Trabajo Práctico 4 'Big Data and Machine Learning'

Integrantes: Andrea Duette, Uriel Masserdotti, Paulo Gonzalez

Resumen Preliminar

Como en los trabajos precedentes, este informe se enmarca en el análisis de la Encuesta Permanente de Hogares. El objetivo en esta oportunidad es predecir la condición de **"desocupado"** y **"salario semanal"** de un individuo a partir de variables organizadas, en previas ocasiones, en dos bases de datos ("respondieron" y "no respondieron") para los años 2004 y 2024 para la región del Noreste Argentino. La Parte 1 detalla la metodología del enfoque de validación, en la que se divide la base "respondieron" en conjuntos de entrenamiento y prueba, creando tablas de diferencias de media. En la Parte 2 estimamos una regresión lineal para predecir salarios semanales utilizando un set de variables explicativas desarrollado en trabajos previos. Por último, la Parte 3 consiste en definir el mejor modelo para estimar la condición de **"desocupado"** entre un modelo de regresión logística y un modelo de k-vecinos cercanos, aplicando el modelo resultante para la base "no respondieron".

Parte 1: Validación de la Partición de Datos

I

Para entrenar y evaluar los modelos de clasificación se dividió la base "respondieron" en un conjunto de entrenamiento (70%) y un conjunto de prueba (30%) con una semilla aleatoria fija en 444 que asegura la reproducibilidad de los datos. Luego, procedimos a crear tablas de diferencia de medias para cada año, con el fin de verificar si esta división generó subconjuntos de datos balanceados que representan de forma correcta a la muestra original.

En general, se observan buenos resultados para ambos años. Los valores de los *t-test* son generalmente bajos y los *p-values* están por encima del 0,05 en casi todas las variables, sugiriendo que no hay diferencias estadísticamente significativas en las medias de los conjuntos creados, lo que es una buena evidencia de que no se generó ningún sesgo a la hora de la partición de los datos.

La 'Tabla 1-1' muestra, a modo de ejemplo, un extracto de la tabla de diferencias de medias para el año 2024:

Variable	Conjunto de Entrenamiento			Conjunto de Prueba			Differ.	
	N train	X media	X desvío	N test	X media	X desvío	*T-test*	*P-value*
CH06	3.177	34,72	21,58	1.362	34,99	21,96	-0,39	0,7
EDUC	3.177	9,11	4,87	1.362	9,02	4,96	0,61	0,54
SALARIO_SEMANAL	1.254	49.278,94	37.746,81	523	52.223,1	47.480,67	-1,26	0,21
PP04A	1.276	1,74	0,45	530	1,72	0,46	0,79	0,43
PP10D	81	1,05	0,69	46	0,87	0,75	1,34	0,18

Tabla 1-1: Extracto de la tabla de diferencia de medias para 2024.

Variables como CH06 **(edad),** PP04A **(sector de la activad en que se trabaja)**, EDUC **(nivel educativo)** presentan diferencias mínimas entre los conjuntos. El salario semanal presenta cierta diferencia, pero sin ser demasiado significativa relativamente. La única variable con una diferencia relativamente más notoria es PP10D **(ha trabajado alguna vez)**, que presenta una menor cantidad de observaciones y una diferencia pequeña en media, aunque el *p-value* sigue siendo no significativo.

Concretamos que la partición de los datos conservó las propiedades estadísticas de la base de datos original, lo que valida la creación y uso de los conjuntos de entrenamiento y testeo en lo que continúa de nuestro informe.

Parte 2: Método de Regresión Lineal

II

Procedemos a estimar un modelo de regresión lineal para predecir la variable **"salario semanal"** (esta es nuestra variable dependiente) de los individuos de la base "respondieron". Las variables explicativas fueron definidas en trabajos anteriores y seleccionadas por su relevancia a la hora de predecir el salario semanal. En la 'Tabla 2-1' se ilustran los resultados de la estimación y se reportan los coeficientes estimados junto a los desvíos estándar correspondientes entre paréntesis.

Var. Dep.: ***salario_semanal***	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Cantidad de Variables	(1)	(2)	(3)	(4)	(6)
Edad	506,2 (77,23)	3.292,5 (449,14)	2.413,18 (430,73)	2.571,6 (420,80)	**2.379,43** **(421,75)**
Edad2		-33,5 (5,32)	-20,8 (5,12)	-22,4 (5,00)	**-20,6** **(5,00)**
Educ			3.585 (222,76)	3.928,6 (219,73)	**3.735,4** **(222,24)**
Mujer				-19.807,2 (1.799,95)	**-20.814,66** **(1.807,73)**
PP04A					**-9.494** **(1.874,98)**
PP03D					**-546,7** **(1.595,40)**
N (Observaciones)	*2.457*	*2.457*	*2.457*	*2.457*	*2.457*
R^2	*0,017*	*0,033*	*0,125*	*0,166*	*0,175*

Tabla 2-1: Modelos de estimación de salarios semanales (regresión lineal) usando conjunto de entrenamiento.
Nota: Todos los *p-values* de los coeficientes reportados son menores que 0,001.

III

Comentando brevemente la 'Tabla 2-1', podemos afirmar que todas las variables son estadísticamente significativas a la hora de estimar el salario semanal. Hablando particularmente de las variables, hay ciertos valores que llaman la atención. Primero, a partir del **"Modelo 2"** parece contradictorio ver que CH06 **(edad)** y EDAD2 **(edad al cuadrado)** tienen signos opuestos. Desde el lado de la lógica puede responderse esta supuesta contradicción afirmando que EDAD2 captura el efecto no lineal (cuadrático) de la edad en el salario semanal, traducido como un ingreso marginalmente decreciente a medida que aumenta la edad. Esto no le quita sentido a una CH06 positiva, pues puede representar la relación positiva entre salario semanal y años de experiencia laboral. Piénsese una función cuadrática cóncava, donde el máximo de la curva puede calcularse de la forma:

$$x = -\frac{b}{2*a}.$$

Trayendo este concepto a nuestro análisis, se puede calcular la edad en la que se alcanza el mayor salario semanal (el pico de la curva) para cada modelo, de la forma:

$$Edad\ pico = -\frac{\beta_{edad}}{2*\beta_{edad2}}.$$

Por ejemplo, el resultado para el **"Modelo 5"** es de 57,8 años.

Luego, hay que resaltar la *dummy* de la variable CH04=2 **(mujer)**, siendo significativamente negativa en ambos modelos en el que se la usa. Presumiblemente, señala que las mujeres tienden a tener un salario semanal menor que los hombres (teniendo presentes las otras variables), lo que refuerza la idea de una brecha salarial de género que perjudica a las mujeres.

IV

Ahora, utilizando el conjunto de prueba, para evaluar el desempeño de los modelos, procedemos a calcular métricas M.S.E., R.M.S.E. y M.A.E. con los coeficientes estimados en los modelos de regresión lineal del apartado anterior. La 'Tabla 2-2' contiene dichas métricas.

Var. Dep.: ***salario_semanal***	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Cantidad de Variables	(1)	(2)	(3)	(4)	(6)
M.S.E. *test*	2.627.690.000	2.587.054.000	2.286.798.000	2.168.626.000	2.134.338.000
R.M.S.E. *test*	51.261,0	50.863,1	47.820,7	46.568,5	46.199,5
M.A.E. *test*	3.176.276	3.139.231	2.903.057	2.810.578	2.783.365

Tabla 2-2: Performance por regresión lineal de la predicción de salarios usando la base de testeo.

Si bien los resultados muestran una tendencia descendiente clara a medida que se van sumando variables al modelo, disminuyendo las métricas de error, confirmando que la inclusión de más variables permite capturar una mayor proporción de la varianza en los salarios. Sin embargo, las mismas magnitudes de las métricas son considerablemente elevadas. Esto se debe a que son salarios en pesos que no están corregidos por inflación.

Para mostrar de forma un poco más ilustrativa estas métricas decidimos tomarnos la libertad de aplicar una transformación logarítmica a las variables debido a las diferencias de escala, principalmente desde el lado del salario semanal, que al no estar ajustado y al poder haber valores atípicos que generen mucha distorsión, eso hace aumentar el error en el modelo de regresión. Además, estandarizamos el resto de variables, que de por sí tienen escalas diferentes.

Var. Dep.: ***Log(salario_semanal)***	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Cantidad de Variables Estandarizadas	(1)	(2)	(3)	(4)	(6)
M.S.E. *test*	**0,65**	**0,64**	**0,53**	**0,49**	**0,48**
R.M.S.E. *test*	**0,81**	**0,80**	**0,73**	**0,70**	**0,69**
M.A.E. *test*	**0,63**	**0,62**	**0,56**	**0,53**	**0,53**

Tabla 2-3: Performance por regresión lineal de las variables con conversiones logarítmicas y estandarizadas.

La ‘Tabla 2-3’ muestra lo descripto anteriormente. Comentando nuevamente el efecto de la variable CH06 y EDAD2, en el “Modelo 2” se puede apreciar que el añadir esta última no aporta mucha capacidad explicativa al modelo. Puede deberse a la colinealidad de estas variables que, obviamente, están muy relacionadas. La variable EDUC, por otro lado, aporta una buena cuota de capacidad explicativa.

Dado que usamos una transformación logarítmica, las métricas están en esta escala. Para interpretarlas en pesos basta con volver a la ‘Tabla 2-2’. Hay que tener en cuenta que, aunque las métricas puedan parecer bajas a priori, la magnitud real de los errores en la predicción puede seguir siendo alta. Para comprobar esto basta con seguir la siguiente formula:

$$e^{(MAE)} - 1 \cong error\ relativo\ aproximado\ (\%)$$

Por ejemplo, si tomamos el M.A.E. del “Modelo 3” el error se aproxima al 70%, un valor significativamente alto que señala como la predicción de los salarios semanales puede desviarse hasta un 70% de la media real. Vale la pena reiterar que esto puede deberse a que los salarios no están ajustados por inflación.

A modo ilustrativo, la ‘Figura 2-1’ refleja las tendencias salariales por edad. Se puede apreciar un crecimiento del salario semanal con la edad en los primeros años de vida laboral, alcanzando un pico alrededor de los 55 y 60 años (lo que concuerda con el pico de edad calculado en el anterior apartado). A partir de estas edades, el salario comienza a disminuir mientras se acerca a edades jubilatorias. Aun así, se nota cierta dispersión en los datos, evidenciando factores que escapan al modelo.

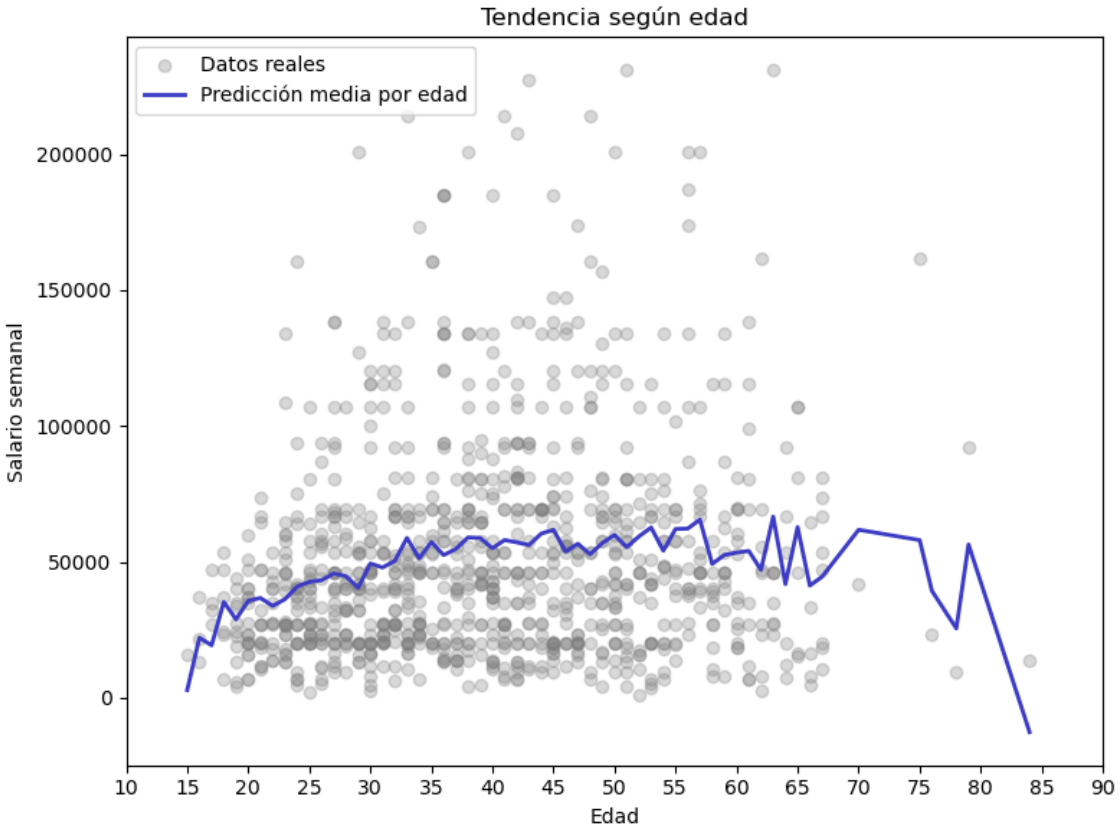


Figura 2-1: Relación entre la edad de los individuos (eje horizontal) y su salario semanal (eje vertical).

Parte 3: Clasificación y Evaluación de Desempeño

V

En este punto comenzamos a aplicar dos métodos de clasificación supervisada: regresión logística y k-vecinos cercanos (k = 5) para predecir la condición **"desocupado"** en la base "respondieron" para ambos años. Los datos se dividieron en conjuntos de entrenamiento y prueba, y se evaluó el desempeño de ambos métodos mediante matrices de confusión, curvas R.O.C., A.U.C., y *accuracy*. Finalmente, se utilizó el mejor método para predecir la desocupación en la base "no respondieron".

Comparando ambos métodos para ambos años (ver figuras debajo), la regresión logística superó al método de k-vecinos cercanos en ambos años, con *accuracies* más altos, sugiriendo mayor capacidad para identificar desocupados. Puede que k-vecinos cercanos se vea limitado por la cantidad de vecinos delimitada por la consigna, pero, en conclusión, se eligió la regresión logística para predecir la desocupación en el próximo apartado.

Si bien se presentan *accuracies* que tienden a 1, esto puede deberse al reducido tamaño de la base "no respondieron" (de únicamente 11 casos totales).

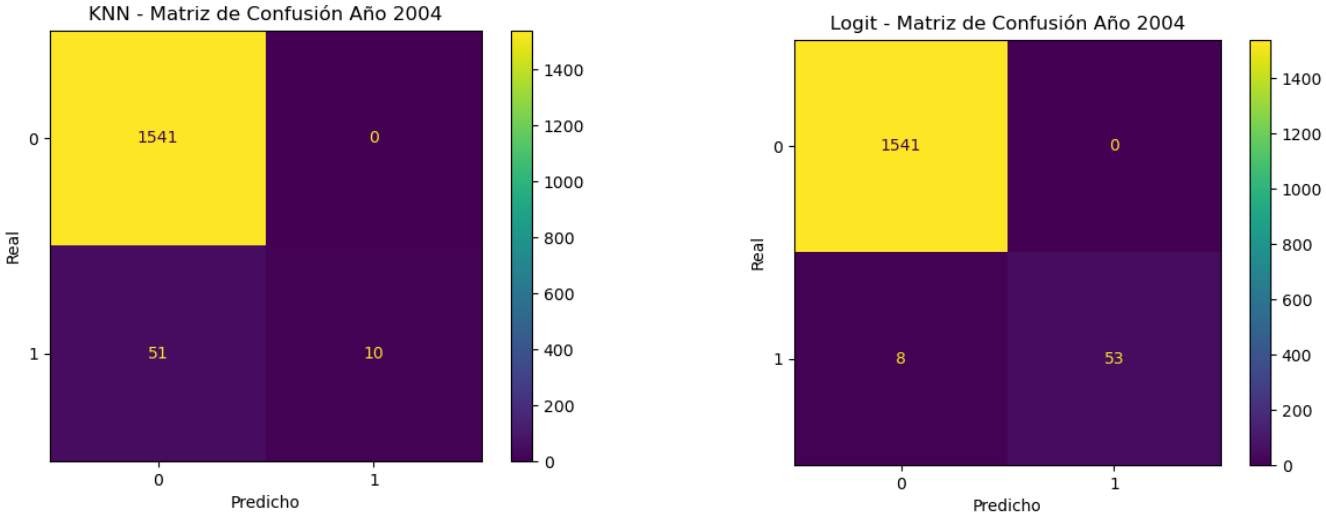


Figura 3-1 y 3-2: Matrices de confusión para los modelos KNN (izquierda) y Logit (derecha) para el 2.004.

En la 'Figura 3-1' y '3-2' se observan las matrices de confusión para ambos modelos. Se ve una diferencia notable a favor del modelo de regresión logística en los falsos negativos y en los verdaderos negativos. De la misma forma, las curvas R.O.C. de la 'Figura 3-3' muestra un A.U.C. sobresalientemente mayor para el modelo de regresión logística. Todo esto para el año 2.004.

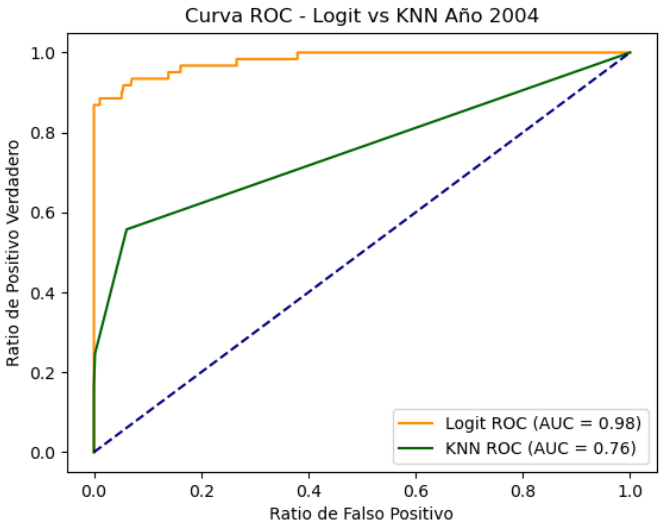
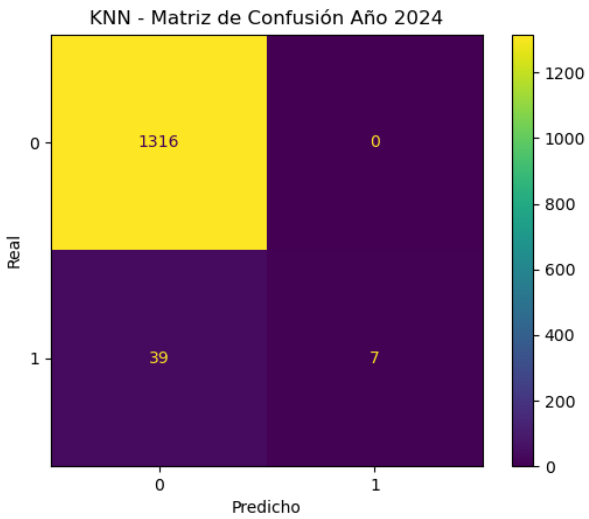


Figura 3-3: Curvas R.O.C. para ambos modelos, Logit (amarillo) y KNN (verde) para el 2.004.



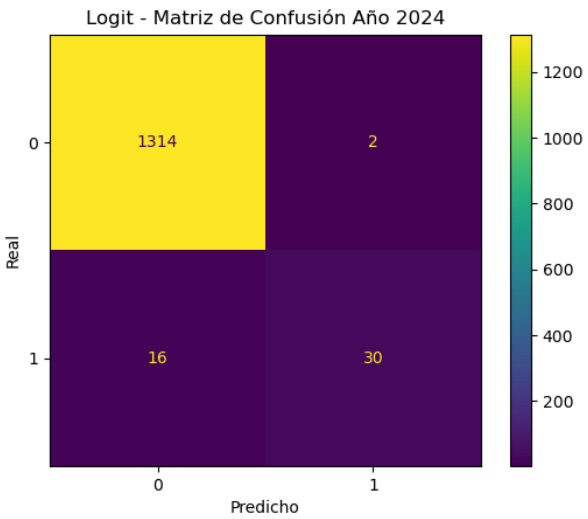


Figura 3-4 y 3-5: Matrices de confusión para los modelos KNN (izquierda) y Logit (derecha) para el 2.024.

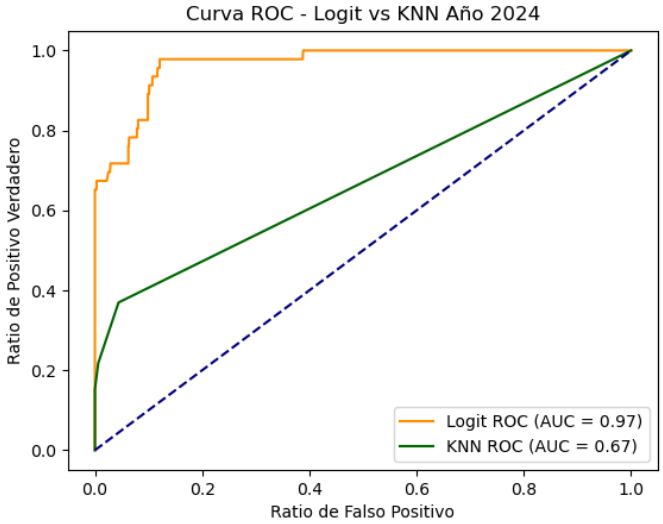


Figura 3-6: Curvas R.O.C. para ambos modelos, Logit (amarillo) y KNN (verde) para el 2.024.

En la 'Figura 3-4' y '3-5' se observan las matrices de confusión para ambos modelos para el año 2.024. Nuevamente se observa mayor robustez de datos, mayor inclusive en este año, para el modelo de regresión logística. Se puede confirmar que esta última superó a k-vecinos cercanos en ambos años. Sin embargo, se genera una pregunta nacida de algo que ya mencionamos antes:

Si el tamaño de la base "no respondieron" es tan limitado, y eso afecta la predicción y el *accuracy* en ambos modelos, ¿cómo se vería una aplicación del mejor modelo (logit) para la base "respondieron"?

VI

Antes de responder esto, procedemos a hacer algunas salvedades. Primero, dividimos la predicción de desocupación por año, entre 2.004 y 2.024, y luego entre las bases "respondieron" y "no respondieron". ¿Por qué decidimos hacerlo así? Porque este enfoque nos permite comparar la proporción de desocupados predichos en ambas bases usando el mismo modelo de regresión logística, lo que es útil para analizar si el modelo entrenado generaliza bien a ambas bases o si hay diferencias significativas. Además, dividirlo por año mantiene cierta consistencia con lo que se viene trabajando hasta ahora en este apartado.

Para 2.004, en "no respondieron" se predijeron 0 desocupados sobre un total de 5 (0%), y en "respondieron" se predijeron 174 desocupados sobre un total de 5.338 (3,26%), mientras que para 2.024, en "no respondieron" se predijeron 0 desocupados sobre un total de 6 (0%), y en "respondieron" se predijeron 97 desocupados sobre un total de 4.844 (2%), sugiriendo que las características de "no respondieron" no se alinean con los patrones de desocupación identificados por el modelo en ninguno de los años, mientras que las predicciones en "respondieron" reflejan proporciones más realistas, con una leve disminución en la tasa de desocupación predicha entre 2.004 y 2.024, lo que podría indicar cambios en las condiciones económicas capturadas por las variables.

Conclusiones

La partición de datos en conjuntos de entrenamiento y prueba resultó aceptable, mostrando subconjuntos representativos sin sesgos significativos. La regresión lineal para predecir salarios reveló una relación no lineal con la edad, alcanzando un pico a los 57,8 años, y una brecha salarial de género que perjudica a las mujeres, con métricas de error descendentes al incluir más variables, aunque las magnitudes reflejan salarios no ajustados por inflación.

En la predicción de desocupación, la regresión logística superó a k-vecinos cercanos, con *accuracies* de 0.9950 (2004) y 0.9867 (2024), y mejor capacidad para identificar desocupados, aplicándose a “no respondieron” (0% de desocupados predichos) y “respondieron” (3,26% en 2.004 y 2% en 2.024), reflejando posibles mejoras económicas entre los años, aunque la predicción en “no respondieron” sugiere limitaciones por el tamaño de la muestra. Estos resultados destacan la utilidad de los modelos supervisados para analizar condiciones laborales, pero también la necesidad de bases más grandes y datos ajustados para mejorar las predicciones.