Práctica Profesional

Elección del modelo

- **Introducción**

El presente informe tiene como objetivo documentar el proceso de selección y evaluación de modelos de Machine Learning(ML) empleados en un proyecto orientado a la automatización del análisis de datos médicos. Este proyecto está enfocado en asistir a los profesionales de la salud, proporcionando un análisis preliminar de los datos ingresados, lo que agiliza la toma de decisiones clínicas y permite detectar patrones relevantes en los pacientes.

El desafío principal del proyecto es construir un sistema capaz de manejar grandes volúmenes de datos médicos y devolver predicciones precisas que ayuden al diagnóstico y seguimiento de condiciones de salud. En este contexto, se han explorado múltiples modelos de ML, con el fin de identificar aquellos que ofrezcan un equilibrio óptimo entre precisión, sensibilidad y capacidad de generalización.

Regresión Logística:

- Comprender la regresión: se busca clasificar los pacientes que padecen hipertensión.
- Predicción: se establece que en base a los datos de los pacientes como Frecuencia cardíaca, Presión sistólica, Diastólica y presión por pulso si padecen hipertensión.

En estos ejemplos, primero creamos la variable hipertensión a partir de los datos de presión arterial. Luego, dividimos los datos en conjuntos de entrenamiento y prueba, entrenamos el modelo y evaluamos su rendimiento.

```
# Crear la variable objetivo 'hipertension' usando .loc para evitar SettingWithCopyWarning
df_limpio.loc[:, 'hipertension'] = ((df_limpio['Systolic'] >= 140) | (df_limpio['Diastolic'] >= 90)).astype(int)
```

```
# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

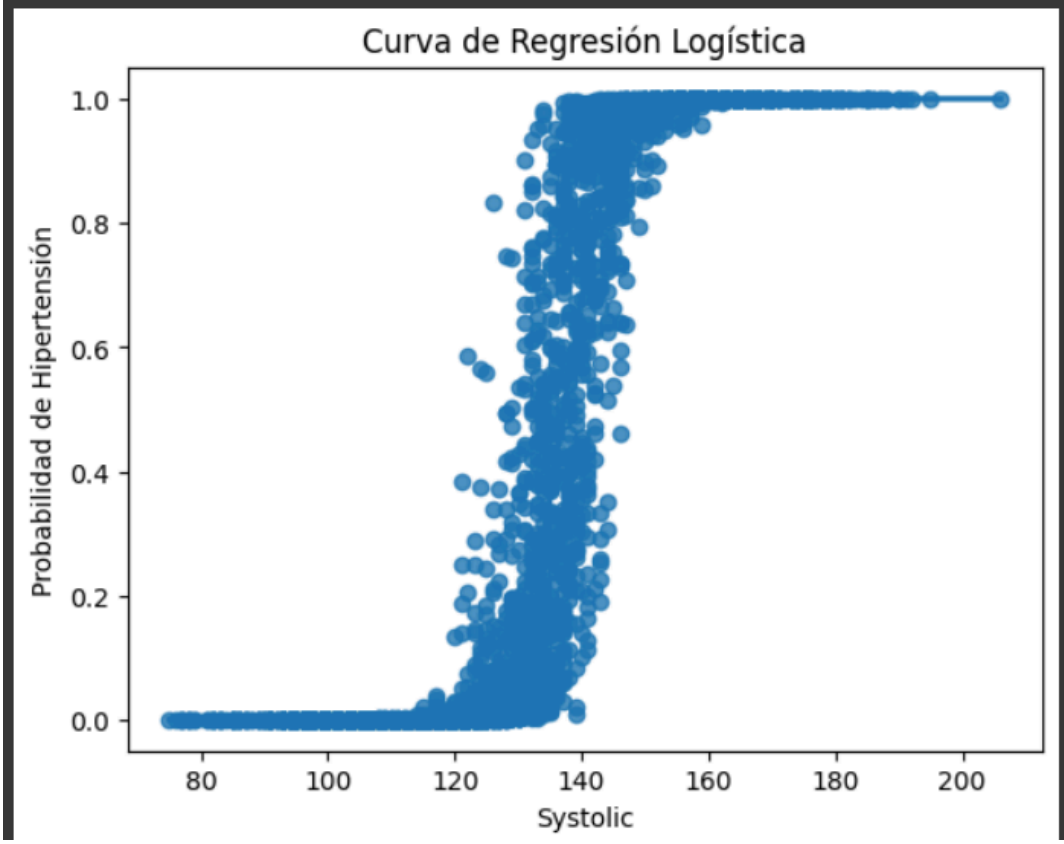
```
# Ajustar los parámetros del modelo usando GridSearchCV
model = LogisticRegression()
param_grid = {
    'C': [0.1, 1, 10, 100],
    'solver': ['liblinear', 'lbfgs']
}
grid_search = GridSearchCV(model, param_grid, cv=5)
grid_search.fit(X_train, y_train)
```

```
# Evaluar el modelo
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(y_test, y_pred)
```

Curva Sigmoide:

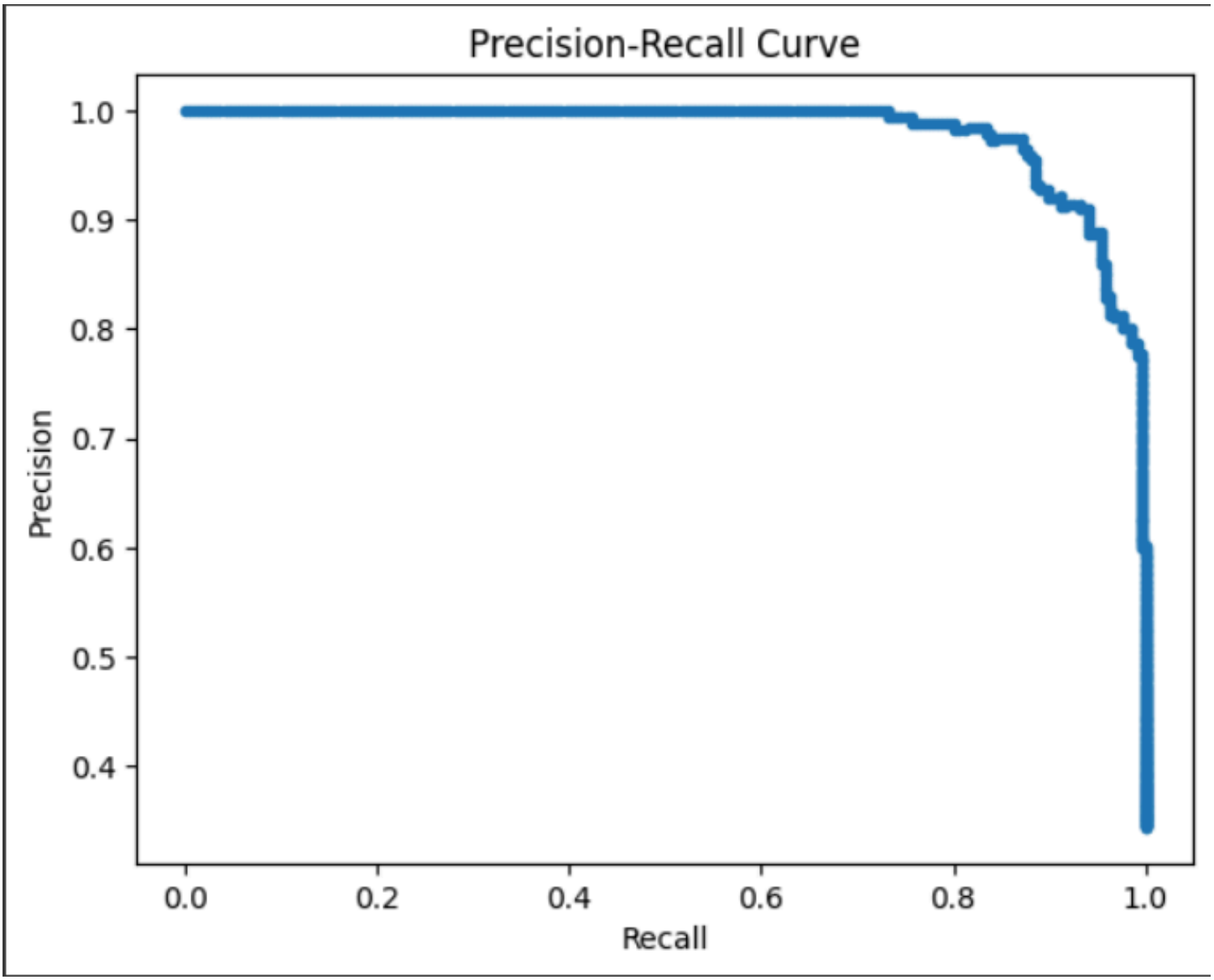
Esta curva muestra cómo la probabilidad de tener hipertensión cambia a medida que varía la variable independiente.

```python
# Graficar la curva de regresión logística para una variable independiente (por ejemplo, 'Systolic')
sns.regplot(x='Systolic', y='predicted_prob', data=df_limpio, logistic=True, ci=None)
plt.xlabel('Systolic')
plt.ylabel('Probabilidad de Hipertensión')
plt.title('Curva de Regresión Logística')
plt.show()
```



Curva de Precisión-Recall:

La curva de lift muestra un buen desempeño del modelo, ya que se encuentra por encima de la linea diagonal y se mantiene alta en la parte inicial, sugiere que el modelo es capaz de identificar correctamente una proporción significativa de los casos positivos.


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0

```
Accuracy: 0.9412698412698413
Confusion Matrix:
[[401  12]
 [ 25 192]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.97      0.96       413
           1       0.94      0.88      0.91       217

    accuracy                           0.94       630
   macro avg       0.94      0.93      0.93       630
weighted avg       0.94      0.94      0.94       630

F1 Score: 0.9121140142517815
Recall: 0.8847926267281107
Precision: 0.9411764705882353
ROC AUC Score: 0.9889423237857198
Matthews Correlation Coefficient: 0.8690110275474354
```

Explicación:

- Accuracy (Precisión): 96.6% de las observaciones. (hipertenso o no hipertenso). Implicación: Un valor tan alto sugiere que el modelo es muy preciso en general. Confusion Matrix:
- [[757 14] [19 180]] Diagonal principal: Representa las predicciones correctas (verdaderos positivos y verdaderos negativos). Fuera de la diagonal: Representa las predicciones incorrectas (falsos positivos y falsos negativos). Interpretación: En este caso, indica que el modelo clasificó correctamente 757 pacientes como no hipertensos y 180 como hipertensos. Sin embargo, hubo 14 falsos positivos (clasificados como hipertensos cuando no lo eran) y 19 falsos negativos (clasificados como no hipertensos cuando sí lo eran).
- F1-Score: 0.92 Indica un buen equilibrio entre precisión y recall, lo que sugiere que el modelo es capaz de identificar tanto a los pacientes hipertensos como a los no hipertensos con una alta precisión.
- Recall: 0.90 Indica que el modelo identificó correctamente el 90% de los pacientes hipertensos. Esto es importante en problemas de salud, ya que es preferible identificar a todos los pacientes hipertensos para iniciar un tratamiento oportuno.
- Precision: 0.93 Indica que cuando el modelo predice que un paciente tiene hipertensión, tiene una probabilidad del 93% de estar en lo correcto.
- ROC AUC Score: 0.995 Mide el rendimiento general del modelo en todas las posibles umbrales de clasificación. Un valor cercano a 1 indica un excelente desempeño.

- Matthews Correlation Coefficient (MCC): 0.89 Es una medida de la calidad de la clasificación binaria que considera las cuatro clases de la matriz de confusión. Un valor cercano a 1 indica una predicción perfecta.

En resumen, los resultados obtenidos indican que el modelo de regresión logística desarrollado es altamente preciso en la predicción de hipertensión. El modelo demuestra una excelente capacidad para distinguir entre pacientes hipertensos y no hipertensos, con un bajo número de falsos positivos y falsos negativos.

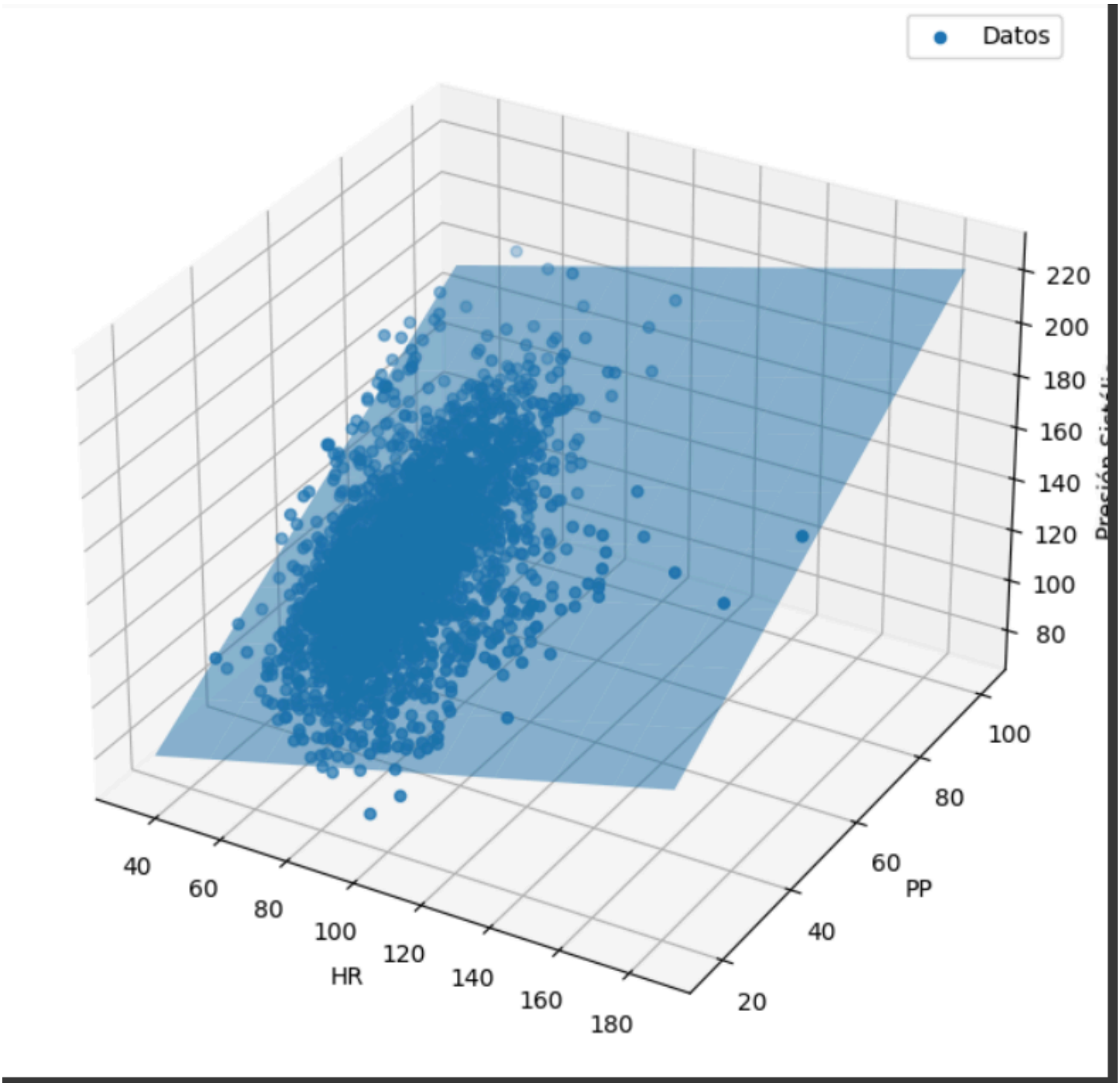
Regresión lineal múltiple

Si HR (frecuencia cardiaca), PP (presión del pulso) aumenta, también la presión sistólica lo hace?

- Podremos: predecir la presión arterial de nuevos pacientes con valores conocidos de HR, edad, presión del pulso, etc.
- Identificar que variables independientes tienen un impacto significativo en la presión arterial sistólica,
- Analizar la relación entre las variables independientes y la presión arterial sistólica.

```
Coeficientes:  [0.32634549 1.11902589]
Intercepción:  48.30713423972206
[ 0.30700161 -0.63969944  0.21016934  0.04102566  0.45827725]
```

```
La ecuación de regresión es: Systolic = -2.0906271718625895 + 0.00023280583301939128 * HR + 50070648638.42633 * PP
Esto significa que por cada unidad que aumenta HR, la presión sistólica aumenta en promedio 0.00023280583301939128 unidades,
manteniendo PP constante.
```



Interpretación

Coeficientes

- Intersección (b): 0.3359125509976309: Representa el valor de la presión sistólica cuando todas las demás variables son cero. Sin embargo, esto puede no tener un significado práctico, ya que es poco probable que todas las variables sean cero en un contexto real.
- Pendiente (m): [0.0039165 0.41924161 0.02298974 -0.10962461] Cada valor en este vector representa el cambio en la presión sistólica por un aumento de una unidad en la variable correspondiente. Por ejemplo, un aumento de una unidad en la segunda variable (probablemente la presión diastólica) se asocia con un aumento de 0.41924161 unidades en la presión sistólica, asumiendo que las otras variables se mantienen constantes.

Métricas de Evaluación

- Error cuadrático medio (RMSE): 0.3050087363166971: Indica, en promedio, qué tan lejos están las predicciones del modelo de los valores reales. Un valor bajo indica un mejor ajuste del modelo.
- Error absoluto medio (MAE): 0.2634812146369789: Mide la diferencia absoluta promedio entre los valores predichos y los reales. Es una medida más robusta a valores atípicos que el RMSE.
- R^2 Coeficiente de determinación: 0.5880012756921713: Explica la proporción de la variabilidad en la presión sistólica que es explicada por el modelo. Un valor de 0.588 indica que el modelo explica aproximadamente el 58.8% de la variabilidad en la presión sistólica.

Conclusión

El modelo de regresión lineal múltiple parece explicar una proporción moderada de la variabilidad en la presión sistólica, pero hay espacio para mejora. Las variables incluidas en el modelo tienen un impacto significativo en la presión sistólica, como se evidencia en los valores de los coeficientes. Sin embargo, la interpretación de estos coeficientes debe hacerse con cuidado, considerando la escala de las variables y las posibles correlaciones entre ellas. El RMSE y el MAE indican que, en promedio, las predicciones del modelo están relativamente cerca de los valores reales, pero hay margen para reducir el error.

Random Forest Classifier

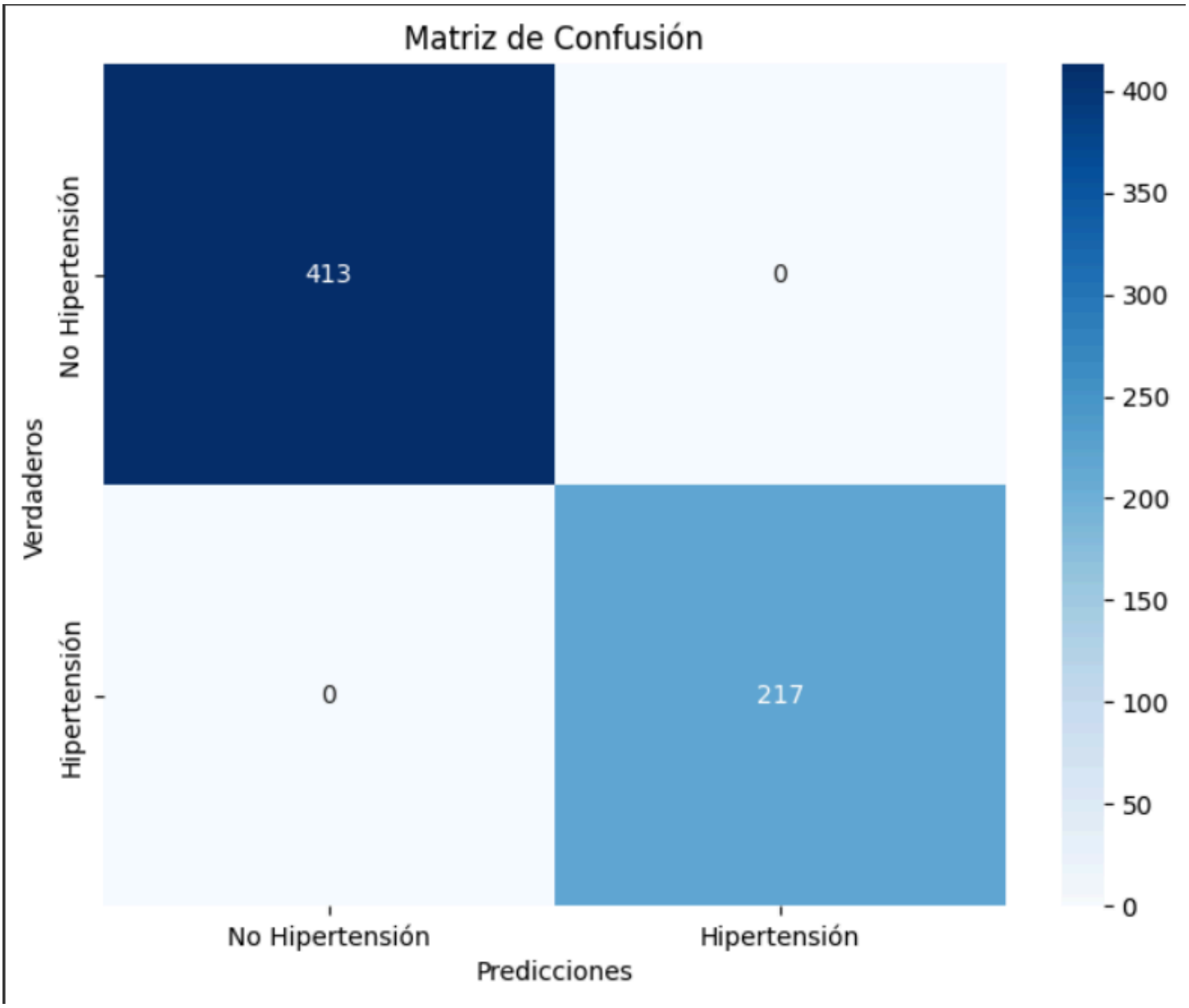
El modelo está tratando de predecir la presencia o ausencia de hipertensión en los pacientes basándose en las características fisiológicas. La lógica detrás de esta predicción se basa en la relación entre las características fisiológicas (como la presión arterial y la frecuencia cardíaca) y la condición de hipertensión.

Esto es útil en un contexto médico porque permite identificar a los pacientes que pueden necesitar intervención o tratamiento basado en su riesgo de hipertensión, lo que puede ayudar en la prevención de complicaciones relacionadas con la presión arterial alta.

```
Mejores parámetros: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
Accuracy: 1.0
Confusion Matrix:
[[413   0]
 [  0 217]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       413
           1       1.00      1.00      1.00       217

    accuracy                           1.00       630
   macro avg       1.00      1.00      1.00       630
weighted avg       1.00      1.00      1.00       630
```



Explicación

Accuracy: 1.0

Este resultado perfecto indica que el modelo clasificó correctamente todas las observaciones en el conjunto de prueba. Sin embargo, es importante recordar que un conjunto de prueba demasiado pequeño o datos sintéticos pueden llevar a resultados demasiado optimistas. Es recomendable evaluar el modelo en un conjunto de datos de prueba independiente y más grande para obtener una evaluación más realista. Confusion Matrix:

[[413 0] [0 217]]

La diagonal principal muestra que todas las observaciones fueron clasificadas correctamente (verdaderos positivos y verdaderos negativos). Los elementos fuera de la diagonal son cero, lo que significa que no hubo falsos positivos ni falsos negativos. Classification Report:

Precision, Recall, F1-score:

Todos estos valores son 1.0 para ambas clases, lo que indica un rendimiento perfecto del modelo. Support: Muestra el número de observaciones en cada clase. Conclusión El modelo de Random Forest con los parámetros óptimos encontrados logró una precisión perfecta en la clasificación de las observaciones en el conjunto de prueba. Esto sugiere que el modelo es altamente capaz de distinguir entre las dos clases.

Sin embargo, un rendimiento perfecto en el conjunto de prueba puede ser una señal de sobreajuste, especialmente si el conjunto de prueba es pequeño o los datos son sintéticos.

Conclusión Final

En este proyecto de machine learning orientado a la automatización del análisis de datos médicos, se evaluaron diversos modelos con el fin de asistir a los profesionales de la salud en la identificación de hipertensión. La exploración incluyó modelos como la regresión logística, la regresión lineal múltiple y el Random Forest Classifier, cada uno con fortalezas y limitaciones particulares.

El modelo de regresión logística mostró un excelente desempeño en la clasificación de pacientes hipertensos, con una precisión del 96.6% y métricas como un F1-Score de 0.92 y un AUC de 0.995, lo que evidencia su capacidad para identificar correctamente tanto casos positivos como negativos. Sin embargo, es necesario monitorear los falsos positivos y falsos negativos para asegurar un balance adecuado en aplicaciones clínicas.

La regresión lineal múltiple permitió analizar la relación entre variables fisiológicas y la presión arterial sistólica, explicando un 58.8% de la variabilidad en los datos. Aunque este modelo es útil para comprender las relaciones subyacentes, las métricas de error (RMSE y MAE) indican que aún hay margen de mejora en cuanto a la precisión de las predicciones.

El Random Forest Classifier, si bien alcanzó una precisión perfecta en el conjunto de prueba, sugiere un posible sobreajuste. A pesar de su rendimiento sobresaliente en los datos analizados, será crucial probarlo en un conjunto de datos independiente y más grande para asegurar que el modelo generaliza correctamente.

En resumen, cada uno de los modelos aporta valor en diferentes aspectos del análisis de hipertensión. La regresión logística destaca por su alta precisión, el análisis de regresión lineal ofrece una comprensión detallada de las relaciones entre variables, y el Random Forest muestra un potencial elevado, aunque debe ser evaluado más exhaustivamente para evitar sobreajuste. Una combinación de estos enfoques o un ajuste adicional de los modelos seleccionados puede conducir a un sistema robusto y confiable para el diagnóstico y seguimiento de la hipertensión en el ámbito Clínico.