Proyecto ABP unificado del primer y segundo cuatrimestre.

Ciencia de Datos II.
Estadística y Exploración de Datos II.

INTEGRANTES CLUSTER-FUSION:

Barbero, Maciel.

Marini, Ian Denis.

Molina, Jonathan Ariel.

Molina, Mauricio Leonel.

Robles, Emilce Lucia Nicole.

Sosa, Sebastian Cristhian.

Virinni, Marco.

Sección Primer cuatrimestre.

NOMBRE DEL PROYECTO: Análisis predictivo de ataques cardíacos con datos clínicos del Hospital Zheen (2019).

TIPO DE PROYECTO: Tecnológico y de investigación. El proyecto integra el uso de herramientas tecnológicas para el procesamiento y análisis de datos reales en el ámbito de la salud, con un enfoque exploratorio para interpretar factores de riesgo en ataques cardíacos.

ESPACIO CURRICULAR O ESPACIOS PARTICIPANTES EN EL MÓDULO:

Procesamiento de datos - Estadística y Exploración de Datos 1.

EJES TEMÁTICOS/RED DE CONCEPTOS: Los ejes temáticos que se trabajan son:

- Análisis exploratorio de datos (EDA).
- Limpieza y validación de datos.
- Estadística descriptiva (media, mediana, desviación estándar, valores atípicos).
- Identificación de valores nulos y registros erróneos.
- Análisis de variables categóricas y numéricas.
- Interpretación de resultados para la toma de decisiones.
- Aplicación de herramientas computacionales (Python, Pandas, Jupyter).

PROBLEMÁTICAS/NECESIDADES: En la actualidad, los ataques cardíacos representan una de las principales causas de muerte a nivel mundial. El acceso a datos médicos permite realizar análisis que podrían ayudar a comprender mejor los factores críticos asociados. Sin embargo, muchas veces estos datos no son aprovechados por falta de habilidades técnicas. Esta situación plantea la necesidad de formar perfiles capaces de

realizar análisis de datos en contextos reales y sensibles, como el de la salud.

FUNDAMENTACIÓN: Este proyecto fue elegido por su potencial para fortalecer competencias técnicas en ciencia de datos aplicadas al ámbito de la salud, utilizando un dataset real recopilado en el Zheen Hospital de Erbil, Iraq, durante los meses de enero a mayo de 2019. La base de datos contiene registros biomédicos de pacientes con sospecha de ataque cardíaco, lo que permite trabajar con información representativa de un problema de salud pública global.

El análisis estadístico y exploratorio busca identificar patrones, relaciones entre variables y validar hipótesis que podrían orientar futuras decisiones clínicas o investigaciones. Esta tarea permite a los estudiantes aplicar conocimientos técnicos (como la limpieza, validación y análisis de datos), éticos (como el manejo responsable de datos sensibles) y científicos (como la interpretación crítica de indicadores biomédicos).

Además, este tipo de análisis puede ser de gran utilidad para el personal sanitario, en especial para médicos cardiólogos, ya que puede servir como una herramienta complementaria para el seguimiento de pacientes y la detección temprana de signos de riesgo, contribuyendo a intervenciones preventivas más efectivas.

El proyecto no solo apunta al desarrollo académico y técnico, sino también promueve la toma de conciencia sobre problemáticas de salud pública y la integración de la tecnología en favor del bienestar social.

VISIÓN DEL PROYECTO: Construir una base sólida de análisis de datos médicos mediante la utilización de herramientas estadísticas y tecnológicas, con el fin de interpretar correctamente variables biomédicas y generar informes útiles que promuevan una toma de decisiones informada en el ámbito de la salud.

DISEÑO DE LOS OBJETIVOS.

OBJETIVO GENERAL: Aplicar técnicas de análisis de datos utilizando Python, Pandas y herramientas estadísticas para identificar patrones relevantes en un dataset médico sobre ataques cardíacos recolectado entre enero y mayo de 2019 en el hospital Zheen de Erbil, Iraq, desarrollando el trabajo durante el primer semestre de 2025.

OBJETIVOS ESPECÍFICOS:

1. Importar, limpiar y preparar un dataset real para análisis estadístico.
2. Analizar las variables numéricas y categóricas en términos de distribución, media, valores faltantes y posibles errores.
3. Interpretar los resultados obtenidos y redactar un informe con fundamentos técnicos y científicos.
4. Crear un entorno reproducible que permita compartir el análisis con la comunidad académica o profesional.

METAS: El proyecto tendrá como meta la presentación de un informe final en formato PDF que refleje un análisis técnico y estadístico riguroso del dataset médico proveniente del Zheen Hospital (Iraq, 2019).

SELECCIÓN DE ACCIONES.

OBJETIVO ESPECÍFICO	ACCIONES
Importar, limpiar y preparar un dataset real para análisis estadístico.	Crear entorno de trabajo en VSCode y/o Colab, importar el CSV, revisar valores nulos y corregir errores

Analizar las variables numéricas y categóricas en términosde distribución, media, valores faltantes y posibles errores.	Usar Pandas y Numpy para describir estadísticamente el dataset y verificar datos erróneos
Interpretar los resultados y redactar informe.	Generar gráficas, detectar correlaciones posibles, extraer conclusiones claras y relevantes
Crear entorno reproducible	Subir el archivo .ipynb al repositorio y generar una presentación o PDF con los hallazgos

CRONOGRAMA:

CRONOGRAMA	SEMANA 1	SEMANA 2	SEMANA 3	SEMANA 4
Objetivo 1: Importar, limpiar y preparar un dataset real para análisis estadístico.	Crear entorno de trabajo en VSCode y/o Colab. Subir el dataset CSV original al repositorio.	Cargar los datos con Pandas. Identificar y tratar valores nulos y registros mal formatea dos	Verificar los tipos de Datos. Normalizar o ajustar formatos si es necesario	Documenta r la limpieza realiza

Objetivo 2: Analizar las variables numéricas y categóricas en términos de distribución, media, valores faltantes y posibles errores.	Definir las variables numéricas y categóricas del dataset Verificar categorías esperadas (0/1, positive/negative, etc.)	Aplicar análisis estadístico con Pandas y Numpy Obtener media, mediana, desvío estándar y conteos	Detectar posibles outliers o inconsistencias Registrar cantidad de nulos o valores inesperados	Consolidar tabla de resultados con insights por valores.
Objetivo 3: Interpretar los resultados y redactar informe.	Comenzar redacción del informe técnico Plantear hipótesis preliminares según patrones visibles	Crear visualizaciones (gráficos de barras, histogramas, etc.) Extraer conclusiones parciales	Redactar versión completa del informe Integrar visualizaciones y referencias	Revisar redacción, formato y ortografía Exportar informe en PDF
Objetivo 4: Crear entorno reproducible.	Configurar notebook limpio y comentado	Subir versión final del archivo .ipynb al repositorio	Generar link compartible de Colab o GitHub	Validar acceso, probar desde otro dispositivo

	Probar funcionamiento de todas las celdas			

PRODUCTO FINAL:

Informe de Análisis de Datos Médicos, Factores Asociados a Ataques Cardíacos.

El presente informe forma parte del proyecto ABP titulado “Análisis predictivo de ataques cardíacos con datos clínicos del Hospital Zheen (2019)”, desarrollado en el marco del espacio curricular Procesamiento de datos - Estadística y Exploración de Datos 1. El objetivo principal del trabajo es aplicar técnicas estadísticas y herramientas de programación (Python, Pandas, ETC) para identificar patrones relevantes en pacientes con y sin diagnóstico de ataque cardíaco, y explorar factores biomédicos que podrían estar asociados a este tipo de eventos.

Variables: ¿Qué son y cuál es su Importancia en un ataque cardíaco?

Variable	¿Qué es/mide?	Importancia
Frecuencia cardíaca	La cantidad de latidos por minuto (lpm) del corazón en reposo. **Valor normal en reposo (adulto)**: 60 - 100 latidos por minuto (lpm).	Una frecuencia muy alta (taquicardia) o muy baja (bradicardia) puede indicar una disfunción cardíaca. Durante un infarto, el corazón puede alterarse y generar ritmos anormales o acelerados.
Presión arterial sistólica	La presión cuando el corazón se contrae y bombea sangre. **Valor normal**: 90 - 120 mmHg. **Hipertensión etapa 1**: 130 - 139. **Hipertensión etapa 2**: ≥ 140.	Valores elevados pueden sobrecargar el corazón y dañar las arterias, aumentando el riesgo de infarto. Un descenso brusco puede ocurrir durante un infarto grave y es señal de colapso circulatorio.
Presión arterial diastólica	La presión cuando el corazón está en reposo entre latidos. **Valor normal**: 60 - 80 mmHg. **Hipertensión etapa 1**: 80 - 89.	Refleja la salud de los vasos sanguíneos. Presiones anormalmente bajas o altas pueden indicar un sistema cardiovascular comprometido.

	Hipertensión etapa 2: ≥ 90.	
Glucosa en sangre	La concentración de glucosa en la sangre (mg/dL). **Valor normal**: en ayunas (8 h sin comer): 70 - 99 mg/dL.	La hiperglucemia es común en pacientes con infarto, incluso sin diagnóstico previo de diabetes. Altos niveles de glucosa durante un evento cardíaco se asocian con peores pronósticos.
Creatina Quinasa MB	Una enzima liberada al torrente sanguíneo cuando hay daño en el músculo cardíaco. **Valor normal**: 0 - 5 ng/mL.	Aumenta unas horas después de un infarto y es un marcador clásico del daño miocárdico. Se usa como prueba de apoyo junto con troponina.
Troponina	Una proteína muy específica del músculo cardíaco que se libera cuando hay daño al corazón. **Valor normal**: Troponina I (TnI): < 0.04 ng/mL. Troponina T (TnT): < 0.01 ng/mL.	Es el marcador más específico y sensible para diagnosticar un infarto. Niveles elevados confirman daño cardíaco, incluso en infartos silenciosos o leves.

Análisis de indicadores.

Comparación de casos positivos vs. casos negativos.

A continuación, se presentan los promedios de los principales indicadores biomédicos según el diagnóstico (presencia o ausencia de ataque cardíaco):

```
--Casos positive--

Masculino: 69.29%
Femenino: 30.71%

Tabla descriptiva para positive - Masculino:
```

	Age	Heart rate	Systolic blood pressure	Diastolic blood pressure	Blood sugar	CK-MB	Troponin
count	564.000000	564.000000	564.000000	564.000000	564.000000	564.000000	564.000000
mean	57.450355	78.338652	127.021277	72.138298	124.578014	23.182356	0.618892
std	13.390408	46.470455	25.031765	13.829667	39.794266	58.482669	1.404175
min	-48.000000	36.000000	65.000000	38.000000	35.000000	0.353000	0.003000
25%	50.000000	63.000000	110.000000	62.000000	98.000000	1.845000	0.018000
50%	60.000000	74.000000	123.000000	72.000000	111.000000	3.585000	0.053000
75%	65.000000	87.000000	140.250000	80.000000	139.000000	11.457500	0.636250
max	100.000000	1111.000000	223.000000	154.000000	247.000000	300.000000	10.300000

```
----------------------------------------

Tabla descriptiva para positive - Femenino:
```

	Age	Heart rate	Systolic blood pressure	Diastolic blood pressure	Blood sugar	CK-MB	Troponin
count	250.00000	250.000000	250.000000	250.000000	250.000000	250.000000	250.000000
mean	61.14400	79.540000	126.004000	72.172000	126.724000	23.522888	0.459036
std	14.41842	67.062822	26.584638	13.913618	41.949384	55.762567	1.345899
min	-25.00000	20.000000	67.000000	40.000000	50.000000	0.452000	0.003000
25%	54.00000	64.000000	110.000000	62.000000	97.000000	1.932500	0.012000
50%	60.00000	74.000000	121.000000	70.000000	111.000000	4.720000	0.030000
75%	70.00000	85.000000	140.000000	80.000000	146.750000	15.722500	0.146500
max	103.00000	1111.000000	223.000000	128.000000	247.000000	300.000000	10.000000

```
--Casos negative--

Masculino: 60.39%
Femenino: 39.61%

Tabla descriptiva para negative - Masculino:
```

	Age	Heart rate	Systolic blood pressure	Diastolic blood pressure	Blood sugar	CK-MB	Troponin
count	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000
mean	50.942122	75.591640	127.971061	72.234727	127.826367	2.659424	0.040090
std	14.002875	15.041356	27.316587	14.762405	42.645221	1.439668	0.567106
min	-20.000000	20.000000	42.000000	40.000000	60.000000	0.345000	0.001000
25%	42.000000	64.000000	110.000000	60.000000	99.000000	1.530000	0.003000
50%	50.000000	74.000000	126.000000	72.000000	111.000000	2.470000	0.006000
75%	61.000000	84.000000	147.500000	82.000000	144.500000	3.465000	0.009000
max	91.000000	132.000000	223.000000	118.000000	246.000000	6.270000	10.000000

```
----------------------------------------

Tabla descriptiva para negative - Femenino:
```

	Age	Heart rate	Systolic blood pressure	Diastolic blood pressure	Blood sugar	CK-MB	Troponin
count	204.000000	204.000000	204.000000	204.000000	204.000000	204.000000	204.000000
mean	53.666667	81.240196	127.651961	72.872549	125.714216	2.416858	0.006544
std	14.835275	73.864449	26.497355	13.655271	42.162764	1.260646	0.003419
min	-3.000000	40.000000	67.000000	44.000000	67.000000	0.321000	0.002000
25%	43.750000	63.000000	109.000000	62.000000	96.750000	1.497500	0.003000
50%	55.000000	76.000000	125.000000	72.000000	111.000000	2.130000	0.006000
75%	65.000000	84.250000	144.000000	82.000000	147.000000	3.202500	0.009000
max	86.000000	1111.000000	220.000000	128.000000	250.000000	7.020000	0.014000

Principales observaciones.

Sexo y edad.

Los datos muestran una mayor proporción de hombres entre los casos positivos, lo que sugiere una mayor incidencia de infarto en varones. Además, los pacientes con diagnóstico positivo tienden a ser de mayor edad en ambos sexos. En los hombres, la edad promedio en casos positivos es de 57 años, frente a 50 años en los negativos. En las mujeres, la media es de 61 años en casos positivos, comparada con 53 años en los negativos. Este patrón reafirma la relación entre edad avanzada y mayor riesgo de eventos cardíacos, y confirma que, en general, las mujeres suelen presentar infartos a edades más tardías que los hombres, una observación consistente con la literatura médica.

Signos vitales y presión arterial.

En relación con los signos vitales, se observa que la frecuencia cardíaca promedio es ligeramente más alta en pacientes positivos, especialmente en hombres, aunque las diferencias no parecen ser clínicamente significativas. En los hombres positivos, la media de frecuencia cardíaca es de 78 latidos por minuto frente a 75. en los negativos, mientras que en mujeres se observa una media de 79 en positivas y 81 en negativas. En cuanto a la presión arterial, los valores promedio de presión sistólica y diastólica son similares entre los grupos, rondando los 127 mmHg y 72 mmHg respectivamente, tanto en hombres como en mujeres. Sin embargo, los pacientes positivos presentan una menor desviación estándar en estas variables, lo que podría sugerir un cuadro clínico más homogéneo, posiblemente asociado a la fase aguda del evento cardíaco.

Biomarcadores clave.

Las diferencias más notables entre los casos positivos y negativos se encuentran en los niveles de los biomarcadores cardíacos CK-MB y troponina. Ambos marcadores muestran elevaciones muy significativas en los pacientes con diagnóstico positivo, lo que respalda su utilidad diagnóstica. En el caso de la troponina, los valores en los percentiles más altos (especialmente el percentil 75 y el valor máximo) son considerablemente mayores en los positivos que en los negativos, tanto en hombres como en mujeres. Esta diferencia es especialmente marcada en los hombres, donde el percentil 75 alcanza 0.63 ng/mL frente a apenas 0.009 ng/mL en negativos. En las mujeres, aunque los niveles también son más altos en positivas (percentil 75: 0.15 ng/mL), siguen siendo más bajos en comparación con los hombres. De forma similar, los niveles de CK-MB son considerablemente más altos en positivos, con valores que superan las 11 U/L en hombres y alcanzan hasta 15.72 U/L en mujeres en el cuartil superior. Estas observaciones confirman la relevancia clínica de estos biomarcadores para la detección de infarto agudo de miocardio y evidencian la buena coherencia y calidad del conjunto de datos analizado.

Paciente con ataque cardíaco: perfil clínico observado.

Indicador	Paciente masculino	Paciente femenino
Edad	50–65 años (percentil 25–75), media 57.45.	54–70 años (percentil 25–75), media 61.14.
Frecuencia cardíaca	63–87 bpm, sin taquicardia significativa.	64–85 bpm.
Presión arterial	Estable, pero algunos con hipotensión (min 65/38 mmHg).	121/72 mmHg promedio, estable.
Glucosa	Elevada en muchos casos (percentil 75: 139 mg/dL, máximo: 300 mg/dL).	Alta variabilidad, frecuente hiperglucemia (percentil 75: 146.75 mg/dL, máximo: 300 mg/dL).
CK-MB	Puede superar 11 U/L en el cuartil superior.	Elevado, con valores frecuentes entre 1.93 y 15.72 U/L.
Troponina	> 0.63 ng/mL en muchos casos.	Elevada aunque en menor grado que en hombres (percentil 75: 0.15 ng/mL).
Perfil	Hombre de mediana edad, con glucosa elevada y biomarcadores cardíacos significativamente alterados.	Mujer de edad más avanzada, con elevación de CK-MB y troponina, aunque menor que en varones. Glucosa elevada también es común.

Interpretación de los indicadores.

Los datos analizados muestran diferencias claras entre los pacientes que sufrieron un ataque cardíaco y aquellos que no. En términos de género, los casos positivos presentan una mayor proporción de hombres que mujeres, lo que determina una mayor incidencia en varones. Asimismo, los pacientes con ataque cardíaco son en promedio más longevos, lo que refuerza la relación entre edad y riesgo cardiovascular.

Si bien las constantes vitales como la presión arterial y el ritmo cardíaco no presentan variaciones tan marcadas entre ambos grupos, los marcadores clínicos específicos muestran diferencias drásticas: los niveles de CK-MB y Troponina —indicadores directos de daño cardíaco— son considerablemente más altos en los pacientes con ataque cardíaco, lo que valida su uso como señales de alerta clínica.

Estos hallazgos reflejan el valor de los datos para identificar patrones de riesgo, y subrayan la importancia de un monitoreo constante en grupos poblacionales con

perfiles similares. El uso consciente de esta información podría permitir mejorar estrategias preventivas, diagnósticas y terapéuticas dentro del sistema de salud.

Vinculación con la comunidad.

La información obtenida a partir del análisis de los datos clínicos permite no solo detectar perfiles de riesgo de infarto agudo de miocardio, sino también pensar estrategias concretas de vinculación con la comunidad. En este sentido, se propone articular acciones con centros de salud barriales, unidades sanitarias móviles y organizaciones sociales con el fin de llevar adelante campañas de prevención y detección temprana.

Acercar la salud a donde vive la gente.

Una estrategia clave para mejorar la prevención y la detección temprana de enfermedades cardiovasculares es fortalecer la articulación entre el sistema de salud y los espacios comunitarios. Esto implica salir de las estructuras tradicionales (hospitales, clínicas) y llevar la atención a los lugares donde las personas desarrollan su vida cotidiana.

Actores clave para la articulación.

- Centros de salud barriales: son el primer nivel de atención en muchos barrios y cumplen un rol fundamental en el acompañamiento territorial. La articulación con ellos permite identificar zonas con mayor prevalencia de factores de riesgo y organizar operativos focalizados.
- Unidades sanitarias móviles: permiten llegar a zonas rurales, barrios periféricos o asentamientos informales donde la población tiene dificultades de acceso al sistema de salud. Estas unidades pueden ofrecer controles básicos, derivaciones y consejería en salud cardiovascular.
- Organizaciones sociales y comunitarias: incluyen comedores, centros culturales, grupos vecinales, parroquias, organizaciones de base y asociaciones civiles. Estos espacios suelen tener alta legitimidad y llegada directa a la población local, lo que los convierte en aliados estratégicos para convocar, difundir y sostener actividades

de salud.

Objetivos de esta articulación

- Ampliar la cobertura del sistema de salud mediante una red territorial integrada.
- Reducir barreras de acceso (geográficas, económicas, simbólicas) al diagnóstico y prevención.
- Fomentar la participación comunitaria y el compromiso colectivo en el cuidado de la salud.
- Adaptar las estrategias de intervención a las características y necesidades concretas de cada territorio.

Estrategias sugeridas

- Reuniones de planificación conjunta entre equipos de salud y referentes comunitarios.
- Calendario rotativo de operativos de salud en diferentes barrios.
- Capacitación a promotores de salud comunitarios en factores de riesgo cardíaco.
- Difusión de campañas a través de radios locales, redes sociales barriales y centros comunitarios.

CONCLUSIÓN.

El análisis de los indicadores clínicos promedio revela diferencias significativas entre los pacientes que han sufrido un ataque cardíaco y aquellos que no. Desde el rol del analista de datos, resulta evidente cómo el procesamiento y comparación de variables como la edad, el género, la glucemia, y especialmente los biomarcadores como la Troponina y la CK-MB, pueden ofrecer información crítica para anticipar situaciones de riesgo. La posibilidad de construir perfiles diferenciales, tanto para varones como para mujeres, permite identificar patrones que, de ser adecuadamente utilizados por los equipos de salud, podrían traducirse en herramientas predictivas de gran valor clínico.

El presente informe no sólo pone de manifiesto el poder del análisis de datos en la toma de decisiones sanitarias, sino que también plantea un desafío ético y social: transformar los datos en conocimiento útil para la prevención. El análisis de datos no debería limitarse a describir lo que ya ocurrió, sino abrir caminos hacia intervenciones más tempranas, personalizadas y efectivas. En un contexto donde la salud pública enfrenta limitaciones de recursos, conocer los perfiles de mayor riesgo es una oportunidad concreta para optimizar esfuerzos y salvar vidas.

SECCIÓN SEGUNDO CUATRIMESTRE.

Transición al Análisis Inferencial y Modelado Predictivo.

El proyecto ABP sobre la predicción de ataques cardíacos, que utilizó datos del Hospital Zheen, mutó radicalmente durante el segundo cuatrimestre. La exploración y descripción de datos fue el foco del primer cuatrimestre, pero en el segundo la labor se centró plenamente en la Estadística Inferencial, buscando identificar las conexiones causales y los verdaderos factores predictivos dentro del dataset. El objetivo final fue modelar, mediante una Regresión Logística, qué variables clínicas son decisivas para un diagnóstico positivo.

Profundización y Actividades Metodológicas.

Evaluación de Impacto con ANOVA de Dos Vías.

La evidencia de aprendizaje N°2 consistió en aplicar el Análisis de Varianza (ANOVA) de Dos Vías. La meta era evaluar, simultáneamente, el impacto del Resultado Final (nuestra variable dependiente) junto con el Género sobre las variables clínicas cuantitativas, como la edad y los biomarcadores. Antes de correr la prueba estadística, tuvimos que hacer una fase de visualización: graficamos boxplots de interacción, una acción crítica para detectar tanto los efectos individuales de cada factor como las interacciones (situaciones donde el efecto de una variable, digamos la edad, variaba entre géneros). Tras ajustar el modelo en Python (usando la librería statsmodels con la Suma de Cuadrados Tipo II), la interpretación se concentró en el p-valor. Al confirmar que la Edad, CK-MB y Troponina mostraban diferencias importantes, procedimos con un análisis *post-hoc* (el método Tukey HSD) para desmenuzar exactamente dónde residían esas diferencias y poder cuantificarlas, un paso indispensable para la selección final de variables en la Regresión.

Correlación Robusta y Preparación del Feature Set.

El paso previo a la construcción del modelo también se llevó a cabo en la evidencia de aprendizaje N°2, en donde se focalizó en consolidar la relación entre las variables. El análisis de Correlaciones no fue un paso menor. Se eligió el coeficiente de Spearman sobre el tradicional Pearson. Esta decisión se fundamentó en la no-normalidad de las distribuciones y la persistencia de valores atípicos, haciendo de Spearman una medida más robusta para cuantificar la relación monotónica entre los parámetros. En esta etapa también se consolidó el manejo de *outliers* iniciado en el primer cuatrimestre (como la imputación de valores extremos de Heart rate con la mediana). Paralelamente, en un mapa de calor (*heatmap*) se plasmó visualmente las correlaciones, permitiéndonos identificar y gestionar la multicolinealidad, un factor que, si es ignorado, compromete seriamente la interpretabilidad y estabilidad de la Regresión.

Rigor Ético y Desarrollo de la Regresión Logística.

La fase final se dedicó al ajuste y la validación del Modelo de Regresión Logística. Se evaluaron tres enfoques principales para la selección de variables en la regresión logística:

Selección hacia adelante.

Se comenzó con un modelo vacío y se fueron incorporando variables una a una, escogiendo en cada paso la variable que proporcionaba la mayor mejora en el criterio Akaike (AIC).

Selección hacia atrás.

Se partió de un modelo con todas las variables candidatas y se fueron eliminando aquellas que menos contribuían a mejorar el modelo según el criterio AIC.

Selección stepwise (hacia adelante y atrás).

Este método combinó ambas estrategias, permitiendo añadir y eliminar variables en cada iteración para optimizar el valor de AIC.

El criterio de información Akaike (AIC) fue el seleccionado para evaluar la calidad y parsimonia de los modelos, buscando un equilibrio entre ajuste y complejidad.

Finalmente, el modelo seleccionado incluyó las variables CK_MB, Troponin, Edad y

Género, todas estadísticamente significativas, con un buen ajuste y capacidad predictiva según el criterio AIC y respaldadas porel ANOVA previo.
La partición de datos (70% para entrenamiento / 30% para prueba) se justificó por una deliberación ética: trabajar con datos de salud implica una responsabilidad mayúscula. Un error de clasificación podría tener repercusiones directas en pacientes, por lo que se hizo imperativo reservar un 30% del dataset para la prueba. Este conjunto de validación más amplio garantizó que las métricas de rendimiento fueran más estables y conservadoras, minimizando el sobreajuste y asegurando una mayor generalización clínica del modelo. A continuación, y solo sobre el set de entrenamiento, se aplicó el escalado estándar de variables. El modelo ajustado reveló que la Troponina era, por lejos, el predictor más potente (OR ≈ 248.3), y el Odds Ratio del género masculino confirmó un incremento de riesgo del 53%. Finalmente, el modelo se validó con las métricas tradicionales y se revisaron sus supuestos fundamentales (VIF y Distancia de Cook), verificando la solidez estructural de los resultados obtenidos.

Conclusiones del segundo cuatrimestre.

El segundo cuatrimestre materializó la transición del conocimiento descriptivo al poder inferencial y predictivo. La implementación rigurosa de herramientas como el ANOVA de Dos Vías y la Regresión Logística Binaria permitió no solo confirmar los factores de riesgo ya conocidos sino también cuantificar su impacto clínico mediante Odds Ratios. La decisión metodológica de emplear un conjunto de prueba del 30% se establece como un protocolo de responsabilidad ética esencial en el manejo de datos de salud, asegurando la robustez y la generalización de la herramienta. Con estas actividades, el proyecto demostró la capacidad de generar un prototipo de apoyo diagnóstico rigurosamente validado, transformando los datos hospitalarios en conocimiento accionable.

BIBLIOGRAFÍA:

- Arteaga, L. (2019). *Estadística para todos: Teoría y práctica con Python y R*.

Ediciones Díaz de Santos.

- NumPy Developers. (2023). *NumPy Documentation*.

  https://numpy.org/doc/ Python Software Foundation. (2024).

  Pandas Documentation. https://pandas.pydata.org/

- McKinney, W. (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media.
- World Health Organization. (2023).

  Cardiovascular Diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvd s)

- Rashid, T. A., & Hassan, B. (2022). *Heart Attack Dataset*. Mendeley Data, V1. https://doi.org/10.17632/wmhctcrt5v.1