Wisconsin Diagnostic Breast Cancer (WDBC)

Docente: Cesar Jesus Lara Avila

INTEGRANTES

- AYDEE ZENAIDA QUISPE RIMACHI
- BRYAN ALEXANDER PEÑALOZA HUAMAN
- RODRIGO ALONSO RODRIGUEZ REATEGUI
- LUIS SANTIAGO ARENAS TORRES
- VICTOR NIKOLAI HUARCAYA PUMACAYO

2023

ÍNDICE

INTRODUCCIÓN

El conjunto de datos que se escogió de las distintas que se mostró fue el Diagnóstico del cáncer de mama en Wisconsin, este dataset contiene información clínica recopilada de muestras de células de cáncer de mama de pacientes en Wisconsin. El objetivo de este conjunto de datos es predecir si una muestra de células es benigna (no cancerosa) o maligna (cancerosa) basándose en diferentes características.

La capacidad de clasificar correctamente los tumores mamarios como benignos o malignos tiene un gran impacto en la atención médica y en la vida de los pacientes. El conjunto de datos del cáncer de mama permite a los investigadores y profesionales del aprendizaje automático abordar este problema y desarrollar modelos que ayuden en la detección temprana, el tratamiento del cáncer de mama,etc.

El conjunto de datos Wisconsin Diagnostic Breast Cancer (WDBC) contiene un total de “X” muestras, cada una con “Y” características. Las características incluyen información sobre el tamaño y la forma de las células, su uniformidad, la cohesión del tejido, entre otros aspectos relevantes para el diagnóstico del cáncer de mama.

El conjunto de datos se divide en conjuntos de entrenamiento y prueba, utilizando una proporción del XX% para el conjunto de prueba. Se utilizan diferentes algoritmos de clasificación para realizar la clasificación de las muestras de células.

Además, se utiliza la importancia de características obtenidas mediante un algoritmo para identificar las características más relevantes en el diagnóstico del cáncer de mama. Estas características proporcionan información valiosa para entender y detectar patrones que puedan ser indicativos de la presencia de cáncer de mama.

OBJETIVO DEL ANÁLISIS

El objetivo de este experimento es evaluar el rendimiento de los diferentes algoritmos de clasificación y determinar cuáles características son más importantes en el diagnóstico del cáncer de mama. Esto permitirá desarrollar modelos predictivos más precisos y brindar información relevante a los profesionales de la salud para la detección temprana y el tratamiento del cáncer de mama.

DISEÑO DEL EXPERIMENTO

-**Descripción del conjunto de datos**
El conjunto de datos es una recopilación bastante amplia de datos de muestras de células de cáncer de mama de pacientes de Wisconsin. Este conjunto en particular consta de un total de 569 datos/ muestras, las cuales se clasifican en benignos(no cancerosos) o malignos (cancerosos). También este conjunto de datos está representado por 30

características que describen diferentes aspectos de las células extraídas de la muestra. En la representación de 30 muestras incluyen tamaño, forma y uniformidades de las células.

-Número y tipo de características (Binarias,discretas,continuas,etc).term

Las caracteristicas estan destribuidas en 2 categorías principales

- Continuas: Estas características son de naturaleza numérica y pueden tomar cualquier valor dentro de un rango específico. En este caso, las características continuas son:

Unset

```
radius (radio)
texture (textura)
perimeter (perímetro)
area (área)
smoothness (suavidad)
compactness (compacidad)
concavity (concavidad)
concave points (puntos cóncavos)
symmetry (simetría)
fractal dimension (dimensión fractal)
```

- Binarias: Son características que pueden tener solo dos valores distintos, generalmente codificados como 0 y 1. En este caso, la característica binaria es:

Unset

```
  class (clase) con dos categorías: "WDBC-Malignant" y
"WDBC-Benign"
  *   Target (1 = malignant, 0 = benign)
```

    - ID numero: variable categorica ordinal

-Número de muestras en los conjuntos de entrenamientos

Como se mencionó anteriormente, en el dataset tiene como un total de 569 de las cuales 398 son muestras de entrenamiento y los 171 son muestras de prueba.

.Metodología:

-Datos faltantes:

En este dataset, no hay un dato faltante y lo podemos comprobar con el código df.isnull().sum().sum() que significa la suma de todos los valores faltantes del dataset, pero al ejecutar el código nos aparece el número 0. Quiere decir que hay 0 datos faltantes.

-Selección y extracción de las características

En el dataset no es necesario la selección y extracción de las características, ya que sus características ya están predefinidas,pero al implementarlo podría mejorar el rendimiento del modelo aprendizaje automático. En el cuaderno, hemos hecho un código con el uso de la técnica PCA (Análisis de componentes Principales).

-Medida de calidad

La medida de calidad que utilizaremos es la precisión (accuracy). ¿Por qué? Es una medida común en Machine Learning que nos ayuda a ver el rendimiento de los modelos que utilizaremos y poder compararlos. Además. su medida es fácil de interpretar, ya que la interpretación es por términos de porcentaje.

-Algoritmos y estrategia para su ajuste

En este proyecto utilizaremos los siguientes algoritmos: embolsado, bosque aleatorio y árboles extra.Luego, Utilizaremos un comando Ajustar (.Fit()) para ajustar los modelos a los datos de entrenamiento.

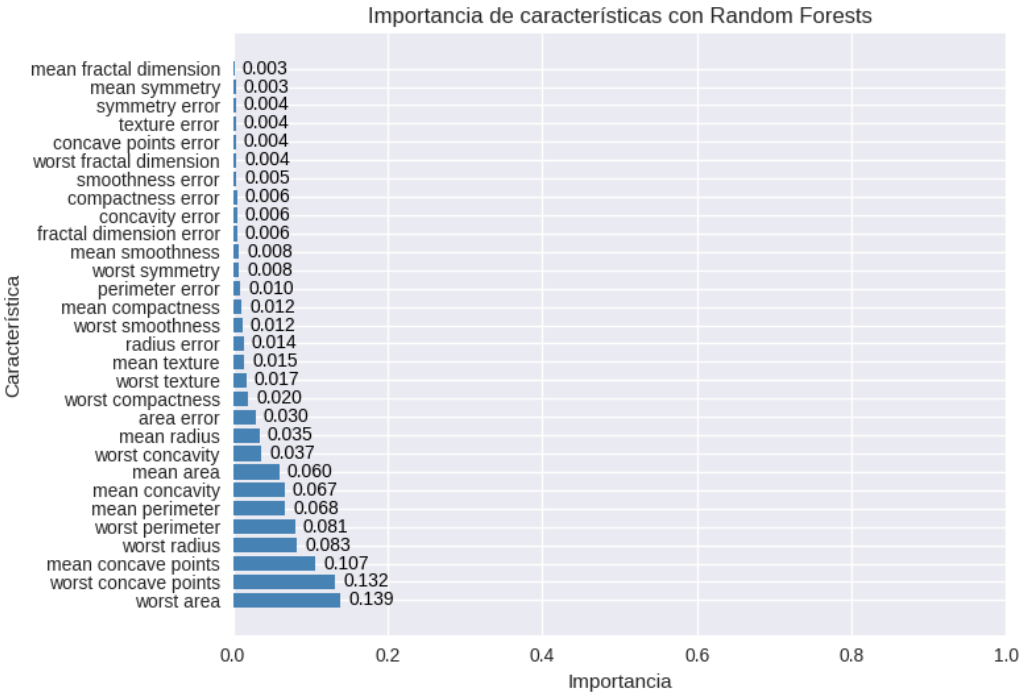
-**Estrategia de Validación a emplear para el ajuste de hiper parámetros si fuese necesario**

La estrategia de Validación que podríamos emplear sería la estrategia de Validación cruzada. La validación cruzada se usa para evaluar y comparar el rendimiento de los modelos para luego elegir el más adecuado.

Experimentación y resultados

-Importación de características con random forest

Esta tabla es el salida(resultado) del código principal de los datos, esta tabla es la representación de los códigos que llaman el dataset que tienen como finalidad el entrenamiento de un modelo bosque aleatorio, calcula las importancias de las características y obviamente la gráfica de barras es el resultado de ello que nos ayuda a visualizar las importancias, la tabla muestra las características ordenadas por importancia, proporcionando una representación visual de la contribución de cada característica en el Modelado bosque aleatorio



mean fractal dimension = dimensión fractal media
mean symmetry = simetría media
symmetry error = error de simetría
textura error = error de textura
concave points error = error de puntos cóncavos
worst fractal dimension = peor dimensión fractal
smoothness error = error de suavidad
compactness error = error de compacidad
concavity error = error de concavidad
fractal dimension error = error de dimensión fractal
mean smoothness = suavidad media
worst symmetry = peor simetría
radius error = error de radio
mean texture = textura media
worst texture = peor textura
worst compactness = peor compacidad
area error = error de área
mean radius = radio medio
worst concavity = peor concavidad
mean area = área media
mean concavity = concavidad media
mean perimeter = perímetro medio
worst perimeter = peor perímetro
worst radius = peor radio
mean concave points = puntos cóncavos medios
worst concave points = peores puntos cóncavos
worst area = peor area

Importancia de características (Random Forests):

Índice	Nombre de Característica	Importancia
1	mean concave points	0.141934
2	worst concave points	0.127136
3	worst area	0.118217
4	mean concavity	0.080557
5	worst radius	0.0779747
6	worst perimeter	0.0742921
7	mean perimeter	0.0600923
8	mean area	0.0538105
9	worst concavity	0.0410796
10	mean radius	0.0323119
11	area error	0.0295384
12	worst texture	0.0187857
13	worst compactness	0.0175391
14	radius error	0.016435
15	worst symmetry	0.0129294
16	perimeter error	0.0117698
17	worst smoothness	0.0117692
18	mean texture	0.0110639
19	mean compactness	0.00921566
20	fractal dimension error	0.00713457
21	worst fractal dimension	0.00692376
22	mean smoothness	0.00622336
23	smoothness error	0.00588079
24	concavity error	0.0058159
25	compactness error	0.00459638
26	symmetry error	0.00400077
27	concave points error	0.00338232
28	mean symmetry	0.00327807
29	texture error	0.00317191
30	mean fractal dimension	0.00314028

- Evaluación del rendimiento de los modelos ensayados y Comparación de línea Base

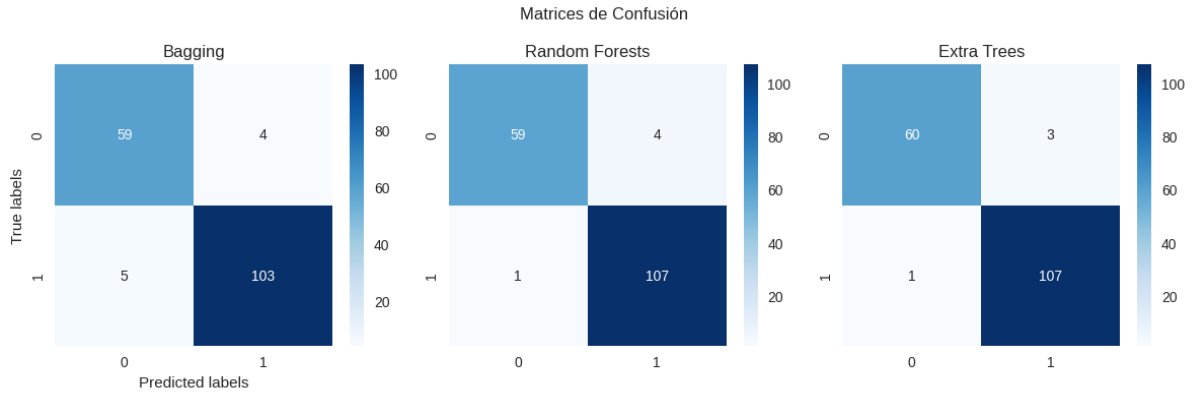
Estadísticas descriptivas de importancia de características:

Estadística	Valor
Mínimo	0.0031
Máximo	0.1419
Media	0.0333
Mediana	0.0123
Desviación estándar	0.0396

El código que se usó para el resumen de clasificación fue obtenido por 3 clasificadores Bagging, Random Forests y Extra Trees, la tabla nos da la precisión de cada clasificador y la precisión obtenida utilizando una línea base(conjunto de evidencias), para ello la La matriz `resumen de clasificacion` es una lista de listas que contiene los nombres de los clasificadores ("Bagging", "Random Forests" y "Extra Trees"), la precisión se refiere a la proporción de muestras correctamente clasificadas sobre el total de muestras.

Resumen de clasificación:

Clasificador	Precisión	Precisión con Línea Base
Bagging	0.9474	0.6316
Random Forests	0.9708	0.6316
Extra Trees	0.9766	0.6316



Discusión

- Interpretación de los resultados obtenidos.

Por un lado Podemos ver que el algoritmo bagging presenta una precision de 0.9474 que quiere decir que tiene una precisión de 94%en la clasificación de .Mientras que los algoritmos bosques aleatorios y arboles extra presentan porcentajes muy cercano a una diferencia de 0.0058. Por otro lado, podemos ver la tabla de estadistica descriptiva de importancia. Donde nos dicen el mínimo valor (la característica menos importante), máximo valor (La característica más importante),etc.

- ¿Cómo podría ser mejorado su sistema?

El sistema Wisconsin Diagnostic Breast Cancer (WDBC) se puede mejorar mediante la recopilación de datos más variados, la exploración de diferentes algoritmos de aprendizaje automático y técnicas de ensamble para mejorar la precisión, la implementación de validación cruzada para una evaluación más precisa, la interpretación de características importantes, la integración de tecnologías emergentes como el aprendizaje profundo, la realización de estudios clínicos para validar y mejorar el sistema, el desarrollo de una interfaz de usuario intuitiva y la colaboración con profesionales de la salud. Estas mejoras multidisciplinarias impulsarían la detección y diagnóstico del cáncer de mama y proporcionarán resultados más confiables y eficaces en el sistema WDBC.

Conclusiones:

En esta investigación, hemos ordenado y mostrado el dataset de nuestra elección que era Wisconsin Diagnostic breast cancer(WDBC), es un conjunto de datos que contiene características digitalizadas.Esta data, contiene la información detallada sobre las características de las células mamarias, como el tamaño, la forma y la textura, estos datos se usan para ver si una muestra es benigna(no cancerosa) o maligna(cancerosa), gracias a ello podemos llegar a las conclusiones que se pueden clasificar el cáncer de mama con una precisión de más del 90% en los algoritmos utilizados (Bosque aleatorio,Extra árboles y embolsado), gracias a ello se clasifica con alta probabilidad las muestras de tejido mamario en benignas o malignas.

Trabajo futuros:

Para trabajos futuros podemos implementar otros algoritmos de clasificación para poder ayudar a buscar soluciones.Otro punto es mejorar los algoritmos existentes para aumentar la precisión de la detección del cáncer de pecho u otras enfermedades.Esto podría involucrar la optimización de hiperparametros. la exploración de técnicas de selección de características,etc. En resumen, nosotros nos enfocaremos más en el perfeccionamiento de los modelos y técnicas que presenta machine learning .

Validación cruzada
https://datascientest.com/es/cross-validation-definicion-e-importancia#:~:text=A%20menudo%20en%20Machine%20Learning,sesgos%20que%20los%20dem%C3%A1s%20m%C3%A9todos.

Datos faltantes:
https://chartio.com/resources/tutorials/how-to-check-if-any-value-is-nan-in-a-pandas-dataframe/

Bagging
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

Random Forest
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Extra tree
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

PCA
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Tabulate
https://www.askpython.com/python-modules/tabulate-tables-in-python

SNS
https://seaborn.pydata.org/generated/seaborn.histplot.html

Accuracy
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

confusion matrix
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

train test split
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
df isnull sum() sum()
https://chartio.com/resources/tutorials/how-to-check-if-any-value-is-nan-in-a-pandas-dataframe/