**UNIVERSIDAD DE BUENOS AIRES - 2025**

***BIG DATA & MACHINE LEARNING***

**Trabajo Práctico N°4**

***grupo 7***

**Integrantes:**

* **Lucas Giardina**
* **yenifer gonzales**
* **franco demir**

1. Enfoque de validación

1)

2004)

Tabla

Descripción generada automáticamente

2024)

Tabla

Descripción generada automáticamente

GRÁFICO TABLA DE MEDIAS DE DIFERENCIAS

Entre 2004 y 2024 se observa una fuerte reducción del salario promedio semanal, que pasó de aproximadamente $5,364 a $2,912, lo que representa una caída cercana al 46%. Además, la dispersión salarial se redujo notablemente en 2024, reflejando un mercado más homogéneo en términos de ingresos.

En cuanto a la educación, se evidencia un leve aumento en los años promedio de escolaridad (de 7.5 a 7.8 años), acompañado por una reducción en la desigualdad educativa, ya que la variabilidad también disminuyó.

Respecto a la variable edad al cuadrado (edad²), su media aumentó, lo que sugiere un posible envejecimiento de la población económicamente activa, aunque la estructura etaria entre los conjuntos de entrenamiento y prueba se mantuvo balanceada.

Por último, en términos del mercado laboral, se destaca una disminución del 15% en las horas trabajadas semanalmente, con una mayor homogeneidad en la jornada laboral de los individuos en 2024 en comparación con 2004.

1. *método Supervisado 1: Modelo de Regresión Lineal*

2) 2004

Tabla

Descripción generada automáticamente

2024

Tabla

Descripción generada automáticamenteconclusion

Elegimos variables 1 y 2 lo que es el Ingreso total familiar y el Ingreso per cápita familiar real para asociarla directamente con el salario semanal, estas variables explicativas nos ayudan a capturar el contexto económico familiar o los ingresos complementarios que pueden afectar al salario con el que cuenta el individuo.

3)

Performance por regresión lineal de la predicción de salario

***AÑO 2004***

| **Año** | **Modelo** | **MSE** | **RMSE** | **MAE** | **Filas\_train** | **Filas\_test** |
| --- | --- | --- | --- | --- | --- | --- |
| 2004 | Modelo 1 | 1.07914e+08 | 10388.2 | 6756.84 | 10206 | 4392 |
| 2004 | Modelo 2 | 9.41855e+07 | 9704.92 | 6149.75 | 10206 | 4392 |
| 2004 | Modelo 3 | 1.14348e+08 | 10693.4 | 7184.79 | 7905 | 3405 |
| 2004 | Modelo 4 | 1.06894e+08 | 10338.9 | 6897.55 | 7905 | 3405 |
| 2004 | Modelo 5 | 9.45734e+07 | 9724.89 | 6529.39 | 7905 | 3405 |

***año 2024***

| **Año** | **Modelo** | **MSE** | **RMSE** | **MAE** | **Filas\_train** | **Filas\_test** |
| --- | --- | --- | --- | --- | --- | --- |
| 2024 | Modelo 1 | 2.962e+07 | 5442.43 | 3787.80 | 9000 | 3897 |
| 2024 | Modelo 2 | 2.40999e+07 | 4909.16 | 3219.67 | 9000 | 3897 |
| 2024 | Modelo 3 | 1.57522e+07 | 3968.90 | 2521.00 | 6179 | 2640 |
| 2024 | Modelo 4 | 1.4592e+07 | 3819.94 | 2433.19 | 6179 | 2640 |
| 2024 | Modelo 5 | 1.3762e+07 | 3709.72 | 2473.47 | 6179 | 2640 |

MSE y RMSE:

Los valores de MSE y RMSE son relativamente grandes en ambas bases (2004 y 2024), indicando que las predicciones tienen errores absolutos importantes. Esto puede deberse a la alta variabilidad y dispersión en los valores de salario semanal.

MAE

El MAE muestra que, en promedio, los errores de predicción oscilan entre $2,400 y $7,200 aproximadamente. En términos prácticos, esto indica que las predicciones suelen desviarse del valor real del salario semanal por miles de pesos.

CONCLUSION

Aunque los modelos mejoran al incorporar más variables, la precisión sigue limitada, probablemente por la alta dispersión del salario semanal y posibles factores no contemplados en los modelos.

PREDICCION

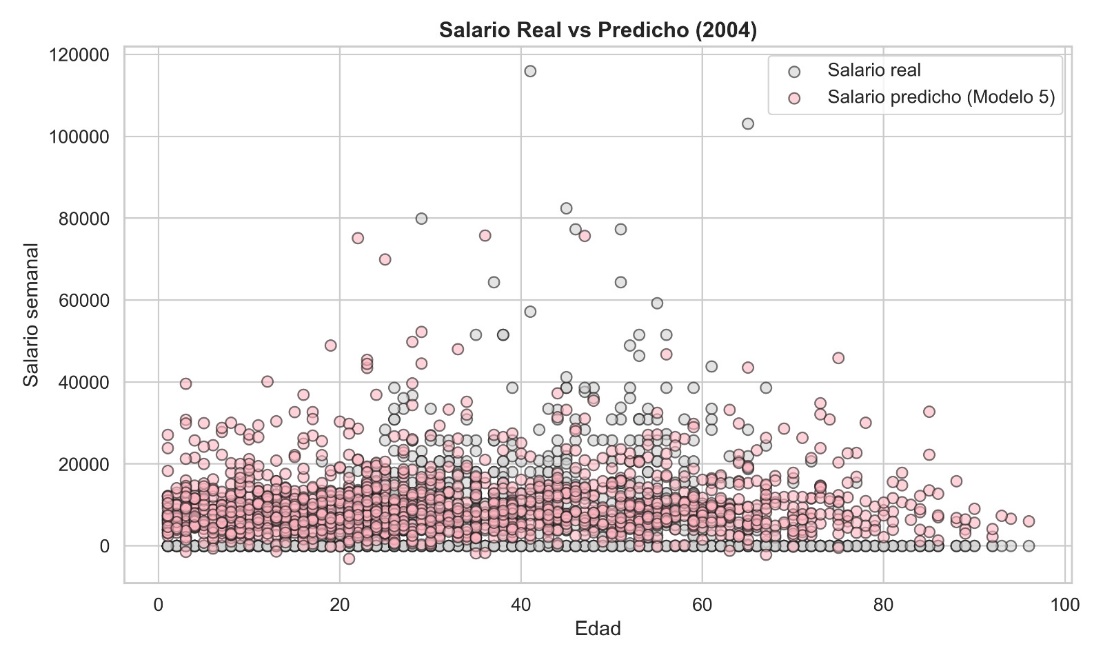
* ACLARACION

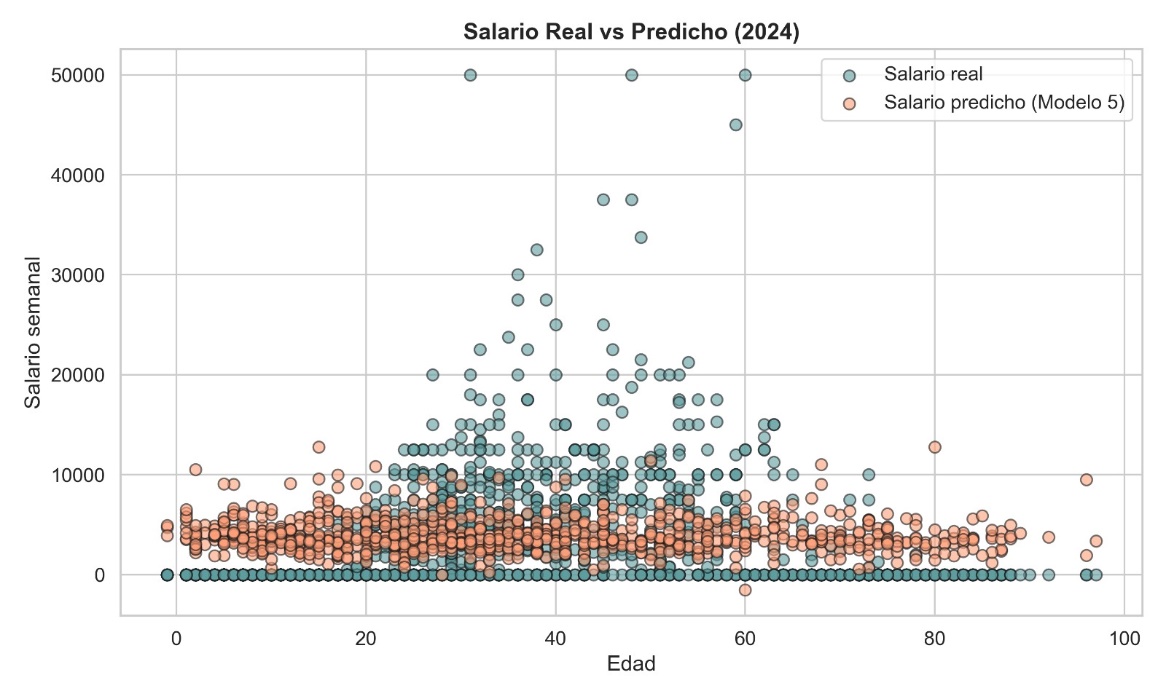
Se trabajó con un archivo Excel que contiene una gran cantidad de observaciones correspondientes al salario semanal real y a los salarios predichos por distintos modelos (Modelo 1 a Modelo 5), utilizando exclusivamente los datos del conjunto de testeo. Si bien el archivo completo incluye una muestra mucho más amplia, en la siguiente tabla se presentan algunas filas seleccionadas con fines ilustrativos, con el objetivo de mostrar cómo se comportaron las predicciones de cada modelo frente a distintos valores reales.

| **Salario semanal** | **Modelo 1** | **Modelo 2** | **Modelo 3** | **Modelo 3 ext** | **Modelo 5** |
| --- | --- | --- | --- | --- | --- |
| 15000 | 2741.42 | 4746.26 | 3966.56 | 4827.03 | 9485.99 |
| 11250 | 3238.09 | 5505.43 |  |  |  |
| 15000 | 3238.09 | 5505.43 |  |  |  |
| 10000 | 3450.95 | 5315.27 | 4377.44 | 3512.32 | 4172.39 |
| 10500 | 2634.99 | 4364.49 | 3675.43 | 4537.49 | 5430.73 |
| 12500 | 3238.09 | 5505.43 |  |  |  |
| 11250 | 2989.76 | 5336.34 | 4110.99 | 4946.76 | 6430.23 |
| 10000 | 3734.77 | 4580.60 | 3192.67 | 4093.49 | 3642.63 |
| 5000 | 3273.57 | 5495.21 | 4524.17 | 3636.34 | 5323.97 |
| 6250 | 3131.67 | 5484.51 | 3917.72 | 4732.25 | 3500.05 |
| 1500 | 3167.14 | 5500.07 | 3928.29 | 2969.35 | 3295.88 |
| 5500 | 2493.09 | 3735.18 | 3194.13 | 4062.03 | 2165.78 |

Estos valores corresponden al salario predicho (𝑠𝑎𝑙𝑎𝑟𝑖𝑜\_𝑠𝑒𝑚𝑎𝑛𝑎𝑙\_𝑡𝑒𝑠𝑡 con sombrero) para cada modelo, calculado utilizando los coeficientes estimados previamente en la etapa de entrenamiento. La comparación entre el salario real y el predicho permite evaluar qué tan bien se desempeñó cada modelo

4)





conclusion

Al comparar los resultados del Modelo 5 para los años 2004 y 2024, se observa un avance significativo en la capacidad del modelo para predecir salarios en función de la edad. En 2004, las predicciones muestran una mayor dispersión y desvío respecto al salario real, especialmente en los rangos más altos. En cambio, en 2024, el modelo logra ajustarse mejor a la distribución real, siguiendo más de cerca la tendencia general del comportamiento salarial.  
Esto sugiere no solo una mejora en el modelo, sino también un mayor entendimiento de los patrons salariales a lo largo del tiempo. Aunque aún hay margen para perfeccionar la precisión en casos extremos, los resultados son bastante alentadores para su aplicación práctica en análisis y proyecciones.

c) *Métodos de Clasificación y Performance*

*5)*

* Base de entrenamiento

| **Métrica** | **Regresión Logística (2024)** | **KNN k=5 (2004)** |
| --- | --- | --- |
| **Accuracy** | 0.9646 | 0.9391 |
| **Precision clase 1** | 0.4615 | 0.7053 |
| **Recall clase 1** | 0.0187 | 0.2036 |
| **F1-score clase 1** | 0.0359 | 0.3160 |
| **Precision clase 0** | 0.9653 | 0.9438 |
| **Recall clase 0** | 0.9992 | 0.9937 |
| **F1-score clase 0** | 0.9820 | 0.9681 |
| **Macro Promedio (F1)** | 0.5089 | 0.6421 |
| **Weighted Prom. (F1)** | 0.9485 | 0.9230 |
| **Soporte clase 1** | 321 | 717 |
| **Soporte clase 0** | 8769 | 9657 |
| **Total de muestras** | 9090 | 10374 |

Los modelos lograron una alta precisión global, con un accuracy de aproximadamente 96.5% para la regresión logística y 93.9% para KNN, lo que indica que ambos funcionan bien en términos generales. sim embargo, los datos están claramente desbalanceados, ya que la clase mayoritaria domina la muestra. Esto hace que el accuracy sea una métrica poco representativa para evaluar el desempeño en la clase minoritaria. La regresión logística apenas detecta casos de la clase minoritaria, reflejado en un recall muy bajo para esa clase, mientras que knn mejora algo la detección, pero sigue siendo insuficiente. Ambos modelos muestran buena precisión para la clase mayoritaria, pero no logran capturar bien la clase minoritaria, lo que afecta negativamente las métricas de recall y f1-score para dicha clase. Por eso, sería recomendable emplear técnicas para balancear las clases o ajustar los umbrales de decisión, y considerar métricas que reflejen mejor el desempeño en la clase minoritaria.

* BASE DE TESTEO

***Regresion logistica (2024)***

***Matriz de confusion***

|  | **Predicho: No desocupado** | **Predicho: Desocupado** |
| --- | --- | --- |
| **Real: No desocupado** | 3762 | 3 |
| **Real: Desocupado** | 129 | 3 |

Accuracy Regresión Logística (2024): 0.9661

AUC Regresión Logística (2024): 0.8435

***KNN(K=5) (2004)***

***Matriz de confusion***

|  | **Predicho: No desocupado** | **Predicho: Desocupado** |
| --- | --- | --- |
| **Real: No desocupado** | 4077 | 52 |
| **Real: Desocupado** | 294 | 24 |

Accuracy KNN k=5 (2004): 0.9222

AUC KNN k=5 (2004): 0.7287

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

La regresión logística presenta una mayor precisión global, mostrando un desempeño general más sólido. sim embargo, tiene dificultades para identificar correctamente a los desocupados, ya que su sensibilidad es baja. por otro lado, el modelo KNN logra una mejor detección de los casos positivos, aunque esto se da a costa de una menor precisión total. la curva roc del modelo de regresión logística muestra un mejor comportamiento general, lo que refuerza su capacidad para distinguir entre clases. En función del objetivo del análisis, Por mas que sea maximizar la precisión o mejorar la detección de desocupados, puede preferirse uno u otro modelo.

¿ Que metodo predice mejor?

Si se prioriza la precisión global y una buena capacidad para separar correctamente las clases, el modelo de regresión logística correspondiente al año 2024 resulta ser el más adecuado. Este modelo muestra una alta exactitud general y una curva roc más cercana al óptimo, lo que indica un buen desempeño discriminativo. en cambio, si el propósito principal del análisis es identificar la mayor cantidad posible de casos positivos, aunque eso implique aceptar un mayor número de falsos positivos, el modelo knn del año 2004 ofrece un mejor rendimiento en ese aspecto. la elección del modelo más apropiado dependerá en última instancia de los objetivos específicos del análisis y del costo relativo asociado a los distintos tipos de error.

6) Utilizamos el modelo de regresión logística con los datos del año 2024, se realizó la predicción sobre la base de personas que no respondieron. De un total de 3897 individuos, el modelo identificó 6 personas como desocupadas. Esto representa una proporción estimada de desocupados del 0.15% entre los no respondientes. Este bajo porcentaje sugiere que, según el modelo, la mayoría de los no respondientes probablemente no se encuentran en situación de desocupación.