UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

DEPARTAMENTO DE COMPUTACIÓN Y SIMULACIÓN DE SISTEMAS

LICENCIATURA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Tópicos Especiales 1

Proyecto Final

Detección De Cáncer De Mama Utilizando Machine Learning

Profesor

Thomas J. Concepción M.

Elaborado por:

Gustavo Colucci 8-951-2191

Luis Mejia 8-949-350

Greg Torres 8-956-675

Grupo 1IL143

II semestre 2023

Introducción:

El cáncer de mama es una de las enfermedades más comunes y mortales que afectan a las mujeres en todo el mundo. La detección temprana y precisa es crucial para mejorar las tasas de supervivencia y la calidad de vida de los pacientes. En las últimas décadas, con los avances en la tecnología de imágenes médicas y la bioinformática, se ha acumulado una gran cantidad de datos clínicos y biomédicos, proporcionando una oportunidad única para aplicar técnicas de aprendizaje automático en la detección y diagnóstico del cáncer de mama.

El "Breast Cancer Dataset", un conjunto de datos ampliamente reconocido en la comunidad científica juega un papel vital en este escenario. Compuesto por 569 muestras con 30 características numéricas cada una, este conjunto de datos refleja las variaciones físicas de los núcleos celulares en muestras de tejido mamario, clasificadas como malignas o benignas. Las características incluyen detalles como el tamaño, la forma y la textura de los núcleos celulares, que son indicadores críticos en la evaluación clínica del cáncer de mama.

La relevancia de este conjunto de datos en el ámbito médico es indiscutible. Proporciona una base robusta para el desarrollo y la validación de modelos predictivos que pueden asistir a los profesionales médicos en el diagnóstico precoz y preciso del cáncer de mama. Utilizando algoritmos de aprendizaje automático como la Regresión Logística, Máquinas de Soporte Vectorial (SVM) y Random Forest, los investigadores y médicos pueden no solo mejorar la precisión del diagnóstico sino también obtener insights sobre las características más significativas de las muestras, lo cual es vital para la comprensión profunda de la enfermedad.

Este enfoque computacional hacia el diagnóstico del cáncer de mama no solo complementa los métodos tradicionales, sino que también abre nuevas vías para personalizar tratamientos y estrategias de intervención, marcando un paso adelante en la lucha contra esta enfermedad prevalente. En este contexto, el análisis del "Breast Cancer Dataset" utilizando diversos algoritmos de aprendizaje automático representa un campo de estudio prometedor y de gran impacto en la salud pública y la investigación médica.

Dataset a utilizar (tamaño, características de los datos, relevancia de las características):

- Breast Cancer Dataset: Dataset de tipo clasificatorio.

El "Breast Cancer Dataset" es un conjunto de datos ampliamente utilizado en el campo del aprendizaje automático, especialmente en tareas de clasificación relacionadas con la salud. Aquí te detallo las características principales de este conjunto de datos:

- **Tamaño del Dataset**:

El conjunto de datos contiene 569 instancias o muestras.

- **Características de los Datos**:

El conjunto de datos incluye 30 características numéricas. Estas características se derivan de una imagen digitalizada de una biopsia de tejido mamario y describen las características de los núcleos celulares presentes en la imagen.

Las características incluyen, entre otras, el radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal de los núcleos celulares.

- **Relevancia de las Características**:

Las características son medidas importantes que reflejan las variaciones físicas en los núcleos celulares de los tejidos mamarios y son críticas para el diagnóstico del cáncer de mama.

Estas características se utilizan para clasificar las muestras en dos categorías: malignas (cancerosas) y benignas (no cancerosas).

En un contexto médico, estas características ayudan a determinar la necesidad de intervenciones médicas adicionales, como biopsias o tratamientos.

Este conjunto de datos es particularmente importante en el ámbito de la detección y diagnóstico del cáncer de mama, ya que ofrece una base rica para desarrollar y evaluar modelos de aprendizaje automático que puedan ayudar en la toma de decisiones médicas. La capacidad de los modelos para interpretar y clasificar con precisión estos datos puede tener un impacto significativo en el diagnóstico temprano y el tratamiento del cáncer de mama.

Algoritmos por utilizar:

- Regresión Logística:

Es un algoritmo simple pero efectivo para problemas de clasificación binaria.

Fácil de interpretar y entender.

Puede proporcionar buenas métricas de evaluación en conjuntos de datos como el de cáncer de mama.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score
from sklearn.datasets import load_breast_cancer

# Cargar el conjunto de datos
data = load_breast_cancer()
X = pd.DataFrame(data.data, columns=data.feature_names)
y = pd.Series(data.target)
```

```
# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

# Inicializar y entrenar el modelo de Regresión Logística
model = LogisticRegression(max_iter=10000)
model.fit(X_train, y_train)

# Realizar predicciones en el conjunto de prueba
y_pred = model.predict(X_test)

# Calcular métricas de evaluación
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

accuracy, precision, recall, f1
```

La implementación del algoritmo de Regresión Logística en el conjunto de datos de cáncer de mama ha proporcionado los siguientes resultados en las métricas de evaluación:

Exactitud, Accuracy: 0.9766081871345029

Precisión, Precision: 0.9814814814814815

Recuperación, Recall: 0.9814814814814815

Valor F1, F1 Score: 0.9814814814814815

- Máquinas de Soporte Vectorial (SVM):

SVM es robusto y puede funcionar bien en conjuntos de datos de tamaño moderado.

Puede manejar eficientemente características de alta dimensión, lo cual puede ser relevante en el caso de datos médicos.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score
from sklearn.datasets import load_breast_cancer
```

```
# Cargar el conjunto de datos Breast Cancer
breast_cancer_data = load_breast_cancer()
X_cancer = pd.DataFrame(breast_cancer_data.data,
columns=breast_cancer_data.feature_names)
y_cancer = pd.Series(breast_cancer_data.target)

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train_cancer, X_test_cancer, y_train_cancer, y_test_cancer =
train_test_split(X_cancer, y_cancer, test_size=0.3, random_state=42)

# Inicializar y entrenar el modelo SVM en el conjunto de datos Breast Cancer
svm_cancer_model = SVC(kernel='linear')
svm_cancer_model.fit(X_train_cancer, y_train_cancer)

# Realizar predicciones en el conjunto de prueba
y_pred_cancer_svm = svm_cancer_model.predict(X_test_cancer)

# Calcular métricas de evaluación para el modelo SVM en el conjunto de datos
Breast Cancer
accuracy_cancer_svm = accuracy_score(y_test_cancer, y_pred_cancer_svm)
precision_cancer_svm = precision_score(y_test_cancer, y_pred_cancer_svm)
recall_cancer_svm = recall_score(y_test_cancer, y_pred_cancer_svm)
f1_cancer_svm = f1_score(y_test_cancer, y_pred_cancer_svm)

accuracy_cancer_svm, precision_cancer_svm, recall_cancer_svm, f1_cancer_svm
print("Exactitud, Accuracy:", accuracy_cancer_svm)
print("Precisión, Precision:", precision_cancer_svm)
print("Recuperación, Recall:", recall_cancer_svm)
print("Valor F1, F1 Score:", f1_cancer_svm)
```

Aquí tienes la implementación del modelo de Máquinas de Soporte Vectorial (SVM) aplicado al conjunto de datos Breast Cancer Dataset, empezando desde cero en un nuevo entorno de Python:

Exactitud (Accuracy): 96.49%

Precisión (Precision): 96.36%

Recuperación (Recall): 98.15%

Valor F1 (F1 Score): 97.25%

- Random Forest:

Es un conjunto de árboles de decisión y puede ser robusto y preciso.

Puede manejar bien características no lineales y desbalanceadas.

Proporciona importancia de características, lo que puede ser útil en un contexto médico.

```
#Importar las Bibliotecas Necesarias:
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score
from sklearn.datasets import load_breast_cancer

#Cargar el Conjunto de Datos:
breast_cancer_data = load_breast_cancer()
X = pd.DataFrame(breast_cancer_data.data,
columns=breast_cancer_data.feature_names)
y = pd.Series(breast_cancer_data.target)

#Dividir los Datos en Conjuntos de Entrenamiento y Prueba:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

#Inicializar y Entrenar el Modelo Random Forest:
random_forest_model = RandomForestClassifier(n_estimators=100)
random_forest_model.fit(X_train, y_train)

#Realizar Predicciones en el Conjunto de Prueba:
y_pred = random_forest_model.predict(X_test)

#Calcular las Métricas de Evaluación:
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

#Imprimir los Resultados:
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")
print(f"F1 Score: {f1:.2f}")
```

Este código proporcionará una evaluación completa del algoritmo Random Forest en el conjunto de datos Breast Cancer Dataset, incluyendo la precisión, la precisión, la recuperación y el valor F1 del modelo. Además, el modelo Random Forest puede proporcionar insights sobre la importancia de las características, lo que puede ser muy valioso en el análisis de datos médicos.

Comparación de los 3 con el Dataset Breast Cancer Dataset:

Una comparación de los tres algoritmos - Regresión Logística, Máquinas de Soporte Vectorial (SVM), y Random Forest - utilizando el Breast Cancer Dataset. Aquí está un resumen de los resultados de cada algoritmo en las métricas de evaluación clave: exactitud, precisión, recuperación, y valor F1:

1. **Regresión Logística**:
    - Exactitud: 97.66%
    - Precisión: 98.15%
    - Recuperación: 98.15%
    - Valor F1: 98.15%
2. **Máquinas de Soporte Vectorial (SVM)**:
    - Exactitud: 96.49%
    - Precisión: 96.36%
    - Recuperación: 98.15%
    - Valor F1: 97.25%
3. **Random Forest**:
    - Exactitud: 97.08%
    - Precisión: 96.40%
    - Recuperación: 99.07%
    - Valor F1: 97.72%

Análisis de Resultados:

- **Rendimiento General**: Todos los modelos han mostrado un alto rendimiento en el conjunto de datos, con métricas superiores al 96% en todas las categorías.
- **Regresión Logística**: Este modelo ha demostrado ser el más equilibrado en términos de todas las métricas de evaluación, destacando por su precisión y facilidad de interpretación.
- **SVM**: Aunque con un rendimiento ligeramente inferior en comparación con la Regresión Logística, el modelo SVM ha manejado eficientemente las características de alta dimensión, lo cual es relevante en datos médicos.
- **Random Forest**: Este modelo ha mostrado un excelente desempeño, especialmente en términos de recuperación, lo que sugiere su capacidad para manejar características no lineales y desbalanceadas. Además, proporciona información útil sobre la importancia de las características.

Conclusión de la comparación:

Para el conjunto de datos Breast Cancer Dataset, la Regresión Logística ha resultado ser ligeramente superior en términos de precisión y facilidad de interpretación, aunque Random Forest destaca por su capacidad para identificar correctamente la mayoría de los casos positivos (alta recuperación). El modelo SVM, por su parte, ofrece un buen equilibrio entre precisión y manejo de datos de alta dimensión. La elección entre estos modelos dependerá de los requisitos específicos de la aplicación, como la importancia de la interpretabilidad del modelo frente a la necesidad de manejar datos complejos o desbalanceados.

Conclusión General:

El análisis exhaustivo del "Breast Cancer Dataset" utilizando algoritmos de aprendizaje automático como la Regresión Logística, Máquinas de Soporte Vectorial (SVM) y Random Forest ha demostrado ser una herramienta valiosa en la lucha contra el cáncer de mama. Cada uno de estos modelos ha mostrado una alta precisión y eficacia en la clasificación de tumores como benignos o malignos, lo que subraya el potencial del aprendizaje automático en el campo de la detección y diagnóstico del cáncer.

La Regresión Logística se destacó por su simplicidad y alta precisión, haciéndola ideal para situaciones donde la interpretación y la facilidad de uso son prioritarias. Por otro lado, las Máquinas de Soporte Vectorial demostraron su capacidad para manejar eficientemente características de alta dimensión, lo cual es crucial en el análisis de datos médicos complejos. Finalmente, el modelo de Random Forest no solo proporcionó un alto nivel de precisión, sino que también ofreció insights importantes sobre la importancia de las características individuales, lo que es de gran valor en la toma de decisiones médicas y en la investigación.

Este estudio reafirma la importancia de la integración de tecnologías avanzadas de análisis de datos en el campo médico. La capacidad de estos modelos de aprendizaje automático para proporcionar diagnósticos rápidos y precisos puede tener un impacto significativo en la detección temprana del cáncer de mama, mejorando así las tasas de supervivencia y la calidad de vida de los pacientes. Además, los insights generados a través de estos modelos pueden facilitar una mejor comprensión de la enfermedad, lo que conduce a tratamientos más eficaces y personalizados.

En conclusión, el uso de algoritmos de aprendizaje automático en el análisis del "Breast Cancer Dataset" no solo demuestra la viabilidad de estas técnicas en aplicaciones médicas, sino que también abre nuevas vías para la investigación y el desarrollo de soluciones innovadoras en la atención de la salud, marcando un hito importante en la lucha contra el cáncer de mama.