Proyecto Data Science

CÁNCER DE MAMA

Introducción al Conjunto de Datos

Coder House.

17/07/2024.

Autor:

Marini, Ian Denis

ÍNDICE

Descripción

El conjunto de datos del Cáncer de Mama Wisconsin (Diagnóstico) contiene características detalladas de las masas mamarias obtenidas a partir de imágenes digitalizadas de aspiraciones con aguja fina (PAAF).

Estas características describen los núcleos celulares presentes en la imagen y se utilizan para predecir si el cáncer es benigno o maligno.

Variables

ATRIBUTOS UTILIZADOS

- ID Número
- Diagnóstico
  (M = maligno, B = benigno)
- Se calculan diez características de valor real para cada núcleo

1. Radio: (Media de las distancias desde el centro a los puntos del perímetro).
2. Textura: (Desviación estándar de los valores de escala de grises).
3. Perímetro.
4. Área.
5. Suavidad: (Variación local en longitudes de radio).
6. Compacidad: (Perímetro^2 / área - 1,0).
7. Concavidad: (Severidad de las porciones cóncavas del contorno).
8. Puntos Cóncavos: (Número de porciones cóncavas del contorno).
9. Simetría.
10. Dimensión Fractal: ("Aproximación de la línea costera" - 1).

IMPORTANCIA DEL CONJUNTO DE DATOS

Este conjunto de datos es crucial para el desarrollo de modelos de aprendizaje automático que pueden ayudar a los médicos a realizar diagnósticos más precisos y rápidos.

CONCLUSIÓN

El análisis y la utilización del conjunto de datos del Cáncer de Mama Wisconsin (Diagnóstico) permite un enfoque más detallado y preciso en la predicción del diagnóstico del cáncer de mama.
La integración de este conjunto de datos con otros, como el SEER, proporciona una herramienta poderosa para mejorar la precisión diagnóstica y potencialmente salvar vidas mediante detecciones tempranas y tratamientos adecuados.

DataSet

DESCRIPCIÓN GENERAL DE LOS DATOS

- Total de muestras: 569
- Diagnósticos: 357 benignos, 212 malignos
- Atributos sin valores faltantes
- Valores recodificados con cuatro dígitos significativos

EXPLORACIÓN DE DATOS

* Media de Radio vs. Textura: Se puede usar un gráfico de dispersión para visualizar la distribución de los diagnósticos.

* Diagramas de dispersión para cada par de características en el conjunto de datos, donde cada punto en el diagrama de dispersión representa una instancia de datos

ANÁLISIS DE DATOS

- Las características de concavidad y puntos cóncavos muestran una relación significativa con la malignidad de las masas mamarias.

- Se utilizan visualizaciones impactantes como gráficos de dispersión y mapas de calor para presentar hallazgos.

Explorando el Diagnóstico del Cáncer de Mama con Datos.

- **Inicio: En el ámbito de la medicina, las decisiones críticas para la salud de las personas pueden revolucionarse con la ayuda de datos precisos y análisis avanzados.**

- **Nudo: Con una variedad de características que describen núcleos celulares en imágenes de aspiraciones, cada atributo podría ser clave para predecir si una masa es benigna o maligna.**

- **Desarrollo: Explorando características relevantes, identificamos relaciones significativas entre atributos como la textura y la concavidad con la malignidad.**

- **Desenlace: Imaginemos un futuro donde, gracias a estos datos, los diagnósticos sean más precisos y tempranos, salvando vidas y ofreciendo mayor tranquilidad a los pacientes.**

Preguntas de Interes

- ¿Existe una relación entre el tamaño medio de la masa mamaria y su diagnóstico de benigno o maligno?
- ¿Las características relacionadas con la textura de la masa mamaria pueden predecir su malignidad?
- ¿Hay alguna correlación entre la suavidad de la masa mamaria y su diagnóstico de cáncer?
- ¿Las masas mamarias con mayor compacidad tienden a ser más malignas?
- ¿La concavidad de la masa mamaria está relacionada con la severidad de su diagnóstico?
- ¿Existen diferencias significativas en la distribución de diagnósticos según las diferentes características?

Análisis

UNIVARIADO

Se centra en examinar y describir una sola variable a la vez.

Esto nos permitirá comprender mejor la distribución de cada variable y cómo se relaciona con el diagnóstico de cáncer benigno o maligno.

Shape: Muestra la forma del DataFrame (número de filas y columnas).

PrettyTable: Define una tabla con las columnas 'Column', 'Type', 'Non-Null', 'Nulls', 'Unique' y 'Example'.

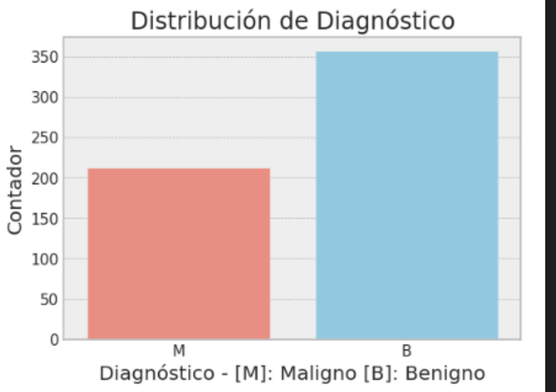
Bucle for: Recorre cada columna del DataFrame y agrega filas a la tabla:

- Column: Nombre de la columna.
- Type: Tipo de datos de la columna.
- Non-Null: Número de valores no nulos en la columna.
- Nulls: Número de valores nulos en la columna.
- Unique: Número de valores únicos en la columna.
- Example: Ejemplo del primer valor no nulo encontrado en la columna.

ESTADÍSTICAS DESCRIPTIVAS

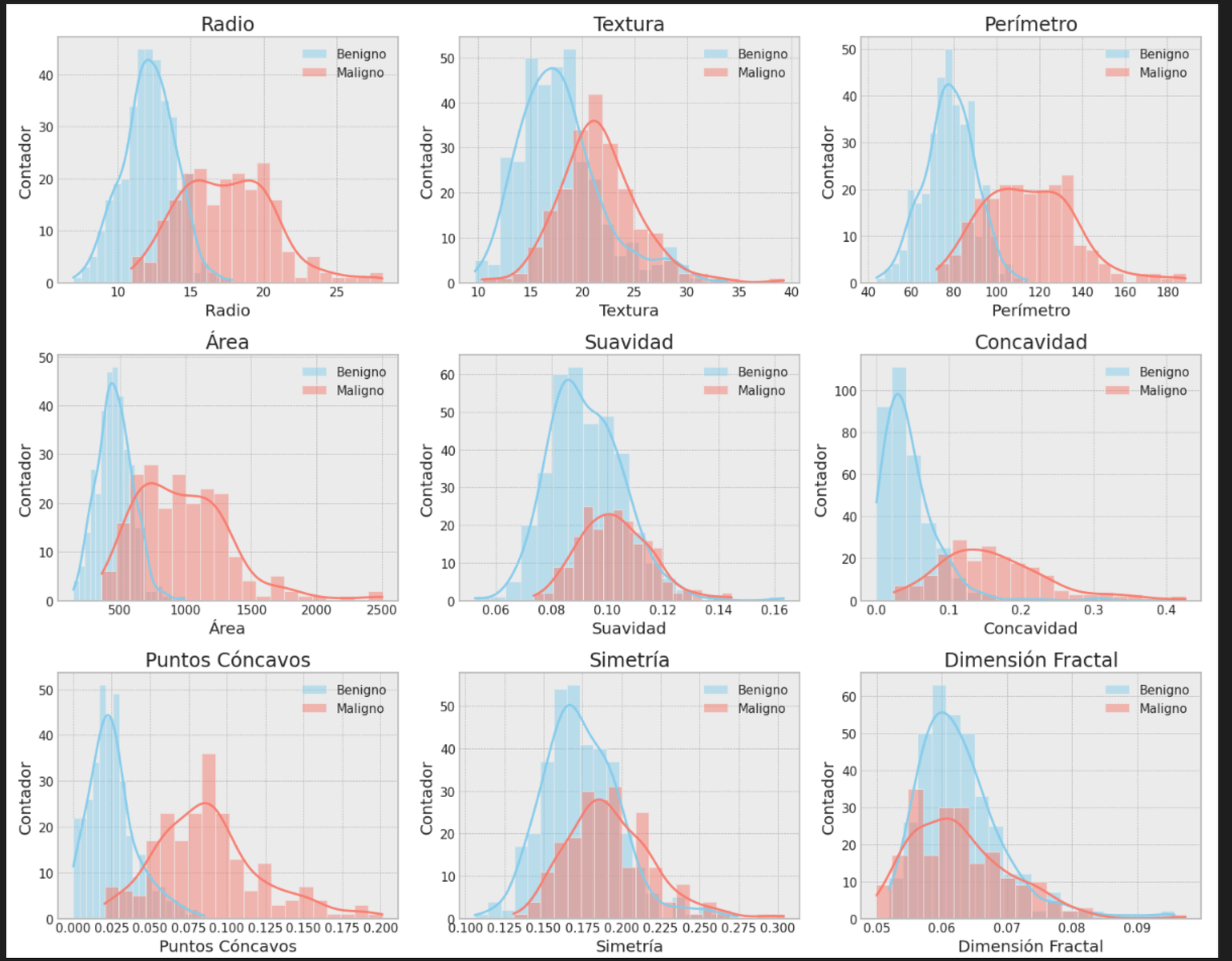
	count	mean	std	min	25%	50%	75%	max
id	569.0	30371831.43	1.250206e+08	8670.00	869218.00	906024.00	8813129.00	9.113205e+08
radius_mean	569.0	14.13	3.520000e+00	6.98	11.70	13.37	15.78	2.811000e+01
texture_mean	569.0	19.29	4.300000e+00	9.71	16.17	18.84	21.80	3.928000e+01
perimeter_mean	569.0	91.97	2.430000e+01	43.79	75.17	86.24	104.10	1.885000e+02
area_mean	569.0	654.89	3.519100e+02	143.50	420.30	551.10	782.70	2.501000e+03
smoothness_mean	569.0	0.10	1.000000e-02	0.05	0.09	0.10	0.11	1.600000e-01
compactness_mean	569.0	0.10	5.000000e-02	0.02	0.06	0.09	0.13	3.500000e-01
concavity_mean	569.0	0.09	8.000000e-02	0.00	0.03	0.06	0.13	4.300000e-01
concave points_mean	569.0	0.05	4.000000e-02	0.00	0.02	0.03	0.07	2.000000e-01
symmetry_mean	569.0	0.18	3.000000e-02	0.11	0.16	0.18	0.20	3.000000e-01
fractal_dimension_mean	569.0	0.06	1.000000e-02	0.05	0.06	0.06	0.07	1.000000e-01
radius_se	569.0	0.41	2.800000e-01	0.11	0.23	0.32	0.48	2.870000e+00
texture_se	569.0	1.22	5.500000e-01	0.36	0.83	1.11	1.47	4.880000e+00
perimeter_se	569.0	2.87	2.020000e+00	0.76	1.61	2.29	3.36	2.198000e+01
area_se	569.0	40.34	4.549000e+01	6.80	17.85	24.53	45.19	5.422000e+02
smoothness_se	569.0	0.01	0.000000e+00	0.00	0.01	0.01	0.01	3.000000e-02
compactness_se	569.0	0.03	2.000000e-02	0.00	0.01	0.02	0.03	1.400000e-01
concavity_se	569.0	0.03	3.000000e-02	0.00	0.02	0.03	0.04	4.000000e-01
concave points_se	569.0	0.01	1.000000e-02	0.00	0.01	0.01	0.01	5.000000e-02
symmetry_se	569.0	0.02	1.000000e-02	0.01	0.02	0.02	0.02	8.000000e-02
fractal_dimension_se	569.0	0.00	0.000000e+00	0.00	0.00	0.00	0.00	3.000000e-02
radius_worst	569.0	16.27	4.830000e+00	7.93	13.01	14.97	18.79	3.604000e+01
texture_worst	569.0	25.68	6.150000e+00	12.02	21.08	25.41	29.72	4.954000e+01
perimeter_worst	569.0	107.26	3.360000e+01	50.41	84.11	97.66	125.40	2.512000e+02
area_worst	569.0	880.58	5.693600e+02	185.20	515.30	686.50	1084.00	4.254000e+03
smoothness_worst	569.0	0.13	2.000000e-02	0.07	0.12	0.13	0.15	2.200000e-01
compactness_worst	569.0	0.25	1.600000e-01	0.03	0.15	0.21	0.34	1.060000e+00
concavity_worst	569.0	0.27	2.100000e-01	0.00	0.11	0.23	0.38	1.250000e+00
concave points_worst	569.0	0.11	7.000000e-02	0.00	0.06	0.10	0.16	2.900000e-01
symmetry_worst	569.0	0.29	6.000000e-02	0.16	0.25	0.28	0.32	6.600000e-01
fractal_dimension_worst	569.0	0.08	2.000000e-02	0.06	0.07	0.08	0.09	2.100000e-01

Univariado



Comentarios

- **Radio: Se observa que el diagnóstico maligno muestra valores más altos en comparación con benigno.**
- **Textura: El diagnóstico benigno tiende a tener valores más concentrados en comparación con maligno.**
- **Perimetro: Existe una marcada diferencia en la distribución entre benigno y maligno.**
- **Área: El área media en diagnóstico maligno muestra una dispersión mayor que en benigno.**
- **Suavidad: La suavidad media muestra diferencias notables entre benigno y maligno.**
- **Concavidad:La concavidad media tiende a ser más pronunciada en el diagnóstico maligno.**
- **Puntos Concavos: Los puntos cóncavos medios son más prominentes en maligno.**
- **Simetría: La simetría media muestra una distribución similar entre benigno y maligno.**
- **Dimencion Fractal: La dimensión fractal media tiene variaciones notables en ambos diagnósticos.**

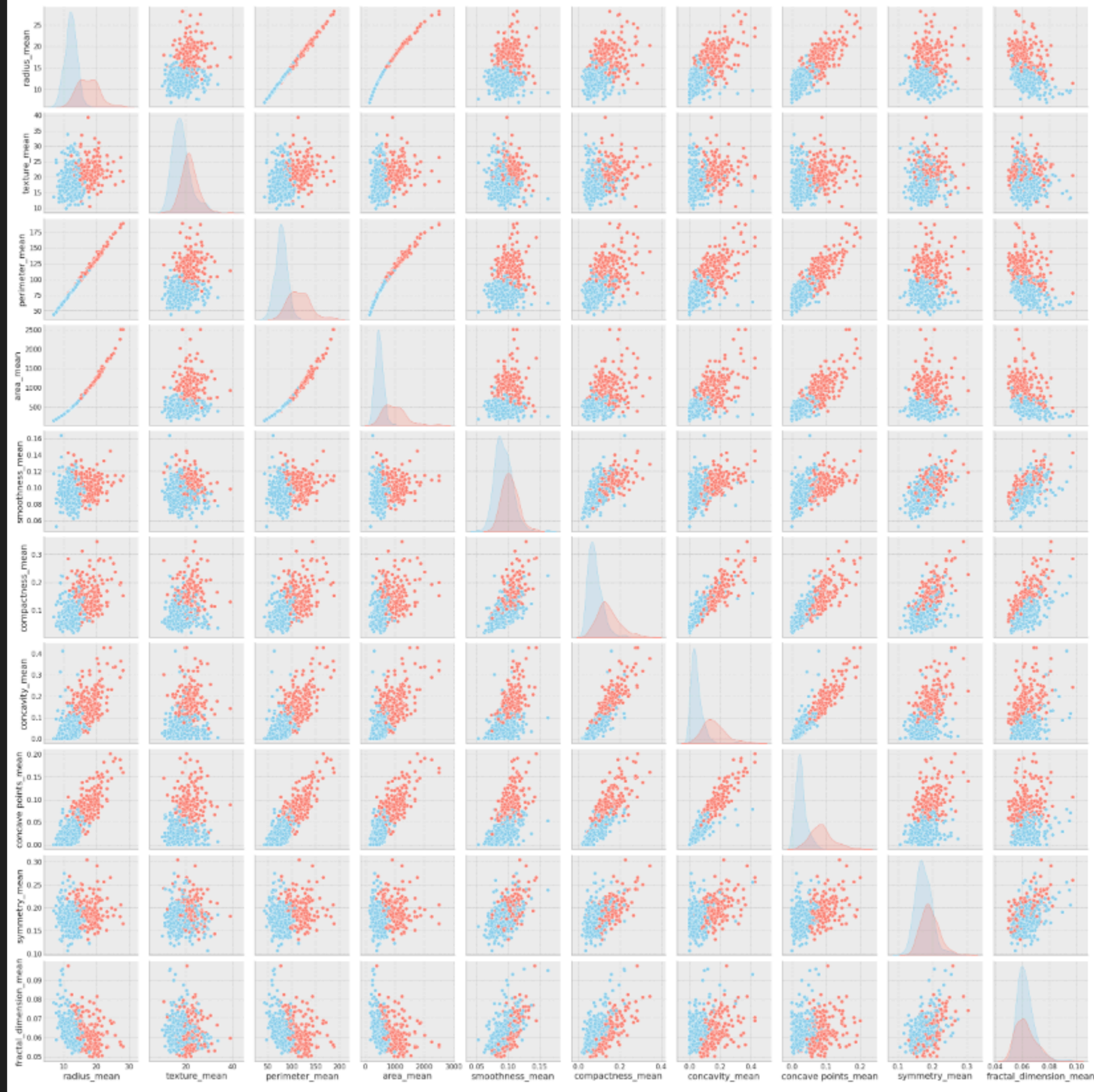


Bivariado

Se enfoca en explorar la relación entre dos variables en un conjunto de datos. En el contexto del conjunto de datos del Cáncer de Mama, podemos realizar un análisis bivariado para entender cómo diferentes características de las masas mamarias se relacionan entre sí y cómo estas relaciones pueden variar según el diagnóstico de cáncer (Benigno o Maligno).

Explicación: Cada fila y columna en la matriz representa una característica en el conjunto de datos. Los diagramas de dispersión en la diagonal principal muestran la distribución de cada característica, mientras que los diagramas de dispersión fuera de la diagonal principal muestran la relación entre pares de características. Los puntos se colorean según el diagnóstico, lo que facilita la identificación de patrones visuales relacionados con el diagnóstico.

Interpretación: Al observar, puedes identificar visualmente cualquier patrón o relación entre las características que pueda estar relacionado con el diagnóstico. Por ejemplo, si ves que los puntos de un determinado par de características están claramente separados por color (Benigno, Maligno), eso sugiere que esas dos características podrían ser útiles para distinguir entre los dos tipos de diagnóstico.



Machine Learning

Algoritmo Utilizado: K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) es un algoritmo de clasificación supervisado. Se utiliza para predecir la categoría a la que pertenece un nuevo dato, basándose en la similitud con datos ya conocidos.

- **Codificación de Variables:**
- **Se transforman las etiquetas categóricas en valores numéricos para facilitar el procesamiento.**
- **Preparación de los Datos:**
- **Separación de Características y Variable Objetivo:**
  - **Las características (X) se separan de la variable objetivo (y).**
- **División en Conjuntos de Entrenamiento y Prueba:**
  - **Los datos se dividen en conjuntos de entrenamiento y prueba.**
- **Normalización de Datos:**
  - **Se escalan los datos para que todas las características tengan una media de 0 y una desviación estándar de 1.**
- **Entrenamiento del Modelo:**
- **Se entrena un modelo KNN con el conjunto de datos escalados.**
- **Evaluación del Modelo:**
- **Se realizan predicciones y se evalúa el modelo utilizando métricas de exactitud, matriz de confusión e informe de clasificación.**
- **Ajuste de Hiperparámetros:**
- **Se prueba con diferentes números de vecinos para encontrar el número óptimo que maximiza la exactitud del modelo.**
- **Se grafica la exactitud del modelo con diferentes números de vecinos.**
- **Reentrenamiento y Evaluación Final:**
- **Se reentrena el modelo con el mejor número de vecinos y se realiza una evaluación final.**
- **Visualización de Características Importantes:**
- **Se crea un gráfico de dispersión para analizar la relación entre algunas características clave.**

Machine Learning

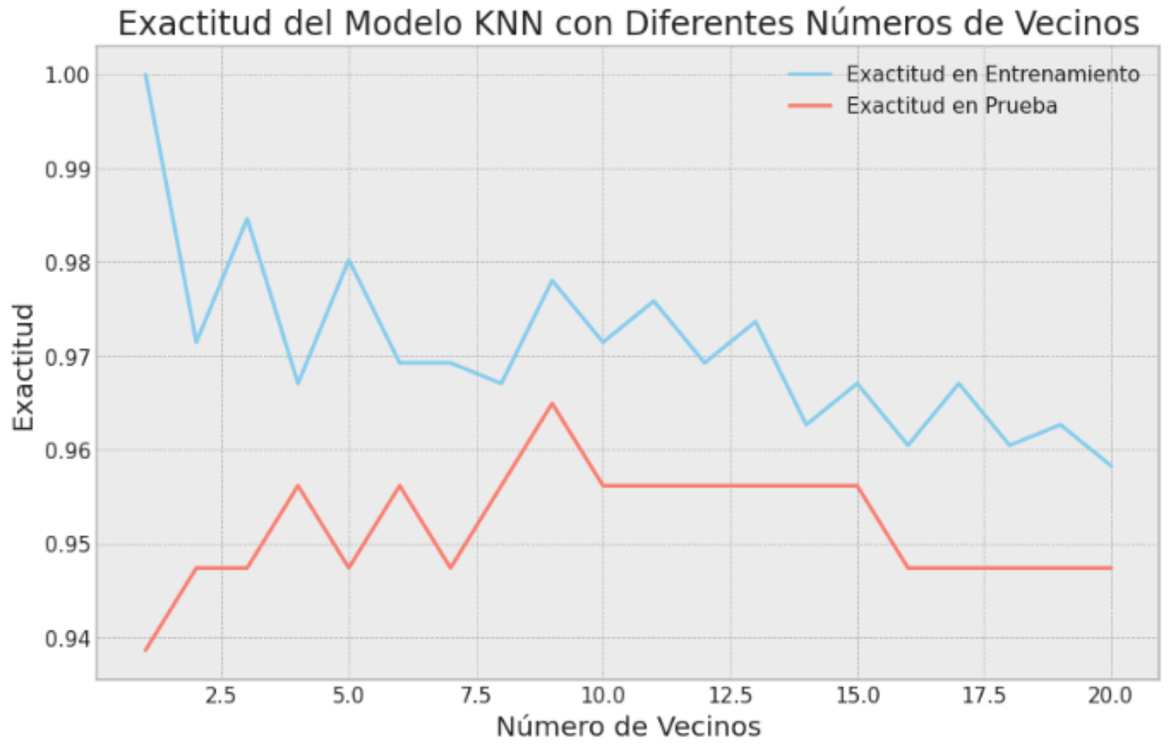
Conclusión

Este código realiza un análisis completo utilizando el clasificador KNN para predecir diagnósticos de cáncer de mama. Desde la carga de datos hasta la evaluación del modelo, cada paso está diseñado para optimizar la comprensión y el rendimiento del modelo en el contexto de un proyecto de ciencia de datos.

```
Exactitud: 0.9473684210526315
Matriz de Confusión:
[[68  3]
 [ 3 40]]
Informe de Clasificación:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        71
           1       0.93      0.93      0.93        43

    accuracy                           0.95       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```

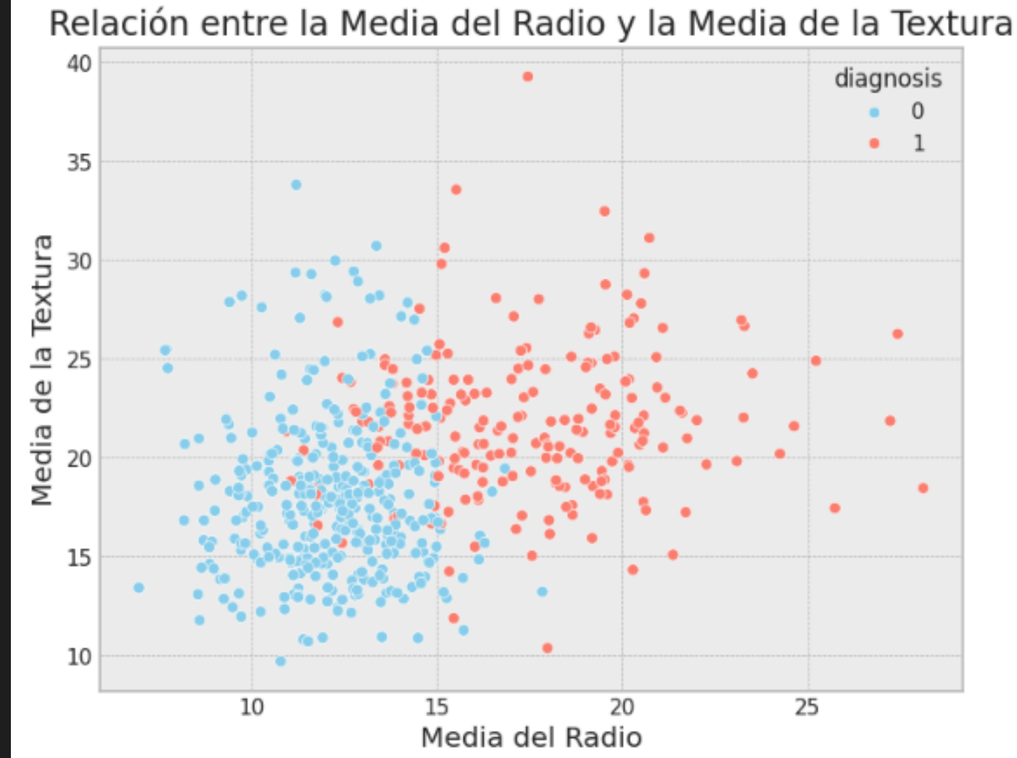



```
Mejor número de vecinos: 9
[[69  2]
 [ 2 41]]
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        71
           1       0.95      0.95      0.95        43

    accuracy                           0.96       114
   macro avg       0.96      0.96      0.96       114
weighted avg       0.96      0.96      0.96       114

Exactitud: 0.9649122807017544
```



Conclusión

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse in mi sed velit lacinia vulputate. Vestibulum dignissim mollis ipsum sed pellentesque.

01

El análisis detallado de los datos relacionados con el cáncer de mama ha proporcionado información valiosa para mejorar el diagnóstico y el tratamiento de esta enfermedad. A través de técnicas avanzadas de data science, se lograron identificar patrones y características clave que distinguen entre tumores benignos y malignos con una precisión significativa.

02

Los modelos predictivos desarrollados demostraron una alta eficiencia en la clasificación de los tipos de cáncer de mama, lo que podría facilitar a los profesionales de la salud en la toma de decisiones informadas y personalizadas para cada paciente. Además, el uso de técnicas de machine learning permitió una mejor comprensión de los factores de riesgo y de los indicadores tempranos de la enfermedad.

03

La integración de estos hallazgos en la práctica clínica tiene el potencial de reducir las tasas de mortalidad y mejorar la calidad de vida de los pacientes. Sin embargo, es crucial continuar con investigaciones adicionales para validar y perfeccionar estos modelos en diversas poblaciones y entornos clínicos.

MUCHAS GRACIAS

Proyecto Final Coder House