Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

Informe 1 - Análisis Exploratorio
Diagnóstico cáncer de mama

El cáncer de mama es una enfermedad que se ha convertido en una de las principales causas de muerte entre mujeres en todo el mundo, poniendo en relevancia la necesidad de un diagnóstico a tiempo y certero. Para tales efectos, se hace imperioso contar no sólo con el conocimiento, sino, además, con las tecnologías adecuadas que permitan prolongar las probabilidades de vida en aquellas mujeres que padezcan la enfermedad y disminuir el margen de error en su diagnóstico. Es este contexto que nos ha motivado a tomar el proyecto "Diagnóstico de cáncer de mama".

Dicho lo anterior, el desglose del informe engloba tres apartados; el primero incluye los objetivos, alcances del proyecto y una breve contextualización de la data. Mientras que el apartado 2, contiene una descripción más detallada de los datos a utilizar (cantidad de datos, de dónde vienen, sus distintos atributos y etiquetas, etc.), con el objetivo de tener un mejor entendimiento de ellos. Y finalmente el último apartado contempla un análisis exploratorio del set de datos.

1. Objetivo, Alcance y Contextualización de los Datos

Objetivos

Construir un modelo predictivo para el diagnóstico de cáncer de mama tumores, para tales efectos, se propone desarrollar un modelo de clasificación binaria para predecir entre un tumor Maligno (M) o Benigno (B).

El presente proyecto cuenta con 569 muestras de las cuales 357 corresponden a imágenes benignas y 212 malignas. Cada muestra posee una identificación (ID number), el diagnóstico (diagnosis) y 30 atributos relacionados a 10 características.

Alcance

El alcance del presente informe es realizar proceso exploratorio del set de datos a fin de entenderlos y visualizar si estos nos permitirán construir un modelo lo suficientemente eficiente y con un margen de error muy bajo de la predicción del diagnóstico.

Contextualización de la Data

La base de datos utilizada en este trabajo corresponde al trabajo publicado en la revista científica *Biomedical Image Processing and Biomedical Visualization* durante Julio de 1993 por el *Dr. William H. Wolberg* del departamento de Cirugía general de la Universidad de Wisconsin, *W. Nick Street* y *Olvi L. Mangasarian* del departamento de Ciencias de computación de la Universidad de Wisconsin.

Los atributos contenidos en la base de datos fueron obtenidos en base a técnicas de procesamiento interactivo de imágenes en conjunto con un algoritmo de clasificación inductivo basado en programación lineal, basado en la definición de contornos tal de caracterizar tamaños, formas y texturas celular (***Figura 1***).

Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

Figura 1: Límites Iniciales Aproximados de los Núcleos Celulares

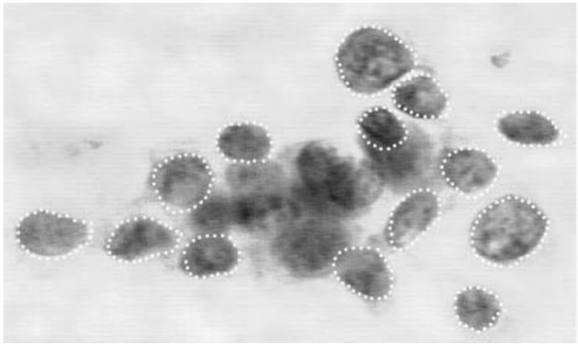


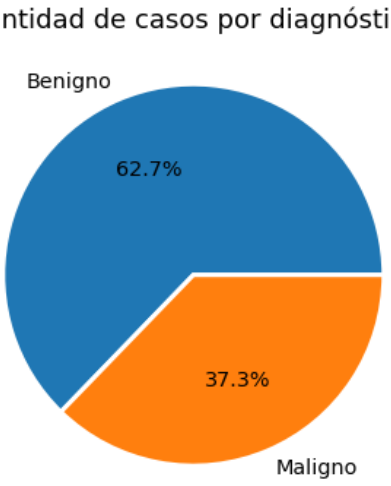
Figure 1: Initial Approximate Boundaries of Cell Nuclei
The user first draws a rough initial outline of some cell nucleus boundaries. Each outline serves as the initial position for a deformable spline which converges to an accurate boundary of the nucleus.

Fuente: Biomedical Image Processing and Biomedical Visualization
doi:10.1117/12.148698

2. Descripción de los Datos

La base de datos utilizada para este trabajo contiene 569 muestras, 357 de ellas benignas y 212 malignas (***Figura 2***).

Figura 2: Cantidad de Casos por Diagnóstico



Fuente: Elaboración propia

Cada una de las muestras posee una identificación (ID number), el diagnóstico (diagnosis) y 30 atributos relacionados a 10 características, las cuales son:

1. ***Radius***: Media de las distancias desde el centro hacia puntos del perímetro.
2. ***Texture***: Desviación estándar de las intensidades en escala de grises.
3. ***Perimeter***: Perímetro
4. ***Area:*** Área
5. ***Smoothness***: Variación local de largo de los radios.
6. ***Compactness***: (Perímetro)^2 / Área - 1.0

7. ***Concavity***: severidad de secciones cóncavas en la delineación.
8. ***Concave points***: Número de secciones cóncavas en la delineación.
9. ***Symmetry***.: Simetría
10. ***Fractal dimension***: Coastline approximation" – 1

Atributos del Dataset

En la base de datos, y como producto del experimento científico del cual proviene la información (*Nuclear features extraction for breast tumor diagnosis*[1]), figuran mediciones de 10 atributos base, registrando valores promedio de las mediciones (mean), medición del valor más alto (worst) y el error estándar de las mediciones de cada atributo (se), los que se describen a continuación:

Tabla 1: Descripción de 10 Atributos Base

	Promedio (mean)	Valor mas alto (worst)	Error estándar (se)
1	radius_mean	radius_worst	radius_se
2	texture_mean	texture_worst	texture_se
3	perimeter_mean	perimeter_worst	perimeter_se
4	area_mean	area_worst	area_se
5	smoothness_mean	smoothness_worst	smoothness_se
6	compactness_mean	compactness_worst	compactness_se
7	concavity_mean	concavity_worst	concavity_se
8	concave points_mean	concave points_worst	concave points_se
9	symmetry_mean	symmetry_worst	symmetry_se
10	fractal_dimension_mean	fractal_dimension_worst	fractal_dimension_se

Fuente: Elaboración propia

Al obtener una descripción rápida de los atributos disponibles, lo primero que podemos identificar es:

- Existencia de atributo *"Unnamed: 32"* que no contiene valores, por lo tanto, será eliminada de la base de dato.
- Atributo *"ID"* que no aporta información para nuestro modelo, por lo tanto, será eliminada de la base de dato.
- Los 31 atributos restantes no tienen valores nulos y son de tipo numéricos

[1] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). .Biomedical Image Processing and Biomedical Visualization. doi:10.1117/12.148698

Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

Luego de realizar la limpieza de datos correspondiente, nos quedamos con el siguiente Dataset:

Tabla 2: Dataset post Procesamiento de la Data (Limpieza)

#	Atributo	Cantidad Nulos	Tipo
0	diagnosis	569 non-null	object
1	radius_mean	569 non-null	float64
2	texture_mean	569 non-null	float64
3	perimeter_mean	569 non-null	float64
4	area_mean	569 non-null	float64
5	smoothness_mean	569 non-null	float64
6	compactness_mean	569 non-null	float64
7	concavity_mean	569 non-null	float64
8	concave points_mean	569 non-null	float64
9	symmetry_mean	569 non-null	float64
10	fractal_dimension_mean	569 non-null	float64
11	radius_se	569 non-null	float64
12	texture_se	569 non-null	float64
13	perimeter_se	569 non-null	float64
14	area_se	569 non-null	float64
15	smoothness_se	569 non-null	float64
16	compactness_se	569 non-null	float64
17	concavity_se	569 non-null	float64
18	concave points_se	569 non-null	float64
19	symmetry_se	569 non-null	float64
20	fractal_dimension_se	569 non-null	float64
21	radius_worst	569 non-null	float64
22	texture_worst	569 non-null	float64
23	perimeter_worst	569 non-null	float64
24	area_worst	569 non-null	float64
25	smoothness_worst	569 non-null	float64
26	compactness_worst	569 non-null	float64
27	concavity_worst	569 non-null	float64
28	concave points_worst	569 non-null	float64
29	symmetry_worst	569 non-null	float64
30	fractal_dimension_worst	569 non-null	float64

Fuente: Elaboración propia

Tabla 3: Resumen de Características del Dataset

Cantidad observaciones	**569**
Característica atributos	Numérico
Número de Atributos	30
Datos faltantes	No
Modelo	Clasificación

Fuente: Elaboración propia

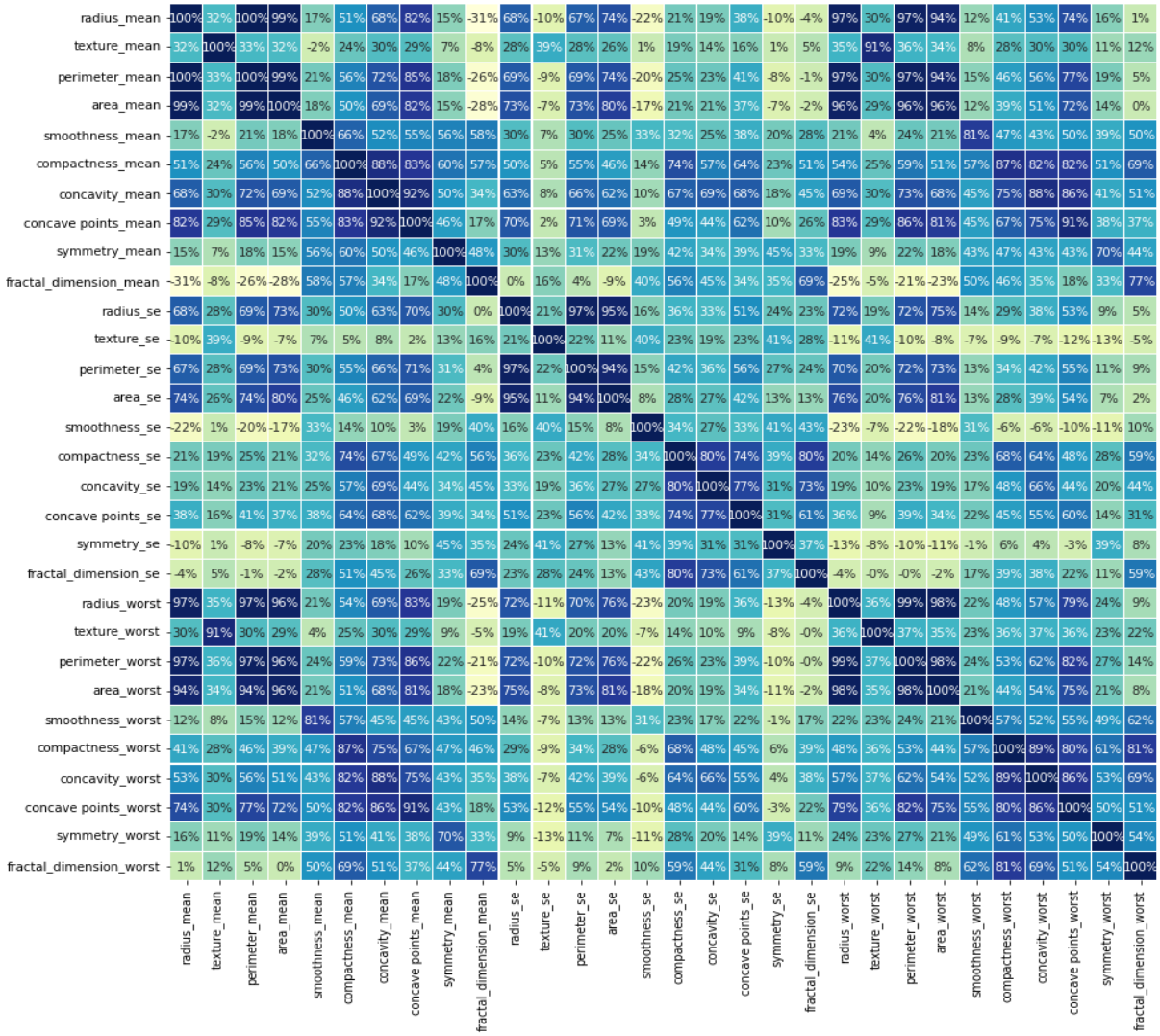
Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

3. Análisis Exploratorio de los Datos

Matriz de correlación

Mediante la siguiente matriz de correlación podremos identificar que atributos nos aportará más información, en función de que tan correlacionadas están entre ellas, por lo tanto, si el valor de correlación es alto indica que los atributos miden la misma característica de la imagen.

Figura 3: Matriz de correlación entre Atributos



Fuente: Elaboración propia

De acuerdo con la figura 3, se puede identificar que algunas variables se encuentran con un alto porcentaje de correlación, por lo cual nos quedaremos solo con una de ellas.

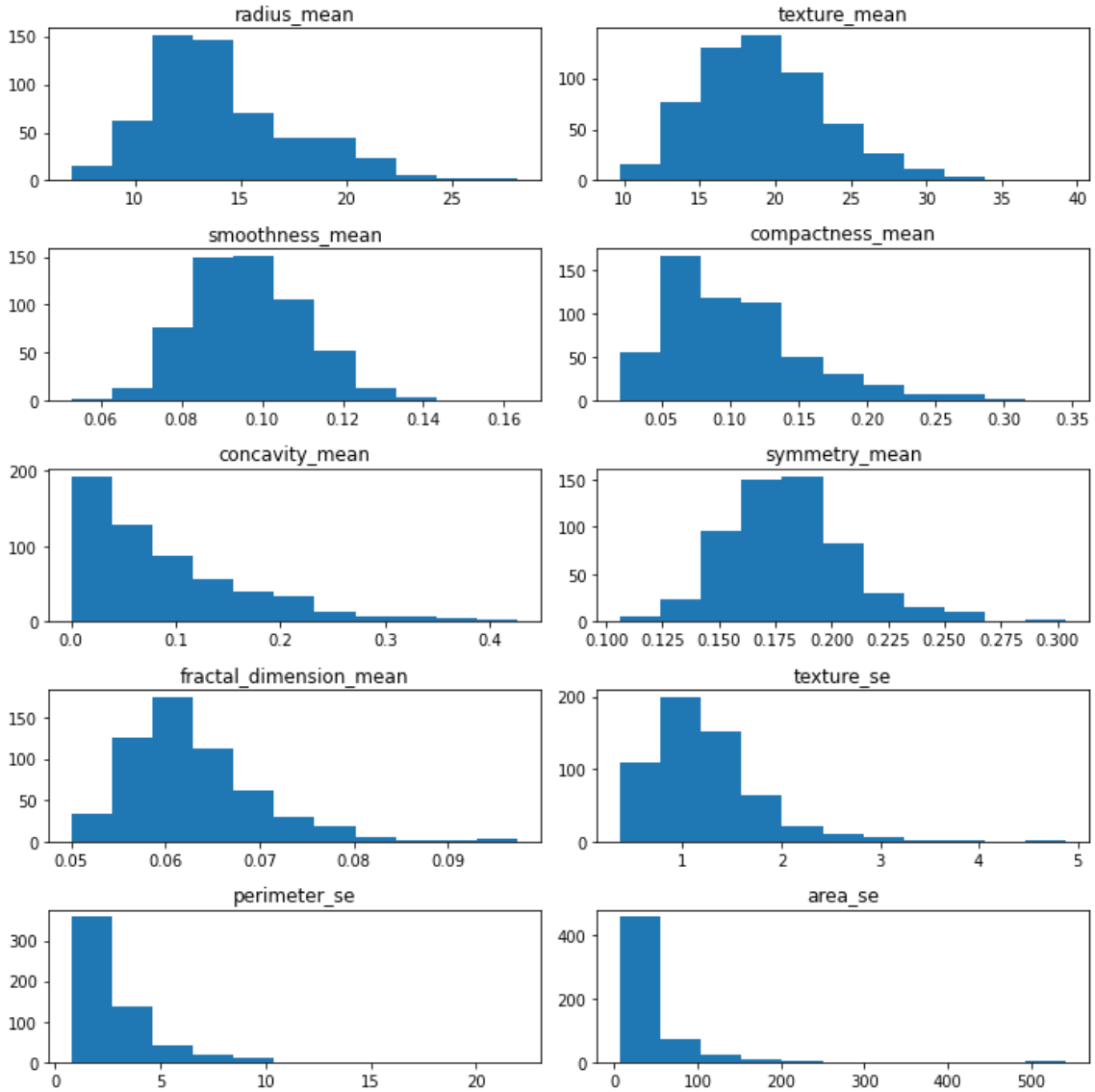
- **"Radius_mean"** tiene una correlación de 100% con **"perimeter_mean"** y 99% con **"area_mean"**, 97% con **"radius_worst"**, 97% con **"perimiter_worst"**, 94% con **"areas_worst"**, por lo tanto, nos quedaremos solo con **"Radius_mean.**

- **"Texture_mean"** tiene una correlación de 91% con **"textura_worst"**, por lo cual nos quedaremos solo con **"texture_mean".**
- **"perimeter_se"**, tiene una correlación de 97% con **"radius_se"** y "áreas_se" tiene una correlación de 95% con "radius_se" por lo cual nos descartaremos **"radius_se".**
- **"concave points_mean"**, tiene una correlación de 92% con "**concave_mean",** por lo cual nos quedamos solo con **"concave_mean".**

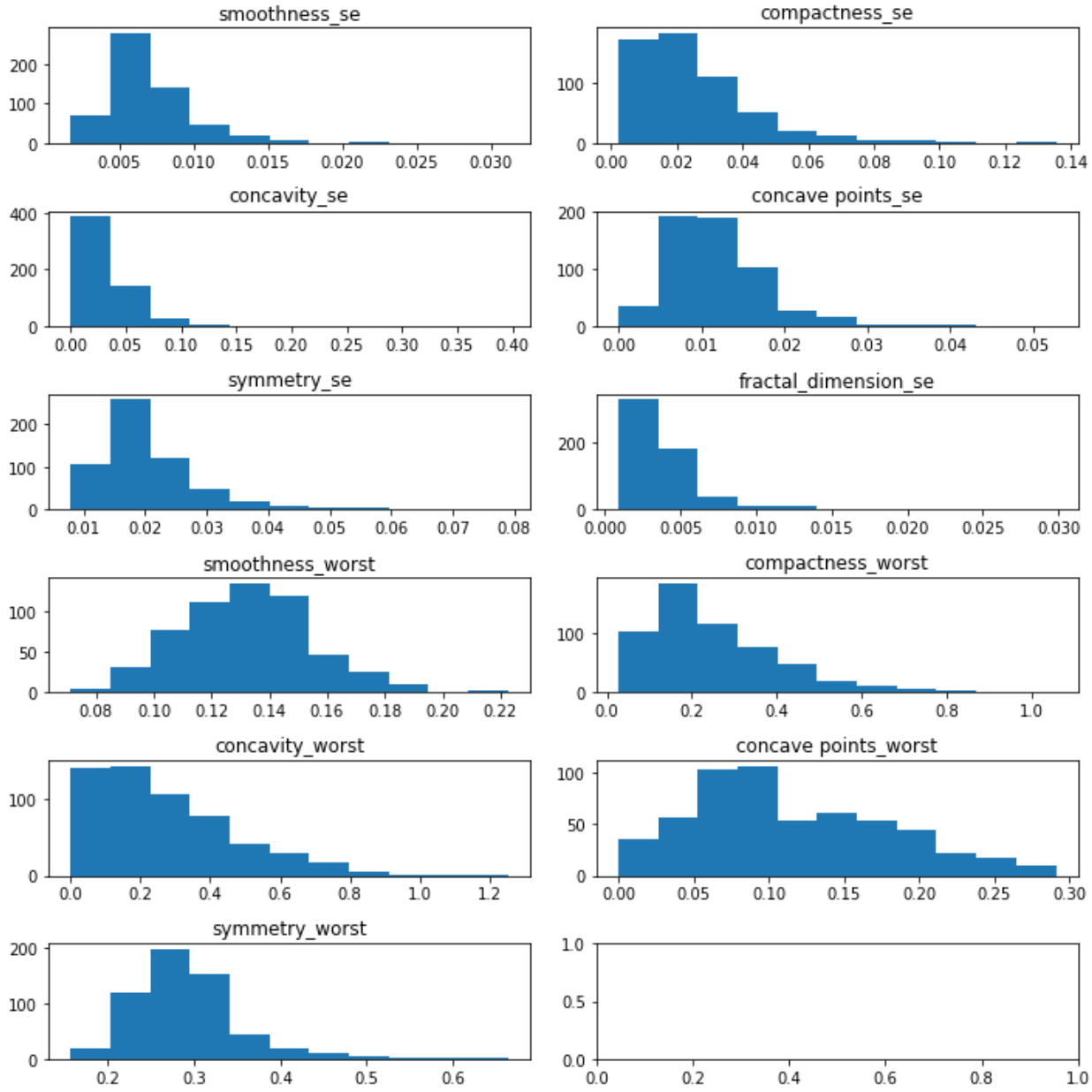
A continuación, presentaremos la distribución que posee cada una de las variables o atributos del Dataset.

Distribución Histogramas

Figura 4: Distribución de cada Atributo



Loreto Mora - Oscar Hermosilla - Mauricio Narváez.



Fuente: Elaboración propia

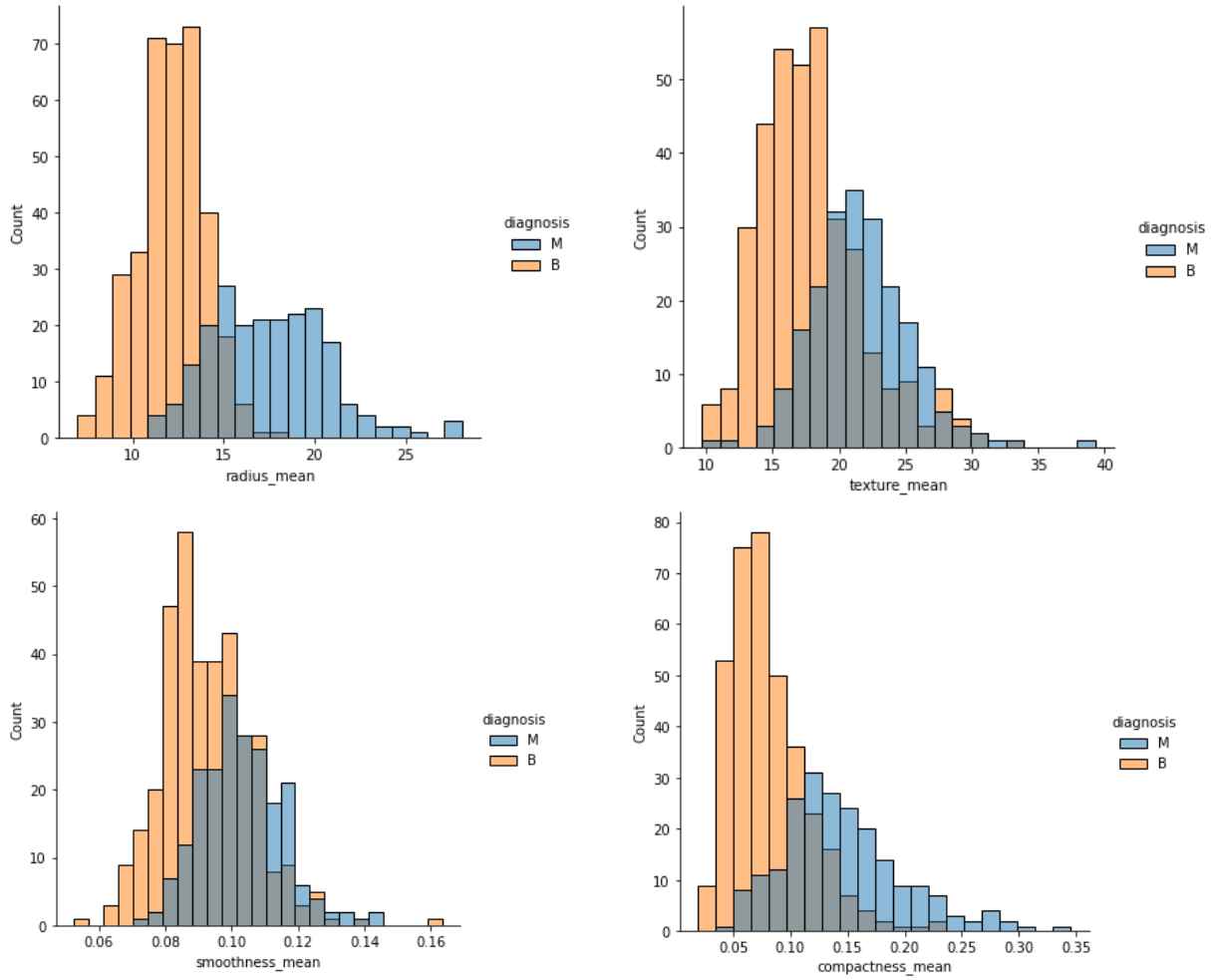
En función de la figura 4, al visualizar las distintas variables o atributos a través de los histogramas, podemos identificar que algunas tienen valores fuera de rango, por lo cual necesitaremos realizar una limpieza sobre ellas para que no generen ruidos al momento de clasificar. Por ejemplo: concavity_mean, texture_se, area_se o perimeter_se.

Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

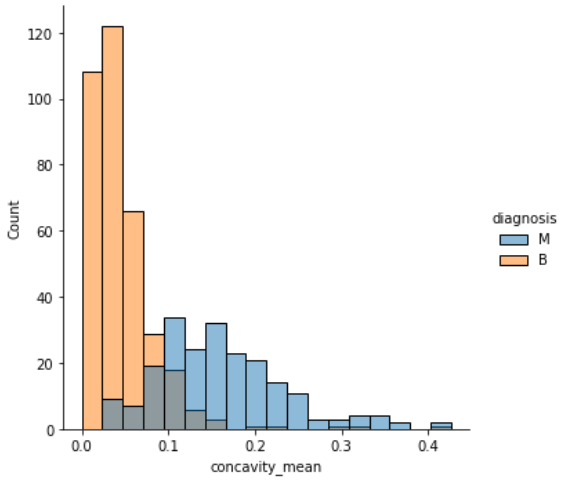
Comparación

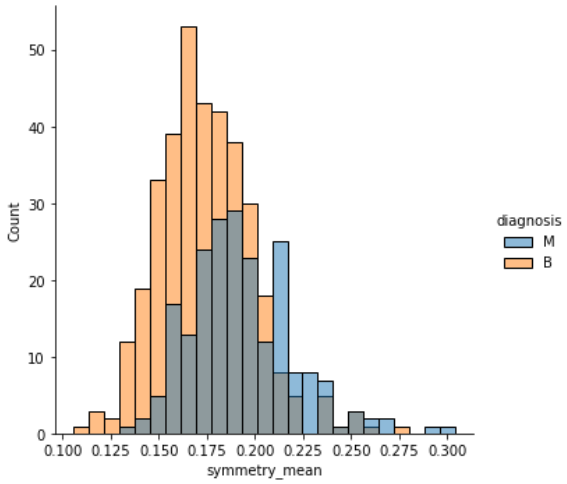
A continuación, compararemos cada una de las variables versus nuestra variable dependiente, con la finalidad de identificar si la variable nos ayudará a clasificar de manera correcta si el diagnóstico es Maligno (M) o Benigno (B)

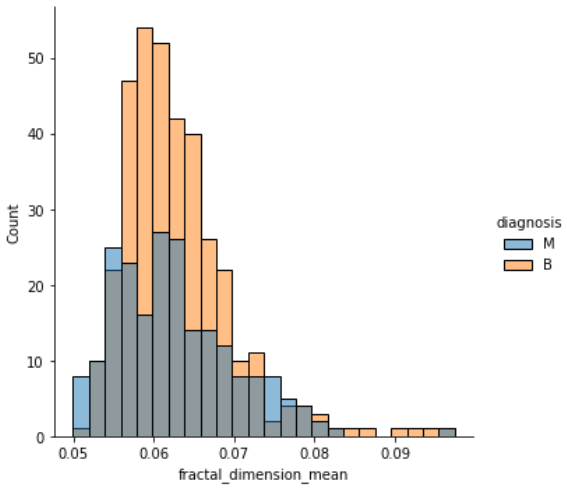
Figura 5: Análisis Comparativo entre los Atributos y la Variable Dependiente

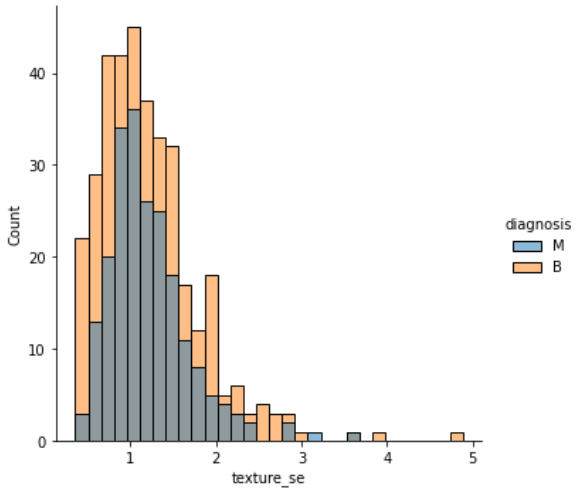


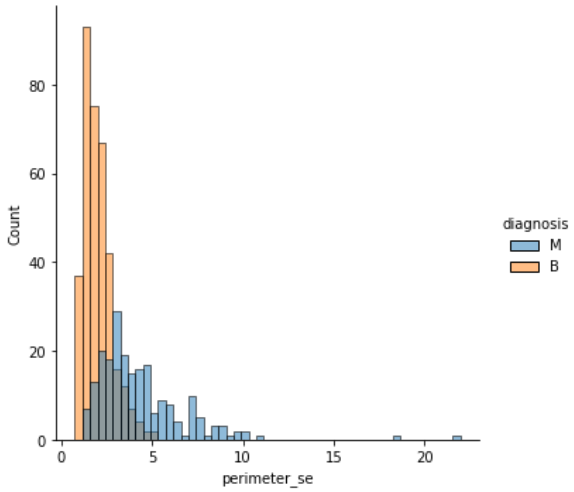
Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

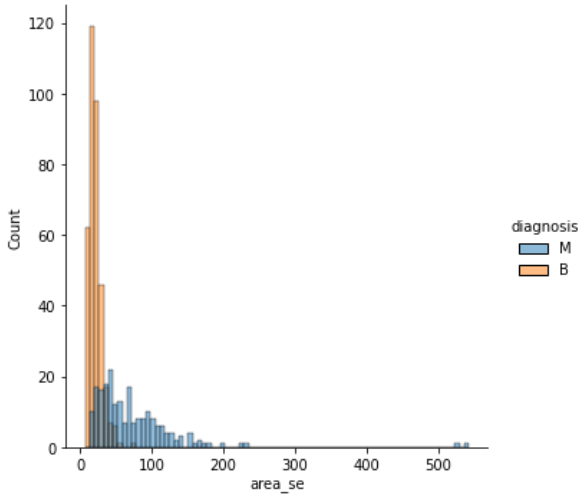




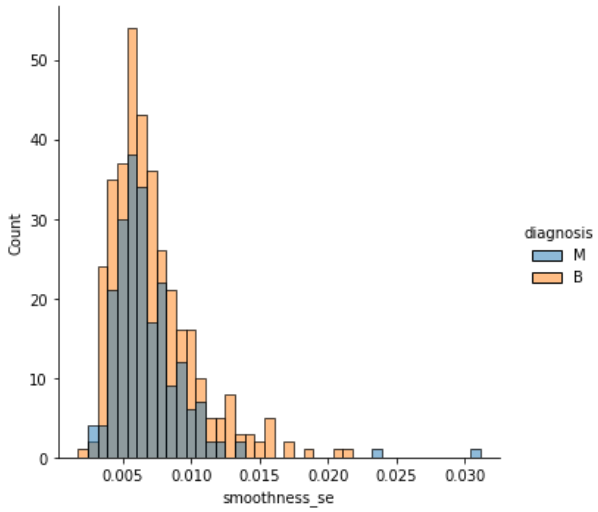


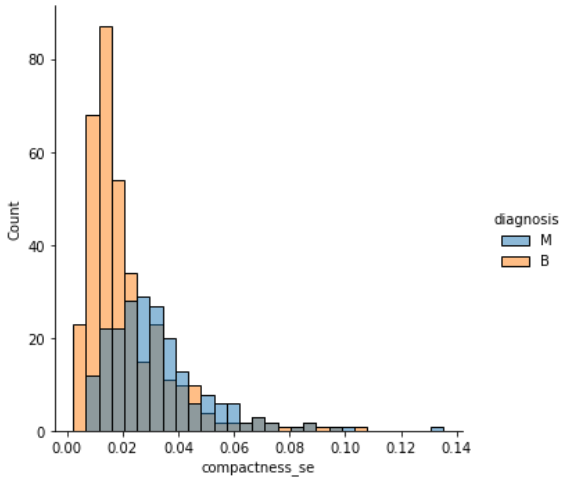


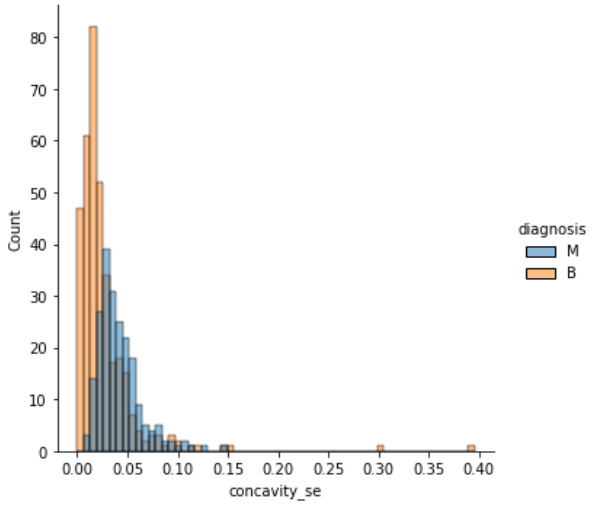


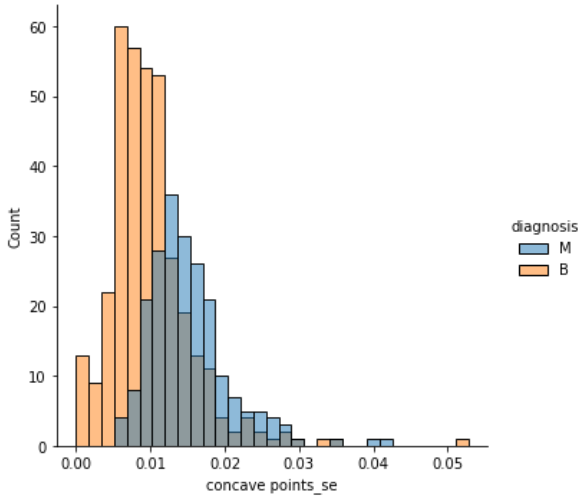


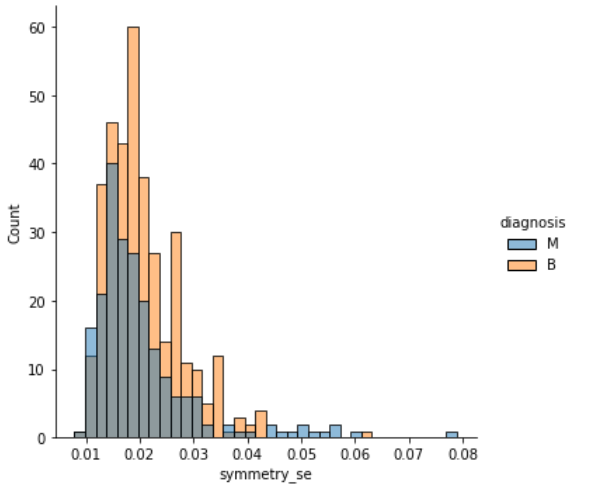
Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

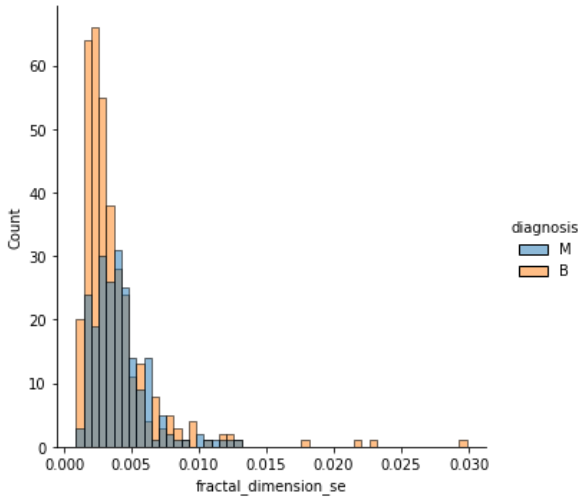




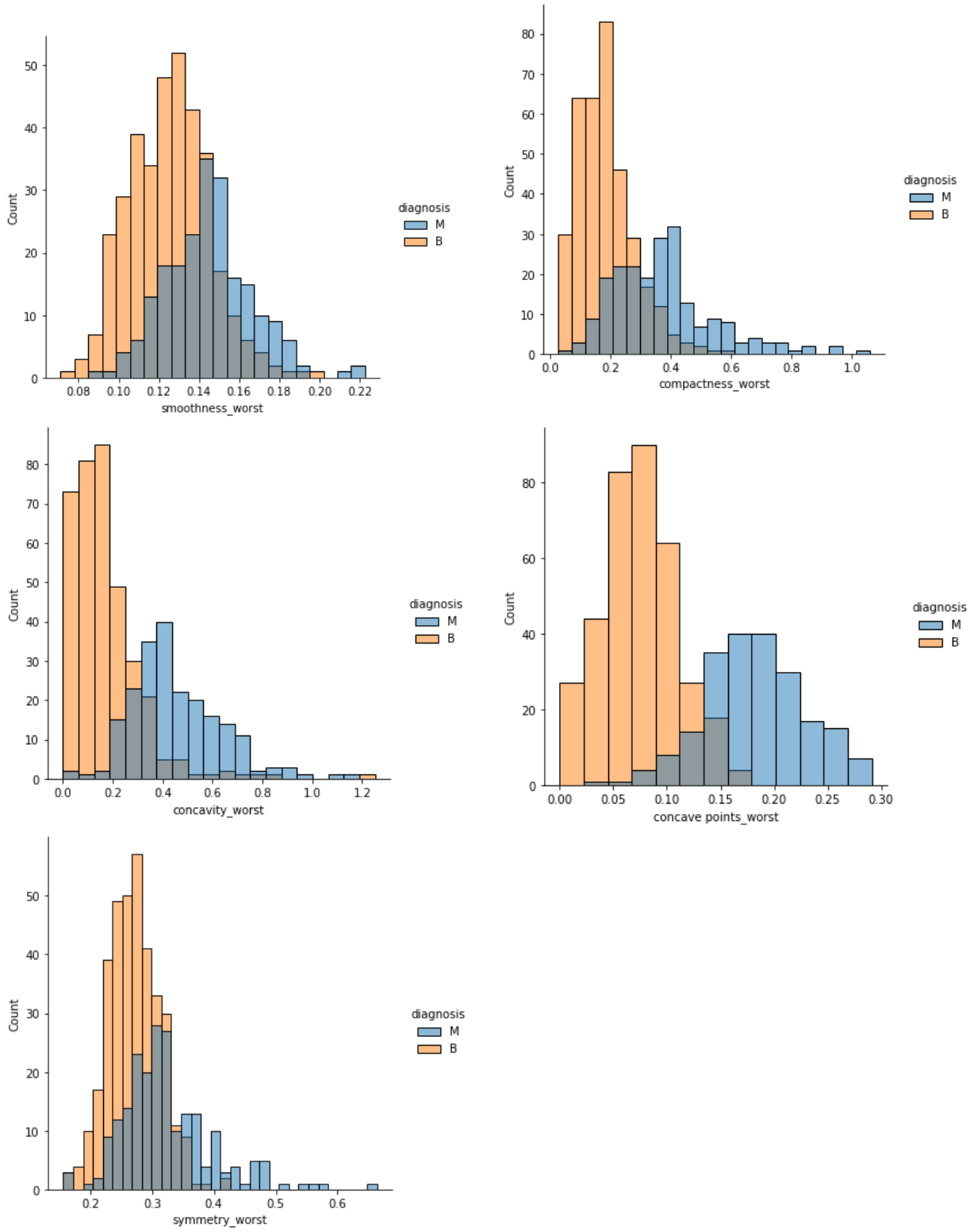








Loreto Mora - Oscar Hermosilla - Mauricio Narváez.



Fuente: Elaboración propia

Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

Al visualizar cada uno de los gráficos presentados en la figura 5, podemos ver que existen variables que nos permitirán clasificar, tales como:

- radius_mean
- compactness_mean
- concavity_mean
- perimeter_se
- area_se
- compactness_worst
- concavity_worst
- concave points_worst
- fractal_dimension_worst

Mientras que otras variables no muestran una preferencia particular de un diagnóstico sobre otro, como son el caso de las siguientes variables:

- texture_mean
- moothness_mean
- symmetry_mean
- fractal_dimension_mean
- texture_se smoothness_se
- compactness_se, concavity_se
- concave points_se, symmetry_se
- fractal_dimension_se, smoothness_worst
- symmetry_worst