Escuela de Ingeniería Industrial

Comorbilidades y COVID-19
Equipo 10

por:

Bárbara Chávez Reveco
20.745.078-2
Daniela Tapia Barrera
21.226.041-k

EII4220

Profesores:
Diego Martinez Cea
Javier Maturana Ross

Noviembre, 2024

PARTE I

Retroalimentación entregada por profesores: “Sigue faltando el uso adecuado de referencias bibliográficas que sustenten los argumentos presentados. La inclusión de datos concretos sobre la prevalencia de las comorbilidades o estudios previos sobre el tema sería de gran valor para reforzar la relevancia del proyecto (consultar con Profe Diego). Evaluar utilizar otra metodología más, además de RL. , algunos de los objetivos propuestos siguen siendo algo vagos, y sería beneficioso reformularlos para hacerlos más concretos y medibles”.

Referencias bibliográficas y argumentos presentados

Como se mencionó en las entregas anteriores, el objetivo de este proyecto es lograr predecir si un paciente portador del virus SARS-CoV-2, o mejor conocido como COVID-19, es de alto riesgo o no, dependiendo de las comorbilidades que éste posea. Para este fin, se seleccionó una base de datos del portal Kaggle, la cual cuenta con diversos atributos, los cuales ya fueron descritos en la entrega 2. Diversos estudios obtenidos de “National Library of Medicine” exponen resultados sobre la relación entre comorbilidades y COVID-19, en donde todos coinciden en una alta correlación existente entre estas variables, lo cual puede conducir a resultados clínicos o mortalidad.

Estudios como “Multisite implementation of a workflow-integrated machine learning system to optimize COVID-19 hospital admission decisions”, “The Association between Presence of Comorbidities and COVID-19 Severity: A Systematic Review and Meta-Analysis”, “Comorbidities' potential impacts on severe and non-severe patients with COVID-19: A systematic review and meta-analysis”, “Comorbidities of COVID-19 Patients”, y otros específicos de cada comorbilidad como “A review on obesity as a risk factor for mortality in COVID-19 disease”, “COVID-19, hipertensión y enfermedad cardiovascular”, entre otros, cumplieron un rol fundamental a la hora de orientarnos y poder determinar aquellas comorbilidades que tienden a estar más relacionadas con los pacientes con diagnóstico severo de COVID-19, el cual se puede entender como un alto grado de intensidad del virus, o bien, si el paciente fue intubado, hospitalizado en cuidados intensivos UCI, entre otros. Dentro de los factores de riesgo para el virus, se encuentran el género, en caso de hombres, edad avanzada y además, las comorbilidades preexistentes. Según estadísticas de “Comorbidities of COVID-19 Patients”, se observó que el 75% de los pacientes hospitalizados por COVID-19 padecen de al menos una comorbilidad, en donde las más comunes son hipertensión, diabetes, cáncer, inmunodeficiencia, enfermedades cardiovasculares, obesidad y enfermedades renales. De una muestra de 99 pacientes, 50 de ellos padecen de comorbilidades, por otra parte, existen comorbilidades que con frecuencia tienden a aumentar el riesgo de mortalidad en pacientes, como cáncer de pulmón, leucemia mieloide crónica e infecciones por VIH. Un estudio realizado en 5700 pacientes con COVID-19 de 12 hospitales en los EE. UU. encontró que la diabetes era la tercera comorbilidad más común (34% de los pacientes), en comparación con la hipertensión (56%) y la obesidad (42%). (Silaghi-Dumitrescu et al., 2023)

Objetivos propuestos

	Entrega 2	Entrega 3
MUST DELIVERABLES	- Realizar diccionario de la Base de Datos, según S.S. Stevens: En este entregable se describirán todas las variables del dataset, para comprender qué propiedades/operaciones posee cada una de ellas. - Realizar limpieza de datos: Se eliminarán o reemplazarán datos que no se consideren relevantes para la predicción. Además se realizará la limpieza para aquellos datos faltantes, en nuestro caso, valores 97, 98 y 99 (caso particular descrito en el desarrollo del diccionario de datos). - Realizar regresión logística: Uno de los últimos pasos, será construir y entrenar a nuestro modelo de regresión logística, para asegurar la validez del trabajo realizado y cumplir con el objetivo de predecir la severidad del COVID-19	- Más que modificar los "must" deliverables, añadir entrega del análisis de resultados una vez aplicado el modelo de regresión logística, además del algoritmo de árbol de decisión. En base a esto se obtendrá un modelo más preciso y con resultados más acertados al tener "dos opciones" de resolverlo. El análisis de resultados es fundamental en Data Science , ya que permite evaluar el desempeño de los modelos aplicados, identificar patrones y/o tendencias, los cuales pueden ser valiosos para la toma de decisiones y definir estrategias futuras, entre otros. Un estudio de los resultados obtenidos asegura que el proyecto tenga un impacto real y relevante, además de que este sea un buen modelo de predicción y a su vez, acertado.
EXPECT DELIVERABLES	- Realizar análisis de correlación: Esto nos servirá para tener una mayor comprensión de aquellas variables que se encuentran relacionadas entre sí - Crear gráficos de resultados: estos nos	Agregar: - Análisis exploratorio de Datos EDA, realizado en entrega 2, el cual muestra un "resumen" del comportamiento de los atributos, comorbilidades en este caso, además de otras clasificaciones. - Entregar métricas de

	permitirán conocer visualmente patrones, outliners, entre otros. Para así tener un poco más de información para tomar las decisiones correctas - Analizar posibles sesgos o patrones: esto para evaluar si existen sesgos ya sea por edad y/o género y proponer soluciones para estas o para poder tratar de mejor manera a ciertos pacientes.	evaluación del modelo, lo cual está previsto para entrega 3, en conjunto con una interpretación del modelo. Estas métricas ayudarán a confirmar la eficacia del modelo y su eventual aplicación en un futuro en un contexto clínico. - Implementar más de un modelo de predicción, en este caso, árboles de decisión y regresión logística, con el fin de tener dos formas distintas y complementarias de analizar y evaluar los resultados obtenidos.
WOULD LIKE DELIVERABLES	- Implementar el modelo en sistemas de salud, ya sean hospitales, clínicas, e incluso, instituciones gubernamentales, con el fin de poder mejorar la distribución de insumos y estar preparados para lo que se vendrá. - Seguir desarrollando capacidades y lograr llegar a probar algoritmos avanzados como redes neuronales, para comparar resultados con la regresión logística. - Extender el análisis no solo para predecir la severidad del COVID-19, sino también para predecir los riesgos de hospitalización o mortalidad en personas con comorbilidades	Agregar: - Implementar una aplicación web para pruebas clínicas, de uso simple, que permita a los médicos ingresar datos específicos de pacientes y obtener predicciones de riesgo de COVID-19 en tiempo real. - Desarrollar un sistema de alertas automáticas que notifique a autoridades sanitarias y médicos en caso de que ocurra un aumento en el riesgo de COVID-19 para pacientes con comorbilidades. - Certificación Internacional y Reconocimientos Médicos: Conseguir que el modelo de predicción obtenga certificaciones de organismos internacionales (por ejemplo, FDA, OMS) para garantizar su seguridad y precisión en contextos clínicos.

PARTE II:

La pandemia de COVID-19 ha impactado de manera desproporcionada a individuos con comorbilidades previas, aumentando su riesgo de complicaciones graves e incluso de mortalidad. Diversos estudios han demostrado que enfermedades como la hipertensión, diabetes, obesidad y condiciones respiratorias pueden exacerbar la gravedad del diagnóstico, debido a cómo estas afecciones afectan la respuesta inmunitaria y la capacidad de los pacientes para recuperarse.

En este contexto este proyecto semestral busca determinar cómo las diferentes comorbilidades pueden llegar a influir en la severidad del diagnóstico de COVID-19.

Para lograr esto se trabajará con una base de datos obtenida de Kaggle, a la cual se le realizó una limpieza en la entrega anterior, quedando sólo con aquellos atributos considerados relevantes para el objetivo final que se espera alcanzar, esto tras la lectura de diversos papers en relación al COVID-19 y patologías como diabetes, obesidad, hipertensión, entre otros.

Inicialmente, el enfoque fue utilizar solo un modelo de regresión logística para predecir la severidad del diagnóstico de COVID-19 en pacientes con comorbilidades. Sin embargo, tomando en consideración la retroalimentación realizada por los profesores, en donde se propone evaluar la utilización de otra metodología, además luego de leer papers sobre estudios en relación a COVID-19 y patologías, se tomó la decisión de realizar dos modelos de manera complementaria, para así obtener un modelo más interpretable y preciso, que nos permita analizar de mejor manera los resultados, esto considerando que su combinación puede conducir a mejores resultados en términos de precisión.

Con el objetivo de comprender mejor el papel que tienen las comorbilidades en la severidad del COVID-19, se realizaron tres modelos de regresión logística independientes. Cada modelo se centró en un *outcome* específico: Intubación (“INTUBED”), ingreso a la UCI (“ICU”) y hospitalizaciones (“PACIENT_TYPE”), cada una de estas variables consideradas como niveles de alto riesgo en caso de que el paciente haya sido intubado, ingresado a la uci, o bien, hospitalizado, esto con el argumento de que pacientes portadores de COVID-19 en un nivel moderado debían cumplir un reposo en sus hogares, con cuidados más ambulatorios, sin supervisión médica constante. El objetivo principal fue identificar las comorbilidades que se asocian significativamente con un mayor riesgo de cada uno de estos outcomes. Los resultados obtenidos permitieron tener una mejor comprensión de los factores de riesgo para la severidad de la enfermedad.

Árbol de decisión y Regresión logística 1 → “ICU” y comorbilidades

A continuación se detallan los pasos y resultados obtenidos de la construcción de cada modelo y se muestran los códigos implementados en Python.

En primer lugar, se seleccionaron las comorbilidades que serán parte de “X”, es decir, la variable independiente y el outcome, en este caso “y”.

```
X =mydata[["PNEUMONIA", "DIABETES", "COPD", "ASTHMA", "INMSUPR", "HIPERTENSION", "CARDIOVASCULAR", "OBESITY", "RENAL_CHRONIC", "TOBACCO"]]
y = mydata["ICU"]
```

Luego se construyó el árbol de decisión con el siguiente código, utilizando las mismas variables dependientes e independientes.

```
dt.fit(X_train, y_train)
```

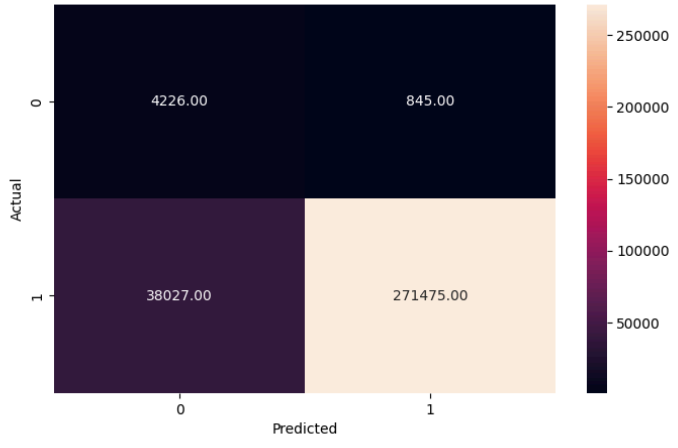
```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=1)
```

A continuación, se evaluó el desempeño del modelo con los datos de prueba, tanto la precisión, recall y F1-Score, además se creó una matriz de confusión, usando el siguiente código

```
def metrics_score(actual, predicted):
    print(classification_report(actual, predicted))
    cm = confusion_matrix(actual, predicted)
    plt.figure(figsize=(8,5))
    sns.heatmap(cm, annot=True,  fmt='.2f', xticklabels=['0', '1'], yticklabels=['0', '1'])
    plt.ylabel('Actual')
    plt.xlabel('Predicted')
    plt.show()

    accuracy = accuracy_score(actual, predicted)
    print(f'Accuracy: {accuracy}')

metrics_score(y_test, y_pred)
```

Obteniendo los siguientes resultados:

```
              precision    recall  f1-score   support

           0       0.10      0.83      0.18      5071
           1       1.00      0.88      0.93    309502

    accuracy                           0.88    314573
   macro avg       0.55      0.86      0.56    314573
weighted avg       0.98      0.88      0.92    314573
```



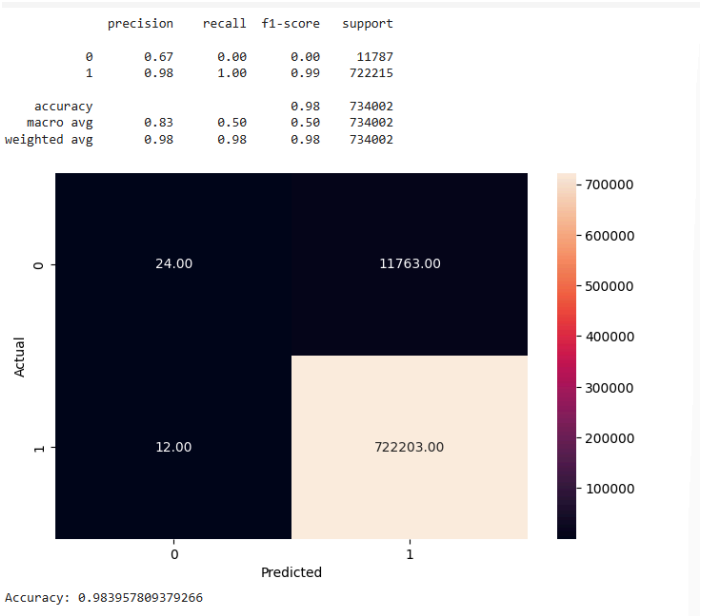

```
Accuracy: 0.8764293184729777
```

La matriz presenta la cantidad de muestras que fueron clasificadas. Por un lado, la diagonal principal, representa las predicciones correctas, mientras más altos sean estos valores, significa que mejor es el desempeño del modelo. Mientras que los valores fuera de la diagonal, indican las predicciones incorrectas.

Por otro lado, con respecto al desempeño del modelo, se puede destacar lo siguiente:

- La precisión para el valor 0 es baja (0.10), lo que indica que cuando el modelo predice muestras que pertenecen al valor 0, por lo general se equivoca.
- El recall mide la proporción de muestras positivas que fueron correctamente identificadas, en donde, para 1, el recall es de 0.88, lo que indica que el modelo es bueno para identificar las muestras que pertenecen a esta clase.
- Accuracy = 0.87, lo cual indica que ha clasificado correctamente el 87% de los datos. Si bien es una medida útil para evaluar el rendimiento general del modelo, es importante complementarlo con otras métricas para obtener una visión más completa.
- Existe un gran desequilibrio entre las clases, con muchas más muestras con el valor 1 que del valor 0. Esto puede influir en el desempeño del modelo, ya que puede tender a favorecer la clase mayoritaria.

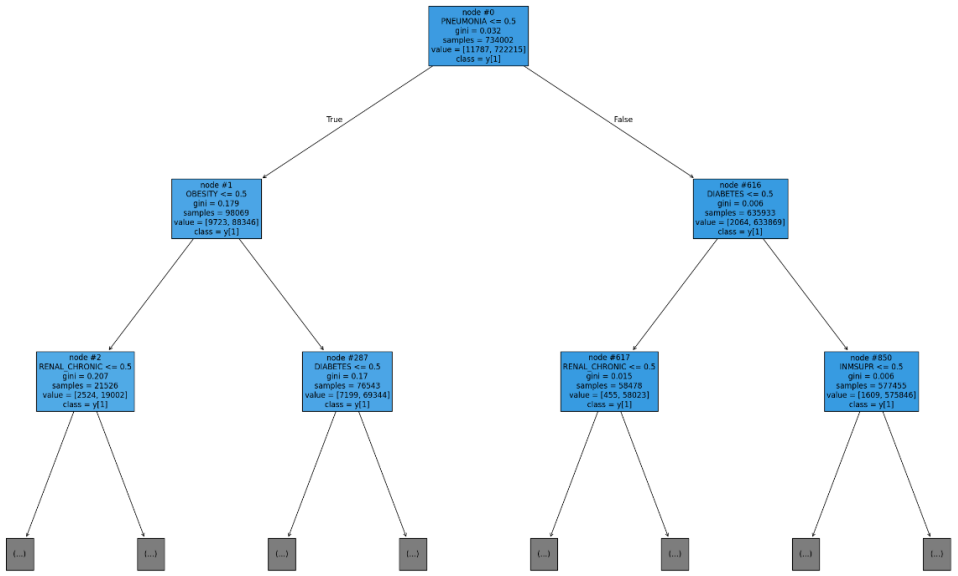
A continuación, se evaluaron los "metrics score" de los datos de entrenamiento. Obteniendo lo siguiente:



Al igual que en el punto anterior, esta matriz y resultados de las métricas, indican que existe un desbalance de los datos. Sin embargo, se decidió visualizar el árbol de decisión de igual forma, esto para entender de forma intuitiva cómo el modelo está tomando decisiones. Además de ver qué variables son las más importantes y cómo se combinan para llegar a una predicción. Es por esto que se utilizó el siguiente código para la creación del árbol de decisión

```
features = list(X.columns)

plt.figure(figsize = (30, 20))

tree.plot_tree(dt, max_depth = 2, feature_names = features, filled = True, fontsize = 12, node_ids = True, class_names = True)
```

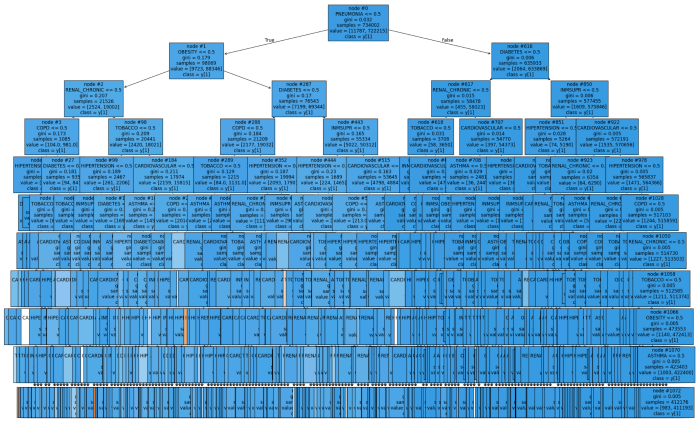
Obteniendo lo siguiente:



Antes de interpretarlo, se presenta una explicación breve de cada elemento:

- Nodos: Puntos de decisión del árbol. Cada nodo representa una pregunta sobre una característica particular
- Ramas: Las líneas que conectan los nodos entre sí. Representan las posibles respuestas.
- Hojas: Son nodos terminales del árbol, donde se toma la decisión final.

Ahora, a pesar de que este árbol nos entrega algo de información, como que PNEUMONIA parece ser la característica más importante para la clasificación, ya que corresponde a la primera pregunta, el árbol parece ser complejo con múltiples niveles de profundidad. Sin embargo al visualizar el árbol completo se observó una alta complejidad, lo que resultó incluso imposible de analizar.



Como señalan Hastie, Tibshirani y Friedman (2009) en su libro '*Elements of Statistical Learning',* los árboles de decisión de gran profundidad pueden volverse extremadamente complejos y difíciles de interpretar. La proliferación de nodos y ramas en un árbol profundo dificulta la identificación de las características más relevantes y las relaciones causales subyacentes. Además, estos modelos son propensos al sobreajuste, lo que limita su capacidad de generalizar a nuevos datos (Hastie, Tibshirani y Friedman, s.f)

Por lo anterior, se optó por truncar el árbol, "sacrificando" precisión para mayor interpretabilidad. No obstante, para obtener una visión más completa y robusta, se complementó el análisis con un modelo de regresión logística, proporcionando una interpretación más sencilla y clara de las relaciones entre las variables.

Para realizar esta regresión logística, se utilizaron las mismas variables que en el árbol de decisión. El primer paso de este segundo algoritmo fue implementar el siguiente código

```
coefficients = pd.DataFrame(log_model.coef_.T, index=X.columns, columns=[f'Clase {i+1}' for i in range(log_model.coef_.shape[0])])
print(coefficients)
```

Luego, dividimos los datos en dos conjuntos: uno para entrenar y otro para probar. La proporción de división que utilizamos fue 70% de entrenamiento y 30% para prueba.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Una vez realizado lo anterior, comenzamos a entrenar el modelo (previamente se importaron las clases necesarias, en este caso *"LogisticRegression"* de sklearn. Durante este proceso, el modelo aprende la relación entre las comorbilidades y la probabilidad de quedar en la UCI ("ICU")

```
from sklearn.linear_model import LogisticRegression

# Crear el modelo de regresión logística
log_model = LogisticRegression()

# Ajustar el modelo a los datos de entrenamiento
log_model.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression()
```

El código presentado a continuación extrae los coeficientes de un modelo de regresión logística, en donde cada uno representa la influencia de una variable predictora, en este caso las comorbilidades, en la probabilidad de ingreso a la UCI. En donde se obtuvieron los siguientes resultados:

	Clase 1
PNEUMONIA	1.168194
DIABETES	0.057854
COPD	-0.010512
ASTHMA	-0.036457
INMSUPR	0.041974
HIPERTENSION	0.026310
CARDIOVASCULAR	0.038363
OBESITY	0.082984
RENAL_CHRONIC	-0.007563
TOBACCO	-0.017550

Los signos positivos en los coeficientes indican que la presencia de la comorbilidad aumenta la probabilidad de ingreso a la UCI, mientras que los signos negativos indican que la presencia de la comorbilidad disminuye la probabilidad de ingreso a la UCI. Por otro lado, los coeficientes indican la fuerza de la asociación, mientras más grande, mayor influencia.

Al interpretar los coeficientes, se pueden realizar las siguientes observaciones

- Neumonía (“PNEUMONIA”) es la comorbilidad que tiene el mayor impacto positivo en la probabilidad de ingresar a la UCI.
- Obesidad (“OBESITY”) también tiene un impacto positivo significativo en la probabilidad de ser ingresado a UCI.
- El resto de las comorbilidades parecen tener un impacto negativo en la probabilidad de ingresar a UCI. Sin embargo este impacto es muy pequeño.

Luego, se evaluó la precisión del modelo de regresión logística en un conjunto de datos de prueba, junto con la “matriz de confusión”, utilizando los siguientes códigos

```
# Predicciones en el conjunto de prueba
y_pred = log_model.predict(X_test)

# Crear el DataFrame con los valores reales (y_test) y predichos (y_pred)
comparison_df = pd.DataFrame({
    'Valor Real (y_test)': y_test,
    'Predicción (y_pred)': y_pred
})

# Añadir una columna que indique si la predicción fue correcta
comparison_df['Es Correcto'] = comparison_df['Valor Real (y_test)'] == comparison_df['Predicción (y_pred)']

# Mostrar las primeras filas del DataFrame
comparison_df.head(100)
```

```
from sklearn.metrics import confusion_matrix

# Calcular la matriz de confusión
conf_matrix = confusion_matrix(y_test, y_pred)
print(f'Matriz de Confusión:\n{conf_matrix}')
```

Obteniendo los siguientes resultados

	Valor Real (y_test)	Predicción (y_pred)	Es Correcto
781974	1	1	True
937737	1	1	True
907828	1	1	True
784628	1	1	True
662460	1	1	True
...	...	...	...
230393	1	1	True
444826	1	1	True
275014	1	1	True
100812	1	1	True
425631	1	1	True

100 rows × 3 columns

```
Matriz de Confusión:
[[     0   5071]
 [     0 309502]]
```

La tabla muestra que cada fila en la que "Es Correcto" es igual a True representa una predicción correcta. Mientras que la Matriz de confusión, indica

- **Verdaderos Positivos:** 0. El modelo no ha clasificado correctamente ningún caso positivo.
- **Falsos Positivos:** 0. El modelo tampoco ha clasificado ningún caso negativo como positivo.
- **Verdaderos Negativos:** 309502. Todos los casos negativos han sido correctamente clasificados como negativos.
- **Falsos Negativos:** 5071. Todos los casos positivos han sido clasificados erróneamente como negativos.

Es decir, el modelo está siempre prediciendo los casos negativos (1), independientemente de los datos de entrada, lo que podría indicar la existencia de datos desbalanceados. Los datos desbalanceados ocurren cuando hay una clase mucho más frecuente que otra, indicando que el modelo podría estar sesgado hacia la clase mayoritaria, esto al igual que ocurrió en el algoritmo del árbol de decisión. En este caso en particular, quiere decir que la predicción del modelo siempre es que el paciente no fue ingresado a la uci (“ICU”).

Sin embargo, en este caso, para equilibrar las clases utilizamos la técnica de “over-sampling”, la cual consiste en aumentar el número de muestras de la clase minoritaria para así reducir el sesgo en el modelo y mejorar su capacidad de predecir correctamente, para el cual se utilizó el siguiente código

```
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix

# ... (tu código para cargar y preprocesar los datos)

# 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)

# Aplicar SMOTE para sobremuestrear la clase minoritaria
smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)

# Crear y entrenar el modelo de regresión logística con los datos remuestreados
log_model = LogisticRegression()
log_model.fit(X_train_resampled, y_train_resampled)


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

# Calcular y mostrar la matriz de confusión
conf_matrix = confusion_matrix(y_test, y_pred)
print(f'Matriz de Confusión:\n{conf_matrix}')
```

Una vez realizado esto, se obtuvo una nueva matriz de confusión presentada a continuación

```
Matriz de Confusión:
[[  4226    845]
 [ 38027 271475]]
```

Al interpretar esta matriz, se puede observar lo siguiente

- **Verdaderos Positivos: 4226:** El modelo clasificó correctamente como positivos a 4226 instancias que realmente eran positivas.
- **Falsos Positivos: 845:** El modelo clasificó erróneamente como positivos a 845 instancias que realmente eran negativas.
- **Verdaderos Negativos: 271475:** El modelo clasificó correctamente como negativos a 271475 instancias que realmente eran negativas.
- **Falsos Negativos: 38027:** El modelo clasificó erróneamente como negativos a 38027 instancias que realmente eran positivas.

Gracias al "over-sampling" el modelo ha logrado identificar correctamente un mayor número de instancias de la clase minoritaria.

Luego se obtuvieron distintas métricas, con el fin de conocer que tan bien está funcionando el modelo.

```
Accuracy: 0.8764293184729777
Precision: 0.9968970329024677
Recall: 0.8771348811962443
F1-score: 0.9331891884459508
```

Estos resultados indican que:

1. El modelo ha clasificado correctamente el 87,64% de las muestras.
2. El modelo tiene una probabilidad del 99,96% de realizar predicciones positivas
3. El modelo identifica correctamente el 87.71% de los casos positivos.
4. El F1-score es una medida balanceada de precisión y recall. Este 0,9332 indica que el modelo tiene un buen equilibrio entre ambos.

Por otro lado, se contaron las predicciones correctas e incorrectas, obteniendo lo siguiente.

```
Es Correcto
True     309502
False      5071
Name: count, dtype: int64
```

Con esta última información se puede decir que, el hecho de haber obtenido valores de recall y F1-score relativamente altos sugiere que el "over-sampling" realizado ha sido efectivo para abordar el desbalance de clases y mejorar el rendimiento del modelo en la clasificación de la clase minoritaria. Al aumentar el número de muestras de la clase minoritaria, el modelo ha

podido aprender mejor sus características y, por lo tanto, ha mejorado su capacidad para identificar correctamente las instancias de esta clase.

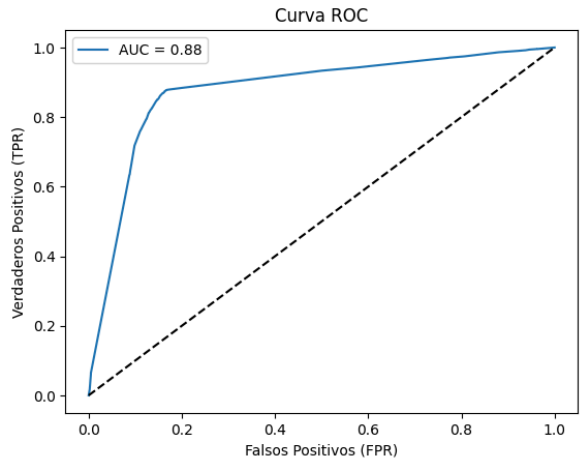
Una vez entrenado el modelo, se realizó una predicción sobre la posibilidad de requerir UCI, con la presencia de diferentes combinaciones de comorbilidades generadas de forma aleatoria, las cuales buscan simular posibles escenarios clínicos.

```python
import numpy as np
import pandas as pd

# Generate random values for the predictor variables
num_predictions = 5  # Number of predictions to generate
X_new = pd.DataFrame({
    'PNEUMONIA': np.random.randint(0, 2, num_predictions),
    'DIABETES': np.random.randint(0, 2, num_predictions),
    'COPD': np.random.randint(0, 2, num_predictions),
    'ASTHMA': np.random.randint(0, 2, num_predictions),
    'INMSUPR': np.random.randint(0, 2, num_predictions),
    'HIPERTENSION': np.random.randint(0, 2, num_predictions),
    'CARDIOVASCULAR': np.random.randint(0, 2, num_predictions),
    'OBESITY': np.random.randint(0, 2, num_predictions),
    'RENAL_CHRONIC': np.random.randint(0, 2, num_predictions),
    'TOBACCO': np.random.randint(0, 2, num_predictions)
})

# Make predictions
predictions = log_model.predict(X_new)

# Get probabilities for each class
probabilities = log_model.predict_proba(X_new)

# Display the results
print("Predictions:", predictions)
print("Probabilities:", probabilities)
```

```
Predictions: [1 1 1 0 0]
Probabilities: [[0.16043159 0.83956841]
 [0.23851661 0.76148339]
 [0.1425001  0.8574999 ]
 [0.91989028 0.08010972]
 [0.56759925 0.43240075]]
```

Finalmente, se utilizó la curva ROC para evaluar el rendimiento del modelo de clasificación. La curva ROC “ideal” corresponde a una línea recta que va de la esquina inferior izquierda a la esquina superior derecha. Esto significa que el modelo puede clasificar perfectamente todos los casos, en donde se obtuvo la siguiente:

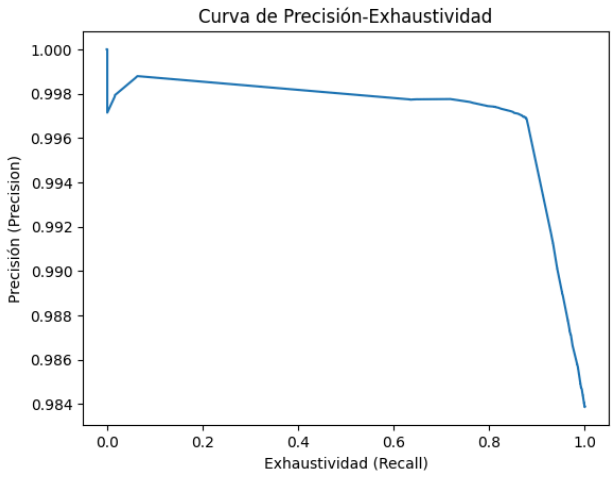
```
AUC-ROC: 0.8772590267245189
```



La curva azul representa el rendimiento del modelo, a medida que nos movemos en la curva, estamos ajustando el umbral para decidir si un caso se clasifica como positivo o negativo. La curva en este caso se encuentra bastante cerca de la esquina superior izquierda, lo cual es bueno, esto significa que el modelo tiene una alta capacidad de identificar correctamente los casos positivos y una alta capacidad de identificar correctamente los casos negativos.

La curva punteada representa el rendimiento de un clasificador aleatorio, si la curva azul está por sobre esta línea, el modelo está funcionando bien. El valor AUC-ROC, 0.87 en este caso resume el rendimiento general del modelo, en donde valores cercanos a 1 indican un buen modelo, mientras que menor a 0.5 indica un modelo sin poder predictivo.En este caso, significa que en promedio, el modelo es capaz de distinguir correctamente entre las dos clases en un 88% de los casos.

Por último, se graficó la curva de precisión-exhaustividad. Esta nos proporciona una visión más detallada del desempeño del modelo, centrándose en la relación entre precisión y exhaustividad.



La curva azul nos indica el rendimiento del modelo al variar el umbral de clasificación, a medida que nos movemos en la curva, estamos ajustando el umbral para decidir si un caso se clasifica como positivo o negativo. En este caso, la curva muestra una tendencia general a disminuir a medida que aumenta la exhaustividad, es decir, refleja la dificultad de lograr una alta precisión y una alta exhaustividad al mismo tiempo, ya que al tratar de aumentar la exhaustividad (identificar más casos positivos), es probable que se incluyan algunos falsos positivos, lo que disminuye la precisión.

El eje X (Exhaustividad) mide la proporción de casos positivos que el modelo identifica correctamente (verdaderos positivos) de todos los casos positivos reales.

El eje Y (Precisión) mide la proporción de casos clasificados como positivos que son realmente positivos.

Ahora, mientras que la curva ROC nos proporciona una visión general del rendimiento del modelo en términos de sensibilidad y especificidad, la curva de precisión-exhaustividad nos permite analizar con más detalle la relación entre la precisión y la exhaustividad del modelo.

Árbol de decisión y Regresión logística 2 → "INTUBED" y Comorbilidades

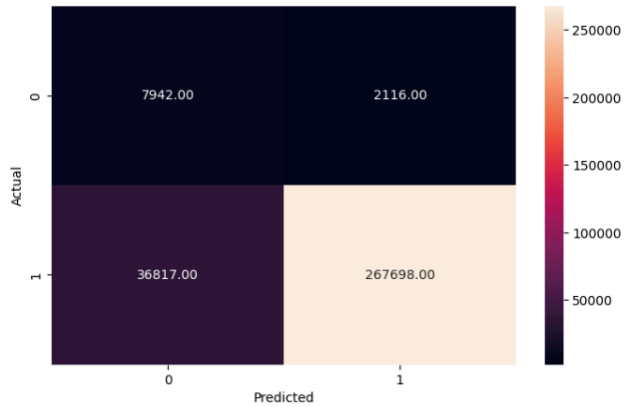
Para el desarrollo tanto del árbol de decisión, como de la regresión logística se aplicó el mismo procedimiento ya descrito en "Árbol de decisión" y "Regresión logística 1", por lo que en este caso, sólo profundizaremos en los valores obtenidos, más que explicar con mayor detalle el código implementado, puesto que como se mencionó, ya fue explicado con anterioridad.

En primer lugar se utilizaron los siguientes datos para la variable dependiente e independiente

```
X =mydata[["PNEUMONIA", "DIABETES", "COPD", "ASTHMA", "INMSUPR", "HIPERTENSION", "CARDIOVASCULAR", "OBESITY", "RENAL_CHRONIC", "TOBACCO"]]
y = mydata["INTUBED"]
```

Al igual que en la primera aplicación del árbol de decisión, se evaluó el desempeño del modelo con los datos de prueba

	precision	recall	f1-score	support
0	0.18	0.79	0.29	10058
1	0.99	0.88	0.93	304515
accuracy			0.88	314573
macro avg	0.58	0.83	0.61	314573
weighted avg	0.97	0.88	0.91	314573



Accuracy: 0.8762354048185954

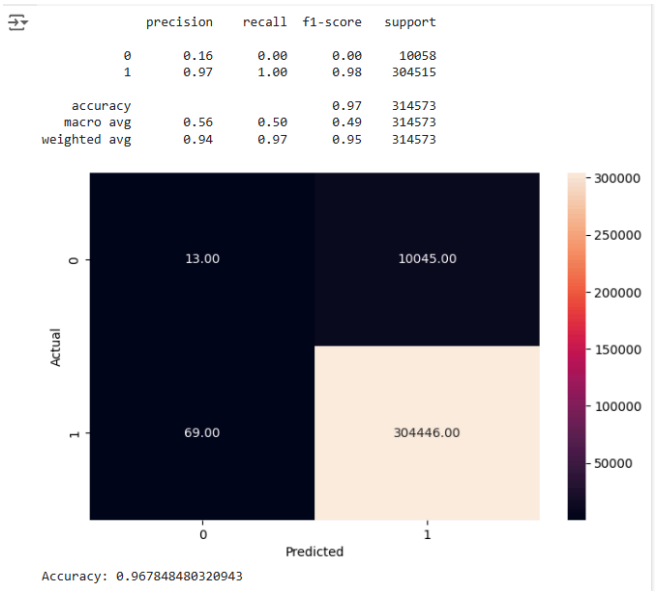
Al interpretar esta matriz, se puede decir

- La precisión para el valor 0 es baja (0.18), lo que significa que muchas de las predicciones positivas para esta clase fueron incorrectas. La precisión para el valor 1 es muy alta (0.99), indicando una alta confianza en las predicciones positivas para esta clase.
- El recall para el valor 0 es alto (0.79), lo que significa que la mayoría de los casos positivos de esta clase fueron identificados correctamente. El recall para el valor 1

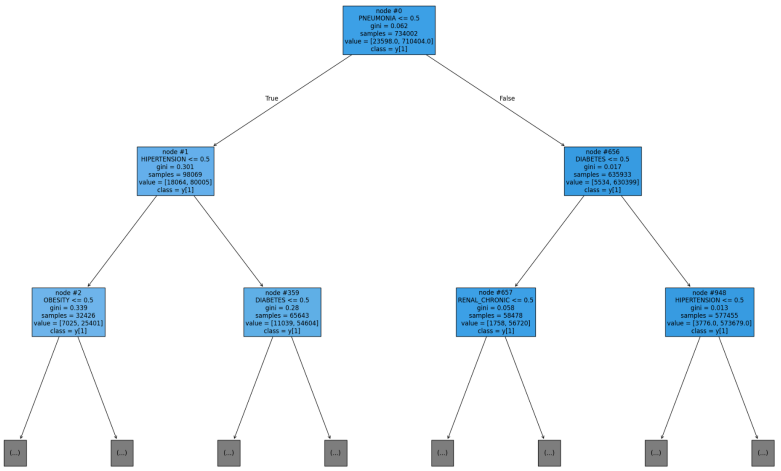
también es alto (0.88), indicando un buen desempeño en la identificación de los casos positivos de esta clase.

- El valor 1 está sobrerrepresentado en los datos, lo que puede influir en el desempeño del modelo. Al igual que en el modelo del árbol de decisión anterior, esto significa que hay un "sesgo" en los valores, lo que provoca que no se realicen predicciones de la clase minoritaria.

De igual forma, se realizó el análisis de los "metric score" para los datos de entrenamiento



Lo indica algo similar a lo interpretado en los datos de prueba, es decir, que existen dificultades para clasificar correctamente la clase minoritaria. De igual forma se decidió visualizar el árbol de decisión, obteniendo lo siguiente



Como se observó en el análisis anterior, el árbol de decisión no resulta adecuado para modelar esta compleja relación entre las variables. Los resultados obtenidos sugieren que un enfoque basado en regresión logística, complementado con un análisis más profundo del árbol de decisión, podría ofrecer una mejor comprensión del fenómeno en estudio. Sin embargo, podemos rescatar, que la primera pregunta que se realiza es sobre la presencia de PNEUMONIA, lo que indica que esta que es la característica más importante del modelo. A continuación se muestran los datos obtenidos con la regresión logística.

Para la aplicación de la regresión logística, se utilizaron las mismas variables dependientes e independientes, mencionadas anteriormente, donde el outcome corresponde a "INTUBED". Además, se usó el 70% de los datos como entrenamiento y de 30% de prueba.

Al aplicar el código que extrae los coeficientes del modelo de regresión logística se obtuvieron los resultados que se presentan a continuación, en donde se concluye que:

- La comorbilidad o variable predictora que tiene mayor influencia en si el paciente es intubado o no, es "PNEUMONÍA", mientras que la que tiene menor incidencia en esta variables es "ASHTMA", por lo que si el paciente COVID-19 es asmático o no, no tiene mayor incidencia en la severidad de su diagnóstico, por otra parte, patologías como "HIPERTENSION" y "DIABETES" también tienen una influencia considerable.

	Clase 1
PNEUMONIA	1.038482
DIABETES	0.116268
COPD	0.011485
ASTHMA	-0.053497
INMSUPR	0.025776
HIPERTENSION	0.151874
CARDIOVASCULAR	0.017876
OBESITY	0.051307
RENAL_CHRONIC	0.021172
TOBACCO	0.006634

Posteriormente, se evaluó la precisión del modelo de regresión logística en un conjunto de datos de prueba, junto con la "matriz de confusión", en donde inicialmente ocurrió el mismo problema presentado anteriormente, la matriz de confusión predecía 0 veces de forma correcta los falsos y verdaderos positivos, tras esto se aplicó nuevamente la técnica de over-sampling, cuyos nuevos resultados fueron:

	Valor Real (y_test)	Predicción (y_pred)	Es Correcto
781974	1	1	True
937737	1	1	True
907828	1	1	True
784628	1	1	True
662460	1	1	True
...	...	...	...
230393	1	1	True
444826	1	1	True
275014	1	1	True
100812	1	1	True
425631	1	1	True

```
Matriz de Confusión:
[[  7942   2116]
 [ 36817 267698]]
```

En donde la interpretación de estos elementos y de la matriz de confusión es la siguiente:

- **Verdaderos Positivos: 267698** → El modelo clasificó correctamente como positivos a 267698 instancias que realmente eran positivas.
- **Falsos Positivos: 2116** → El modelo clasificó erróneamente como positivos a 2116 instancias que realmente eran negativas.
- **Verdaderos Negativos: 7942** → El modelo clasificó correctamente como negativos a 7942 instancias que realmente eran negativas.
- **Falsos Negativos: 36817** → El modelo clasificó erróneamente como negativos a 36817 instancias que realmente eran positivas.

Las métricas calculadas con el fin de saber qué tan preciso es el modelo implementado fueran las siguientes:

```
Accuracy: 0.8762354048185954
Precision: 0.9921575603934563
Recall: 0.8790962678357388
F1-score: 0.9322113283501269
```

- Acurracy (exactitud) indica el porcentaje de predicciones correctas del modelo sobre el total de predicciones realizadas, con un valor de 87,6%, indicando que predice de forma correcta la clase en app el 87,6% de las veces.
- Precisión del modelo, el valor de 99,2% indica que cuando el modelo predice la clase positiva, tiene una alta probabilidad de estar en lo correcto.
- Recall (exhaustividad o sensibilidad) hace referencia a que el modelo identifica correctamente alrededor del 87,9% de los casos positivos reales.
- F1-score (puntaje F1) proporciona un balance entre precisión y recall, un puntaje de F1 alto como 0,932 indica un buen rendimiento general del modelo, considerando su capacidad de identificar correctamente casos positivos así como también, su capacidad de evitar falsos positivos.

En el conteo de predicciones positivas y negativas, se obtuvo lo siguiente:

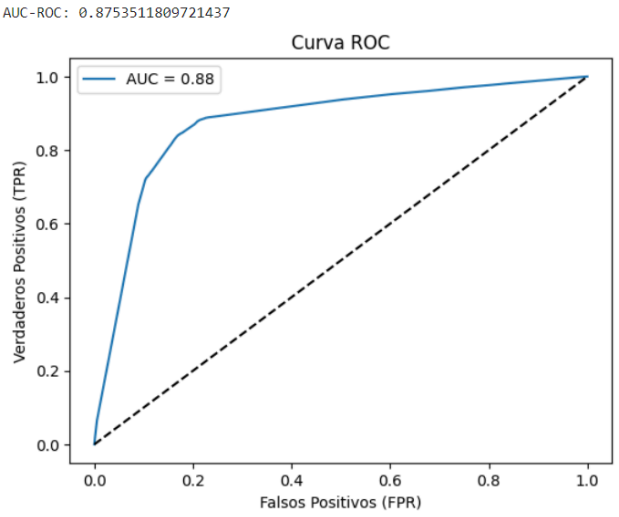
```
Es Correcto
True     304515
False     10058
Name: count, dtype: int64
```

El modelo realizó 304515 predicciones correctas, mientras que las predicciones incorrectas fueron de 10058, de un total de 314573 predicciones realizadas.

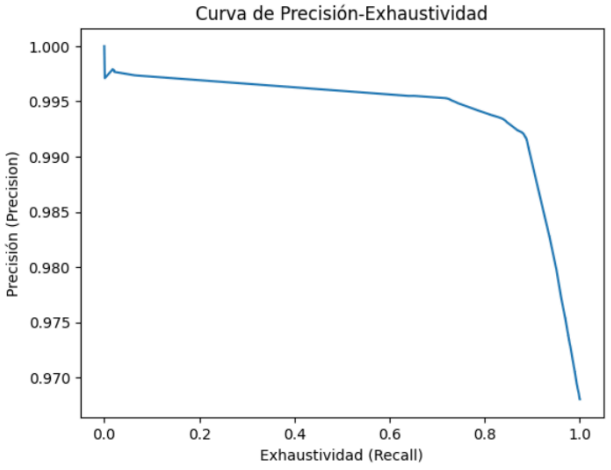
Al igual que en la regresión anterior, una vez entrenado el modelo, se realizó una predicción sobre la posibilidad de requerir INTUBACIÓN, con la presencia de diferentes combinaciones de comorbilidades generadas de forma aleatoria, las cuales buscan simular posibles escenarios clínicos, obteniendo lo siguiente, en donde se presenta la condición médica de 5 pacientes hipotéticos, mostrando las predicciones y probabilidades asociadas.

```
Predictions: [0 0 0 1 0]
Probabilities: [[0.98335653 0.01664347]
 [0.59648934 0.40351066]
 [0.97654879 0.02345121]
 [0.31479846 0.68520154]
 [0.95521098 0.04478902]]
```

En cuanto a la curva ROC, cuya utilidad es la misma ya mencionada en la predicción 1, el valor AUC-ROC, 0.875351… ≈ 0.88 resume el rendimiento general del modelo, en donde valores cercanos a 1 indican un buen modelo, mientras que menor a 0.5 indica un modelo sin poder predictivo. En este caso, significa que en promedio, el modelo es capaz de distinguir correctamente entre las dos clases (positiva, paciente de alto riesgo y negativa en caso contrario) en un 88% de los casos.



Otro elemento implementado en esta regresión logística consistió en la elaboración de la curva de precisión - exhaustividad, centrada en la relación entre esas métricas. En esta curva, Idealmente, se espera que se acerque lo más posible a la esquina superior derecha del gráfico, lo cual significa alta precisión y alta exhaustividad, características que cumple la curva obtenida.



Árbol de decisión y Regresión logística 3 → "PATIENT_TYPE" y Comorbilidades

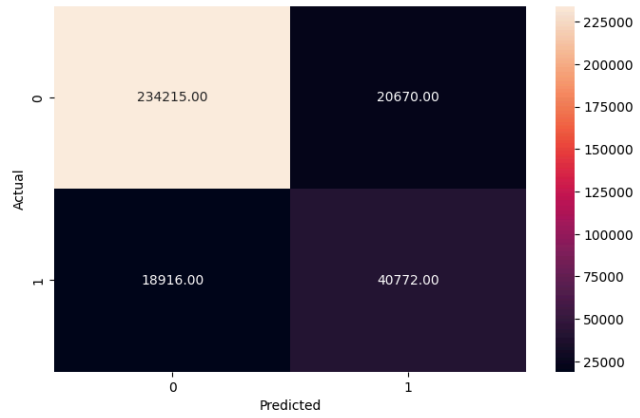
Para llevar a cabo el modelo de árbol de decisión y de regresión logística, se siguió el procedimiento ya descrito. En esta ocasión, al igual que en el modelo anterior, nos centraremos en los valores obtenidos,

Primero, se emplearon los siguientes datos para definir la variable dependiente y la independiente

```
X =mydata[["PNEUMONIA", "DIABETES", "COPD", "ASTHMA", "INMSUPR", "HIPERTENSION", "CARDIOVASCULAR", "OBESITY", "RENAL_CHRONIC", "TOBACCO"]]
y = mydata["PATIENT_TYPE"]
```

A continuación, se evaluó el rendimiento del modelos usando los datos de prueba

	precision	recall	f1-score	support
1	0.93	0.92	0.92	254885
2	0.66	0.68	0.67	59688
accuracy			0.87	314573
macro avg	0.79	0.80	0.80	314573
weighted avg	0.88	0.87	0.87	314573



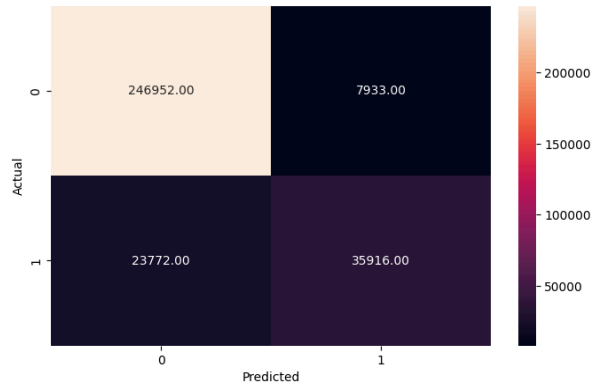
Accuracy: 0.8741595750429948

Al interpretar esta matriz, se puede concluir

- En este caso, a diferencia de los analizados anteriormente, la precisión para ambas clases (0 y 1) es similar y relativamente alta (alrededor del 85%), lo que sugiere que el modelo confía en sus predicciones positivas.
- El recall también es similar para ambas clases (alrededor del 80%), indicando que el modelo está identificando una buena proporción de los casos positivos de cada clase.
- En este caso, la precisión general es del 87%, lo que indica un buen desempeño.
- El hecho de que ambas clases tengan un soporte similar y que el modelo obtenga resultados similares para ambas clases sugiere que el modelo está generalizando bien a ambos grupos.

Asimismo, se analizaron los "metric scores" en el conjunto de datos de entrenamiento.

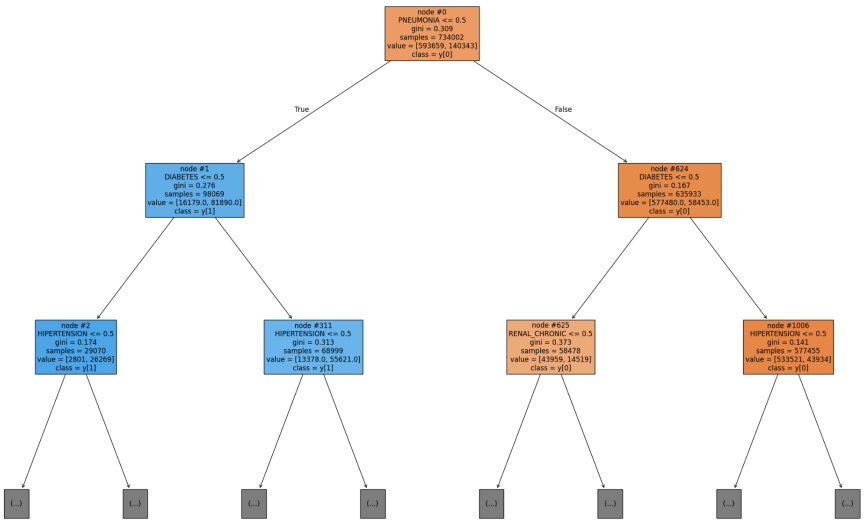
	precision	recall	f1-score	support
1	0.91	0.97	0.94	254885
2	0.82	0.60	0.69	59688
accuracy			0.90	314573
macro avg	0.87	0.79	0.82	314573
weighted avg	0.89	0.90	0.89	314573



Accuracy: 0.8992125834067132

En este caso, en comparación con los modelos anteriores, se observa un equilibrio en los valores 1 y 0, lo que indica que no existe un desbalance o un “sesgo” hacía alguno de los dos valores. Sin embargo, si bien el modelo tiene una alta precisión, el recall para la clase 0 es ligeramente menor que para los valores 1. Esto sugiere que el modelo podría estar teniendo dificultades para identificar algunos casos positivos de la clase 0.

Al visualizar el árbol de decisión, se obtuvo lo siguiente.



Como en el análisis previo, se observa que el árbol de decisión no es adecuado para modelar la relación compleja entre las variables. Los resultados sugieren que un enfoque con regresión logística, junto con un análisis adicional del árbol de decisión, podría proporcionar una

comprensión más clara del fenómeno en estudio. A continuación, se presentan los resultados obtenidos con la regresión logística.

Luego tras la implementación del modelo de regresión logística sobre las variables descritas anteriormente para el árbol de decisión, donde el outcome es "PATIENT_TYPE" , se obtuvieron los resultados que serán expuestos y explicados a continuación.

En primer lugar, se usó el 70% de los datos como entrenamiento y de 30% de prueba, luego se calcularon los coeficientes, los cuales permiten visualizar el grado de incidencia de cada comorbilidad en la atención médica que recibe el paciente contagiado de COVID-19, obteniendo lo siguiente

	Clase 1
PNEUMONIA	-3.806219
DIABETES	-0.931471
COPD	-1.256782
ASTHMA	0.100335
INMSUPR	-1.291894
HIPERTENSION	-0.683086
CARDIOVASCULAR	-0.640443
OBESITY	0.050438
RENAL_CHRONIC	-1.236784
TOBACCO	0.172353

A partir de esto, se evidencia coeficiente negativo en varios atributos, lo cual concluye que patologías como neumonía, diabetes, enfermedad pulmonar obstructiva (COPD), inmunosupresión, enfermedades cardiovasculares y renales crónicas no, disminuyen la probabilidad de que un paciente contagiado del virus en cuestión sea hospitalizado. En caso contrario, comorbilidades como asma, obesidad y fumador (tobacco) aumentan esta probabilidad.

Luego, se aplicaron códigos para poder determinar la precisión del modelo de regresión logística y junto con esto, calcular la matriz de confusión, en donde se obtuvo lo siguiente:

	Valor Real (y_test)	Predicción (y_pred)	Es Correcto
781974	1	1	True
937737	1	1	True
907828	1	1	True
784628	1	1	True
662460	1	1	True
...	...	...	...
230393	1	1	True
444826	2	2	True
275014	1	1	True
100812	1	1	True
425631	1	1	True

100 rows × 3 columns

```
Matriz de Confusión:
[[234215  20670]
 [ 18916  40772]]
```

Pudiendo concluir a partir de esto:

- **Verdaderos Positivos: 40772** → El modelo clasificó correctamente como positivos a 40772 instancias que realmente eran positivas.

- **Falsos Positivos: 20670** → El modelo clasificó erróneamente como positivos a 20670 instancias que realmente eran negativas.
- **Verdaderos Negativos: 234215** → El modelo clasificó correctamente como negativos a 234215 instancias que realmente eran negativas.
- **Falsos Negativos: 18916** → El modelo clasificó erróneamente como negativos a 18916 instancias que realmente eran positivas.

Por otra parte, continuando con la precisión del modelo implementado, las métricas obtenidas de este son las siguientes:

```
Accuracy: 0.8741595750429948
Precision: 0.9252718947896543
Recall: 0.9189046040371148
F1-score: 0.9220772574092154
```

- Acurracy: del total de predicciones realizadas por el modelo, el 87,4% de estas fueron correctas.
- Precisión: del total de predicciones positivas, clasificando a pacientes como positivos (de alto riesgo), el 92,5% fueron realmente positivos.
- Recall: de los casos realmente positivos, el 91,9% fueron identificados correctamente por el modelo.
- F1 - score = 0,922, este valor indica un buen equilibrio entre precisión y exhaustividad del modelo, mientras más cercano a 1 es el valor, mejor.

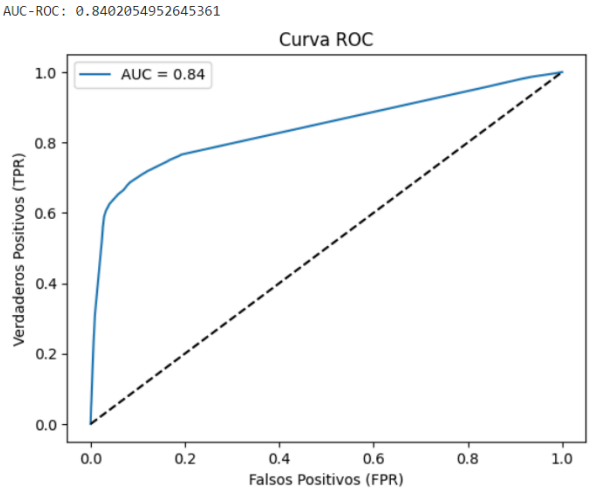
En conclusión, el modelo tiene un buen rendimiento general, con una alta exactitud, precisión y exhaustividad, logrando predecir correctamente la mayoría de los casos, tanto positivos como negativos, teniendo una buena capacidad para evitar falsos positivos. Además, se realizó un conteo de las predicciones positivas y negativas realizadas, en donde los resultados fueron:

```
Es Correcto
True     304515
False     10058
Name: count, dtype: int64
```

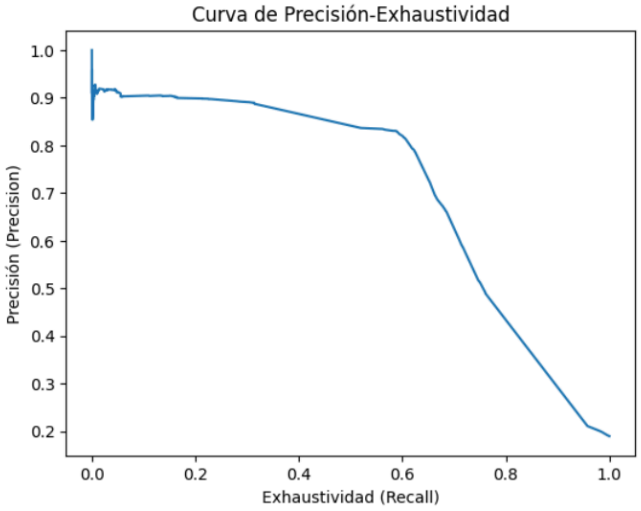
Se realizó una predicción sobre la posibilidad de requerir HOSPITALIZACIÓN tras contraer el virus, con la presencia de diferentes combinaciones de comorbilidades generadas de forma aleatoria, las cuales buscan simular posibles escenarios clínicos, obteniendo lo siguiente, en donde se presenta la condición médica de 5 pacientes hipotéticos, mostrando las predicciones y probabilidades asociadas.

```
Predictions: [2 2 2 2 2]
Probabilities: [[0.21995895 0.78004105]
 [0.00294109 0.99705891]
 [0.06580436 0.93419564]
 [0.00136723 0.99863277]
 [0.02523742 0.97476258]]
```

Al momento de realizar la curva ROC se obtuvo el siguiente gráfico, en donde el valor AUC - ROC = O,84020… ≈ 0,84, indica que el modelo tiene una probabilidad de 84% de poder distinguir entre una instancia positiva elegida al azar y una instancia negativa elegida al azar. La curva ROC como tal, refiere que el modelo está funcionando bien para predecir si el paciente covid es de alto riesgo al padecer alguna comorbilidad, sin embargo, aún existe un margen de mejora, pudiendo mejorar la especificidad del modelo.



Finalmente, en la curva de precisión - exhaustividad se obtuvo el siguiente gráfico, en donde a medida que aumenta la exhaustividad (identificando a más pacientes de alto riesgo que requieren de hospitalización), disminuye la precisión, lo cual se traduce en que al intentar capturar a todos los pacientes que necesitan hospitalización, el modelo también podría clasificar erróneamente a algunos pacientes sanos como necesitados de hospitalización.



Conclusiones y recomendaciones para el sistema en estudio

Los resultados obtenidos tras la implementación de algoritmos de data science, como lo son árboles de decisión y regresión logística, permiten concluir que la presencia de ciertas comorbilidades incrementa de manera significativa el riesgo de que un paciente con COVID-19 desarrolle una condición severa que requiera hospitalización, intubación o ingreso a la UCI. A partir de la aplicación de los diferentes algoritmos, se logró identificar que comorbilidades como neumonía, obesidad, hipertensión y diabetes tienen una gran influencia en este diagnóstico. A continuación, se presenta un resumen del impacto de estas variables en función de cada outcome:

- Ingreso a la UCI por contagio de COVID-19 → Neumonía y obesidad son las patologías que más influyen en la probabilidad de ingreso a UCI, especialmente la primera mencionada, que aumenta de forma considerable la posibilidad de que un cuadro severo de COVID-19 requiera de cuidados intensivos.
- Intubación tras COVID-19 → Se repite la neumonía como patología clave en la predicción de intubación por COVID-19, sin embargo, también influyen factores como hipertensión y diabetes.
- Paciente que requiere hospitalización por COVID-19 → Además de neumonía y obesidad, también influyen variables como asma y antecedentes de tabaquismo, las cuales elevan el riesgo de hospitalización de pacientes portadores del virus SARS - CoV- 2.

Por otra parte, los árboles de decisión, aunque son efectivos en la identificación de patrones, presentaron una alta complejidad y tendencia a sobre ajustarse, lo cual los volvió más limitados en cuanto a interpretación en comparación con la regresión logística, sin embargo, a pesar de esto, refuerzan los hallazgos en cuanto a las comorbilidades más populares, entregando preguntas claves en los nodos superiores del árbol. Los resultados concuerdan con el artículo de Nature, en donde por ejemplo, el cual asocia la hipertensión con un mayor riesgo de complicaciones severas, y sugiere que los medicamentos IECA y BRA, comúnmente usados en su tratamiento, podrían jugar un papel en la evolución grave del COVID-19 (McFarlane et al., 2022).

La neumonía es el predictor más fuerte para la probabilidad de ingreso a la UCI, intubación y hospitalización, lo cual confirma lo expuesto en el estudio “Improving the early identification of COVID-19 pneumonia: a narrative review”, donde se destaca cómo la neumonía agrava la inflamación pulmonar y contribuye a insuficiencia respiratoria severa al promover la formación de microtrombos y empeorar la oxigenación pulmonar (Goyal et al., 2021). También, ambos modelos reflejan que la diabetes es un factor que aumenta el grado de riesgo de COVID-19, particularmente en la intubación y hospitalización. Este hallazgo está alineado con los resultados del estudio “Cell Metabolism”, que observa que la diabetes tipo 2 aumenta la susceptibilidad de los pacientes al COVID-19 grave, y estos representan una proporción considerable de los ingresos en UCI en comparación con pacientes sin esta comorbilidad (Zhu et al., 2020).

Como se puede ver, ambos algoritmos de predicción brindan una herramienta sólida para predecir la necesidad de cuidados intensivos, intubación y hospitalización, contribuyendo a anticipar el manejo de pacientes con comorbilidades y así potenciar el desarrollo de herramientas de soporte clínico, como podría serlo la implementación de un sistema de predicción en tiempo real que asista a los médicos en la toma de decisiones rápidas y precisas sobre el nivel de cuidado que requiere cada paciente, esto dependiendo de su gravedad (de alto riesgo o no), permitiendo optimizar y a la vez priorizar los recursos médicos, siendo asignados de forma eficiente.

Con respecto a las recomendaciones para el sistema en estudio, se proponen las siguientes:

- Optimización del modelo para manejar el desequilibrio de las clases y mejorar el rendimiento del mismo, aplicando otras técnicas de balance adicionales al over-sampling, con el fin de experimentar, como por ejemplo, aplicar la técnica de under - sampling o bien, otros algoritmos específicos para datos con estas características. También, se considera ajustar hiper parámetros de los árboles de decisión y regresión logística, mejorando la interpretabilidad del modelo.
- Integrar herramientas de apoyo clínico, como la ya descrita unos párrafos más arriba. Este sistema puede ser utilizado al momento de anticipar el riesgo de ingreso a la UCI, intubación u hospitalización de un paciente COVID-29, tras ingresar si padece o no de alguna comorbilidad, y cuál es esta en caso de que sí. Esto con el fin de optimizar procesos y recursos médicos.
- Capacitación de personal en la interpretación de predicciones y resultados obtenidos, logrando que estos identifiquen las comorbilidades claves y cómo estos factores afectan el riesgo del paciente.
- Ampliación del estudio a otras patologías, que no solo esté enfocado en la predicción de los riesgos de severidad, si no también, a la relación de comorbilidades y otras enfermedades o virus, lo cual podría ser útil en caso de futuras pandemias o enfermedades de alta complejidad.
- Finalmente, una gran recomendación para el sistema estudiado es un monitoreo y evaluación continua del rendimiento del modelo de predicción en tiempo real, así se puede identificar cualquier pérdida de precisión y poder realizar los ajustes pertinentes, además, puede ser útil recopilar feedbacks de los usuarios del modelo en cuanto a su utilidad en el área de salud y poder realizar correcciones en base a esto.

Se considera que la implementación de estas recomendaciones, o de al menos alguna de ellas, puede mejorar la precisión, confiabilidad y aplicabilidad del modelo.

PARTE III: Reflexiones

1. ¿Creen que es necesario hacer cambios en los roles del equipo? (por ejemplo, en el spokeperson).

Como equipo consideramos que ambas hemos llevado a cabo nuestros roles de la forma correcta. Al ser sólo dos integrantes en este equipo, en comparación a algunos que se conforman de tres estudiantes, consideramos que desde un comienzo lo mejor sería repartir las responsabilidades de forma equitativa y acorde a las capacidades y/o habilidades de cada una. Considerando esto, ocasionalmente hemos compartido o bien intercambiado los diferentes roles, cuidando no pasar a llevar a alguna de nosotras y logrando desempeñar de forma exitosa los roles ya asignados, nos hemos entendido bien como equipo de trabajo y todo lo que esto conlleva. Dicho esto, creemos que no es necesario realizar cambios en los roles de equipo, sin embargo, consideramos sumamente importante ir revisando de forma constante que cada una de nosotras esté cumpliendo con su rol, de esta forma nos aseguramos de que cosas como la comunicación tanto interna como externa esté siendo efectiva, además de ir logrando un óptimo y eficiente desarrollo de nuestro proyecto, llegando de esta forma a cumplir los objetivos esperados.

2. Reflexionen sobre cómo afectó la incertidumbre en un proyecto como este y cómo la enfrentaron, junto con qué decisiones tomaron y qué aprendieron de la experiencia.

Como equipo consideramos que la incertidumbre ha estado presente a lo largo del desarrollo de todo el proyecto y en cada una de las entregas, sin embargo, esta no ha estado presente en igual magnitud siempre. En la primera clase al momento de los profesores dar las instrucciones de este trabajo, con mi compañera sentimos mucha confusión e incertidumbre al respecto, puesto que todo el mundo de Data Science es sumamente nuevo para ambas, por lo que en nuestras cabezas habían mil dudas e inseguridades sobre cómo afrontar este gran desafío. En la medida en que íbamos avanzando con las clases de la asignatura y junto con esto en el desarrollo del proyecto, además de apoyarnos bastante en nuestros profesores, fuimos adquiriendo mayores nociones sobre el objetivo de este trabajo y a la vez, herramientas para poder desarrollarlo, lo cual fue fundamental para poder afrontar la incertidumbre y también, poder disminuirla. En cuanto a las decisiones que tomamos, también estuvo muy presente en cada una de ellas la confusión, en donde muchas veces sentimos que no sabíamos si lo que estábamos implementando/decidiendo estaba bien o era lo mejor para nuestro propósito planteado, creemos que tuvimos más errores que aciertos antes que llegar a los resultados finales de cada entrega o bien, ver verdaderos avances y una mayor claridad en nuestro proyecto de predicción. En cuanto a la experiencia, creemos que definitivamente fue y ha sido muy enriquecedora, a pesar de la incertidumbre y frustración que en muchas ocasiones llegamos a sentir. El comenzar sin saber bien qué hacer ni tener las ideas claras, para luego ir viendo cómo todo comenzaba a tomar forma y adentrarnos más en el mundo de la ciencia de datos ha sido gratificante a más no poder, descubriendo una nueva área de la ingeniería que a ambas nos llama la atención, haciendo de esta experiencia aún más provechosa. Consideramos también que este desafío nos ha servido para poder confiar más en

nosotras mismas, en nuestras habilidades y capacidades, además, ha sido una buena aproximación para lo que será nuestro futuro profesional, en donde quizás no necesariamente nos desarrollemos en esto, pero sí nos ha contribuido en el ámbito de que probablemente tendremos que enfrentarnos a situaciones similares, en donde se nos estén asignando tareas completamente nuevas y fuera de nuestra zona de confort que eventualmente podrían generar la misma (o más) incertidumbre, así como lo ha sido este proyecto semestral.

3. ¿Cómo afectó la incertidumbre para el desarrollo del proyecto la elección de la base de datos? Reflexione al respecto de esto.

En relación a la elección de la base de datos, la incertidumbre también estuvo muy presente. Esto ya que como mencionamos anteriormente, la ciencia de datos para ambas era un mundo nuevo, por lo que desconocíamos todo lo que se podía lograr mediante la utilización y empleo de ella. Es por esta misma razón que, así como se nos hizo confuso y dificultoso el comprender este proyecto, aún más lo fue el poder seleccionar una base de datos y así poder asegurarnos de saber que llevar a cabo nuestro proyecto sería posible. En este proceso de búsqueda y elección nos encontramos con distintas bases de datos, las cuales se trataban de temas muy opuestos entre sí, por ejemplo, encontramos algunas bases de datos sobre vuelos de aerolíneas, y finalmente optamos por escoger aquella que hablaba sobre COVID-19 y comorbilidades asociadas, decisión impulsada por lo interesante que es este tema para ambas además de ser una situación reciente que afectó a todo el mundo, por lo que nos motivaba el saber más. Un segundo desafío fue identificar qué se podía lograr por medio de la base de datos, ya que tuvimos muchas ideas en mente, pero luego descubríamos que quizás el data set no era apropiado para esos fines, por lo que tras varias lecturas de papers, logramos definir el objetivo y propósito final adecuado, junto con una BD consolidada que nos sirviera para esto.

Bibliografía

Epidemiology of COVID-19. (s/f). Lww.com. Recuperado el 8 de noviembre de 2024, de https://journals.lww.com/jrms/fulltext/2021/26000/Epidemiology_of_COVID_19__An_updated_review.82.aspx

Goyal, D., Inada-Kim, M., Mansab, F., Iqbal, A., McKinstry, B., Naasan, A. P., Millar, C., Thomas, S., Bhatti, S., Lasserson, D., & Burke, D. (2021). Improving the early identification of COVID-19 pneumonia: a narrative review. *BMJ Open Respiratory Research*, *8*(1). https://doi.org/10.1136/bmjresp-2021-000911

McFarlane, E., Linschoten, M., Asselbergs, F. W., Lacy, P. S., Jedrzejewski, D., Williams, B., & on Behalf of the CAPACITY-COVID Consortium. (2022). The impact of pre-existing hypertension and its treatment on outcomes in patients admitted to hospital with COVID-19. *Hypertension Research: Official Journal of the Japanese Society of Hypertension*, *45*(5), 834–845. https://doi.org/10.1038/s41440-022-00893-5

Silaghi-Dumitrescu, R., Patrascu, I., Lehene, M., & Bercea, I. (2023). Comorbidities of COVID-19 patients. *Medicina (Kaunas, Lithuania)*, *59*(8), 1393. https://doi.org/10.3390/medicina59081393

The Elements of Statistical Learning Data Mining, Inference, and Prediction. (s/f). Upenn.edu. Recuperado el 8 de noviembre de 2024, de https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

Zhu, L., She, Z.-G., Cheng, X., Qin, J.-J., Zhang, X.-J., Cai, J., Lei, F., Wang, H., Xie, J., Wang, W., Li, H., Zhang, P., Song, X., Chen, X., Xiang, M., Zhang, C., Bai, L., Xiang, D., Chen, M.-M., … Li, H. (2020). Association of blood glucose control and outcomes in patients with COVID-19 and pre-existing type 2 diabetes. *Cell Metabolism*, *31*(6), 1068-1077.e3. https://doi.org/10.1016/j.cmet.2020.04.021

Cheng, S., Zhao, Y., Wang, F., Chen, Y., Kaminga, A. C., & Xu, H. (2021). Comorbidities' potential impacts on severe and non-severe patients with COVID-19. *Medicine*, *100*(12), e24971. https://doi.org/10.1097/md.0000000000024971

Goyal, D., Inada-Kim, M., Mansab, F., Iqbal, A., McKinstry, B., Naasan, A. P., Millar, C., Thomas, S., Bhatti, S., Lasserson, D., & Burke, D. (2021). Improving the early identification of COVID-19 pneumonia: a narrative review. *BMJ Open Respiratory Research*, *8*(1), e000911. https://doi.org/10.1136/bmjresp-2021-000911

Hinson, J. S., Klein, E., Smith, A., Toerper, M., Dungarani, T., Hager, D., Hill, P., Kelen, G., Niforatos, J. D., Stephens, R. S., Strauss, A. T., & Levin, S. (2022). Multisite implementation of a workflow-integrated machine learning system to optimize COVID-19 hospital admission decisions. *Npj Digital Medicine*, *5*(1). https://doi.org/10.1038/s41746-022-00646-1

Honardoost, M., Janani, L., Aghili, R., Emami, Z., & Khamseh, M. E. (2021). The Association between Presence of Comorbidities and COVID-19 Severity: A Systematic Review and Meta-Analysis. *Cerebrovascular Diseases*, *50*(2), 132-140. https://doi.org/10.1159/000513288

McFarlane, E., Linschoten, M., Asselbergs, F. W., Lacy, P. S., Jedrzejewski, D., & Williams, B. (2022). The impact of pre-existing hypertension and its treatment on outcomes in patients admitted to hospital with COVID-19. *Hypertension Research*, *45*(5), 834-845. https://doi.org/10.1038/s41440-022-00893-5

Salazar, M., Barochiner, J., Espeche, W., & Ennis, I. (2020). COVID-19, hipertensión y enfermedad cardiovascular. *Hipertensión y Riesgo Vascular*, *37*(4), 176-180. https://doi.org/10.1016/j.hipert.2020.06.003

Silaghi-Dumitrescu, R., Patrascu, I., Lehene, M., & Bercea, I. (2023a). Comorbidities of COVID-19 Patients. *Medicina*, *59*(8), 1393. https://doi.org/10.3390/medicina59081393

Tenorio-Mucha, J., & Hurtado-Roca, Y. (2020). Revisión sobre obesidad como factor de riesgo para mortalidad por COVID-19. *ACTA MEDICA PERUANA*, *37*(3). https://doi.org/10.35663/amp.2020.373.1197