

# Breast Cancer (Diagnosis) Prediction using ML

EDWIN MOLINA  
FRANCISCO PORRAS



# Tabla de Contenidos

A close-up photograph of two hands, one from the left and one from the right, holding a thin, red, translucent ribbon. The hands are positioned as if they are about to tie the ribbon into a knot or are in the process of doing so. The background is a dark, textured surface, possibly a wall or a backdrop. The lighting is soft, highlighting the texture of the skin and the sheen of the ribbon.

## **Parte 1:**

Descripción de dataset e investigación



## **Parte 2:**

Presentación de trabajo



## **Parte 3:**

Presentación de resultados



## **Parte 4:**

Conclusiones y resumen



**En todo el mundo, el cáncer de mama es el tipo de cáncer más frecuente en las mujeres y el segundo en cuanto a tasas de mortalidad. El diagnóstico del cáncer de mama se realiza cuando se encuentra un bulto anormal (a partir de una autoexploración o una radiografía) o se ve una pequeña mota de calcio (en una radiografía). Tras encontrar un bulto sospechoso, el médico realizará un diagnóstico para determinar si es canceroso y, en caso afirmativo, si se ha extendido a otras partes del cuerpo.**

DATASET

**Este conjunto de datos sobre el cáncer de mama se obtuvo de los Hospitales de la Universidad de Wisconsin, Madison, del Dr. William H. Wolberg.**

[HTTPS://WWW.KAGGLE.COM/MERISHNASUWAL/BREAST-CANCER-PREDICTION-DATASET](https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset)

Descripción de dataset

Filas
569
Columnas
6

# Exploración de variables

## Mean\_radius

Es el promedio de las distancias entre el centro a los puntos del perímetro.



## Mean\_texture

Es la desviación estandar de los valores de la escala de grises en la imagen.



## Mean\_perimeter

Es el promedio del tamaño del núcleo del tumor.



# Exploración de variables

## Mean\_area

Es la media del área del núcleo del tumor.



## Mean\_smoothnes

Es la media de la variación local de las longitudes del radio del tumor.

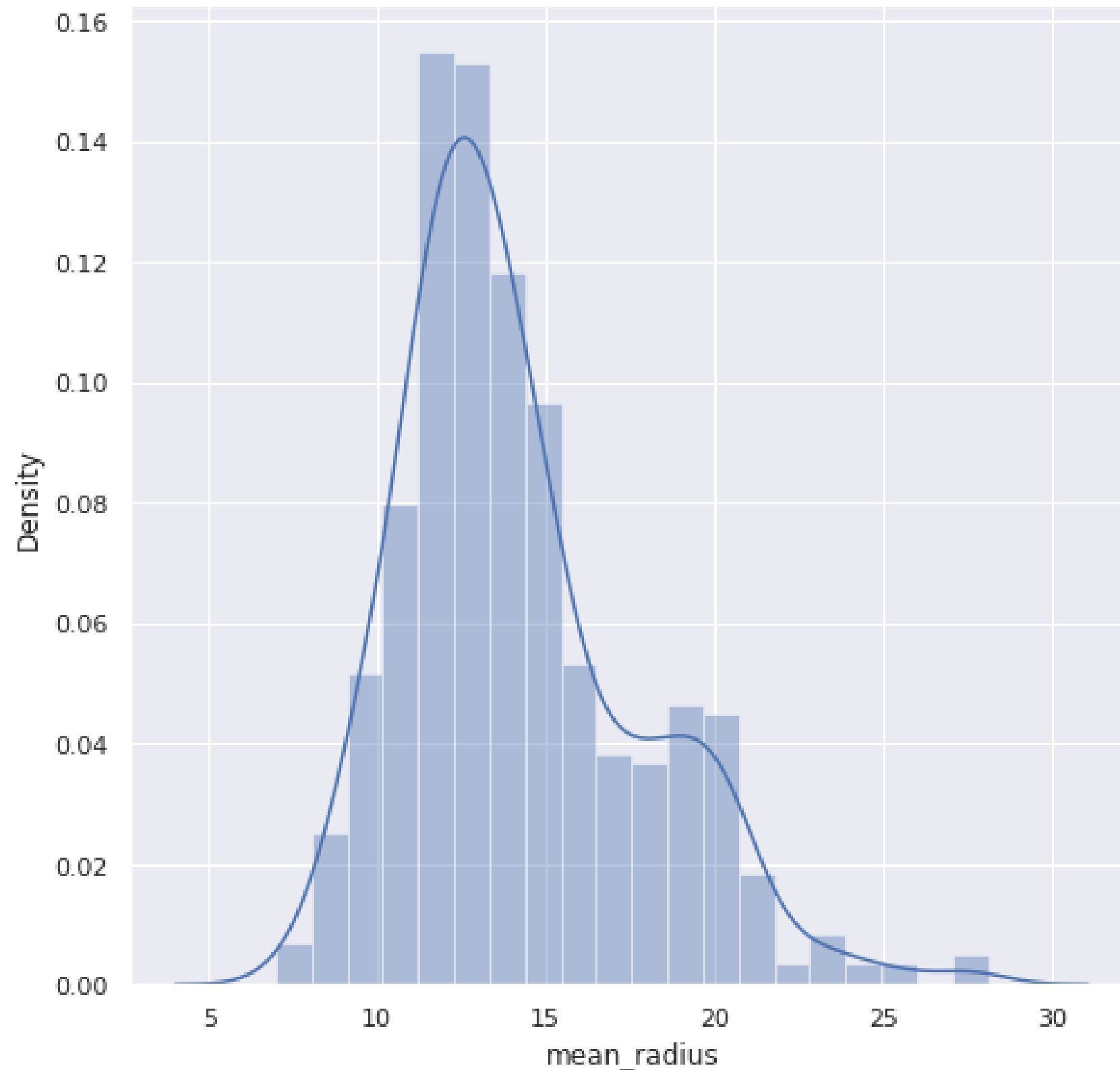


## Diagnosis

Es el diagnóstico del tejido mamario. 1 es maligno, 0 es benigno, donde maligno significa que la enfermedad es peligrosa.



# mean\_radius



Moda: 12.34

Media: 14.13

Mediana: 13.37

Primer Cuartil: 11.7

Segundo Cuartil: 13.37

Tercer Cuartil: 15.78

Mínimo: 6.981

Máximo: 28.11

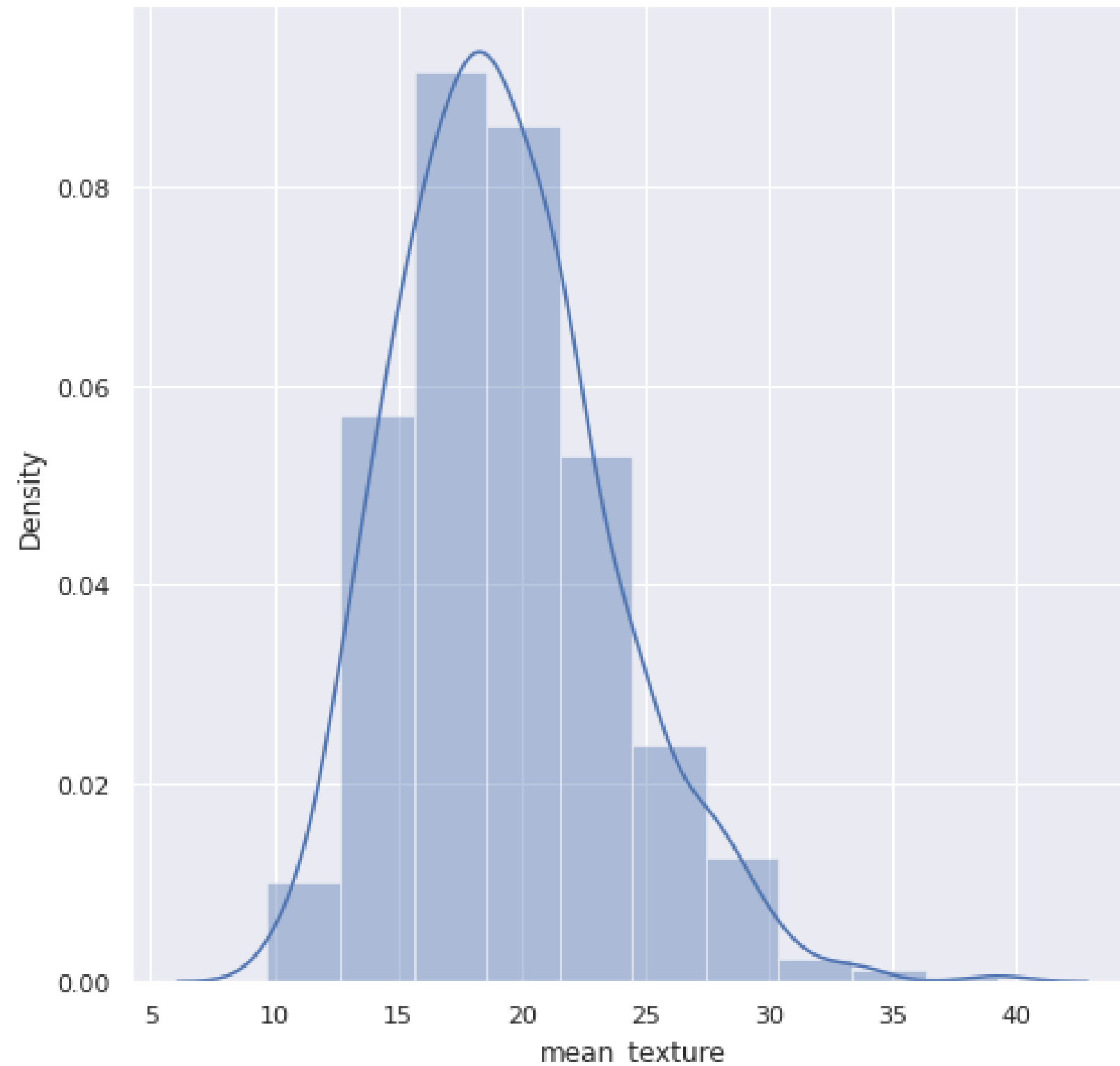
Varianza: 12.42

Desviación Estandar: 3.52 con respecto a la media.

Esto puede indicar si fuera una distribución normal, que el 68% de las compañías van desde los 10.60 y los 17.65 en este índice y que el 95% de las compañías van desde los 7.08 y los 21.18 . Asimetría: La curva de frecuencia cuenta con una leve asimetría positiva



# mean\_texture



Modas: 14.93, 15.7, 16.84, 16.85, 17.46, 18.22, 18.9  
19.83, 20.52

Media 19.29

Mediana: 18.84

Primer Cuartil: 16.17

Segundo Cuartil: 18.84

Tercer Cuartil: 21.8

Mínimo: 9.71

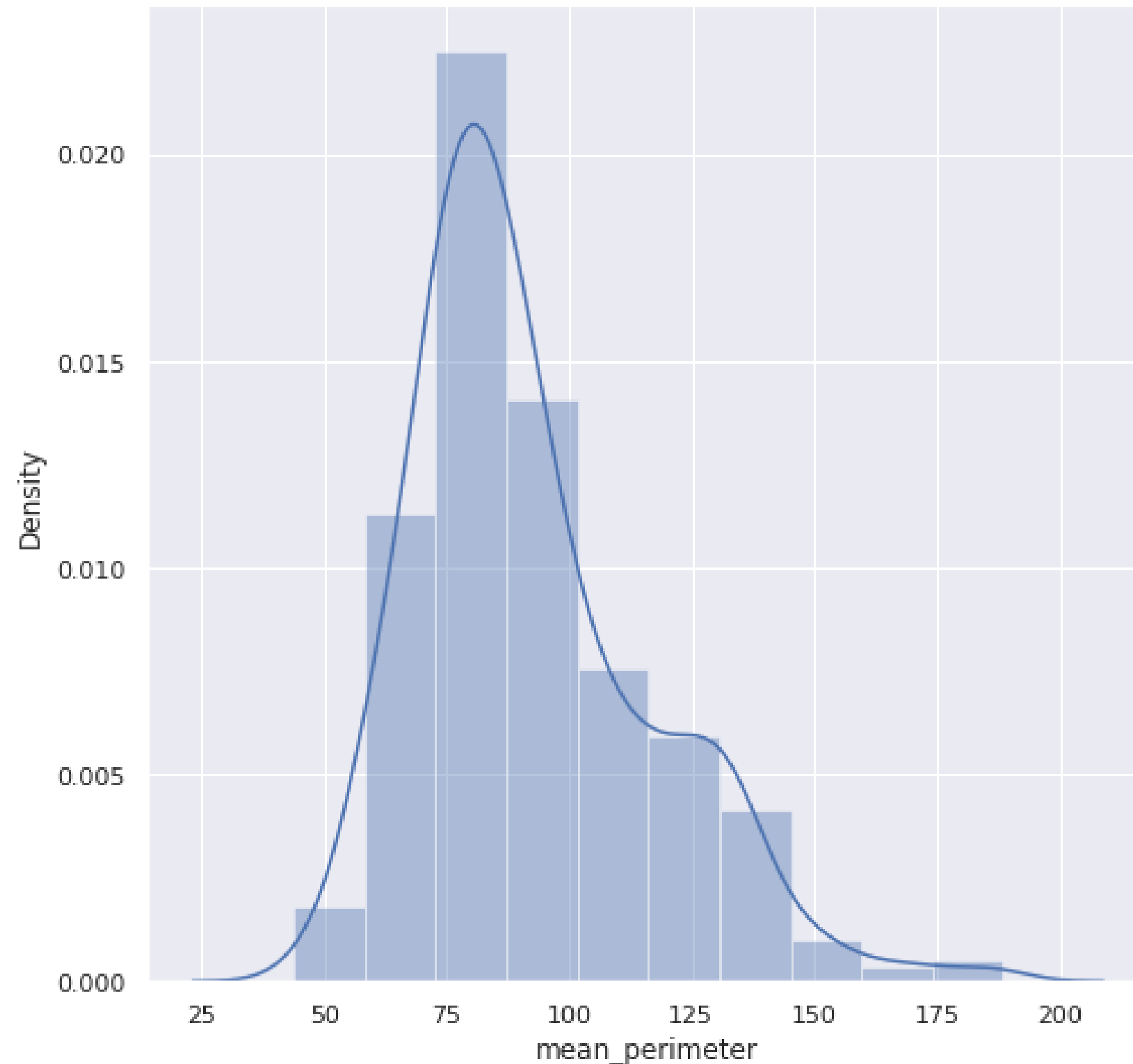
Máximo: 39.28

Varianza:18.50

Desviación Estandar: 4.30 con respecto a la media.

Esto puede indicar si fuera una distribución normal, que el 68% de las compañías van desde los 14.99 y los 23.59 en este índice y que el 95% de las compañías van desde los 10.69 y los 27.89 . Asimetría: La curva de frecuencia cuenta con una leve asimetría positiva.

# mean\_perimeter



Modas: 82.61, 87.76, 134.7

Media: 91.96

Mediana: 86.24

Primer Cuartil: 75.17

Segundo Cuartil: 86.24

Tercer Cuartil: 104.1

Mínimo: 43.79

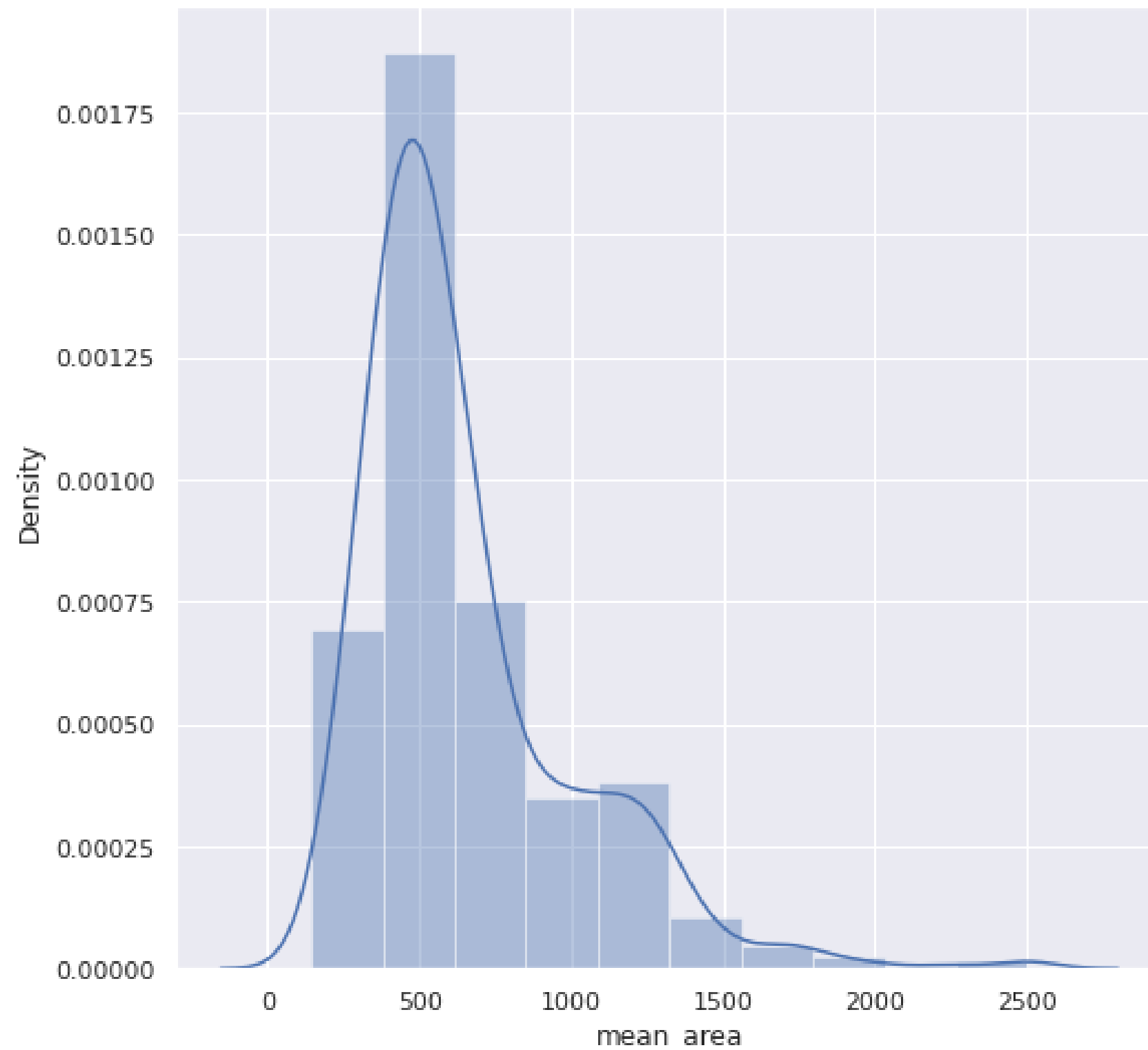
Máximo: 188.5

Varianza: 590.44 con respecto a la media

Desviación Estandar: 24.30 con respecto a la media.

Esto puede indicar si fuera una distribución normal, que el 68% de las compañías van desde los 67.67 y los 116.27 en este índice y que el 95% de las compañías van desde los 43.37 y los 140.57 . Asimetría: La curva de frecuencia cuenta con una leve asimetría positiva

# mean\_area



Moda: 512.2

Media: 654.88

Mediana: 551.1

Primer Cuartil: 420.3

Segundo Cuartil: 551.1

Tercer Cuartil: 782.7

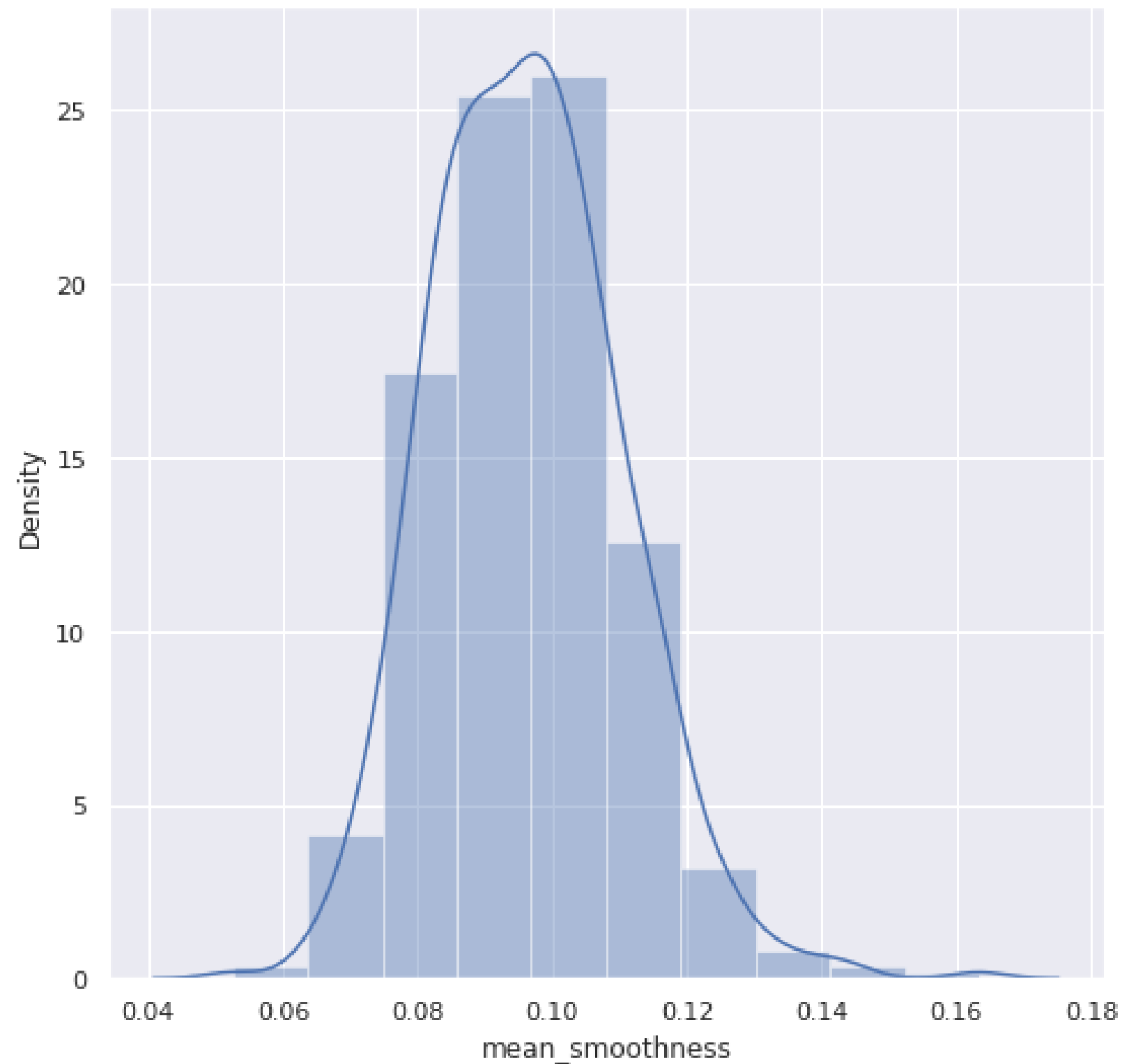
Mínimo: 143.5

Máximo: 2501.0

Desviación Estandar: 351.91 con respecto a la media.

Esto puede indicar si fuera una distribución normal, que el 68% de las compañías van desde los 302.97 y los 1006.80 en este índice y que el 95% de las compañías van desde los 48.94 y los 1358.72 . Asimetría: La curva de frecuencia cuenta con una asimetría positiva muy marcada

# mean\_smoothness



Moda: 0.1007

Media: 0.096

Mediana: 0.095870

Primer Cuartil: 0.08637

Segundo Cuartil: 0.0958700

Tercer Cuartil: 0.1053

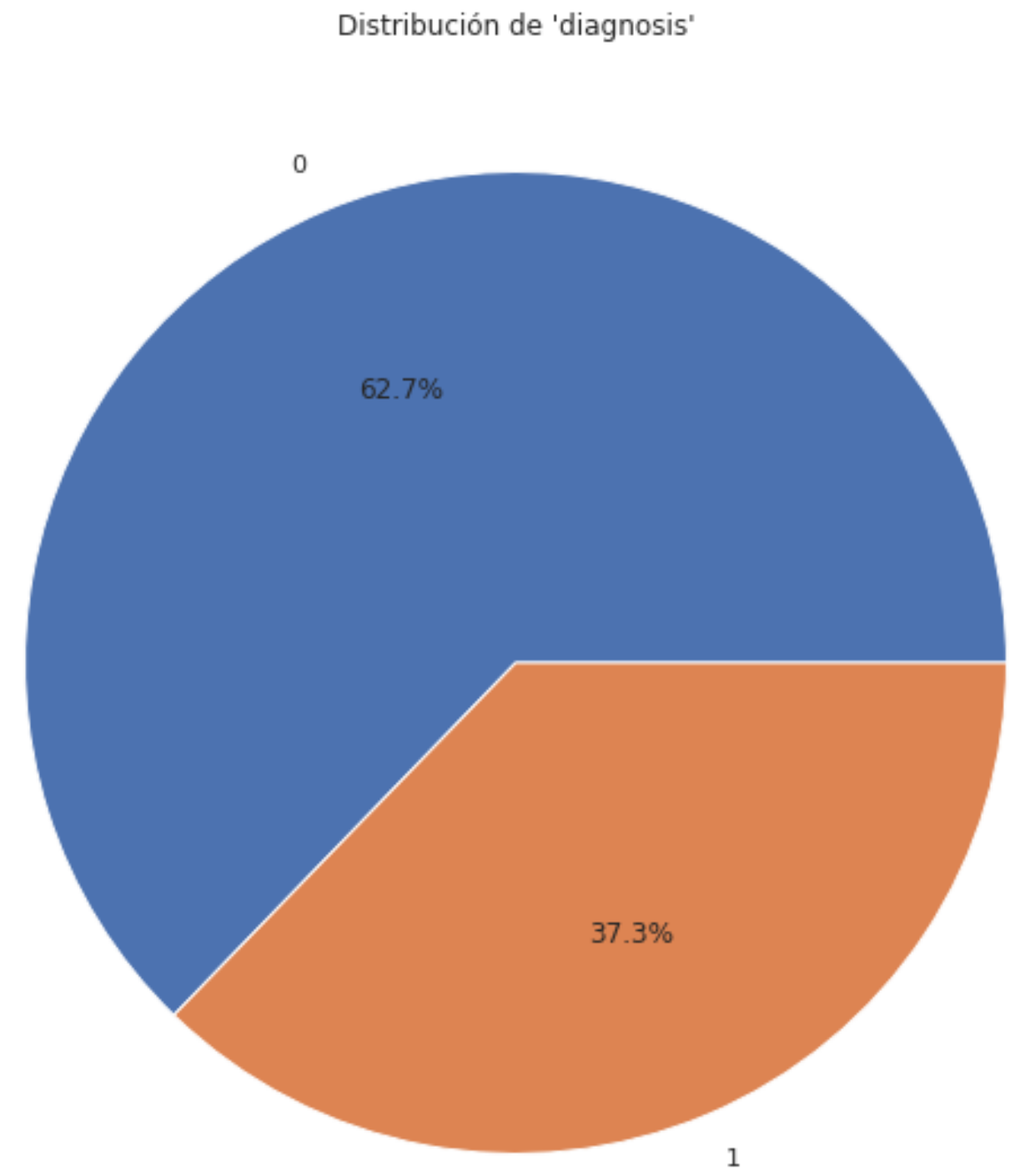
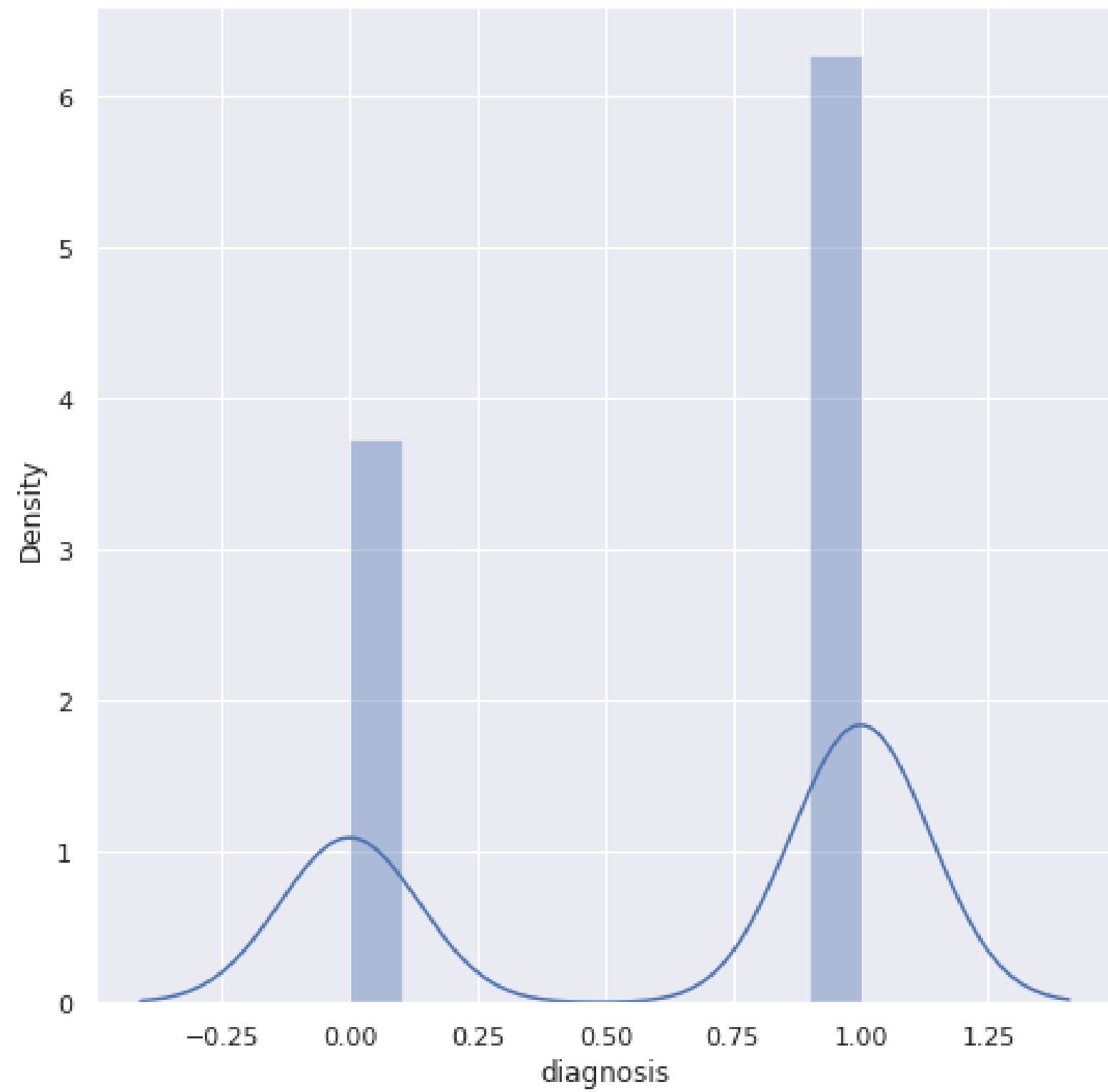
Mínimo: 0.052629

Máximo: 0.1634

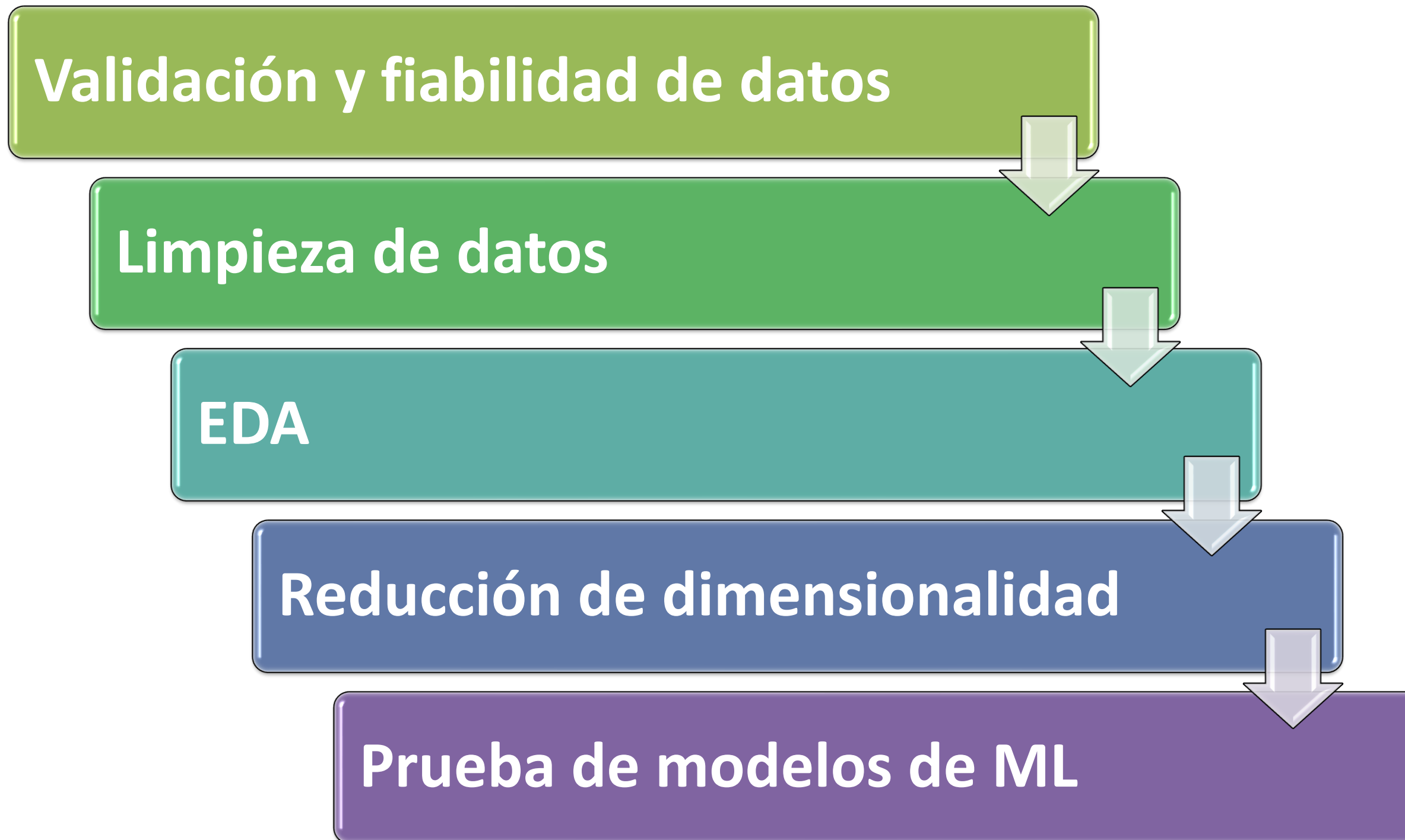
Desviación Estandar: 0.01 con respecto a la media.

Esto puede indicar si fuera una distribución normal, que el 68% de las compañías van desde los 0.08 y los 0.11 en este índice y que el 95% de las compañías van desde los 0.07 y los 0.12 . Asimetría: La curva de frecuencia cuenta con una leve asimetría positiva

# Diagnosis

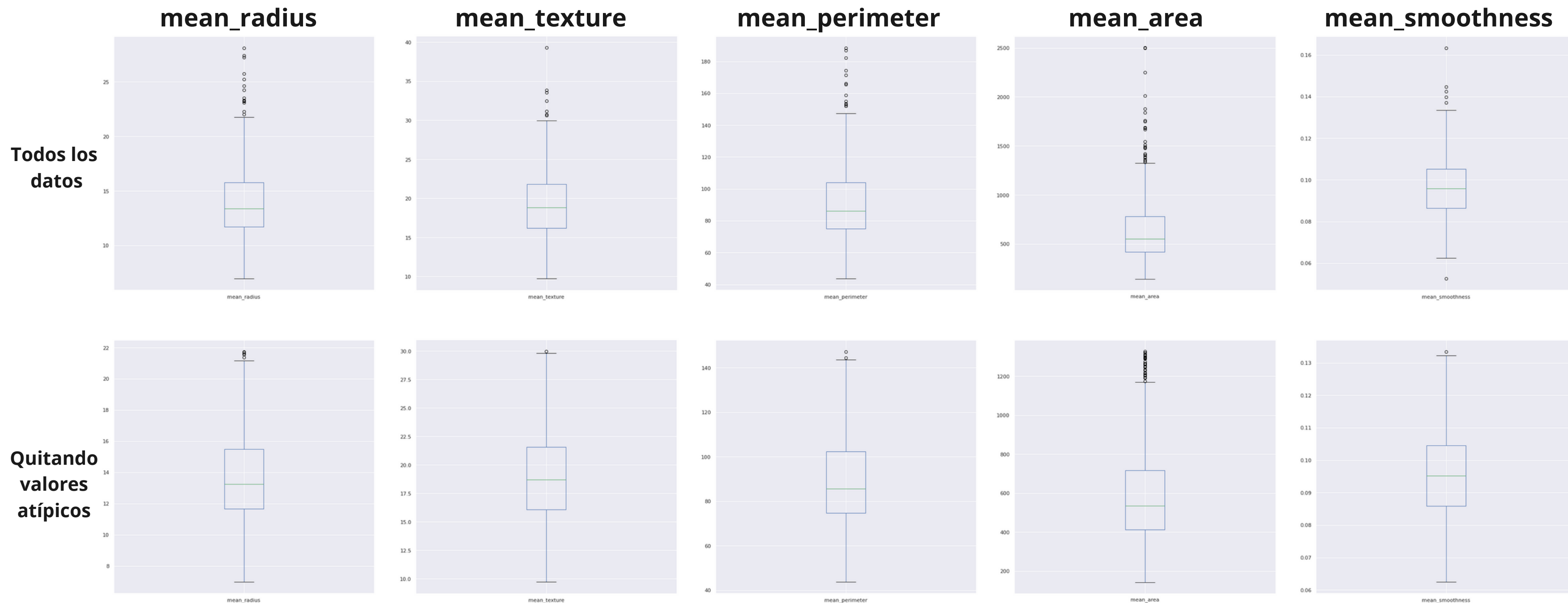


# Proceso

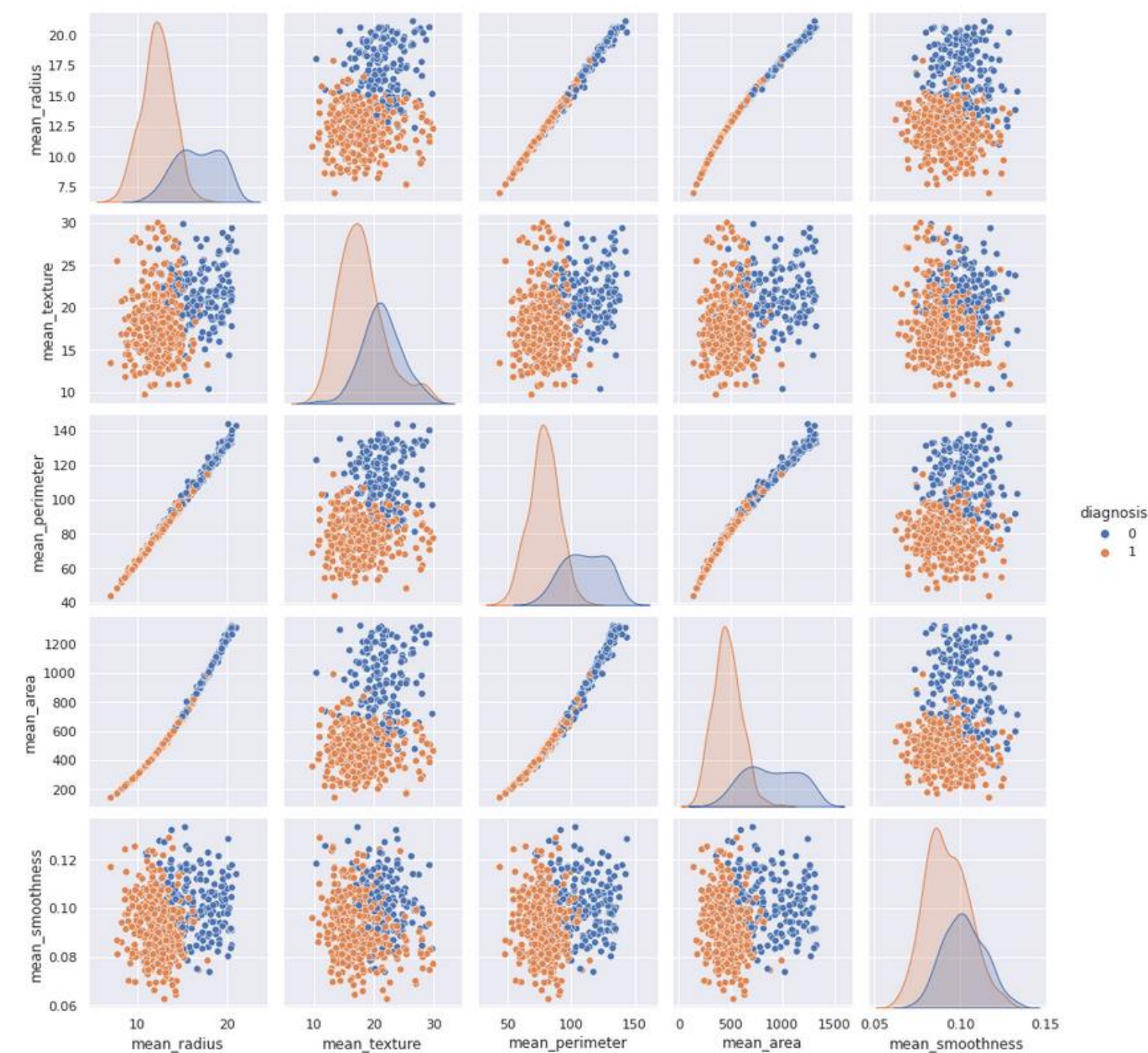
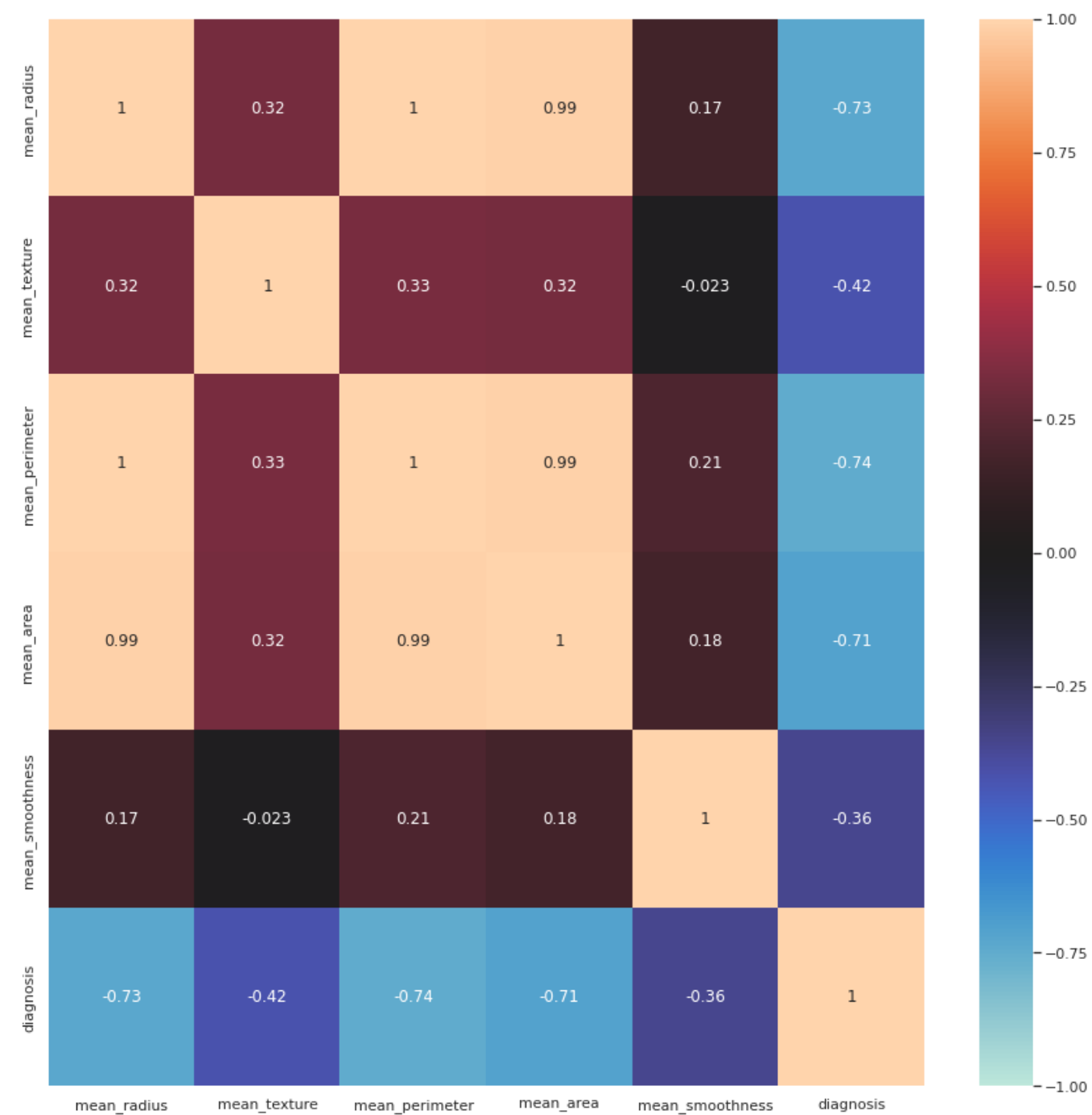


# Preparación de los datos

## Transformación y limpieza

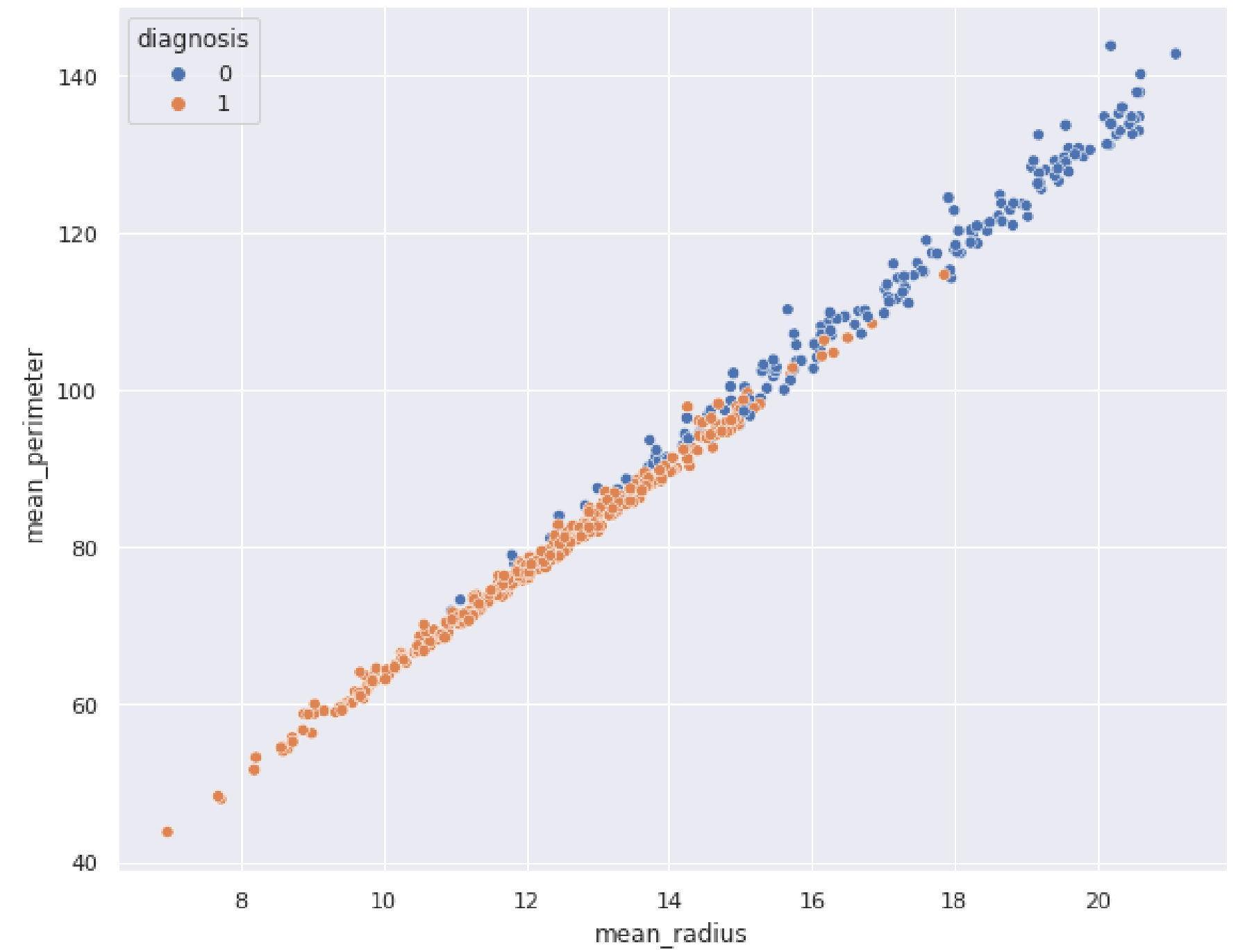
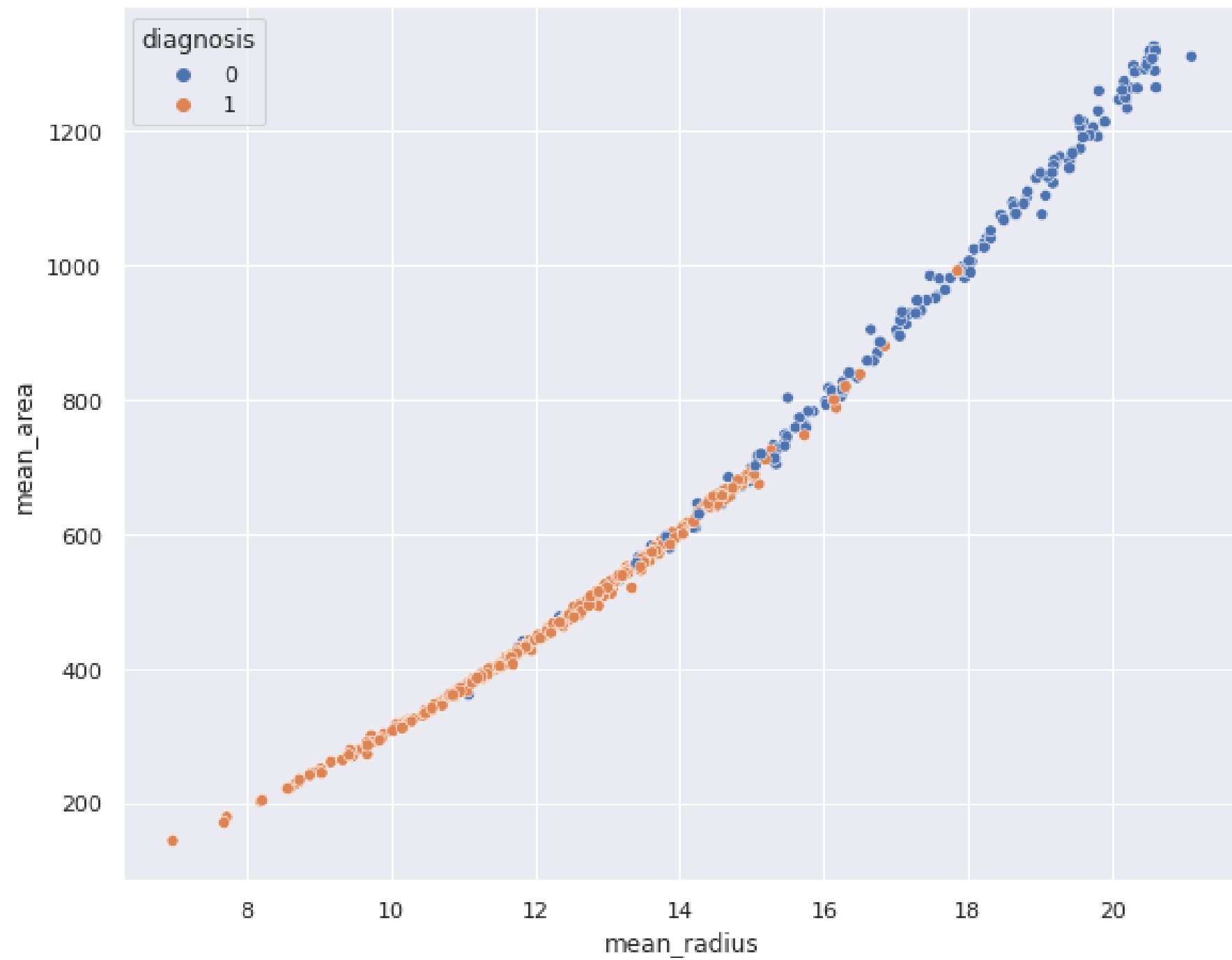


# Evaluacion de datos





# Evaluacion de datos



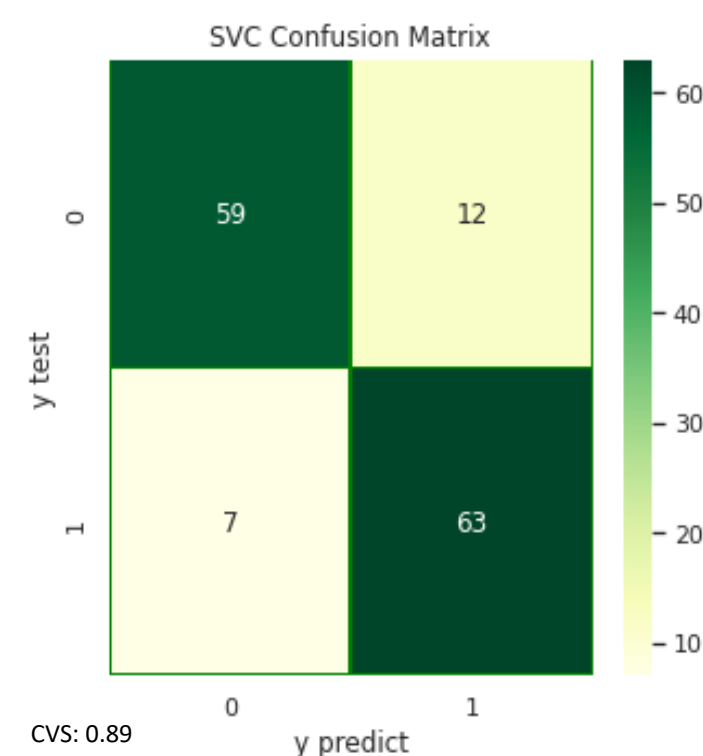
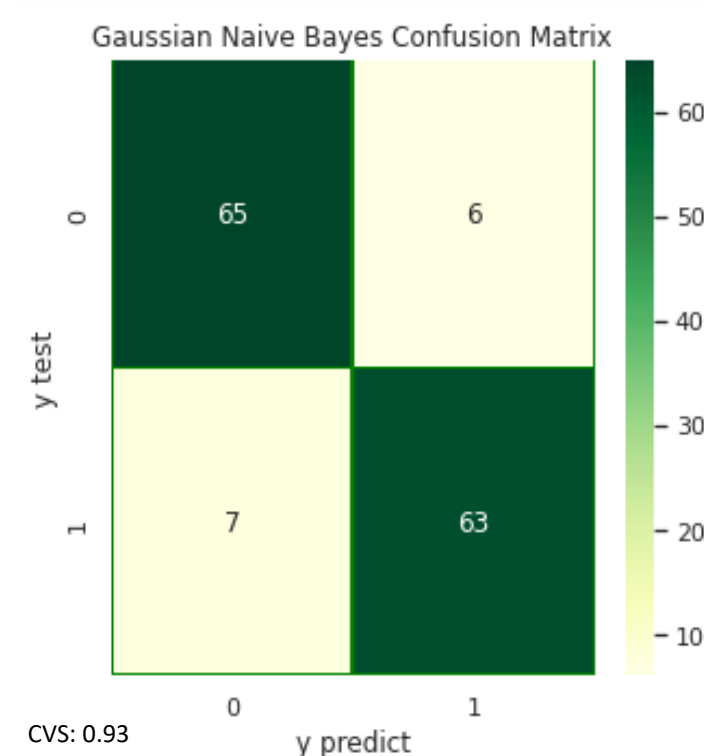
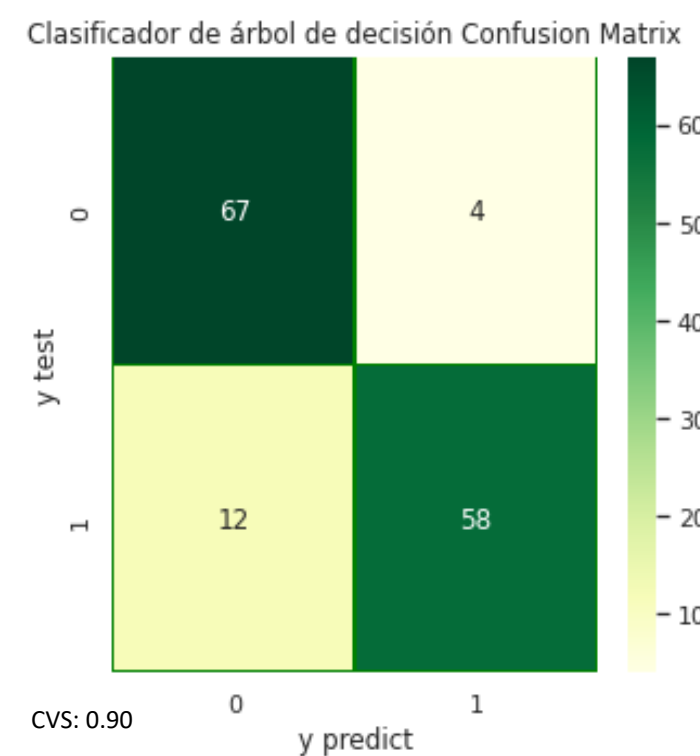
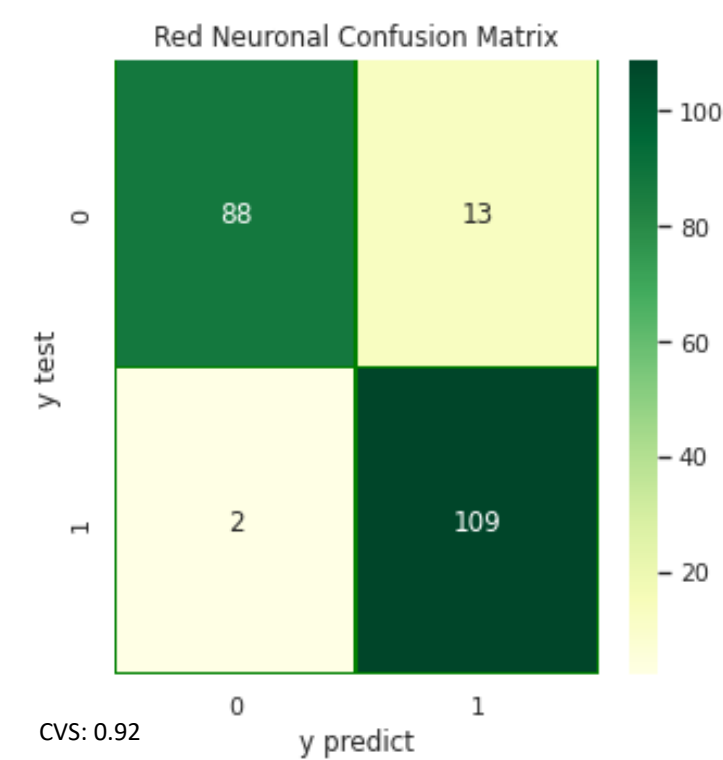
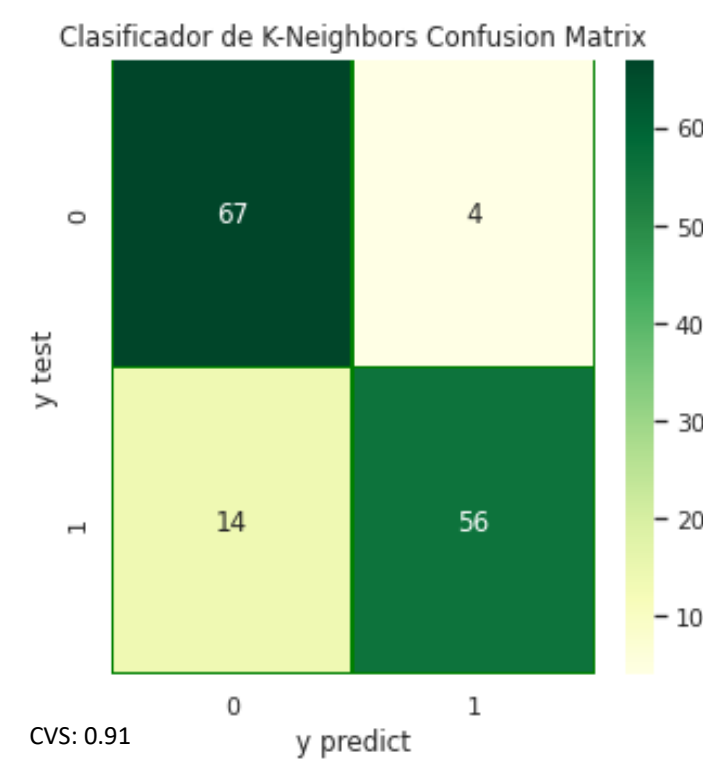
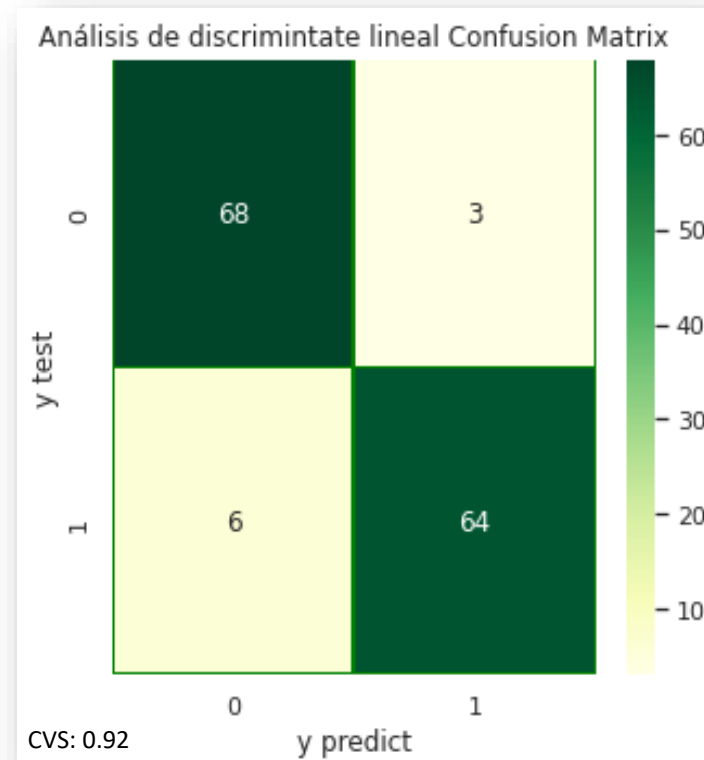
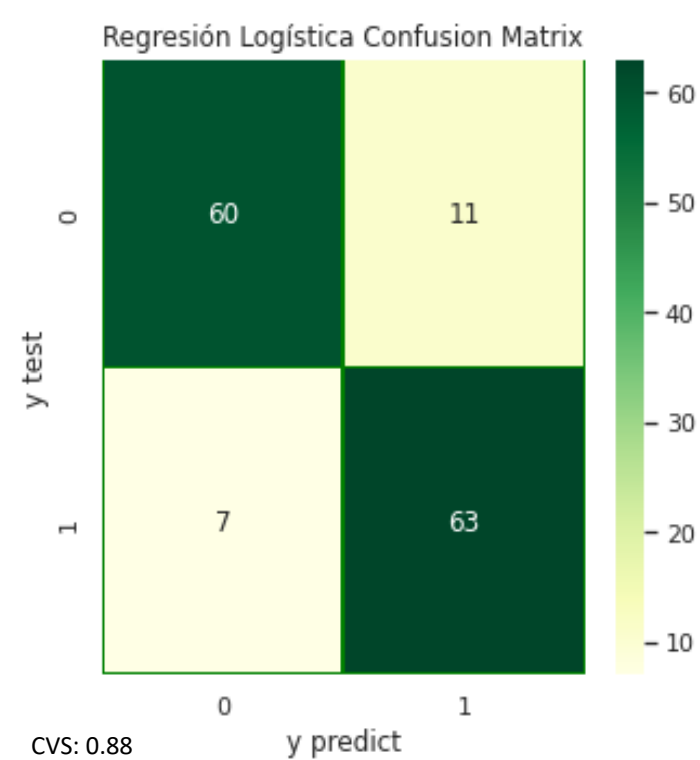
# Modelos Supervisados

## MACHINE LEARNING

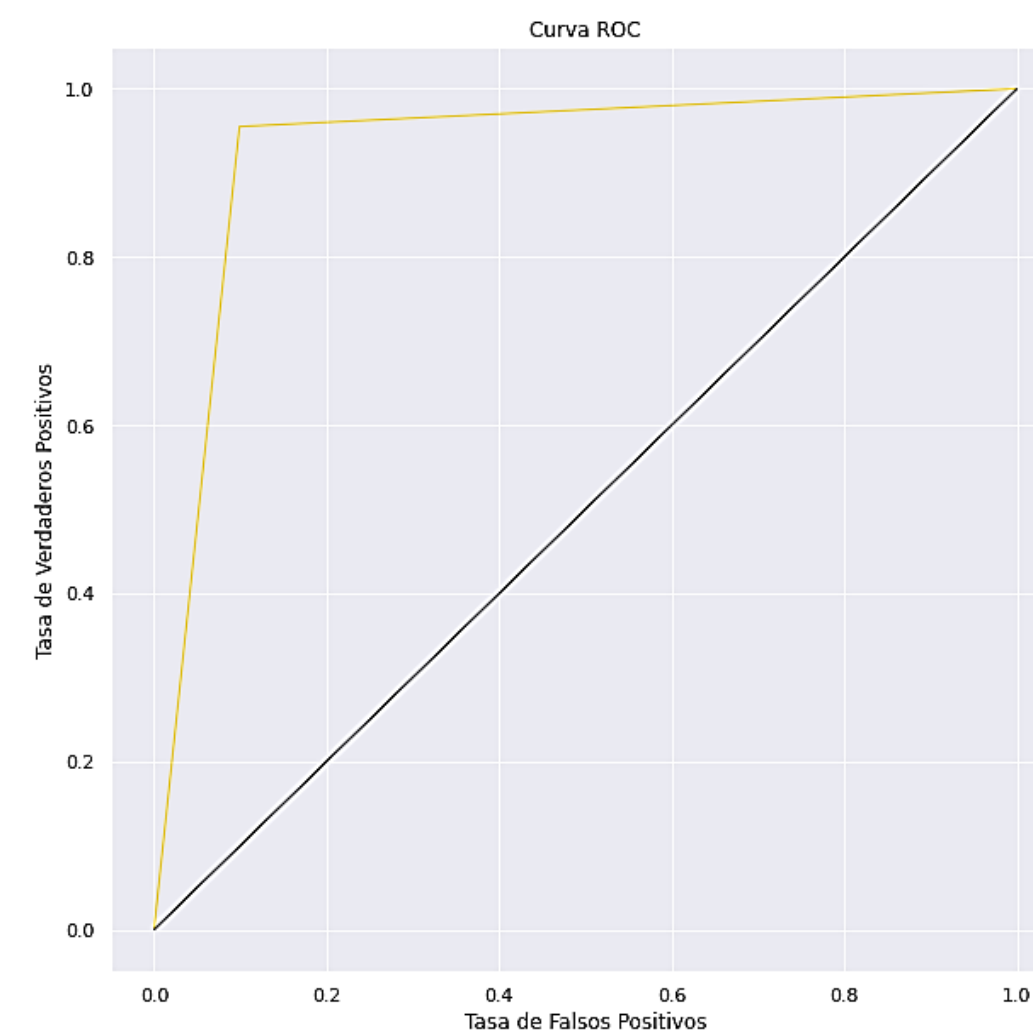
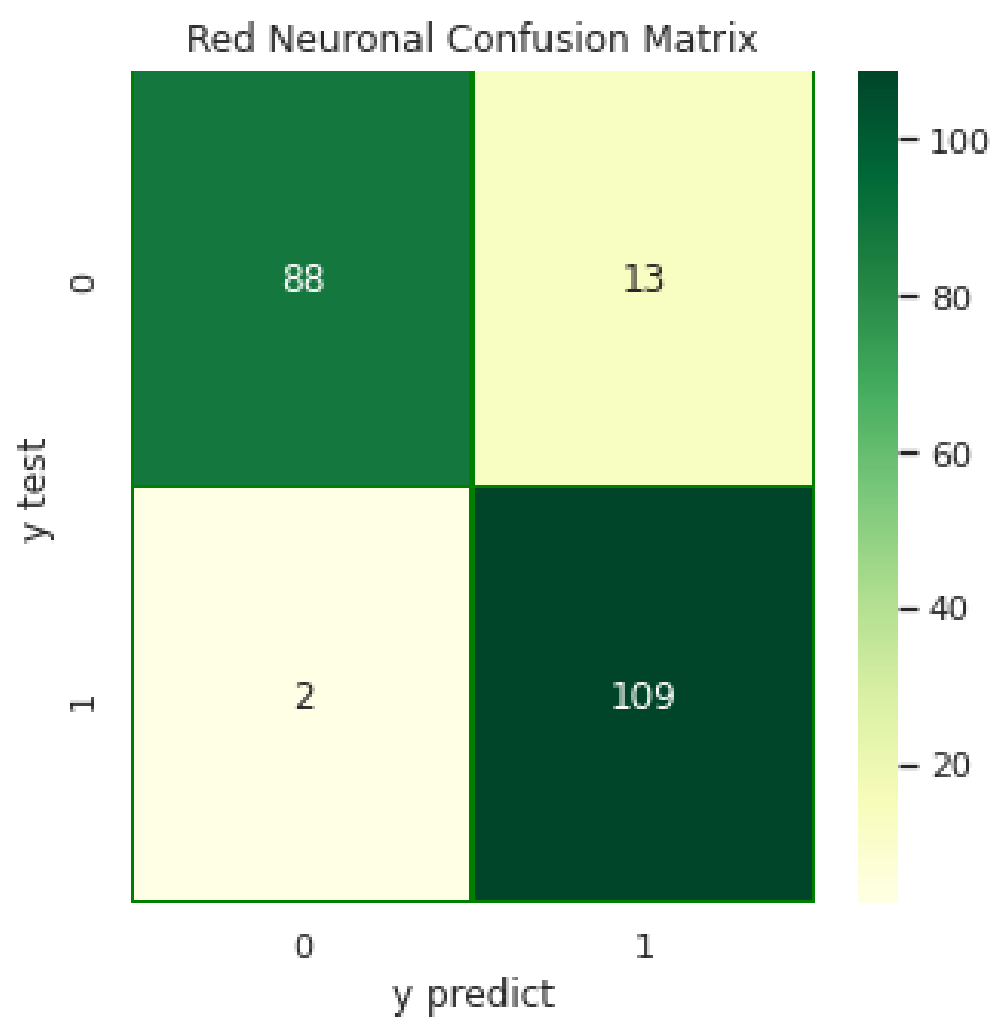
Modelos supervisados

- Se realizó limpieza de datos atípicos
- Se utilizó upsampling mediante el método SMOTE
- Se utilizaron las variables independientes:
  - Mean\_textura
  - Mean\_perimeter
  - Mean\_smothness
- Se utilizó la variable independiente:
  - diagnosis
- Se compararon los siguientes modelos:
  - Regresión logística
  - Análisis de discriminante lineal
  - Clasificador de K-Neighbors
  - Clasificador de Árbol de Decisión
  - Gaussian Naive Bayes
  - SVC
  - Red Neuronal

# Resultados de modelos



# Red neuronal & curva ROC



# Mejores resultados

Predicción de cancer de mama

01

Gaussian Naive Bayes  
93%

02

Análisis de  
discriminante lineal  
92%

03

Clasificador de K-  
Neighbors 91%

A través de la implantación de modelos supervisados de ML, se logra identificar la clasificación de las categorías en estudio en cuanto a la predicción del cáncer de mama.

Al utilizar los modelos de ML se obtienen los siguientes resultados:

- Regresión Logística: 0.880949 (0.062688)
- Análisis de discriminante lineal: 0.924980 (0.058848)
- Clasificador de K-Neighbors: 0.914387 (0.047660)
- Clasificador de árbol de decisión: 0.900395 (0.065743)
- Gaussian Naive Bayes: 0.928538 (0.054154)
- SVC: 0.889802 (0.059542)

# Conclusiones

- Se notó una mejoría significativa al balancear los datos usando un método de upsampling.
- Encontramos que la prueba con mejores resultados en la matriz de confusión es el modelo de Análisis de discriminante lineal. Este modelo parece seleccionar con bastante precisión si el tumor es benigno o maligno, lo cual puede ayudar a los médicos y pacientes a tener un diagnóstico más temprano.



Muchas gracias.