignos. Finalmente, se aplicó SVM para predecir la presencia de la enfermedad basado en PCA. La precisión de este modelo fue del 95%. Ambos modelos son útiles para la clasificación de tumores.



¡¡MUCHAS GRACIAS!!