Loreto Mora - Oscar Hermosilla - Mauricio Narváez.

Proyecto
Diagnóstico cáncer de mama

El presente documento abarca los avances solicitados, durante el desarrollo del Curso: Fundamentos en Ciencia de Datos, asociados a la elaboración y la implementación progresiva de los conocimientos adquiridos en el Curso a través de la ejecución del proyecto Diagnóstico de cáncer de mama.

Dicho lo anterior, el desglose del informe engloba dos apartados; ***Informe 1*** incluye una breve introducción sobre la temática a abordar, seguido del objetivo del proyecto, su alcance y una contextualización del Dataset lo que permite la posterior descripción más detallada del conjunto de datos (cantidad de datos, de dónde vienen, sus distintos atributos y etiquetas, etc.), a fin de obtener un mejor entendimiento de ellos, lo que finalmente será conducente al análisis exploratorio del Dataset. Todo lo anterior facilitará la elaboración del presente ***Informe 2***, correspondiente a la descripción, análisis y/o evaluación del modelo predictivo propuesto, finalizando con las respectivas conclusiones y posibles trabajos futuros.

Informe 1
01 Octubre 2021

En el siguiente apartado se procede a presentar introducción, objetivo y alcance del proyecto, contextualización y descripción detallada del Dataset, el que será conducente al análisis exploratorio del mismo.

Introducción

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 motiva a desarrollar el presente proyecto.

Objetivo

Construir un modelo predictivo para el diagnóstico de cáncer de mama, para tales efectos, se propone trabajar con distintos modelos 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 Dataset a fin de entenderlos y visualizar si estos nos permitirán construir un modelo de predicción del diagnóstico de cáncer de mama lo suficientemente eficiente y con bajo margen de error. Posteriormente

extrapolar distintos modelos predictivos de clasificación para finalmente seleccionar aquel modelo que entregue la mejor predicción con el menor error.

Contextualización del Dataset

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***).

Figura 1: Límites Iniciales Aproximados de los Núcleos Celulares

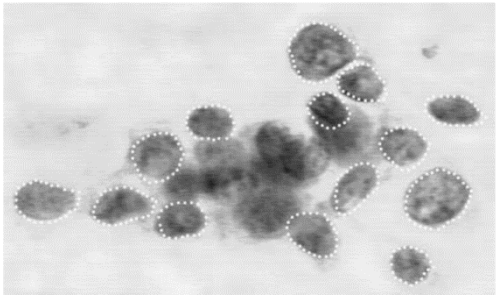


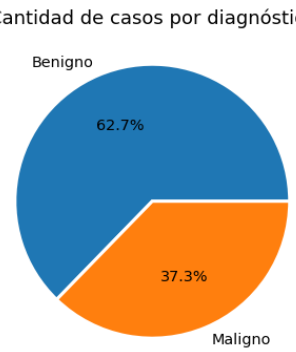
Figure 1: Initial Approximate Boundaries of Cell Nuclei
The user first draws a rough initial outline of some cell nucleus boundaries. Each outline serves as the initial position for a deformable spline which converges to an accurate boundary of the nucleus.

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

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 se presentan en la tabla 1,

Tabla 1: Características del Dataset

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

Fuente: Elaboración propia

Atributos del Dataset

En el Dataset, 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 2: 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

Luego de realizar la limpieza de datos correspondiente, nos quedamos con el siguiente Dataset, ver tablas 3 y 4,

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

Tabla 3: 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 4: Resumen de Características del Dataset

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

Fuente: Elaboración propia

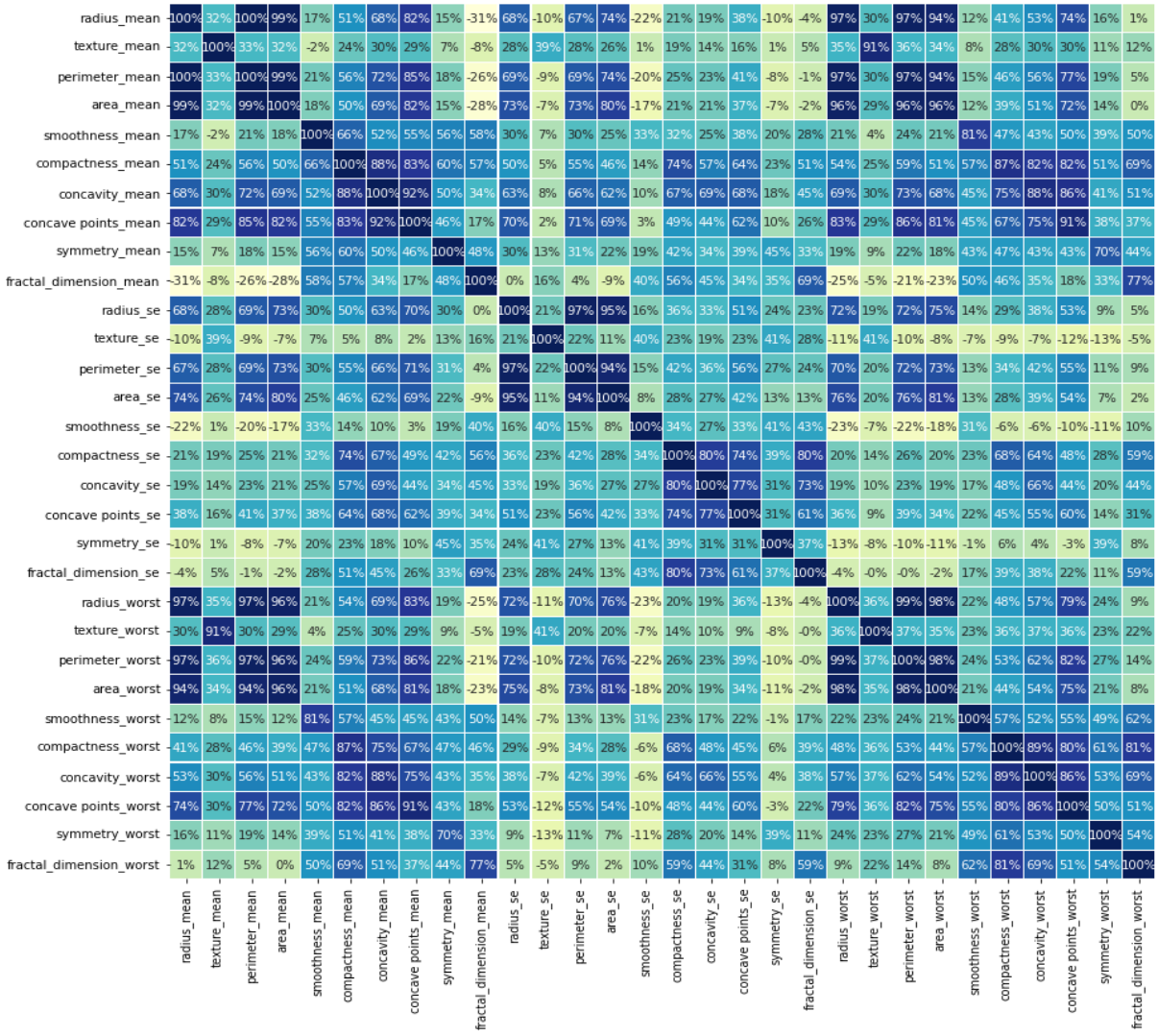
Ya terminada la descripción de los datos con los que cuenta el Dataset, a continuación procedemos a realizar el análisis exploratorio del mismo.

Análisis Exploratorio del Dataset

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

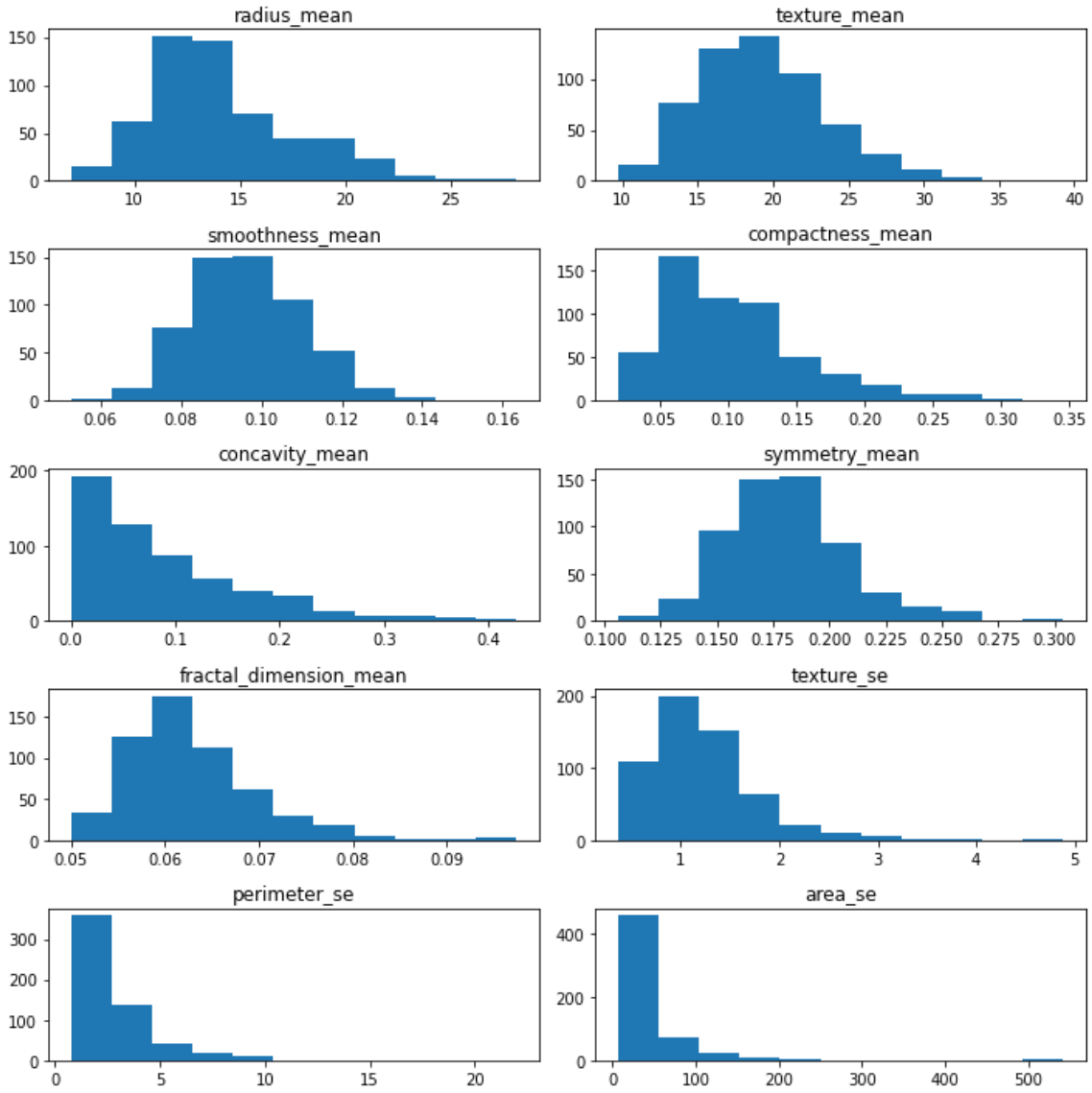
De acuerdo con la figura 3, se puede identificar que algunas variables se encuentran con un alto porcentaje de correlación, por lo cual nos quedaremos solo con una de ellas.

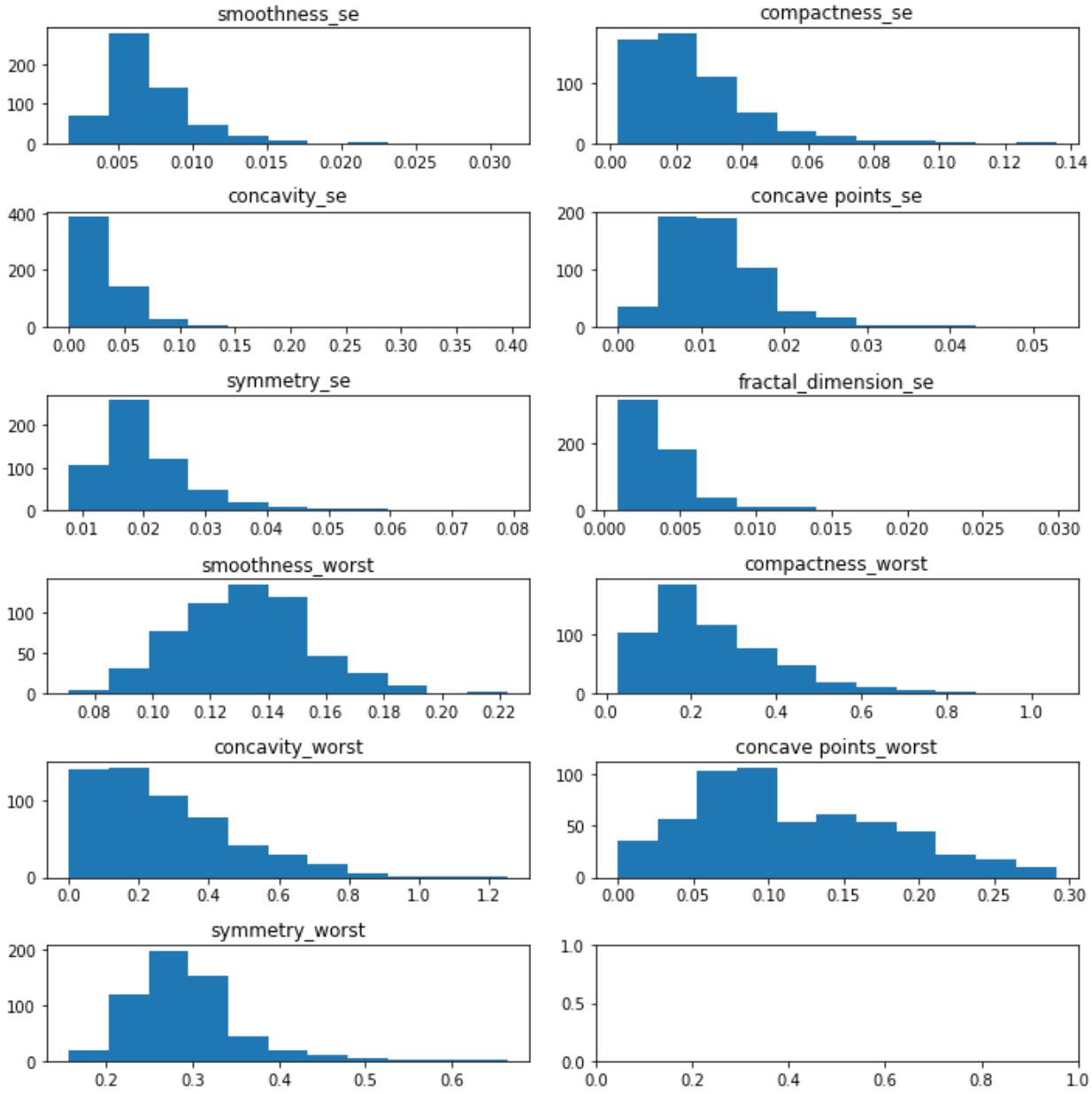
- "radius_mean" tiene una correlación de 100% con "perimeter_mean" y 99% con "area_mean", 97% con "radius_worst", 97% con "perimiter_worst", 94% con "area_worst", por lo tanto, nos quedaremos solo con **"radius_mean".**
- "texture_mean" tiene una correlación de 91% con "textura_worst", por lo cual nos 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 quedaremos con **"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





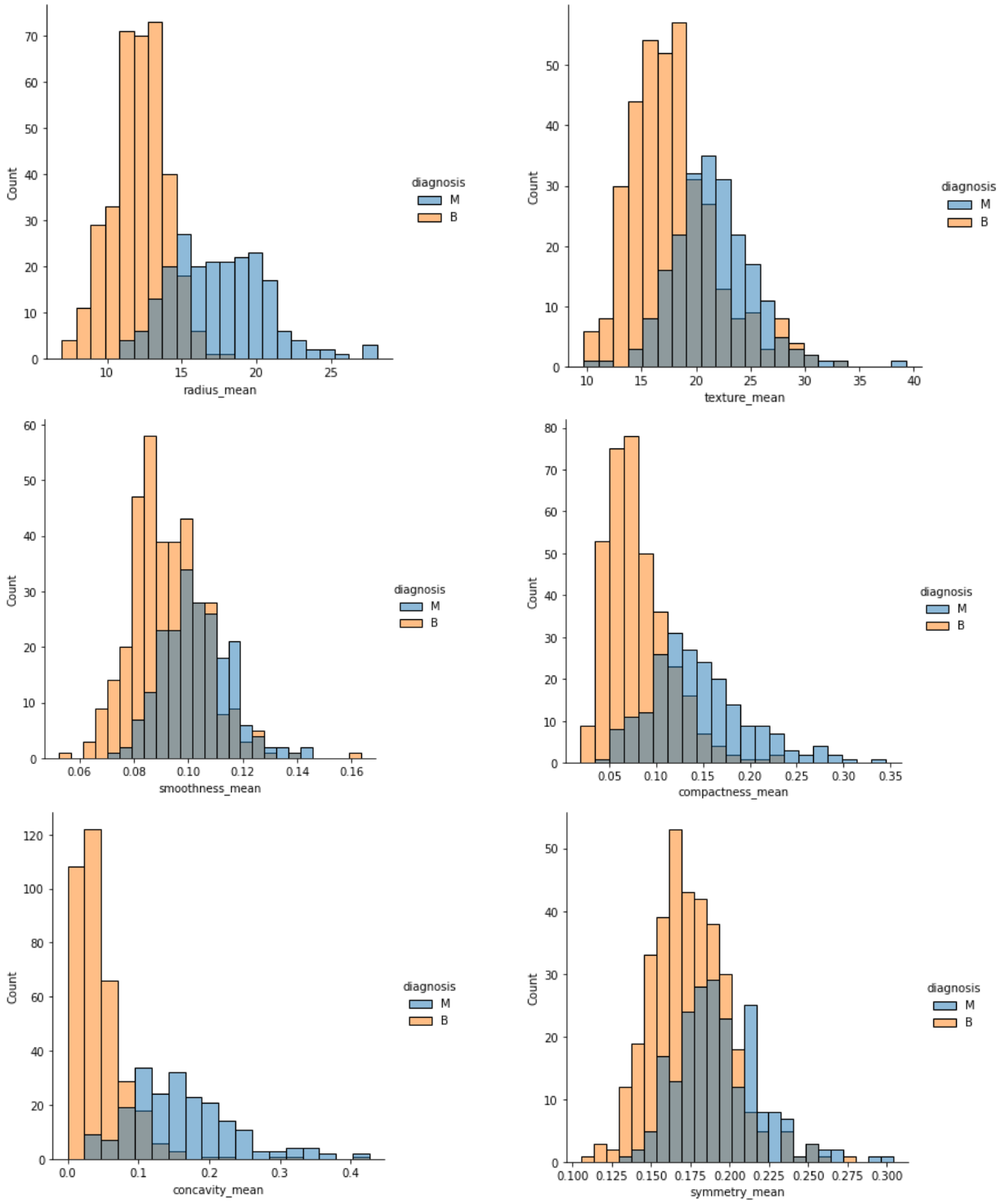
Fuente: Elaboración propia

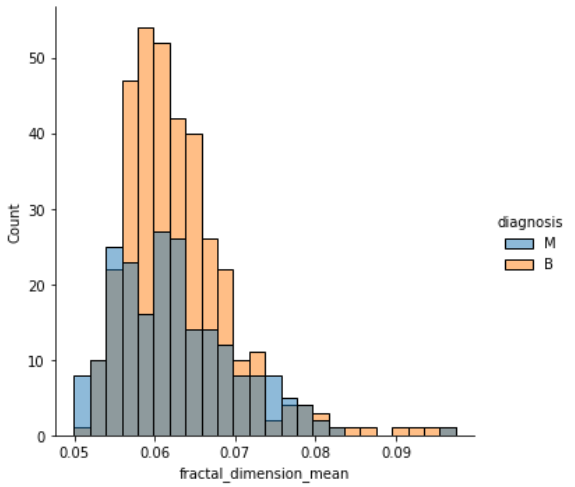
En función de la figura 4, al visualizar las distintas variables o atributos a través de los histogramas, podemos identificar que algunas tienen valores fuera de rango, por lo cual necesitaremos realizar una limpieza sobre ellas para que no generen ruidos al momento de clasificar. Por ejemplo: “concavity_mean”,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

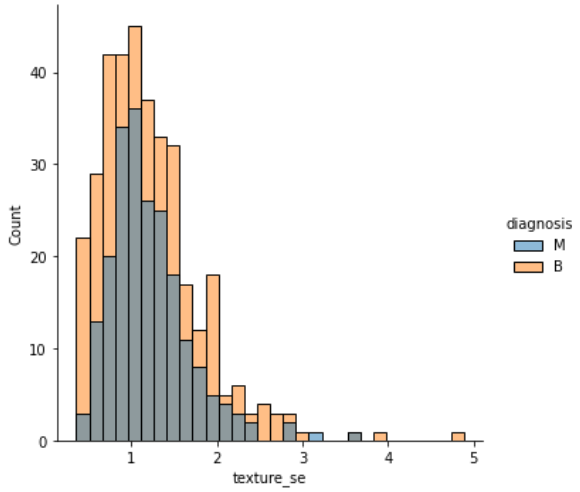
Comparación

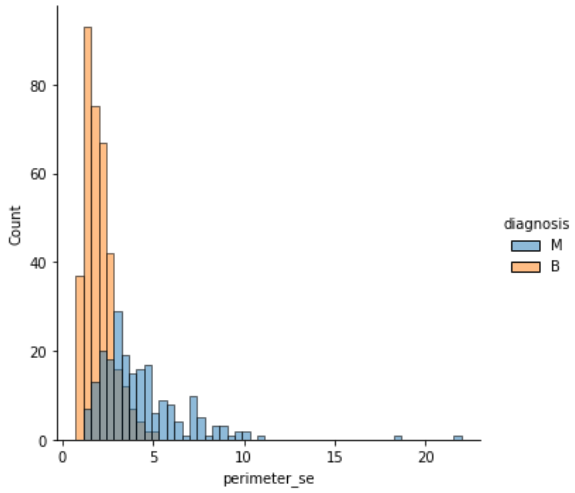
A continuación, compararemos cada una de las variables versus nuestra variable dependiente, con la finalidad de identificar si la variable nos ayudará a clasificar de manera correcta si el diagnóstico es Maligno (M) o Benigno (B)

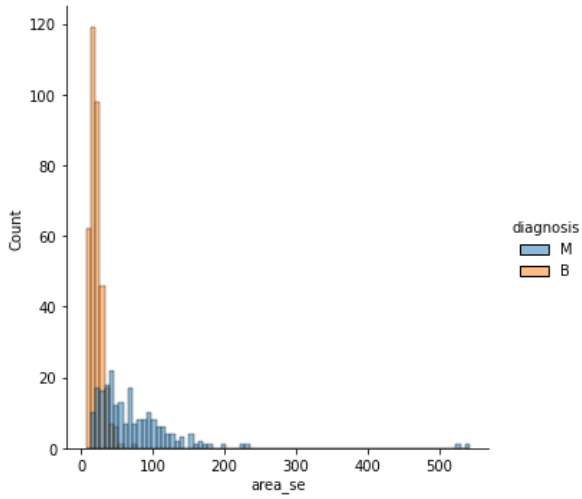
Figura 5: Análisis Comparativo entre los Atributos y la Variable Dependiente

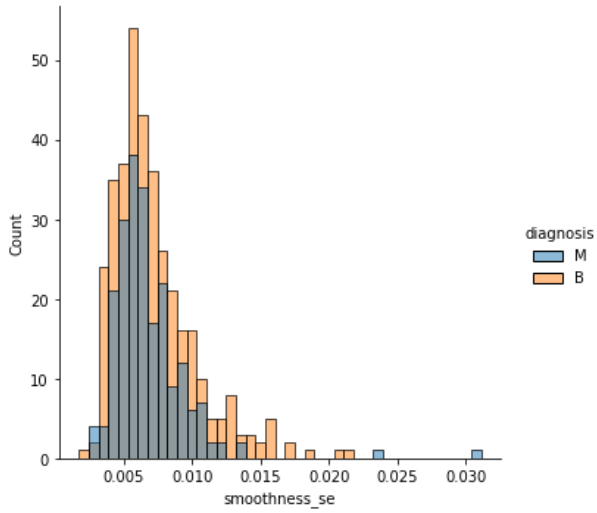


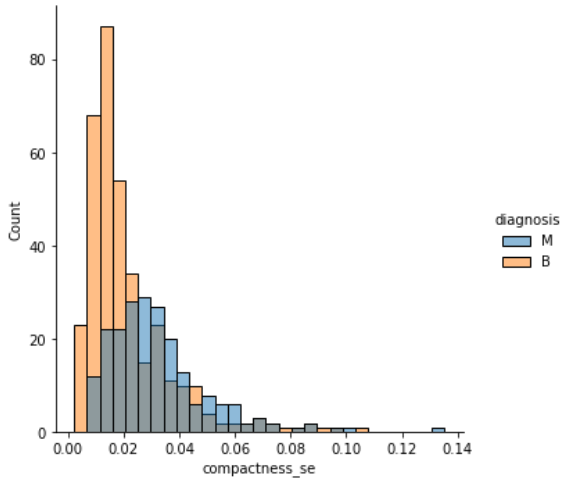

Count
fractal_dimension_mean
diagnosis
M
B

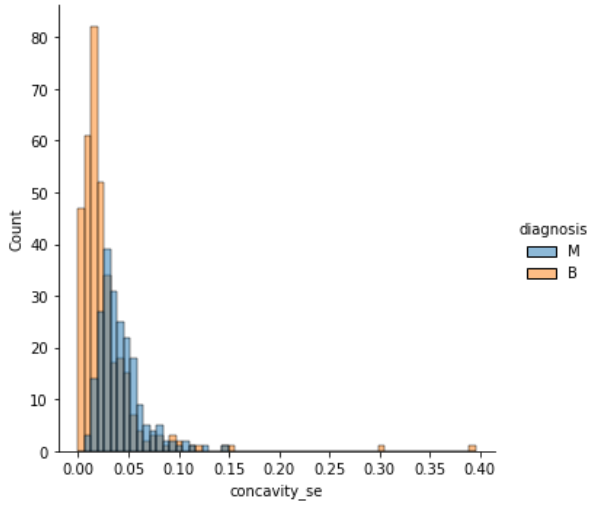

Count
texture_se
diagnosis
M
B

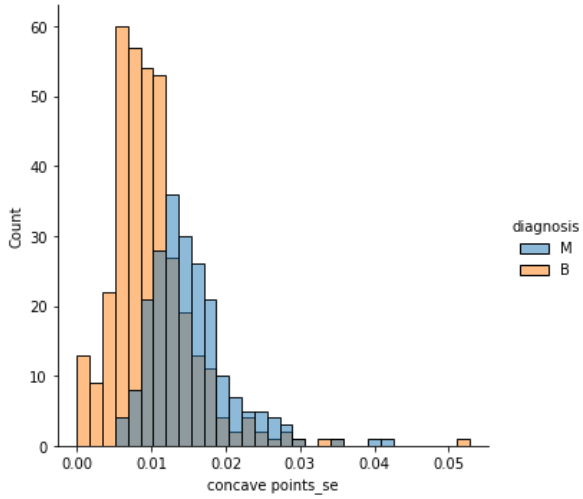

Count
perimeter_se
diagnosis
M
B

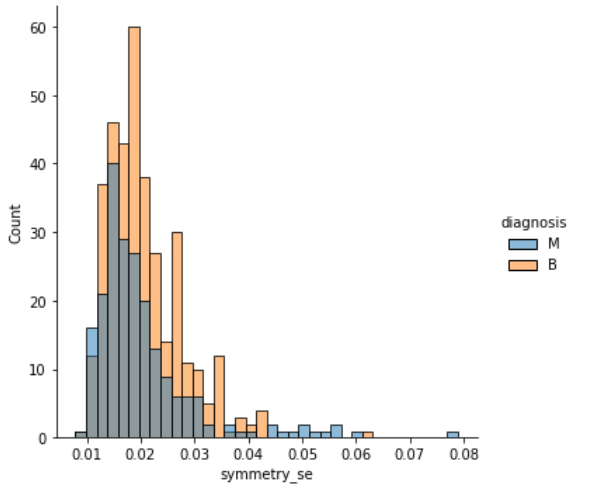

Count
area_se
diagnosis
M
B

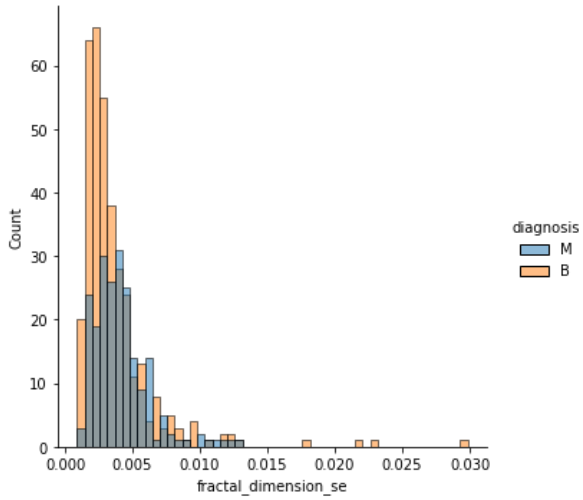

Count
smoothness_se
diagnosis
M
B

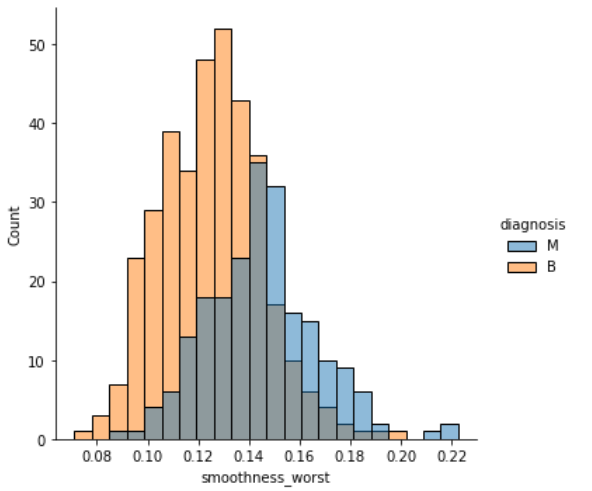

Count
compactness_se
diagnosis
M
B

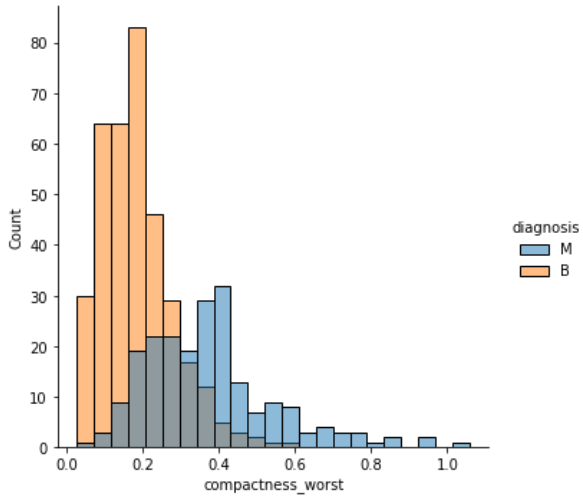

Count
concavity_se
diagnosis
M
B

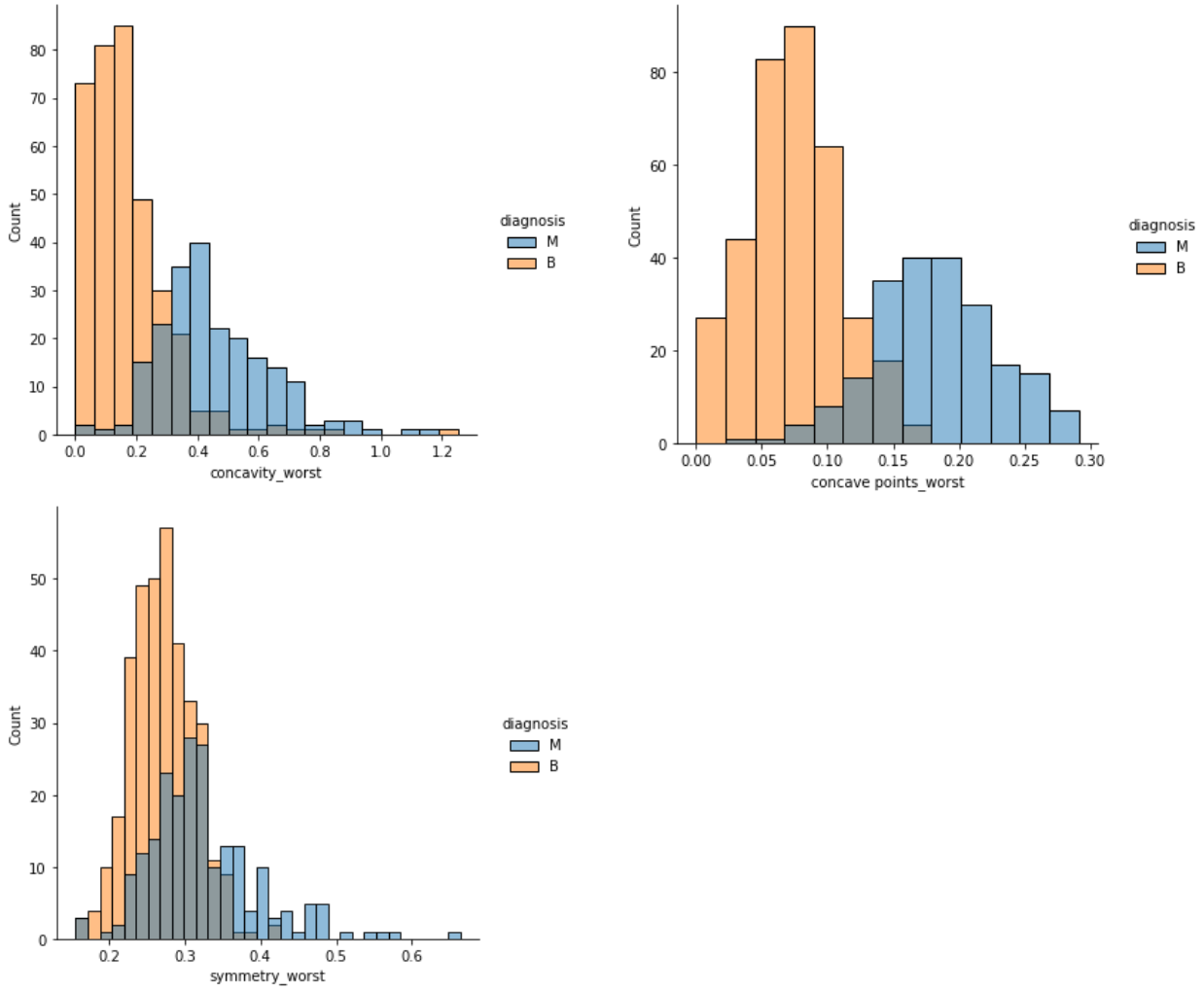

Count
concave points_se
diagnosis
M
B


Count
symmetry_se
diagnosis
M
B


Count
fractal_dimension_se
diagnosis
M
B


Count
smoothness_worst
diagnosis
M
B


Count
compactness_worst
diagnosis
M
B



Fuente: Elaboración propia

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
- 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
- smoothness_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

Informe 2
23 Octubre 2021

El siguiente apartado entrega la propuesta del o los modelos predictivos con su respectiva descripción, análisis y/o evaluación asociada a cada uno.

Los modelos predictivos propuestos para el conjunto de datos Diagnóstico de cáncer de mama son de tipo Clasificación.

Antes de definir el o los modelos a implementar se realizan algunos pasos previos; primero se hace una preparación del Dataset, de acuerdo a los resultados obtenidos del analisis exploratorio previamente realizado con el objetivo de definir las variables a utilizar en los modelos. (ver tabla 5)

Tabla 5: Variables seleccionadas del Dataset a utilizar en modelos

#	Variable
0	diagnosis
1	radius_mean
2	compactness_mean
3	concavity_mean
4	radius_se
5	compactness_worst
6	concavity_worst
7	concave points_worst
8	fractal_dimension_worst

Fuente: Elaboración propia

Luego la tabla 6 muestra la separación del Dataset entre el conjunto de variable dependiente (y) con el conjunto de datos de entradas (X).

Tabla 6: Conjunto de dato de variable dependiente (Y) y conjunto de datos/variables de entrada (X)

Conjunto	Filas	Columnas
X	569	8
y	569	

Fuente: Elaboración propia

A continuación, se debe dividir el Dataset en Conjunto de Datos de **Entrenamiento** (muestras utilizadas para entrenar el modelo, ajustar los hiperparámetros y evaluar los modelos) y de **Test** (para probar el modelo entregando muestras nuevas nunca vistas antes por el modelo). En este contexto, (ver tabla 7) para efectos del presente proyecto primero se procede a dividir el Dataset bajo el criterio de asignación de un *75% entrenamiento* y *25% test*.

Tabla 7: División del Dataset para entrenar y evaluar el o los modelos

Conjunto	Filas	Columnas
X_train	426	8
y_train	426	
X_test	143	8
y_test	143	

Fuente: Elaboración propia

Finalmente se **escalan** los datos de ambos conjuntos definidos anteriormente (entrenamiento /test) a fin de dejarlos en una escala **entre 0 y 1**.

Una vez terminada la división del Dataset anteriormente descrita, se procede a entrenar diferentes tipos de modelos para finalmente hacer el análisis comparativo y seleccionar el mejor.

Modelamiento y Resultados

Entrenamiento

Con el objetivo de encontrar la mejor clasificación de datos de manera tal que se logre la mejor representatividad de la estructura intrínseca de los datos, se optó por entrenar los datos con cuatro modelos distintos y así seleccionar aquel que mejor predijera respecto a la variable dependiente de entrenamiento. Para tales objetivos, para cada modelo seleccionado se generan métricas que permiten evaluar el desempeño de cada de ellos en la tarea de predecir. En este contexto, se aplica la técnica de validación cruzada con 5 “folds” para cada uno de los modelos de clasificación, lo que se presentan a continuación.

Modelo - Regresión Logística

Tabla 8: LogisticRegression (random_state = 0)

Accuracy	**94,13%**
Cross Validation Score 1	90,69%
Cross Validation Score 4	94,11%
Cross Validation Score 3	91,76%
Cross Validation Score 4	94,11%
Cross Validation Score 5	95,29%
Cross Validation Mean	**90,69%**
Cross Validation Std	1.69%

Fuente: Elaboración propia

De acuerdo con el resultado presentado en la tabla 8, se observa que el valor “Accuracy” como la media de la validación cruzada son altos, por lo cual podemos decir que es un modelo que nos permitiría realizar una predicción optima.

Modelo - Árbol de Decisión

Tabla 9: DecisionTreeClassifier (criterion = 'entropy', random_state = 0)

Accuracy	**100%**
Cross Validation Score 1	91,86%
Cross Validation Score 4	95,29%
Cross Validation Score 3	90,58%
Cross Validation Score 4	90,58%
Cross Validation Score 5	87,05%
Cross Validation Mean	**91,07%**
Cross Validation Std	2,64%

Fuente: Elaboración propia

Se observa en la tabla 9 que el valor del "Accuracy" evidencia un sobre ajuste del modelo dado que la precisión disminuye a un 91.07%.

Modelo - Support Vector Machine

Tabla 10: SVC (kernel='linear', C=1, random_state=0)

Accuracy	**93.66%**
Cross Validation Score 1	91,86%
Cross Validation Score 4	92,94%
Cross Validation Score 3	92,94%
Cross Validation Score 4	94,11%
Cross Validation Score 5	92,94%
Cross Validation Mean	**92,96%**
Cross Validation Std	0.07%

Fuente: Elaboración propia

En este caso se puede apreciar en la tabla 10 que ambos valores, del "Accuracy" y de la media de la validación cruzada, son altos y muy similares, información que nos lleva a suponer que el modelo – SVM podría ser un buen candidato para seleccionar para la predicción.

Modelo - Random Forest

Tabla 11: RandomForestClassifier (n_estimators = 100, criterion = 'entropy', random_state = 0, min_samples_split=25, max_depth=5, max_features=2)

Accuracy	**96%**
Cross Validation Score 1	88,37%
Cross Validation Score 4	96,47%
Cross Validation Score 3	90,58%
Cross Validation Score 4	95,29%
Cross Validation Score 5	89,17%
Cross Validation Mean	**93.2%**
Cross Validation Std	2,87%

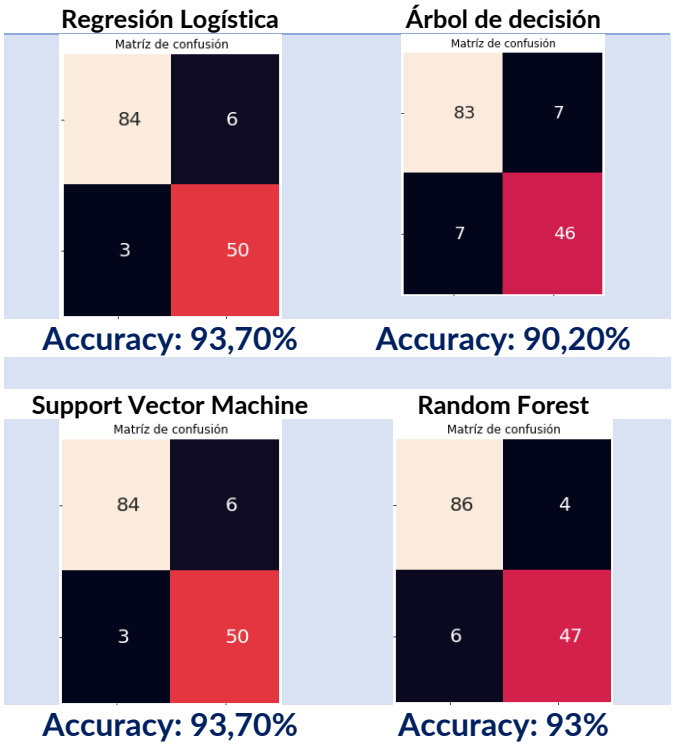
Fuente: Elaboración propia

De acuerdo con la tabla 11, el modelo Random Forest posee un "Accuracy" alto al igual que la media de la validación cruzada, lo que es coincidente con el anterior modelo.

Predicción

Al comparar los resultados obtenidos por las métricas de los cuatros modelos, el último de ellos es el modelo que mejor desempeño presenta en la predicción. Para confirmar tal hipótesis, se procede a realizar la predicción con los datos de test, cuyos resultados de despeño se visualizan en la tabla 12 resumen:

Tabla 12: Resumen del desempeño de cada modelo en la predicción



Fuente: Elaboración propia

Según los resultados presentes en la matriz de confusión se puede concluir que el Modelo Árbol de Decisión es el que peor desempeño presenta, mientras que en los casos del Modelo de Regresión Logística y el Modelo Support Vector Machine, ambos presentan un Accuracy del 93,7% siendo el valor más alto alcanzado entre todos los modelos. No obstante, el Modelo Random Forest un Accuracy muy cercano a los valores obtenidos en los dos últimos modelos recientemente mencionado.

Conclusiones y Trabajo Futuro

En términos de predicción y resultado obtenido por cada uno de los modelos propuestos, se puede concluir que los *Modelos de Regresión Logística* y el *Support Vector Machine*, ambos logran cumplir con el objetivo de predecir el cáncer de mamas con un porcentaje de predicción del ***93.7%.*** que permite genera alertas preventivas o bien un diagnóstico a tiempo que posibilita el éxito de los tratamientos de los pacientes. Sin embargo, el proceso de preparación y limpieza de los datos (realizar imputación de los valores fuera de rango o cambiar algunos parámetros en los modelos) aún debe ser depurada para que el modelo pueda entregar una mejor precisión en la clasificación predictiva para un diagnóstico.