PRÁCTICA
PROFESIONAL

ISPC
INSTITUTO SUPERIOR
POLITÉCNICO CÓRDOBA

INFORME FINAL SYSTOCARE

2024 | ISPC

AGUILAR GIOVANY
BAROZZI, EUGENIA
BATTAUZ, JULIETA
GURREA, FEDERICO
MOLINA, MARCELO
PAGANO, DANTE JAVIER

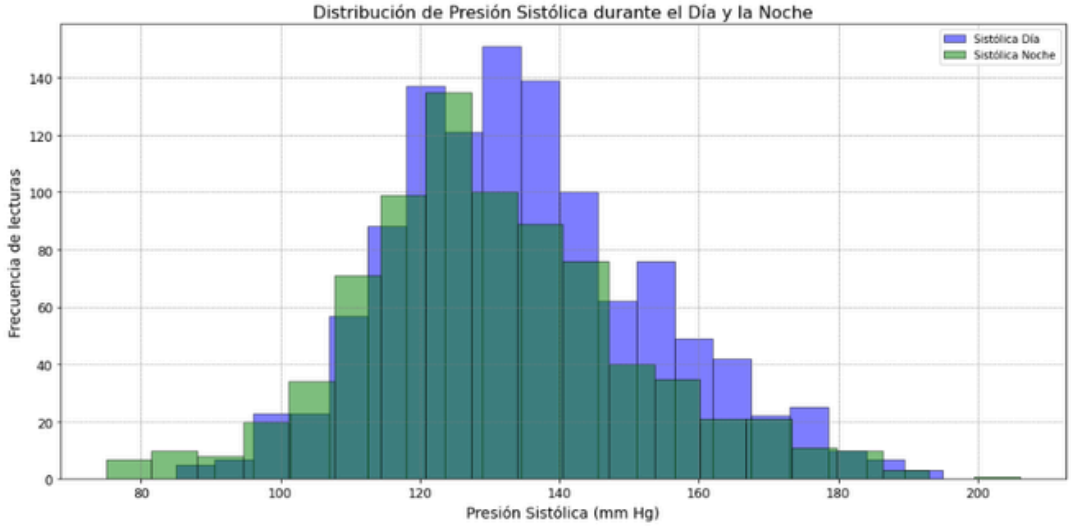
índice

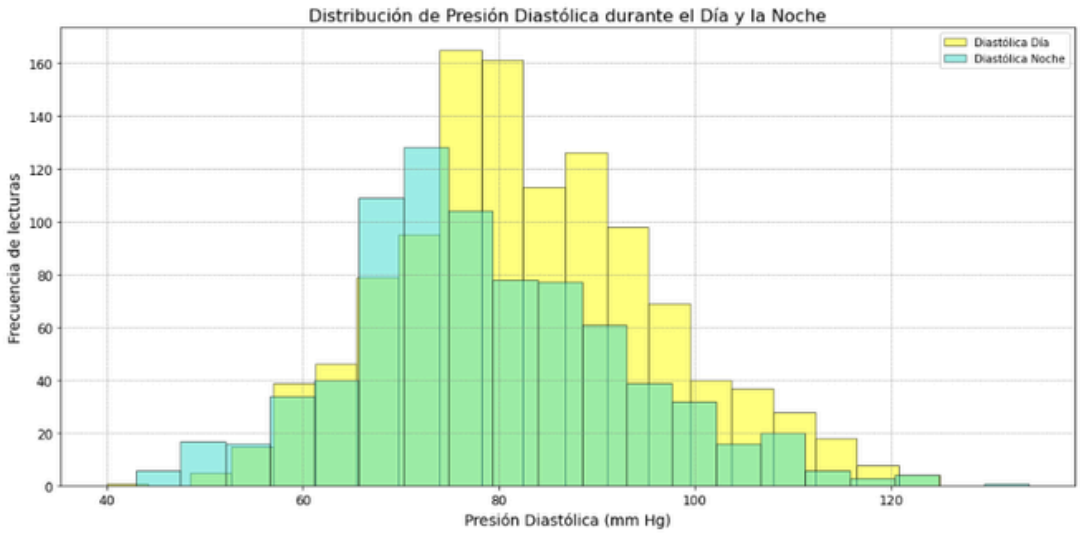
Resumen Ejecutivo:

Propuesta de negocio:

La hipertensión es una condición médica común que puede llevar a graves problemas de salud, como enfermedades cardíacas y accidentes cerebrovasculares. Sin embargo, muchas personas no son conscientes de su condición debido a la falta de monitoreo regular y análisis de sus datos de presión arterial.
Nuestra plataforma utilizará los datos proporcionados de los pacientes y se realizará una limpieza y un análisis exhaustivo. Generaremos gráficos y reportes personalizados que muestran las tendencias, riesgos de padecer hipertensión , patrones y tendencias importantes permitiendo a los usuarios y/o médicos especialistas monitorear continuamente sus presiones sistólicas y diastólicas.

A continuación se presenta un un ejemplo de nuestro análisis en donde podemos visualizar la distribución de la presión sistólica y diastólica durante las horas del día y la noche, en que se mantuvo el examen médico:

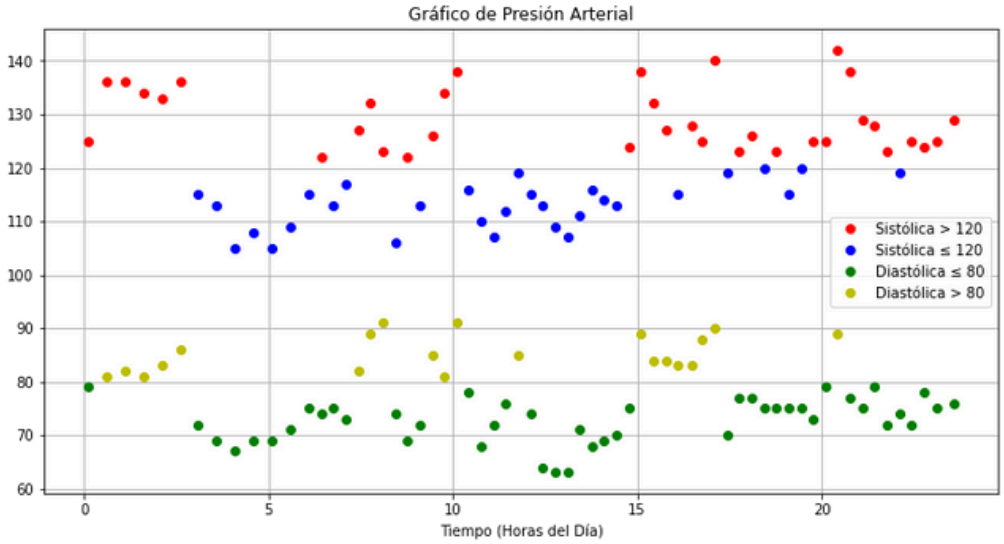




Los beneficios esperados abarcarían estos aspectos:

Objetivo	Características	Beneficio
Monitoreo Personalizado de la Salud	Los usuarios pueden monitorear sus presiones sistólicas y diastólicas de manera continua y personalizada.	Permite a los usuarios tomar decisiones informadas sobre su salud y detectar posibles problemas de hipertensión a tiempo.
Análisis Detallado y Visualización de Datos	La plataforma ofrece gráficos y reportes detallados que muestran las tendencias de presión arterial según la edad y el género (como se observa en el grafico extraído de nuestro análisis).	Facilita la comprensión de los datos de salud, haciendo que la información sea accesible y útil tanto para pacientes como para profesionales de la salud.
Modelos Predictivos de Riesgo:	Utilización de modelos predictivos para anticipar riesgos de hipertensión basados en datos históricos.	Proporciona alertas tempranas y recomendaciones personalizadas, mejorando la prevención y el manejo de la hipertensión.
Mejora Continua de la Salud del Paciente	La plataforma permite un seguimiento continuo y ajustes basados en los datos más recientes.	Ayuda a los pacientes a mantener un control constante sobre su salud, promoviendo hábitos saludables y reduciendo riesgos a largo plazo.
Valor Añadido para Profesionales de la Salud:	Los profesionales de la salud pueden acceder a análisis detallados y visualizaciones claras de los datos de sus pacientes.	Mejora la calidad de la atención médica al proporcionar información precisa y actualizada, facilitando diagnósticos y tratamientos más efectivos.
Diferenciación en el Mercado:	DataVista Analytics se posiciona como líder en soluciones de análisis de datos personalizados para la salud.	Aumenta la competitividad de la empresa al ofrecer una solución innovadora y de alto valor añadido en el mercado de la salud.

Otro ejemplo de la distribución de la presión arterial que se extrajo de los datos para su visualizacion se refleja en este gráfico que nos permite observar cómo varía la presión arterial sistólica y diastólica a lo largo de las 24 horas del día por paciente. Las líneas de presión proporcionan una visión clara de los cambios a lo largo del tiempo, mientras que las líneas horizontales punteadas sirven como referencia para los límites recomendados de presión arterial. de cada paciente



En conclusión de la propuesta

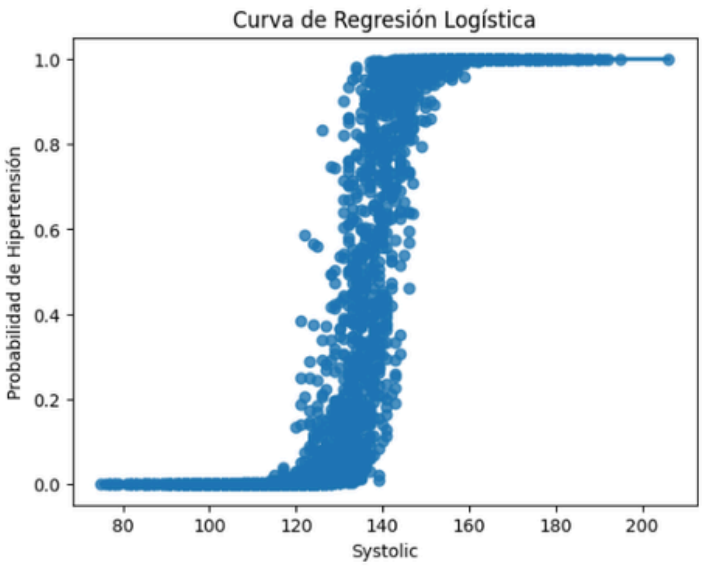
La implementación de esta idea no solo mejorará la calidad de vida de los usuarios al proporcionarles herramientas para monitorear y gestionar su salud, sino que también posicionará a DataVista Analytics como una empresa pionera en el uso de la ciencia de datos para soluciones de salud personalizadas.
Los objetivos propuestos son el proporcionar a los usuarios una herramienta eficaz para monitorear y gestionar su presión mejorando así la salud del paciente. esto reducirá los casos de hipertensión no diagnosticada y mejorara la calidad de sus vidas.

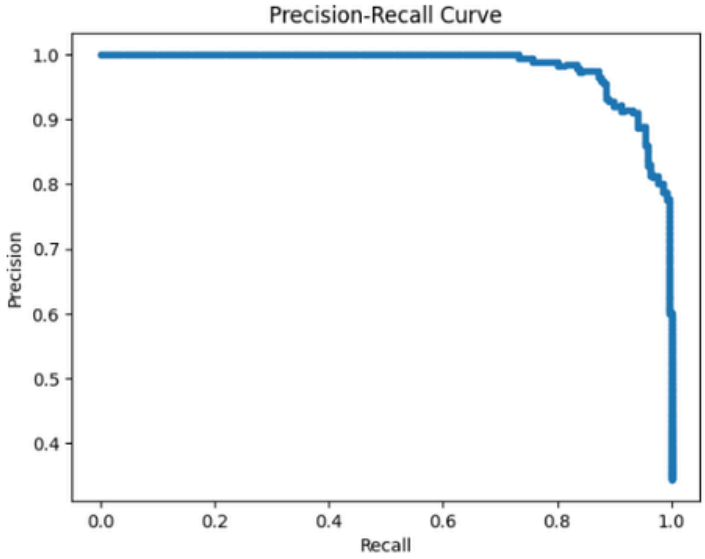
Modelo predictivo y la solución implementada

El presente informe tiene como objetivo documentar el proceso de selección y evaluación de un modelo de Machine Learning 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, la elección fue el modelo de regresión logística

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 poder comprobar si padecen hipertensión.

Entre los gráficos obtenidos de el análisis de regresión Logística podemos destacar la curva sigmode que nos permite visualizar como cambia la probabilidad de padecer hipertensión según la característica fisiológica. En este caso se uso el ejemplo de la presión sistólica y a la izquierda la Curva de Precisión-Recall que 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.





Resultados obtenidos

Resumen Ejecutivo:
El modelo de regresión logística desarrollado ha demostrado una alta precisión en la predicción de hipertensión. Este modelo es capaz de distinguir eficazmente entre pacientes hipertensos y no hipertensos, manteniendo una baja incidencia de falsos positivos y falsos negativos.
Métricas Clave de Desempeño:

- Precisión: 96.6%
- F1-Score: 0.92
- AUC: 0.995

Estas métricas subrayan la robusta capacidad del modelo para identificar correctamente tanto casos positivos como negativos. La alta precisión del 96.6% indica que el modelo es confiable en la clasificación de pacientes hipertensos. El F1-Score de 0.92 refleja un equilibrio adecuado entre precisión y sensibilidad, mientras que el AUC de 0.995 evidencia una excelente capacidad para discriminar entre pacientes hipertensos y no hipertensos.

Es importante destacar que, aunque el modelo muestra un excelente desempeño, es esencial monitorear continuamente los falsos positivos y falsos negativos. Esto garantizará que el modelo mantenga un equilibrio adecuado y sea confiable en aplicaciones clínicas, minimizando el riesgo de diagnósticos incorrectos.
En conclusión, el modelo de regresión logística no solo es altamente preciso, sino que también es una herramienta valiosa para la identificación temprana y precisa de la hipertensión en pacientes. La implementación de este modelo en entornos clínicos puede mejorar significativamente la toma de decisiones médicas y la gestión de la salud de los pacientes

Informe Técnico:

Metodología empleada

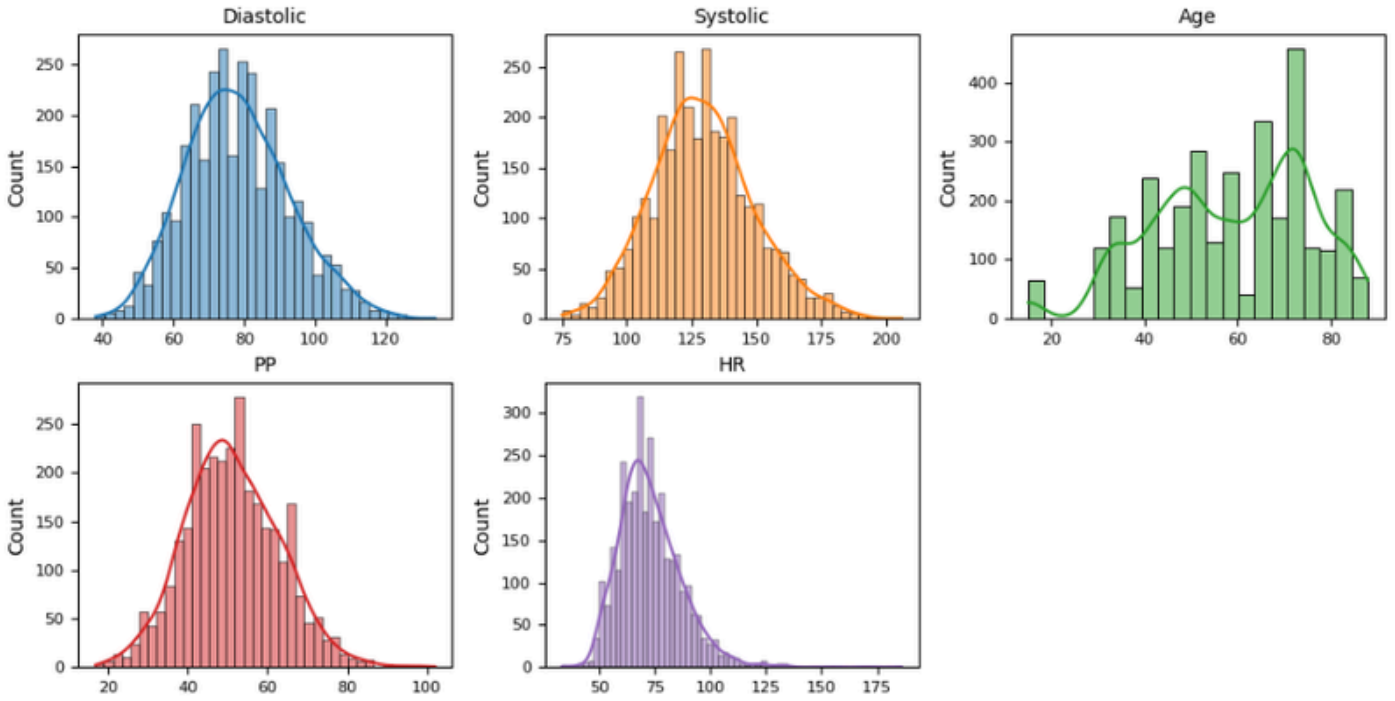
Se propone utilizar la metodología de trabajo de Scrum, en la cual nos organizamos fijando objetivos de acuerdo a los plazos de entrega de los sprints, realizando un seguimiento de los avances de cada tarea de forma semanal mediante reuniones virtuales a través de Google meet (weekly). Los daily se realizarán a través de mensajería instantánea (grupo de Whatsapp) cuando sea necesario. El proyecto quedará descripto y estructurado como Project de GitHub, el cual tiene un modelo Kanban, donde se irán describiendo y asignando las tareas, a demás de conectar cada avance con la documentación del repositorio del proyecto. En el mismo repositorio se crearán determinados hitos (o milestones en GitHub) a los que estarán fijadas ciertas tareas del proyecto. Mediante estas herramientas todos los miembros del proyecto podrán colaborar en el avance del mismo y estar al tanto de las actualizaciones que realiza el resto del grupo.

Análisis realizados

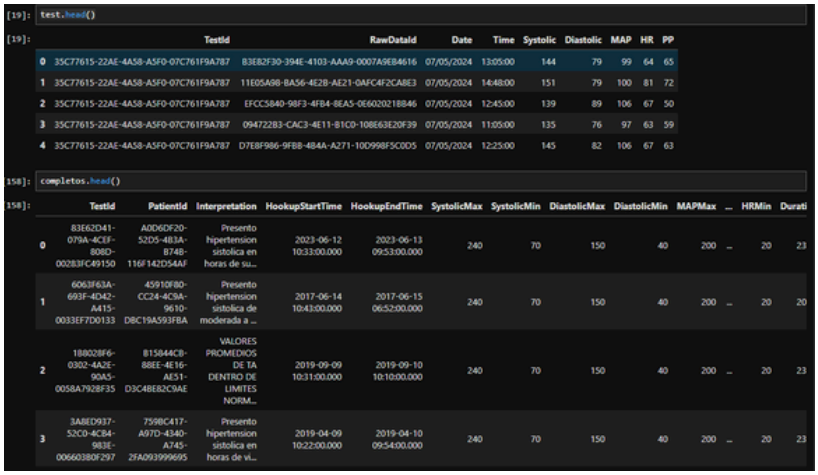
Introducción

a. **Contexto del Dataset:** El conjunto de datos utilizado en este análisis corresponde a mediciones clínica de pacientes, que incluye información sobre presiones arteriales, edad, género, y otros factores de salud. Estos datos fueron recolectados como parte de un estudio longitudinal con el objetivo de identificar factores de riesgo en la hipertensión
b. **Características del Dataset** : "El dataset contiene 4854 registros de pacientes, con un total de 32 variables. Estas variables incluyen mediciones numéricas como presión sistólica, presión diastólica y edad, además de variables categóricas como género y condición médica (por ejemplo, presencia de hipertensión)
c. **Objetivo del Análisis**: "El objetivo de este análisis exploratorio es comprender la distribución de las presiones arteriales en la población estudiada, identificar posibles outliers que puedan indicar errores en la recolección de datos o casos extremos, y analizar la relación entre las variables clínicas.

A continuación se presenta un gráfico que muestra la distribución de variables numéricas permitiendo una comprensión visual rápida y efectiva de los datos de las presiones diastólicas, sustólicas, presión por pulso, la frecuencia cardíaca y las edades, obtenidos de los pacientes



Una vez cargado el dataset, se revisó su estructura utilizando el método .**info()** de pandas. El dataset cuenta con 3.148 registros y 9 columnas. A continuación, se muestran los nombres de las columnas y los tipos de datos asociados a cada una. A su ves se presenta una vista preliminar de los primeros 5 registros del conjunto de datos utilizando el método **.head()**. Se observan las primeras mediciones de presión arterial y las características asociadas.



```
[19]: test.head()
```

[19]:

	TestId	RawDataId	Date	Time	Systolic	Diastolic	MAP	HR	PP
0	35C77615-22AE-4A58-A5F0-07C761F9A787	83E82F30-394E-4103-AAA9-0007A9EB4616	07/05/2024	13:05:00	144	79	99	64	65
1	35C77615-22AE-4A58-A5F0-07C761F9A787	11E05A98-BA56-4E2B-AE21-0AFC4F2CA8E3	07/05/2024	14:48:00	151	79	100	81	72
2	35C77615-22AE-4A58-A5F0-07C761F9A787	EFCC5840-98F3-4FB4-8EA5-0E6020218B46	07/05/2024	12:45:00	139	89	106	67	50
3	35C77615-22AE-4A58-A5F0-07C761F9A787	094722B3-CAC3-4E11-81C0-108E63E20F39	07/05/2024	11:05:00	135	76	97	63	59
4	35C77615-22AE-4A58-A5F0-07C761F9A787	D7E8F986-9FB8-484A-A271-10D998F5C0D5	07/05/2024	12:25:00	145	82	106	67	63

```
[158]: completos.head()
```

[158]:

	TestId	PatientId	Interpretation	HookupStartTime	HookupEndTime	SystolicMax	SystolicMin	DiastolicMax	DiastolicMin	MAPMax	...	HRMin	Durati
0	83E62D41-079A-4CEF-808D-002B3FC49150	A0D6DF20-52D5-4B3A-B748-116F142D54AF	Presento hipertension sistolica en horas de su...	2023-06-12 10:33:00.000	2023-06-13 09:53:00.000	240	70	150	40	200	...	20	23
1	6063F63A-693F-4D42-A415-0033EF7D0133	45910F80-CC24-4C9A-9610-D8C19A593FBA	Presento hipertension sistolica de moderada a ...	2017-06-14 10:43:00.000	2017-06-15 06:52:00.000	240	70	150	40	200	...	20	20
2	188028F6-0302-4A2E-90A5-0058A7928F35	815844CB-88EE-4E16-AE51-D3C4BE82C9AE	VALORES PROMEDIOS DE TA DENTRO DE LIMITES NORM...	2019-09-09 10:31:00.000	2019-09-10 10:10:00.000	240	70	150	40	200	...	20	23
3	3A8ED937-52C0-4CB4-983E-00660380F297	759BC417-A97D-4340-A745-2FA093999695	Presento hipertension sistolica en horas de vi...	2019-04-09 10:22:00.000	2019-04-10 09:54:00.000	240	70	150	40	200	...	20	23

Se realizo la unión de los conjuntos de datos “test y completos” mediante la columna "TestId", que está presente en ambos datasets, con 4854 registros y 32 campos . El método de unión utilizado es "right join", por lo que se mantienen todas las filas del dataset completos,

Limpieza de datos :

Se lleva a cabo la limpieza de datos en archivos CSV para facilitar el análisis, asegurando la integridad y precisión de la información. Este informe detalla el proceso de limpieza realizado en los archivos de datos obtenidos, como “**Completos test.csv** y **Test.csv**”. se ha decidido realizar una revisión adicional utilizando bibliotecas de Python para garantizar la integridad y precisión de los datos.

Se mejoro la calidad del conjunto de datos al abordar los siguientes problemas, facilitando decisiones basadas en información confiable:

```
[252]: # filtrando valores para tener una dataframe limpio
       # Ver duplicados
       duplicados = df_limpio[df_limpio.duplicated()]

       # Eliminar columnas irrelevantes
       df_limpio = df_limpio.drop_duplicates()

       # Corregir valores fuera de rango
       df_limpio = df_limpio[(df_limpio['Age'] >= 0) & (df_limpio['Age'] <= 120)]

       # Otro ejemplo: presión arterial sistólica (90-180)
       df_limpio = df_limpio[(df_limpio['Systolic'] >= 90) & (df_limpio['Systolic'] <= 180)]

       # Eliminar columnas innecesarias
       df_limpio = df_limpio.drop(['RawDataId', 'SysDipping', 'DiaDipping','MapDipping' ], axis=1)

       # Normalizar los datos
       df_limpio = df_limpio.apply(lambda x: x.astype(int) if x.dtype == 'float' else x)

       # Crear una columna categórica de grupos de edad
       df_limpio['grupo_edad'] = pd.cut(df_limpio['Age'], bins=[0, 18, 35, 50, 65, 100],
                                        labels=['Infantil', 'Joven', 'Adulto joven', 'Adulto', 'Senior'])

       # Eliminar filas con cualquier valor nulo en cualquier columna
       df_limpio = df_limpio.dropna()
       # esportamos el archivo a un csv.
       df_limpio.to_csv("df_limpio.csv", index=False,sep=';')
       # verificamos cantida de Registros y campos
       df_limpio.shape
```

A continuación presentamos una tabla que contiene **valores estándar** de presiones sistólica y diastólica por edad y sexo, usados como referencia para detectar **variaciones** en los datos que estamos analizando. Con estos valores normales, puedes comparar las presiones de los pacientes

VALORES NORMALES DE LA TENSIÓN ARTERIAL SEGÚN LA EDAD

EDAD	PRESIÓN SISTÓLICA		PRESIÓN DISTÓLICA	
	HOMBRE	MUJER	HOMBRE	MUJER
16 a 18	105 - 135	100 - 130	60 - 86	60 - 85
19 a 24	105 - 139	100 - 130	62 - 88	60 - 85
25 a 29	108 - 139	102 - 135	65 - 89	60 - 86
30 a 39	110 - 145	105 - 139	68 - 92	65 - 89
40 a 49	110 - 150	105 - 150	70 - 96	65 - 96
50 a 59	115 - 155	110 - 155	70 - 98	70 - 98
60 o más	115 - 160	115 - 160	70 - 100	70 - 100

- El Análisis Univariado

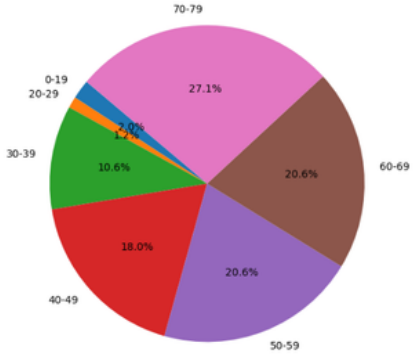
se analizo cada variable por separado para entender su distribución, características y posibles anomalías.
EL método describe proporciona un resumen estadístico de las columnas numéricas de los datos obtenidos. Este resumen incluye varias estadísticas descriptivas univariadas, tales como:**Count**: El número de valores no nulos.

- **Mean**: La media aritmética de los valores.
- **Std**: La desviación estándar, que mide la dispersión de los valores respecto a la media.
- **Min**: El valor mínimo.
- **25%**: El primer cuartil, que es el valor por debajo del cual se encuentra el 25% de los datos.
- **50%:** La mediana o segundo cuartil, que es el valor central.
- **75%**: El tercer cuartil, que es el valor por debajo del cual se encuentra el 75% de los datos.
- **Max**: El valor máximo.

```
df_limpio.describe()
```

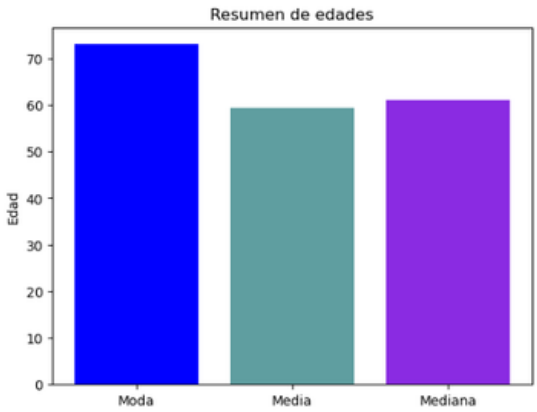
	Systolic	Diastolic	MAP	HR	PP	SuccessfullReading	Age
count	3148.000000	3148.000000	3148.000000	3148.000000	3148.000000	4845.000000	4845.000000
mean	129.380241	78.152795	95.597205	72.770330	51.227446	58.155624	59.297007
std	19.632757	14.632187	15.551684	14.130255	11.899259	7.861107	16.777024
min	75.000000	38.000000	54.000000	33.000000	17.000000	1.000000	9.000000
25%	116.000000	68.000000	85.000000	63.000000	43.000000	54.000000	47.000000
50%	128.000000	77.000000	95.000000	71.000000	50.000000	60.000000	61.000000
75%	141.000000	88.000000	105.250000	81.000000	59.000000	64.000000	73.000000
max	206.000000	134.000000	155.000000	186.000000	102.000000	74.000000	123.000000

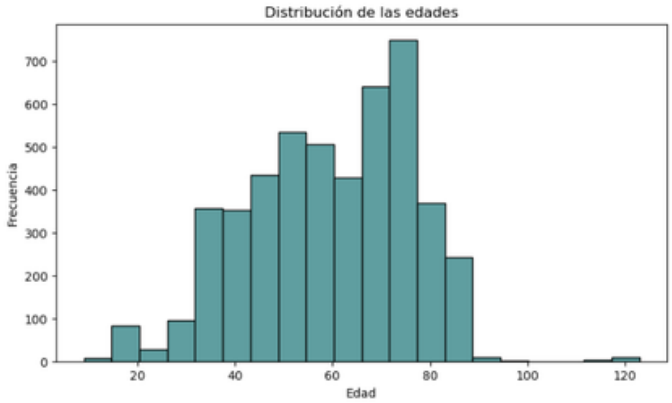
visualizamos la distribución porcentual de los rangos etarios de los pacientes.



En Este Gráfico Utilizando **matploit** y buscamos mostrar la **Moda**, **Media** y **Mediana** de los datos de E**dades**.

- La moda es el valor que aparece con mayor frecuencia en un conjunto de estos datos.
- La media muestra la suma de todos los valores en un conjunto de datos edades dividida por el número de valores en ese conjunto.
- La mediana es el valor medio de un conjunto de estos datos. EL proximo Gráfico muestra lo mismo pero en un gráfico de Histograma.



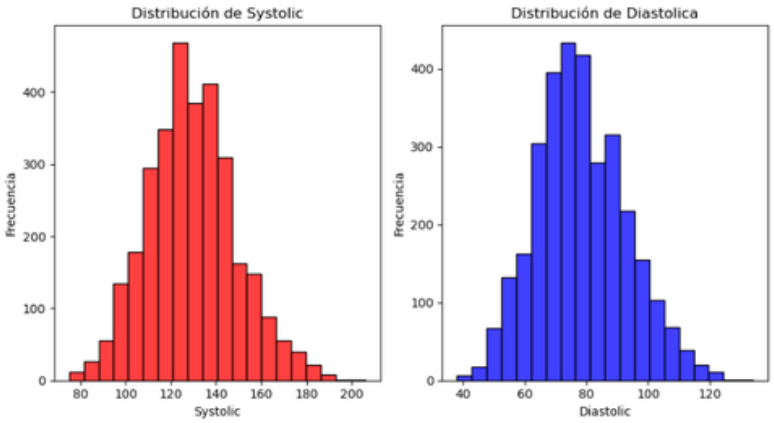


Histograma: Distribución de las variables Systolic y Diastolic

En un histograma de frecuencia, Este gráfico presenta dos histogramas en un diseño de subplots, proporcionando una visualización clara de la distribución de las variables Systolic (presión sistólica) y Diastolic (presión diastólica) en tus datos.

- Histograma de Systolic (izquierda): Muestra la frecuencia de los valores de presión sistólica agrupados en intervalos definidos. Cada barra representa la cantidad de datos que caen dentro de un rango específico de presión sistólica. Este histograma te permite observar cómo se distribuyen los valores de presión sistólica, identificar posibles picos (rangos de valores con alta frecuencia) y asymetrías en la distribución.

- Histograma de Diastolic (derecha): Similar al histograma de presión sistólica, este gráfico ilustra la frecuencia de los valores de presión diastólica en intervalos de rango. Ayuda a visualizar la distribución de los valores diastólicos, permitiéndote ver la forma general de la distribución, incluyendo cualquier sesgo o anomalías.



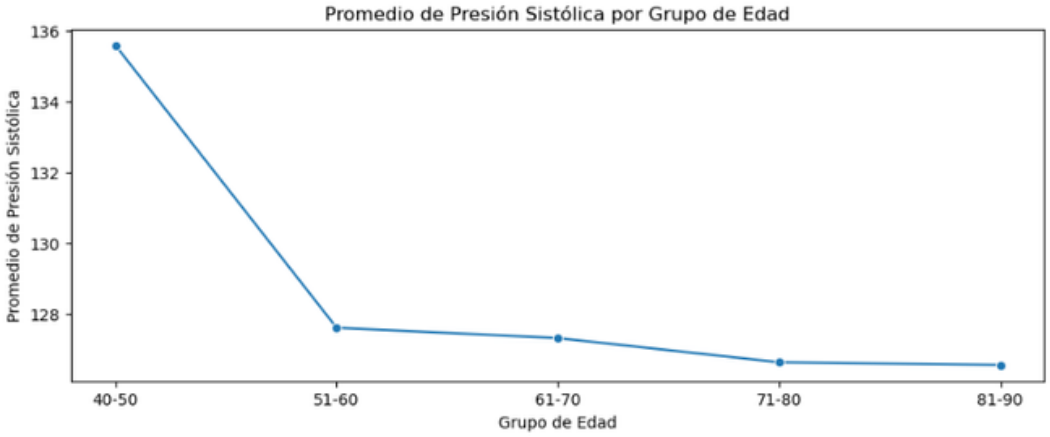
- **Análisis Bivariado**

Es una técnica estadística que se utiliza para investigar la relación entre dos variables. A diferencia del análisis univariado, que se centra en una sola variable, el análisis bivariado examina cómo dos variables interactúan entre sí.

Gráfico de líneas

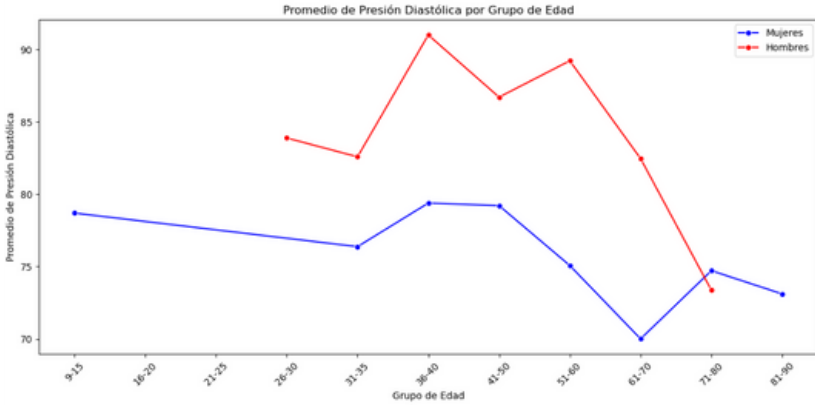
Examina la relación entre dos variables numéricas: la edad y la presión sistólica. De manera general, en este caso, se agrupan las edades en rangos y se calcula la media de la presión sistólica en cada grupo de edad, lo que permite visualizar cómo cambia la presión sistólica en función de la edad.

La línea conecta los puntos que representan el promedio de la presión sistólica en cada grupo de edad, lo permite observar cómo la presión sistólica varía a medida que las personas envejecen. Tendencia: Si la línea tiene una pendiente ascendente, indica que a medida que la edad aumenta, el promedio de la presión sistólica también aumenta. Si es descendente, significaría que la presión sistólica disminuye con la edad. Relación: Este gráfico muestra de manera visual la relación entre la variable edad (agrupada) y la variable presión sistólica, destacando las tendencias promedias.



Este gráfico lineal muestra el promedio de presión **Sistólicas y Diastólicas** por grupo de edad para hombres y mujeres. La línea azul indica el promedio de presión para mujeres, y la línea roja para hombres.

Este gráfico facilita la comparación de cómo varía la presión Arterial promedio con la edad entre los dos sexos, mostrando claramente cualquier tendencia o diferencia significativa en la presión a través de los grupos de edad. Ambos gráficos proporcionan una visión clara de cómo las presiones sistólica y diastólica cambian con la edad y permiten realizar comparaciones entre hombres y mujeres.

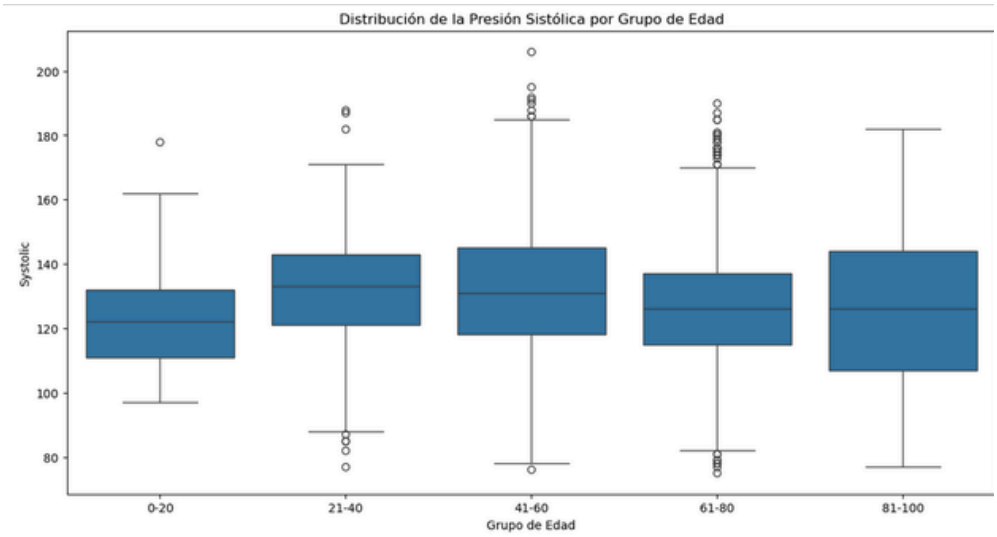


Boxplot

Muestra la distribución, como varía la presión sistólica en diferentes grupos de edad. Los datos se agrupan en rangos de edad (0-20, 21-40, 41-60, etc.), y para cada grupo.

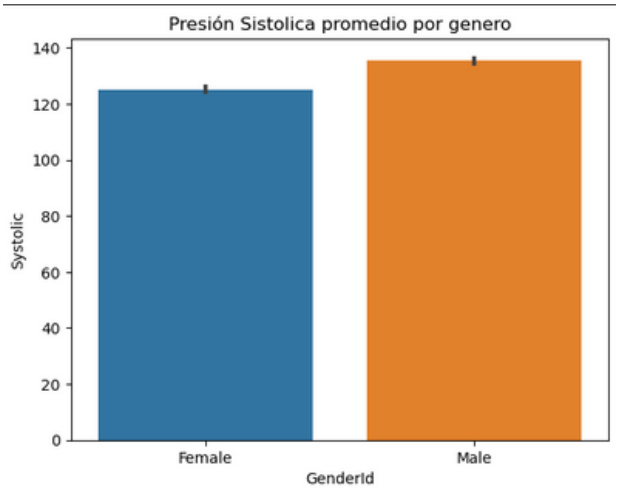
La mediana (línea dentro de la caja), El rango intercuartílico (Q1 a Q3, la caja), Valores atípicos (puntos fuera de los "bigotes"). Los "bigotes" representan la dispersión de los datos sin considerar los outliers.

Puedes ver si la presión tiende a aumentar en grupos de mayor edad, o si hay mucha variabilidad dentro de un grupo.



Presión Sistólica y Diastólica Promedio por Género

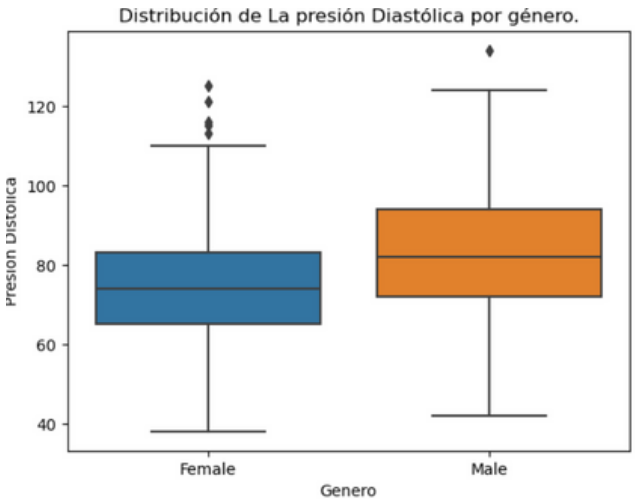
El gráfico te permite visualizar de forma rápida si existe alguna diferencia en los promedios de la presión sistólica entre los géneros. Si una barra es visiblemente más alta que la otra, indicará que uno de los géneros tiene una presión sistólica promedio mayor que el otro. Este gráfico es útil para comparar de manera sencilla los promedios de presión sistólica entre hombres y mujeres, y puede revelar posibles diferencias significativas entre ambos grupos.



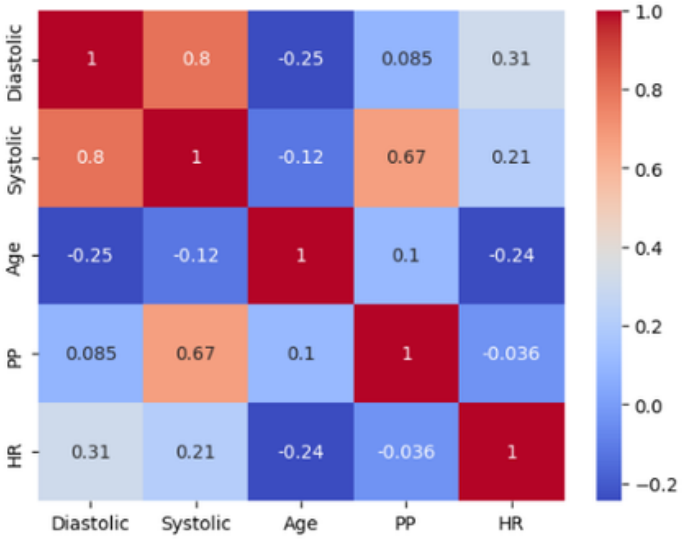
Boxplots: Comparación de Género y Presión Diastólica

Este gráfico utiliza **boxplots** para comparar la distribución de la presión diastólica entre diferentes categorías de género. Cada boxplot representa la distribución de los valores de presión diastólica para cada categoría de género, proporcionando una visión clara de cómo se comparan estos valores entre los grupos.

- **Mediana**: La línea dentro de la caja representa la mediana de la diastólica por cada género.
- **Cuartiles**: Las cajas representan el rango intercuartílico, que contiene el 50% de los datos.El límite inferior de la caja es el primer cuartil (Q1) y el límite superior es el tercer cuartil (Q3). La altura de la caja muestra la variabilidad de la presión diastólica dentro de cada categoría de género.
- **Valores atípicos**: los puntos fuera pueden indicar outliers, o puede indicar una distribución más dispersa.

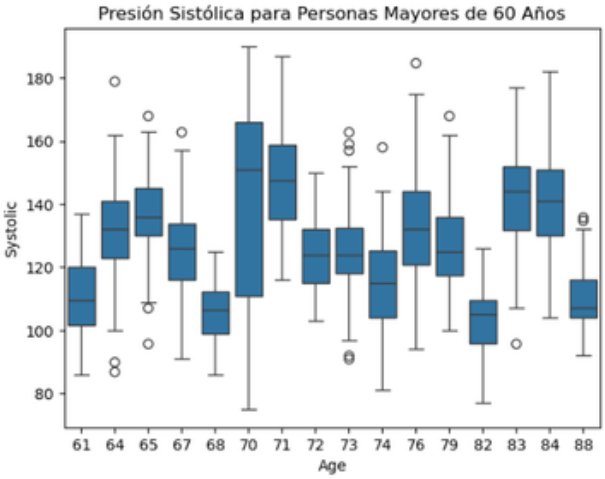


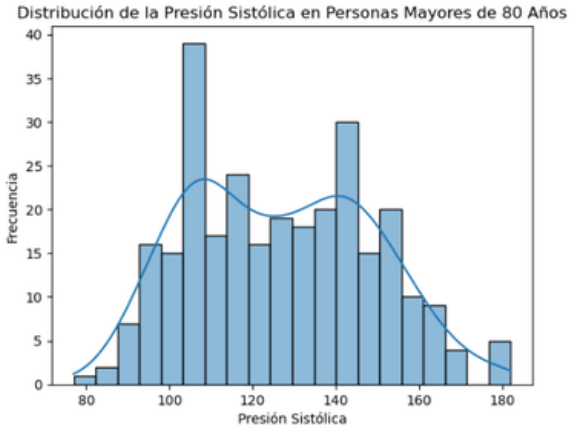
Calculando Matriz de correlación:
muestra cómo las variables se relacionan entre sí mediante coeficientes de correlación, que varían entre -1 y 1. Un valor cercano a 1 indica una **correlación positiva fuerte**, mientras que un valor cercano a -1 indica una **correlación negativa fuerte**. Un valor cercano a 0 indica una **correlación débil** o inexistente.



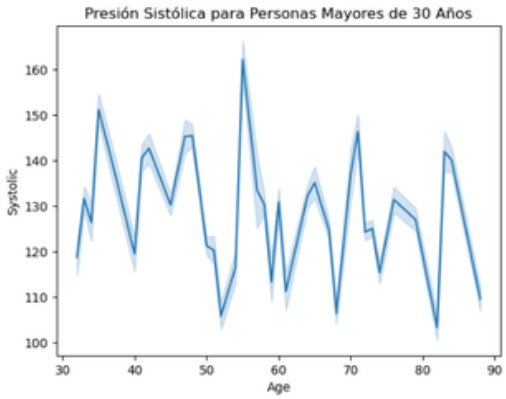
Los Gráficos que vemos a continuación son más Análisis de edades y presiones sistólicas. En donde se está examinando la relación entre dos variables: la edad (Age mayor a 30, 60 y 80 años) y la presión sistólica (Systolic). Este análisis nos ayuda a entender cómo una variable puede influir o estar relacionada con otra. Serán de utilidad en el futuro a medida que recopilemos más datos, estos gráficos se volverán más específicos y precisos, especialmente cuando se analicen según diferentes grupos de edad. Esto nos permitirá identificar patrones y tendencias más detalladas, mejorando así nuestra capacidad para realizar diagnósticos y tratamientos personalizados

Usamos gráficos de Boxplot, Histplot y Lineplot



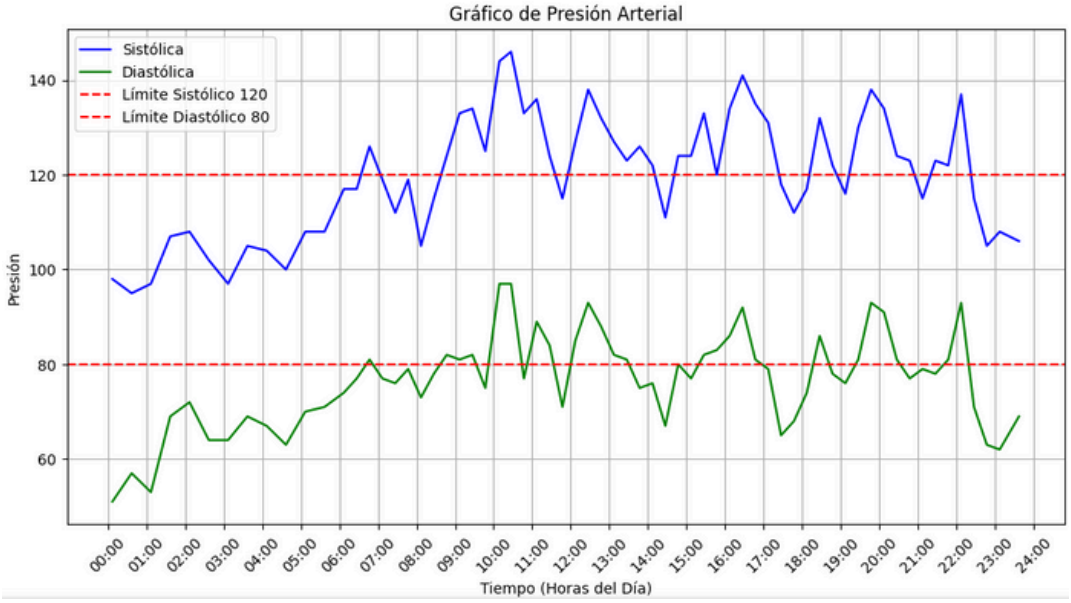


El primer Gráfico muestra las presiones sistólicas de personas mayores a 60 años usando boxplot, El segundo gráfico muestra la distribución de la presión sistólica para personas mayores de 80 años. El histograma, junto con la curva de densidad, ilustra cómo se distribuyen los valores de presión sistólica en este grupo de edad, permitiendo identificar patrones y concentraciones de datos en distintos rangos de presión.



El gráfico de líneas muestra la variación de la presión sistólica en función de la edad para las personas mayores de 30 años. La línea permite observar tendencias y patrones en cómo cambia la presión sistólica a medida que aumenta la edad en el grupo filtrado.

El siguiente gráfico nos permite observar cómo varía la presión arterial sistólica y diastólica a lo largo de las 24 horas del día por paciente. Las líneas de presión proporcionan una visión clara de los cambios a lo largo del tiempo, mientras que las líneas horizontales punteadas sirven como referencia para los límites recomendados de presión arterial. de cada paciente



Desarrollo del modelo

Modelo de Regresión Logística
El script implementa un modelo de regresión logística para predecir la probabilidad de hipertensión en pacientes, usando datos como la presión sistólica, la frecuencia cardíaca, el pulso y la presión diastólica. También incluye técnicas para imputar valores faltantes y evaluar el modelo a través de métricas estándar y visualización.

1. **Carga de Librerías** Se importan las librerías necesarias para el proceso, incluyendo pandas para la manipulación de datos, LogisticRegression de sklearn para el modelo, SimpleImputer para manejar valores faltantes, y statsmodels para un análisis adicional del modelo.
2. **Creación de la Variable Objetivo** "hipertensión: definimos hipertensión, la variable objetivo, que es binaria (0 o 1). Se clasifica como hipertensión (hipertensión = 1) si la presión sistólica es igual o superior a 140 mmHg o la presión diastólica es igual o superior a 90 mmHg.
3. **Definición de Variables Independientes y Dependiente.** Las variables independientes (X) seleccionadas incluyen los factores que podrían predecir hipertensión (presión sistólica, frecuencia cardíaca, presión de pulso y presión diastólica). La variable dependiente (y) es hipertensión
4. **Imputación de Valores Faltantes.** Para manejar valores faltantes en las variables independientes, usamos SimpleImputer con la estrategia mean, que rellena los valores NaN con el promedio de la columna
5. **Estandarización**: Se normalizan las características para que tengan una media de 0 y una desviación estándar de 1, lo que ayuda a que el modelo funcione mejor, especialmente con algoritmos como la regresión logística.
6. **División de Datos en Conjuntos de Entrenamiento y Prueba.** Se dividen los datos en entrenamiento y prueba (80% y 20%, respectivamente) para evaluar el modelo en datos no vistos.

7. **Ajuste del Modelo usando GridSearchCV:** se utiliza la regresión logística y se ajustan los parámetros del modelo utilizando GridSearchCV, que busca los mejores hiperparámetros (como el valor de C y el solver) mediante validación cruzada. Se obtienen los mejores parámetros del modelo a partir de la búsqueda.
8. **Predicción**: Se realiza una predicción sobre el conjunto de prueba (X_test) utilizando el mejor modelo encontrado. Este es el primer conjunto de predicciones, donde se clasifica a cada individuo como hipertenso o no.
9. **Métricas de Evaluación**: Se evalúa la precisión del modelo utilizando varias métricas: **Accuracy**: Proporción de predicciones correctas, **Confusion Matrix**: Matriz que muestra el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, **Classification Report**: Informe que incluye precisión, recall y f1-score para cada clase.
10. **Métricas Adicionales**: Se calculan otras métricas como el F1 Score, Recall y Precision. **ROC AUC Score**: Se evalúa el área bajo la curva ROC, que es una medida de la capacidad del modelo para distinguir entre las clases.
11. **Resumen del Modelo (result. summary)**: Se utiliza statsmodels para obtener un resumen estadístico del modelo de regresión logística, que incluye los coeficientes y sus pruebas de significancia.
12. **Resumen General:** En total, el código hace dos tipos de predicciones: la primera mediante el método predict de LogisticRegression sobre el conjunto de prueba, y la segunda a través de predict_proba para calcular el ROC AUC. Se hace una única división de los datos en conjuntos de entrenamiento y prueba, pero se realizan múltiples evaluaciones y métricas para validar el rendimiento del modelo. La primera predicción (y_pred) se usa para calcular métricas de rendimiento, y el uso de predict_proba permite evaluar la confianza del modelo en sus predicciones.

Resultados de las métricas

```
Accuracy: 0.9721
Confusion Matrix:
[[753  13]
 [ 14 189]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       766
           1       0.94      0.93      0.93       203

    accuracy                           0.97       969
   macro avg       0.96      0.96      0.96       969
weighted avg       0.97      0.97      0.97       969

F1 Score: 0.9333
Recall: 0.9310
Precision: 0.9356
ROC AUC Score: 0.9958
Optimization terminated successfully.
         Current function value: 0.079681
         Iterations 10
```

Métricas

Precisión del Modelo

- Accuracy: 0.9660: El modelo tiene una precisión del 96.60%, lo que indica que aproximadamente el 96.6% de las predicciones del modelo son correctas. Este es un resultado excelente y sugiere que el modelo está funcionando bien en general.

Matriz de Confusión

- Verdaderos Negativos (TN): 757 (predicciones correctas para la clase 0)
- Falsos Positivos (FP): 14 (predicciones incorrectas para la clase 1)
- Falsos Negativos (FN): 19 (predicciones incorrectas para la clase 0)
- Verdaderos Positivos (TP): 180 (predicciones correctas para la clase 1)

La matriz de confusión muestra que el modelo tiene un buen desempeño en ambas clases, con un alto número de verdaderos positivos y verdaderos negativos en comparación con los falsos.

Métricas Adicionales

- F1 Score: 0.9160: Es un promedio ponderado de precisión y recall. Un valor de 0.9160 indica un buen equilibrio entre precisión y recall, especialmente para la clase 1.
- Recall: 0.9045: Este es un indicador de la capacidad del modelo para identificar correctamente los casos positivos (hipertensión en este caso). Un valor de 0.9045 es muy bueno.
- Precision: 0.9278: Esto indica que cuando el modelo predice un caso positivo, hay un 92.78% de probabilidad de que sea realmente positivo.
- ROC AUC Score: 0.9954: Este valor indica un rendimiento excepcional del modelo en la clasificación. Un AUC cercano a 1 implica que el modelo tiene una gran capacidad para distinguir entre las clases positivas y negativas.

Información General del Modelo y Coeficientes

```
Optimization terminated successfully.
         Current function value: 0.079681
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:           hipertension   No. Observations:                 3876
Model:                          Logit   Df Residuals:                     3871
Method:                           MLE   Df Model:                            4
Date:                Fri, 01 Nov 2024   Pseudo R-squ.:                  0.8488
Time:                        05:56:44   Log-Likelihood:                -308.84
converged:                       True   LL-Null:                       -2042.6
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.8859      0.228    -21.389      0.000      -5.334      -4.438
x1             0.4187      0.115      3.633      0.000       0.193       0.645
x2            11.6839      0.572     20.437      0.000      10.563      12.804
x3             0.5909      0.118      5.021      0.000       0.360       0.822
x4            -3.0354      0.188    -16.137      0.000      -3.404      -2.667
==============================================================================

Possibly complete quasi-separation: A fraction 0.34 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Variable Dependiente

- Dep. Variable (Variable Dependiente): hipertension
  - La variable que el modelo intenta predecir es hipertensión, la cual indica la probabilidad de que un individuo sufra de hipertensión. Este modelo utiliza varias variables independientes, como Systolic, Diastolic, HR, entre otras, para hacer esta predicción.

Modelo

- Modelo: Logit (Regresión Logística)
  - El método utilizado es una regresión logística, que estima la probabilidad de un evento binario (en este caso, hipertensión) al aplicar el método de máxima verosimilitud (MLE). Este enfoque es adecuado para clasificaciones binarias.

Métricas del Modelo

- Pseudo R-squared (Pseudo Coeficiente de Determinación): 0.8488
  - Este valor indica que el modelo explica el 84.88% de la variabilidad en la probabilidad de hipertensión, lo cual sugiere un ajuste fuerte. Aunque no se interpreta igual que el R-squared de una regresión lineal, un valor cercano a 1 sigue siendo un indicador positivo.
- Log-Likelihood (Log-Verosimilitud): -308.84
  - La log-verosimilitud es una medida del ajuste general del modelo; un valor menos negativo representa un mejor ajuste.
- LL-Null (Log-Verosimilitud Nula): -2042.6
  - Es el log-likelihood de un modelo sin predictores, que solo tiene el término constante. La diferencia con el Log-Likelihood indica que los predictores ayudan a mejorar el ajuste.
- Convergencia: True
  - El modelo logró converger exitosamente, lo que significa que el algoritmo de optimización encontró una solución estable para los parámetros del modelo.

Parámetros del Modelo

- Cada parámetro tiene un impacto específico en la probabilidad de hipertensión. Estos coeficientes estiman cómo cambia la probabilidad de hipertensión al variar una unidad de cada predictor, manteniendo los demás constantes.

Errores y Pruebas de Significancia

- Error Estándar (std err):
  - Cada coeficiente tiene un error estándar asociado que mide la precisión de la estimación. Valores bajos de error estándar sugieren una estimación más precisa.
- Z (Estadístico Z):
  - Cada coeficiente se divide entre su error estándar para calcular el estadístico Z, utilizado para evaluar la significancia estadística del predictor.
- P>|z| (Valor p):
  - Valores p menores a 0.05 indican que el predictor es estadísticamente significativo en el modelo. En este caso, todos los predictores tienen un valor p de 0.000, lo que sugiere que son significativos para predecir la hipertensión.
- Intervalo de Confianza (0.025 - 0.975):
  - Estos valores definen el intervalo dentro del cual el coeficiente se espera que caiga con un 95% de confianza. Un intervalo que no incluye el cero implica que el efecto del predictor es significativo.

Estadísticas del Modelo

- Número de Observaciones (No. Observations): 3876
  - El modelo se ha ajustado utilizando un total de 3876 observaciones.
- Grados de Libertad del Modelo (Df Model): 4
  - Representa el número de predictores en el modelo.
- Grados de Libertad Residuales (Df Residuals): 3871
  - Los grados de libertad residuales resultan de restar el número de predictores al total de observaciones.
- LLR p-value: 0.000
  - Este valor p evalúa la significancia global del modelo; un valor bajo indica que el modelo, como un todo, tiene un poder predictivo significativo.

Pruebas de Supuestos del Modelo

- Quasi-Separation Warning:
  - Possibly complete quasi-separation: A fraction 0.34 of observations can be perfectly predicted.
  - Esta advertencia de cuasi-separación sugiere que algunas observaciones son separables de manera perfecta, lo que puede dificultar la identificación de ciertos parámetros en el modelo.

Resultados obtenidos

El modelo de regresión logística desarrollado para predecir la hipertensión en base a variables como la presión arterial sistólica (Systolic), frecuencia cardíaca (HR) y otros factores relevantes ha mostrado un rendimiento sólido y significativo. Con un pseudo R-squared de 0.8488, el modelo explica una gran parte de la variabilidad en la probabilidad de sufrir hipertensión, lo que sugiere que las variables seleccionadas son predictivas y relevantes para esta condición.

Todos los parámetros del modelo resultaron estadísticamente significativos, lo que indica que cada uno de ellos tiene un efecto notable en la predicción de la hipertensión. Los coeficientes positivos y negativos obtenidos ofrecen información valiosa sobre cómo cada variable influye en la probabilidad de hipertensión, permitiendo identificar factores de riesgo y guiar intervenciones preventivas.

Sin embargo, la advertencia sobre la posible cuasi-separación sugiere que existe una fracción de observaciones que puede ser perfectamente predicha, lo que puede limitar la capacidad del modelo para generalizar en ciertos casos. Esto implica que, aunque el modelo es efectivo en su estado actual, se deben considerar ajustes adicionales o el uso de técnicas más complejas para abordar esta cuestión y mejorar la robustez del modelo.

En general, este análisis resalta la importancia de una evaluación continua y el ajuste de modelos predictivos en la identificación y manejo de condiciones de salud como la hipertensión, contribuyendo así a la promoción de la salud pública y el bienestar en la población.

Recomendación para futuras mejoras:

1. **Refinamiento del Modelo**: Implementar técnicas adicionales para abordar la cuasi-separación observada, como el uso de regularización (por ejemplo, Lasso o Ridge) o modelos más complejos como los árboles de decisión o redes neuronales.
2. **Validación Cruzada**: Realizar una validación cruzada más exhaustiva para evaluar la robustez del modelo y su capacidad de generalización a diferentes subconjuntos de datos.
3. **Incorporación de Nuevas Variables**: Explorar la inclusión de nuevas variables predictivas que puedan mejorar la precisión del modelo, como factores genéticos, hábitos de vida, o datos socioeconómicos.
4. **Evaluación de Interacciones**: Analizar posibles interacciones entre las variables existentes para identificar combinaciones que puedan tener un impacto significativo en la predicción de la hipertensión.
5. **Actualización de Datos**: Mantener el modelo actualizado con datos recientes para asegurar que las predicciones sigan siendo relevantes y precisas en el tiempo.
6. **Implementación Práctica**: Desarrollar herramientas prácticas basadas en el modelo que puedan ser utilizadas por profesionales de la salud para la identificación temprana y manejo de la hipertensión en pacientes.
7. **Investigación Continua**: Fomentar la investigación continua para identificar nuevas tendencias y factores de riesgo emergentes que puedan ser incorporados en futuros modelos predictivos.