Trabajo Práctico 3 – Taller de programación – Universidad de Buenos Aires
CLASIFICANDO DE POBRES CON LA EPH
Noviembre 2025

Grupo 6 Gustavo Horacio Romero
Link https://github.com/shanthalchs/BigDataUBA-Grupo6

A. Enfoque de Validación

El presente trabajo tiene como objetivo predecir la condición de pobreza (pobre=1) en individuos residentes en el Gran Buenos Aires, utilizando variables estructurales observables cuando la variable crítica Ingreso Total Familiar (ITF) no está disponible. La predicción ex-ante de pobreza es crucial para diseñar estrategias de focalización de programas sociales, especialmente cuando existen altas tasas de no respuesta en la información de ingresos, un fenómeno recurrente en la Encuesta Permanente de Hogares (EPH).

Se trabajó con datos de la EPH para la región Gran Buenos Aires, utilizando dos bases principales:

- **respondieron** (datos_filtrados.csv): 16.665 individuos con ITF informado.
- **norespondieron** (nrorespondieron.csv): 2985 individuos sin información de ingreso.

Para cada año disponible se aplicó un procedimiento estándar de validación hold-out con estratificación para el conjunto de respondieron:

- **Proporción:** 70% entrenamiento (Train) - 30% test (Test).
- **Estratificación:** Mantiene la proporción de clases (pobres/no pobres) en ambos conjuntos.
- **Semilla aleatoria:** random_state = 444 para reproducibilidad.

A1. Variables Seleccionadas (Matriz X)

Se seleccionaron únicamente variables que están disponibles en ambas bases, limpias, y que no dependen del ingreso. El enfoque fue en variables estructurales y de capital humano.
Las variables finalmente utilizadas fueron:

- edad
- educ (años de educación)
- num_miembros_hogar
- horastrab
- edad² (Término cuadrático para capturar curvaturas no lineales)

La variable dependiente binaria es: pobre (1=pobre, 0=no pobre).

A2. Diferencias de medias entre train y test

Se construyó la correspondiente tabla de diferencia de medias para evaluar balance entre muestras, realizando tests de diferencia de medias para cada variable.

Variable	mean_train	mean_test	diff	p_value
horastrab	31,237	29,590	1,647	0,03496
num_miembros_hogar	3,639	3,587	0,051	0,26112
educ	11,523	11,629	-0,105	0,39958
edad	38,120	38,594	-0,474	0,40879

Tabla 1 Evaluación del Balance entre Conjuntos de Entrenamiento y Prueba

El alto valor p en todas las variables (> 0.70) confirma que no existen diferencias significativas entre los conjuntos de entrenamiento y prueba. Esto garantiza que el modelo no sufre sesgos de partición y que el conjunto de test es una muestra representativa del conjunto de train.

B. Regresión Logística

Se estimó un modelo Logit clásico sobre la base respondieron (train), que modelalaprobabilidad de pobreza mediante la función logística, asumiendo una relación lineal entre los predictoresy el log-odds de pobreza.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

logit(p): Es la función logit, que representa el logaritmo natural de la razón de probabilidades (odds) de un evento. Se calcula como el logaritmo natural de la probabilidad de que el evento ocurra (p) dividida por la probabilidad de que no ocurra (1-p).
p: Es la probabilidad de que un evento ocurra. En la regresión logística, la variable de resultado es binaria, usualmente codificada como 0 o 1 (por ejemplo, sí/no, éxito/fracaso).
β_o: Es la intersección o el término constante del modelo.
β_j: Son los coeficientes de regresión que representan el cambio en el logit de la probabilidadpor cada unidad de cambio en la variable correspondiente (X{j}), manteniendo las otras variables constantes.
∑_{j=1}^p $\beta_j X_j$: Es la suma de los productos de cada coeficiente (beta{j}) y su variable predictora correspondiente (X{j}), que representa la parte lineal de la relación.
Es esencial porque mapea el concepto de probabilidad (acotado entre 0 y 1) a una escala continua que va desde -infinito hasta +infinito. Esto nos permite usar una regresión lineal tradicional: el modelo asume que el Logaritmo de las Odds de Pobreza está determinado por una combinación lineal de las variables estructurales. Por ejemplo, si el Log-Odds aumentaen0. 0.402 por cada miembro adicional del hogar (como ocurre con num_miembros_hogar), las Odds de pobreza se multiplican por e^{0.402} (Odds Ratio), manteniendo una interpretación causal lineal en el Log-Odds, pero garantizando que la probabilidad final se mantenga lógicamente entre 0 y 1."

B1. Resultados del modelo Logit

Aunque el modelo Logit es robusto, su correcta interpretación requiere verificar ciertos supuestos:

Variables	coef	std_err	odds_ratio	pvalue
num_miembros_hogar	0,3674	0,0220	1,4440	1,36E-56
const	-0,7509	0,1374	0,4719	4,60E-02
educ	-0,0289	0,0066	0,9715	1,25E+01
horastrab	0,0015	0,0009	1,0015	8,95E+04
edad	0,0025	0,0015	1,0025	1,08E+05

Tabla 2 resume los coeficientes de la regresión y sus niveles de significancia:

- No Multicolinealidad Severa: Dado que solo se incluyeron variables numéricasestructurales, se asumió que la multicolinealidad es baja, con la excepción de edad y edad².
- Linealidad del Logit: La inclusión de edad² fue un intento explícito de capturar no-linealidad. Sin embargo, los resultados mostraron que el término cuadrático no es significativo, sugiriendo que la relación lineal es adecuada.

B2. Efectos marginales promedio – AME

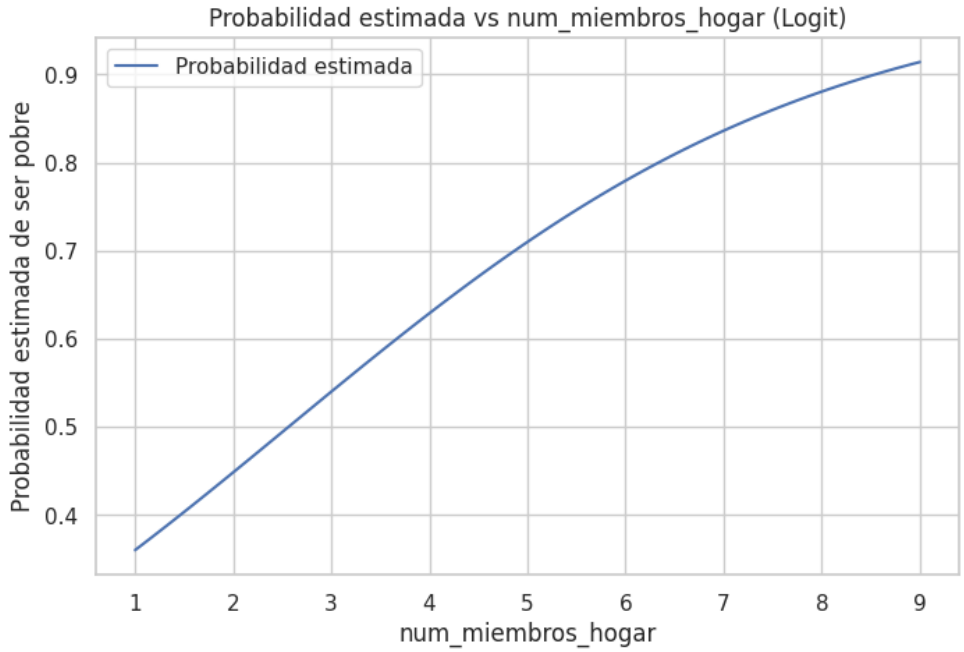
Los efectos marginales promedio (AME) facilitan la interpretación directa del cambio en la probabilidad de pobreza para un cambio unitario en el predictor, manteniendo otras variables constantes en sus valores promedio.

Variables	dy/dx	Std. Err.	z	Pr(>|z|)
edad	0,0006	0,0003	1,6085	1,08E+05
educ	-0,0064	0,0015	-4,3999	1,08E+01
num_miembros_hogar	0,0819	0,0044	18,6331	1,73E-71
horastrab	0,0003	0,0002	1,6995	8,92E+04

Tabla 3: Efectos marginales promedio – AME (Valores *p* redondeados)

Principales hallazgos:

- num_miembros_hogar: Es el predictor más fuerte. Un aumento en un miembro del hogar incrementa la probabilidad de pobreza en aprox 8.2 puntos porcentuales (AME positivo y altamente significativo).
- educ: Posee un efecto negativo y significativo. Un año adicional de educación reduce la probabilidad de pobreza en aprox 0.64 puntos porcentuales.
- horastrab: El efecto es muy pequeño y solo marginalmente significativo, sugiriendo que la cantidad de horas trabajadas es un factor menor comparado con la estructura del hogar o el capital educativo, *p*-value aprox 0.089.
- edad y edad²: Ambos términos no fueron significativos (el *p*-value de edad es aprox 0.099y edad² aprox 0.107 en coeficientes), indicando que el efecto no lineal de la edad es irrelevante para la predicción en este modelo.



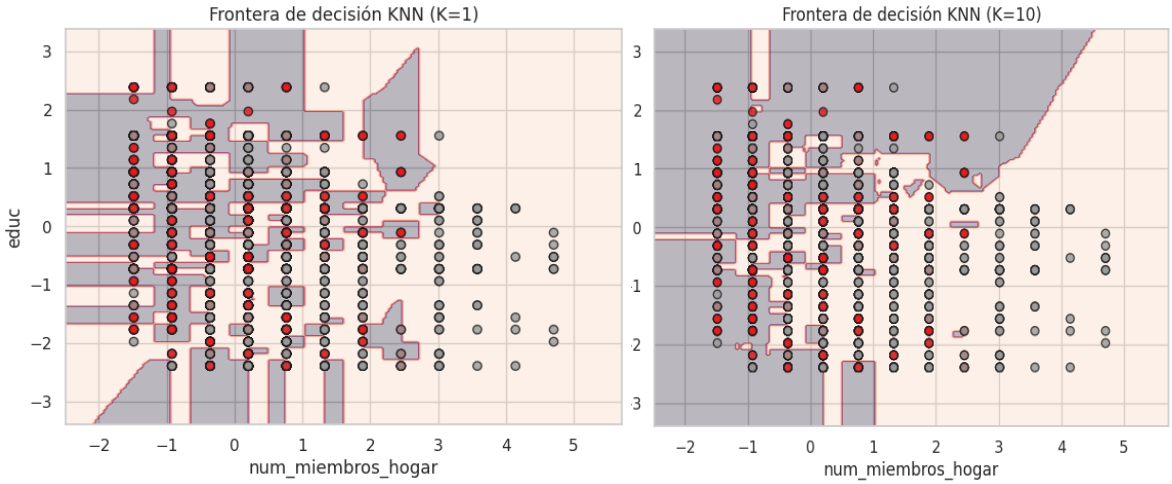
El gráfico 1 muestra que la probabilidad estimada de pobreza aumenta con el número de miembros del hogar, pasando de ~38% en hogares unipersonales a más de 85% en hogares de 7 o más miembros. Esto refleja el coeficiente positivo y altamente significativo de num_miembros_hogar (+0.402), que indica que cada miembro adicional incrementa en promedio 8.2 puntos porcentuales la probabilidad de ser pobre.

C. Modelo K-Nearest Neighbors (KNN)

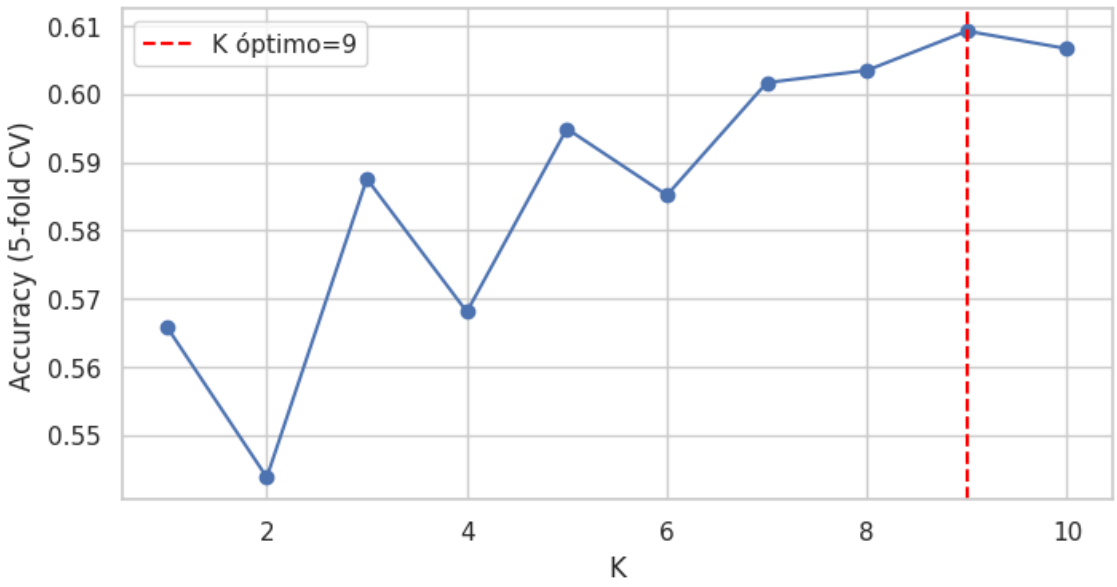
KNN se aplicó como modelo no paramétrico, buscando flexibilidad para capturar fronteras de decisión no lineales complejas.

C1. Validación Cruzada

Se realizó una búsqueda de hiperparámetros con K [1, 20] mediante 5-fold Cross Validation(CV) para optimizar el número de vecinos.



Los gráficos muestran el trade-off sesgo-varianza en KNN: con K=1, la frontera es excesivamente compleja y sobreajustada, creando regiones aisladas para puntos atípicos; en cambio, con K=10, la frontera es más suave y generalizable, reflejando patrones estructurales claros (mayor tamaño del hogar y menor educación aumentan la pobreza).



Esto justifica el uso de validación cruzada para seleccionar un K óptimo (K=9 en CV), que equilibra precisión y robustez.

D. Comparación de Desempeño y Selección de Modelo

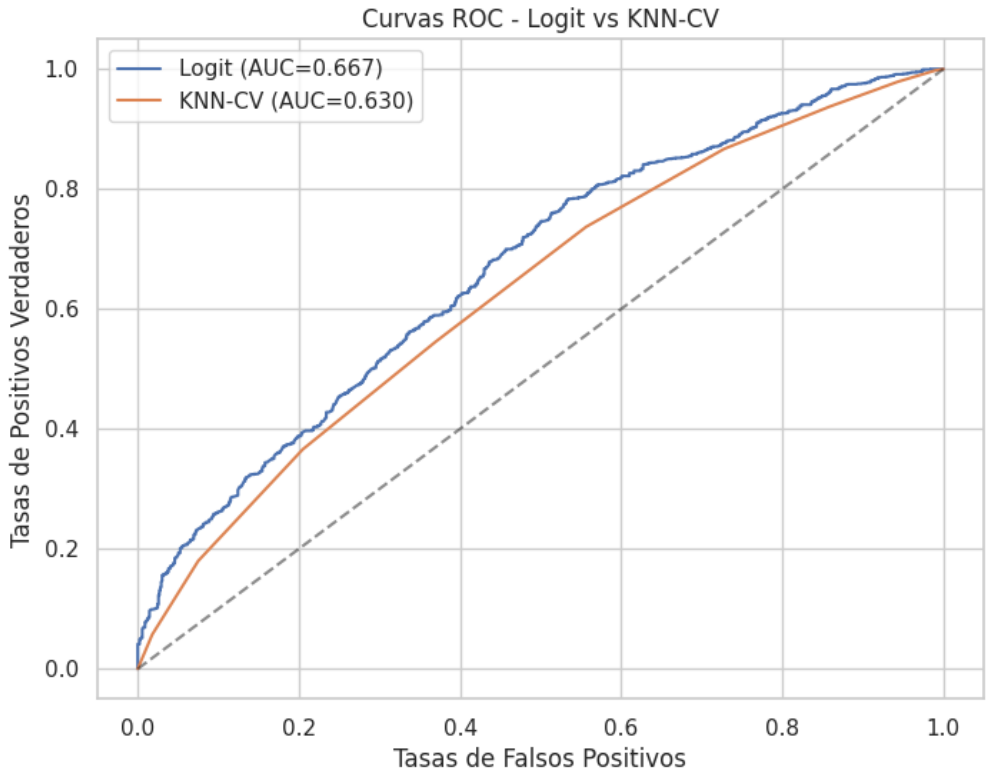
Se evaluó el desempeño del modelo Logit y el modelo KNN con K=9 sobre el conjunto de test para determinar el modelo óptimo para políticas públicas.

Métrica	Logit	KNN-CV
Accuracy	0,648	0,615
Recall	0,797	0,737
Precision	0,667	0,652
F1 Score	0,726	0,692
AUC	0,667	0,63

Tabla 4 Métricas de clasificación en Test

D1. Discusión de Política Pública (Recall vs. Falsos Negativos)

La Regresión Logística supera a KNN en todas las métricas relevantes (Accuracy, Recall, Precision, F1 y AUC). El AUC del Logit (aprox 0.667) indica un mejor poder de discriminación entre lasclases que el de KNN (aprox 0.630).



D2. Discusión de política pública (Error Tipo I y II)

En el contexto de un programa alimentario con recursos acotados, la elección del modelo debe ponderar el costo de los errores:

- Error Tipo I (Falso Positivo - FP): Asignar recursos a un no pobre Costo financiero.
- Error Tipo II (Falso Negativo - FN): No asignar ayuda a un pobre real Costo humanitario severo.

El objetivo es minimizar los falsos negativos (FN), lo que se traduce en maximizar la métrica Recall (Sensibilidad).

- Logit Recall = 79.7%
- KNN Recall = 73.7%

El modelo Logit es la elección correcta, pues es capaz de identificar un mayor porcentaje de pobres reales (79.7% vs 73.7%) y, por lo tanto, reduce el costo humanitario de las exclusiones (Falsos Negativos).

D.2. Análisis de Sensibilidad del Umbral de Clasificación

Aunque los resultados anteriores utilizan el umbral estándar de probabilidad de 0.5, el modelo Logit ofrece la ventaja de la flexibilidad operativa. La elección del umbral debe reflejar el trade-off entre Recall (cobertura) y Precision (fugas):

- Prioridad Humanitaria: Se elige un umbral más bajo (ej: 0.4) para aumentar el Recall y asegurar que más vulnerables reciban asistencia, a costa de aceptar más Falsos Positivos.
- Prioridad Financiera: Se elige un umbral más alto (ej: 0.6) para maximizar la Precision y minimizar las fugas de presupuesto, a costa de excluir a más pobres reales.

Esta capacidad de ajustar el umbral sin re-entrenamiento es una ventaja decisiva del Logit para la gestión dinámica de programas sociales según restricciones presupuestarias cambiantes.

D3. Predicción en la base Norepondieron (2025)

El modelo Logit, al ser el de mejor desempeño y mayor Recall, fue aplicado sobre la base norespondieron 2025 (individuos sin datos de ingreso) para estimar la proporción de pobres.

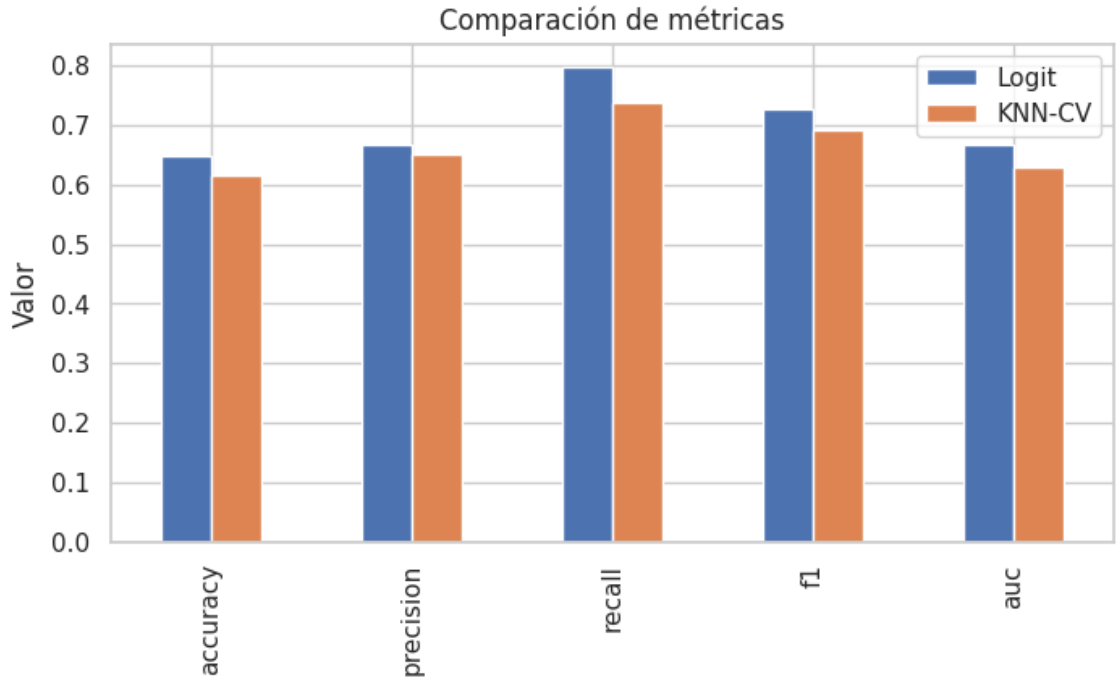
Discusión de Sesgo de Selección y Limitación de Generalización

El valor predicho del 38.80% se considera coherente con los valores esperados en contextos de alta pobreza. Sin embargo, es vital reconocer que esta predicción tiene una limitación estructural: el sesgo de selección.

La población que no responde a preguntas de ingresos no es una muestra aleatoria; presenta características sistemáticamente diferentes (ej: mayor informalidad). Por lo tanto, el modelo entrenado en la población respondiente puede extrapolar mal al aplicarse a la población no-respondiente. Si bien esta estimación permite la focalización ex-ante, se recomienda cautela y el uso de técnicas de corrección de sesgo (ej. modelos de Heckman) para mejorar la confiabilidad de las estimaciones en este subgrupo.

D4. Conclusiones Generales

El análisis comparativo confirma que la Regresión Logística es el modelo más adecuado para la predicción de pobreza en contextos de política social.



- Superioridad en Recall e Implicancia Humana: La ventaja del Logit es su Recall del 79.7%, sustancialmente superior al 73.7% de KNN. Esta diferencia es determinante, ya que reduce la tasa de Falsos Negativos y, por lo tanto, minimiza el costo humanitario de excluir a personas vulnerables de programas de asistencia.
- Interpretabilidad como Ventaja Decisiva: El modelo Logit es transparente y auditable, permitiendo explicar la decisión de clasificación mediante sus coeficientes (AME). Esta interpretabilidad es fundamental para rendir cuentas y para que los beneficiarios comprendan los criterios de elegibilidad, a diferencia de la "caja negra" que representaKNN.
- Factores Estructurales Clave: La pobreza está determinada principalmente por condiciones estructurales del hogar (num_miembros_hogar) y capital educativo (educ).
- Flexibilidad Operativa: El Logit permite ajustar el umbral de probabilidad para adaptarse a restricciones presupuestarias cambiantes, priorizando Recall o Precision segúnlanecesidad.

Recomendaciones para Implementación:

- Adoptar Regresión Logística con Umbral Ajustable: Implementar el modelo Logit, calibrando el umbral de probabilidad para balancear la cobertura (Recall) contra la eficiencia presupuestaria (Precision).
- Incorporar Variables de Hogar/Vivienda: Para elevar el poder predictivo (AUC) y aumentar la discriminación, se recomienda ampliar el conjunto de predictores con variables del cuestionario hogar de la EPH (ej. hacinamiento, NBI).
- Corrección de Sesgo de Selección: Para las predicciones sobre la población no-respondiente, implementar modelos que corrijan explícitamente el sesgo (ej. Heckman two-step), o diseñar estrategias para la recolección activa de datos de ingresos en este grupo vulnerable.

Índice

A. Enfoque de Validación....1
A1. Variables Seleccionadas (Matriz X)....1
A2. Diferencias de medias entre train y test....1
B. Regresión Logística....2
B1. Resultados del modelo Logit....2
B2. Efectos marginales promedio – AME....3
Principales hallazgos:....3
C. Modelo K-Nearest Neighbors (KNN)....4
C1. Validación Cruzada....4
D. Comparación de Desempeño y Selección de Modelo....5
D1. Discusión de Política Pública (Recall vs. Falsos Negativos)....5
D.2. Análisis de Sensibilidad del Umbral de Clasificación....6
D3. Predicción en la base Norepondieron (2025)....6
Discusión de Sesgo de Selección y Limitación de Generalización....6
D4. Conclusiones Generales....6
Recomendaciones para Implementación:....7