Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad 3.2 (Regresión No Lineal)

Analítica de datos y herramientas de inteligencia artificial II (Gpo 501)

Alfredo García Suarez

EQUIPO 2

André Calmus González A017333529

Diego Sánchez Márquez A01734778

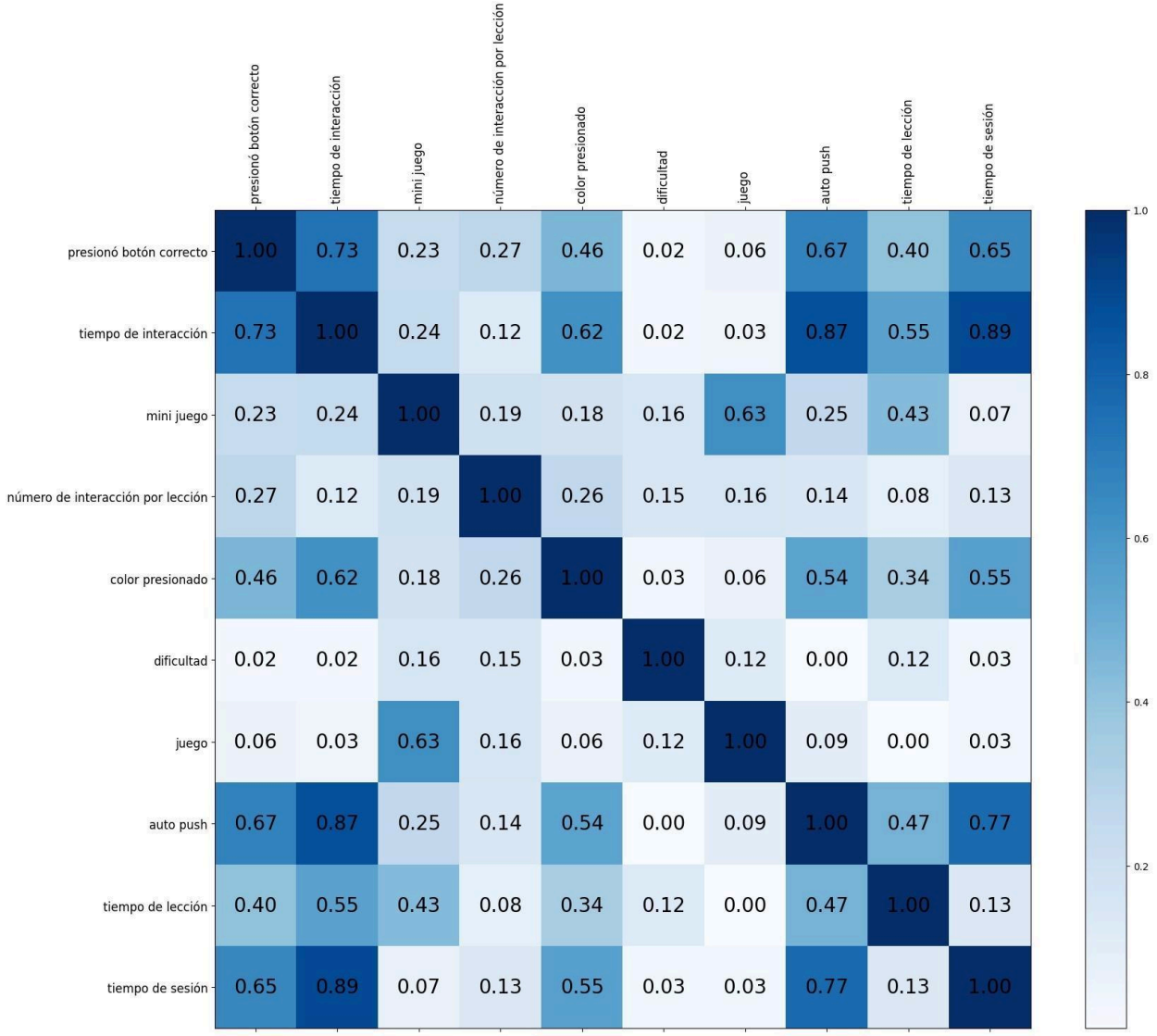
Emilio Rugerio Pastrana A01737819

Ximena Italya Jimenez Huerta A01277667

Tecnológico de Monterrey

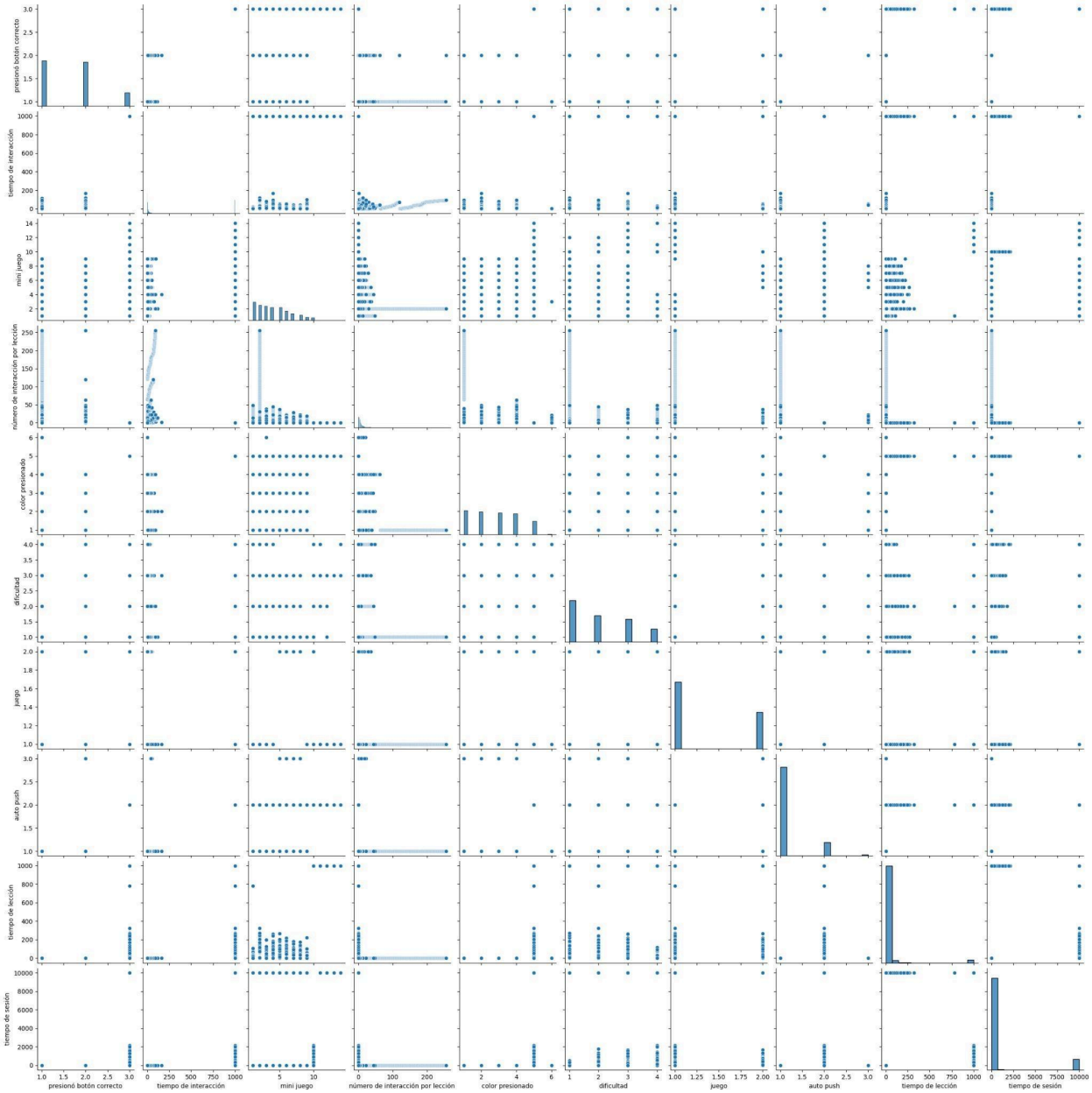
12 de Abril del 2025

Para poder realizar un análisis no lineal de los datos de Wuppi, primero tuvimos que eliminar las columnas que no eran realmente necesarias para el estudio (fecha, administrador y usuario). Se seleccionó la columna objetivo que es la de más importancia de acuerdo con estudios lineales previos que es ‘presionó botón correcto’, sacamos la correlación de la información adquirida por el socio anteriormente, lo que ayudó a la elaboración del Heatmap.



Tecnológico de Monterrey

De esta misma información ‘limpia’ obtuvimos modelos lineales de la dispersión entre todas las variables:



Con esto ya sabemos que todas las columnas son numéricas; modificadas anteriormente.

Obtuvimos el modelo lineal simple que dió como resultado lo siguiente:

La mejor variable correlacionada con presionó botón correcto es tiempo de interacción con una correlación de 0.73

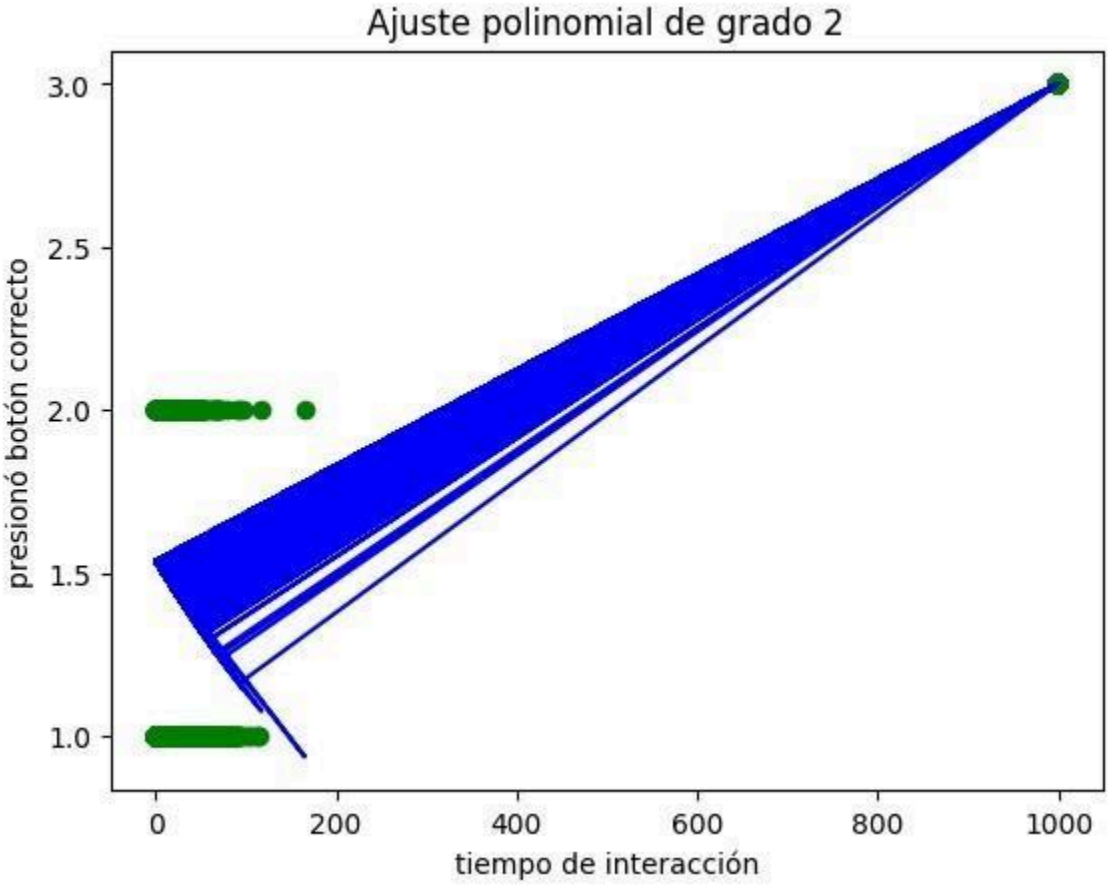
También se calculó un modelo no lineal:

R^2 del modelo polinomial de grado 2: 0.55

R^2 de la correlación lineal simple: 0.54

Lo que demuestra que el modelo no lineal supera por .01 el modelo lineal; por lo tanto, la relación entre ‘tiempo de interacción’ y ‘presionó botón correcto’ no es perfectamente lineal.

Arrojando la siguiente gráfica:



La forma de abanico indica que la varianza cambia con el tiempo de interacción. Al inicio hay muchos datos agrupados; conforme el tiempo aumenta, los datos se dispersan más.

Como conslución, se observó que la correlación lineal simple entre "tiempo de interacción" y "presión botón correcto" presentó un R^2 de 0.54. Posteriormente, al ajustar un modelo polinomial de grado 2, se obtuvo un R^2 de 0.55, evidenciando una ligera mejora en la capacidad de ajuste del modelo no lineal respecto al lineal.
La dispersión de los datos sugiere una relación no estrictamente lineal, apoyada visualmente por la forma de abanico en la gráfica, lo cual justifica el pequeño incremento en el desempeño al emplear un modelo de mayor complejidad.

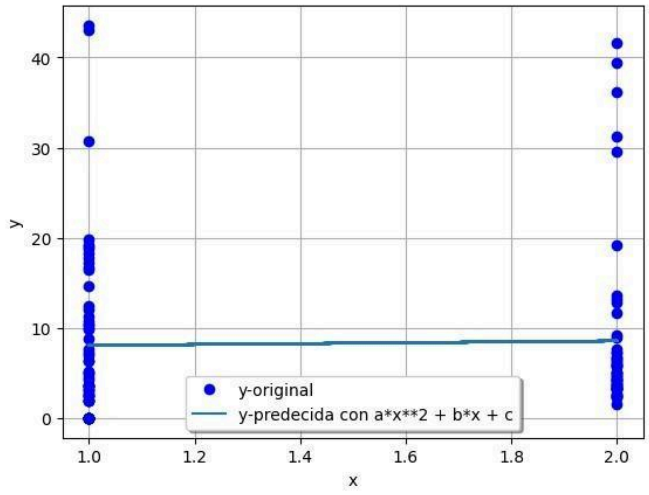
Después pudimos hacer filtros por usuarios para tener una análisis más profundo y personalizado.

Usando como ejemplo a Erick Osvaldo usando:

```
y = ax^2 + bx + c    ("Función cuadrática")
```

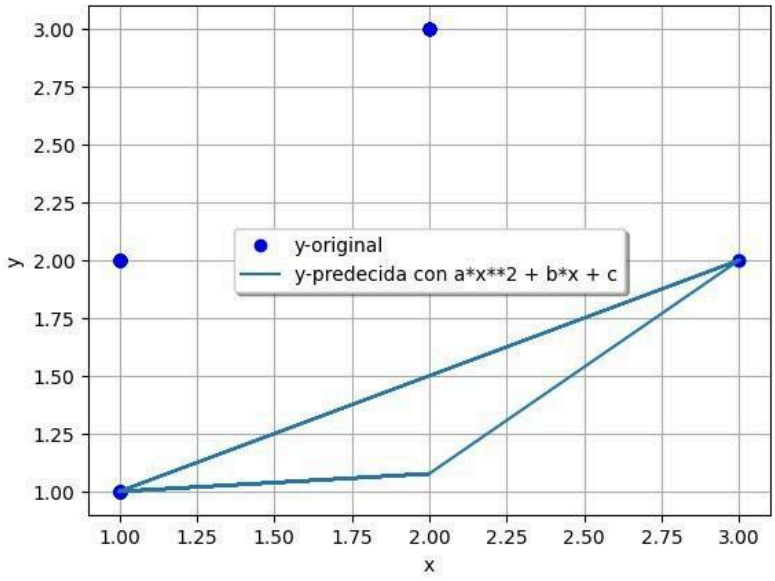
Propusimos 4 modelos con la misma función para conocer la correlación que existe entre las columnas objetivo.

Modelo 1 x = preionó botón correcto (datos nulos eliminados) | y = tiempo de interacción



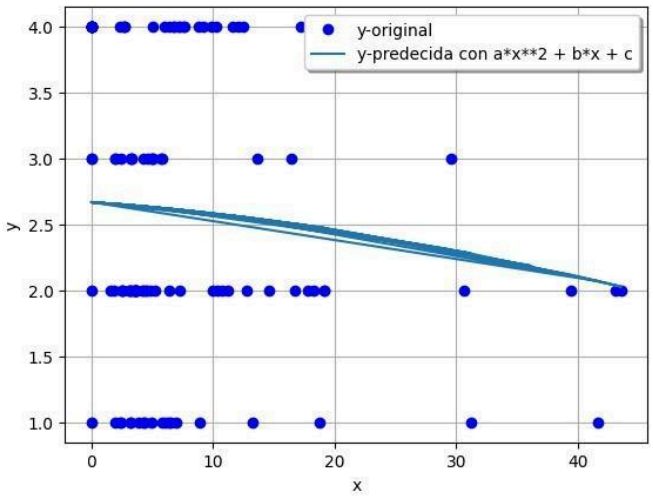
El coeficiente de correlación es 0.02

Modelo 2 x = auto push (datos nulos eliminados)| y = preionó botón correcto



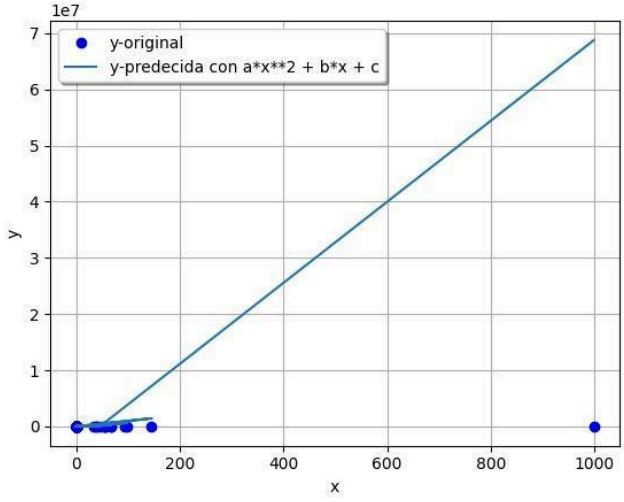
El coeficiente de correlación fue nulo, lo que significa que no existe una dependencia entre los datos para su comportamiento.

Modelo 3 x = tiempo de interacción | y = colore presionado (en ambos se eliminaron los datos nulos)



Con un coeficiente de correlación de 0.11

Modelo 4 x = tiempo de lección | y = dificultad (Estos datos nulos no fueron eliminados)

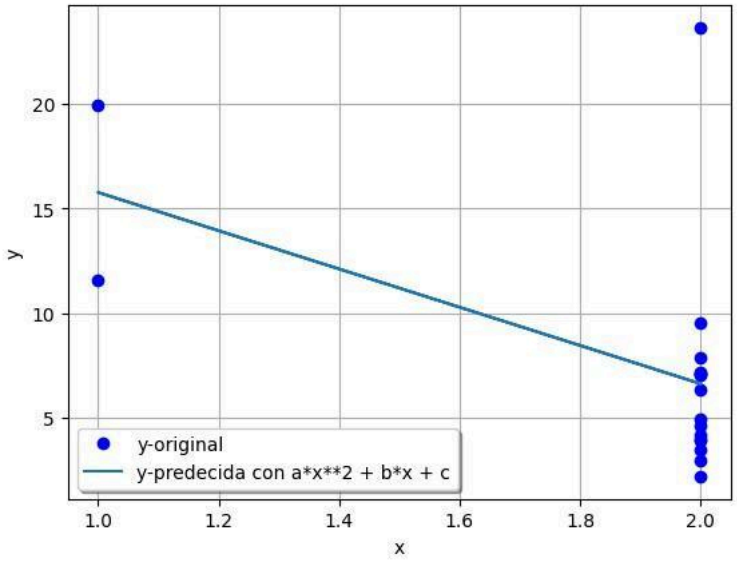


El coeficiente de correlación es nulo.

Posteriormente realizamos los modelos para Esmeralda, utilizando la misma función para los 4 modelos:

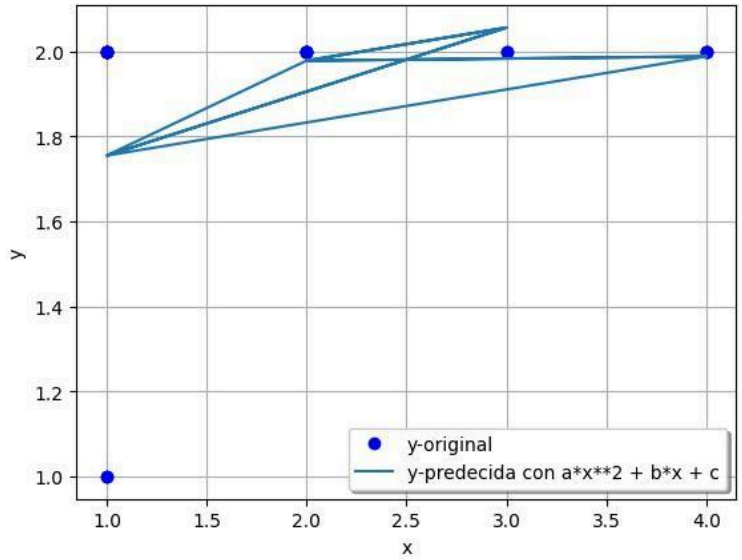
```
y = ax^2 + bx + c    ("Función cuadrática")
```

Modelo 1 x = preionó botón correcto (datos nulos eliminados) | y = tiempo de interacción



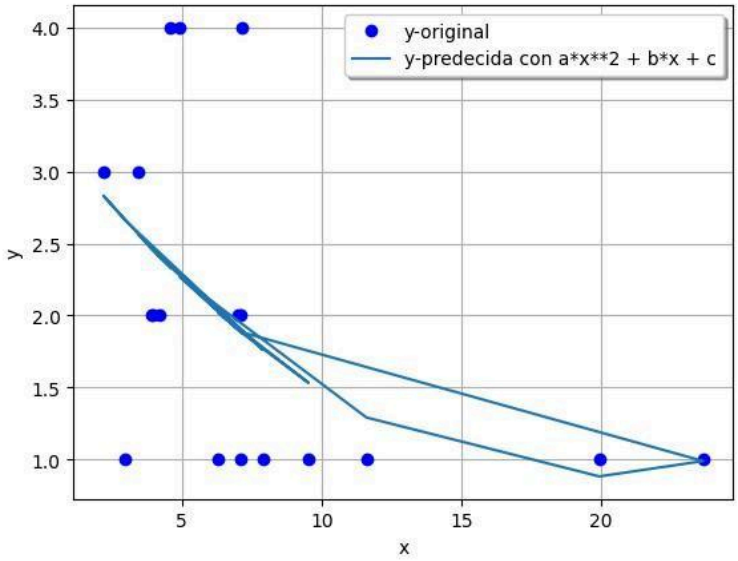
El coeficiente de correlación es 0.517 y el de determinación de 0.268. Esto nos indica un rendimiento bajo del modelo para predecir la variable objetivo.

Modelo 2 x = color presionado | y = preionó botón correcto (datos nulos eliminados)



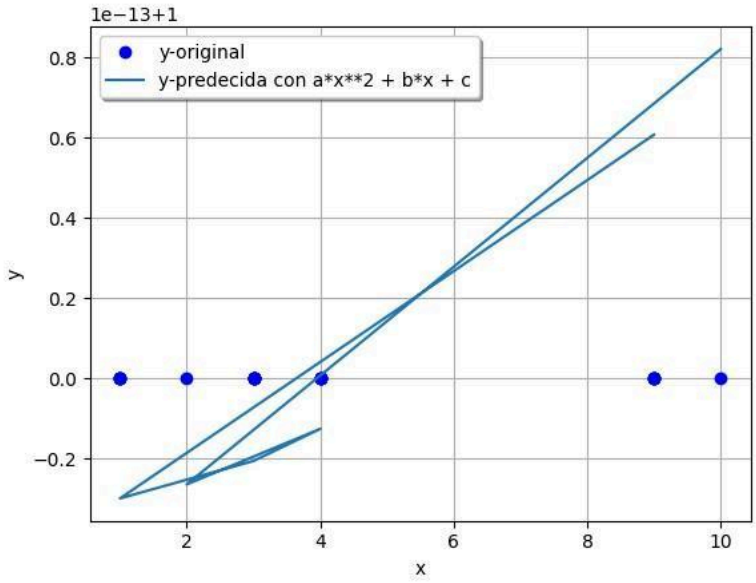
El coeficiente de correlación fue 0.38 y el de determinación fue de 0.15. Esto nuevamente nos indica un rendimiento bajo del modelo para predecir la variable objetivo.

Modelo 3 x = tiempo de interacción | y = color presionado (en ambos se eliminaron los datos nulos)



El un coeficiente de correlación fue de 0.49 y el de determinación de 0.24. Esto también nos indica rendimiento bajo del modelo para predecir la variable objetivo.

Modelo 4 x = mini juego | y = dificultad (Estos datos nulos no fueron eliminados)

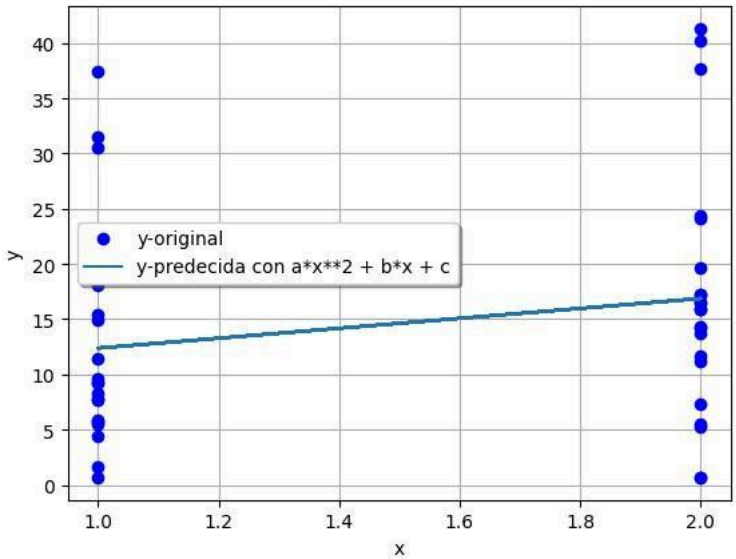


El coeficiente de correlación es 0. Por lo que no hay relación entre las variables.

También realizamos los modelos para Ingrid, donde volvimos a utilizar la misma función para los 4 modelos:

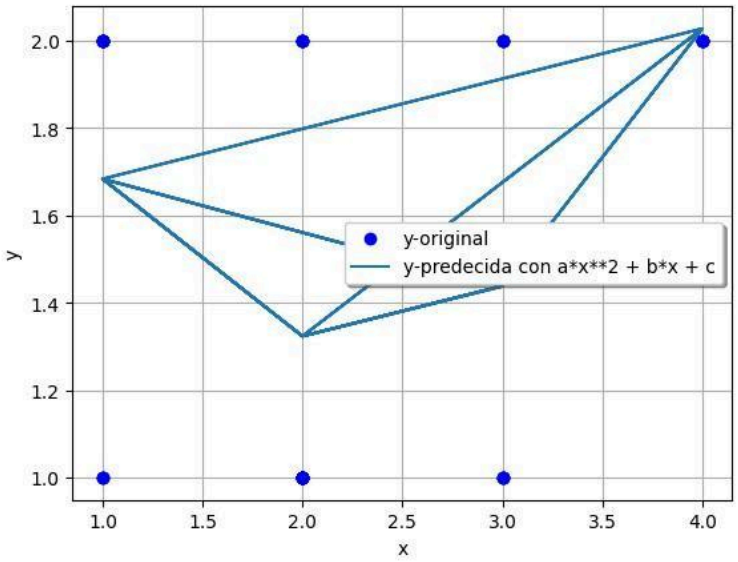
```
y = ax^2 + bx + c    ("Función cuadrática")
```

Modelo 1 x = preionó botón correcto (datos nulos eliminados) | y = tiempo de interacción



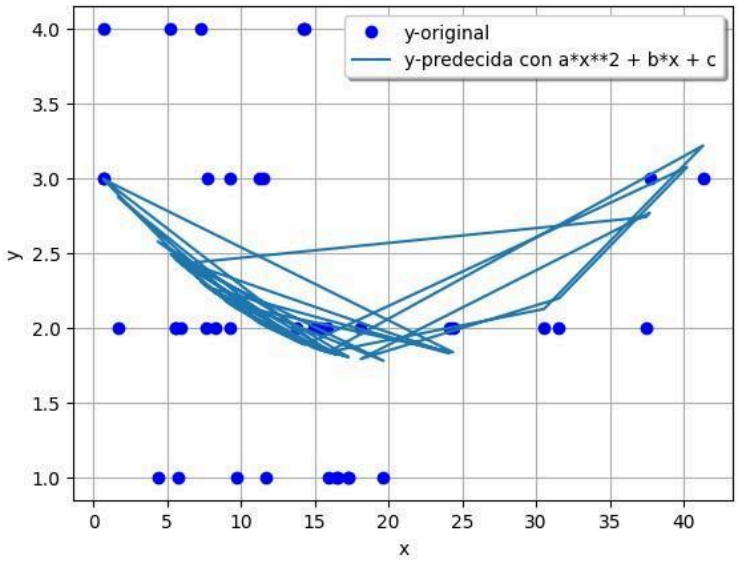
El coeficiente de correlación es 0.20 y el de determinación de 0.43. Esto nos indica un rendimiento y correlación prácticamente nulos, por lo que el modelo no es lo suficientemente significativo para predecir la variable objetivo.

Modelo 2 x = color presionado | y = preionó botón correcto (datos nulos eliminados)



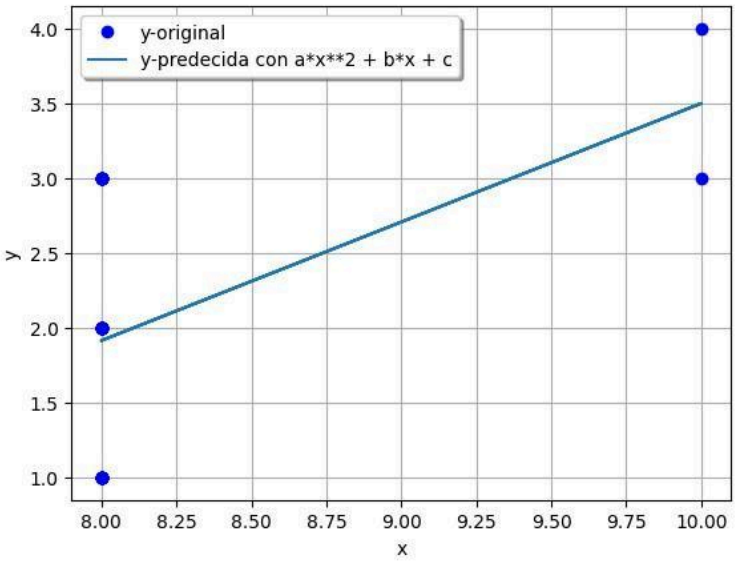
El coeficiente de correlación fue 0.42 y el de determinación fue de 0.18. Esto nuevamente nos indica un rendimiento bajo del modelo para predecir la variable objetivo.

Modelo 3 x = tiempo de interacción | y = color presionado (en ambos se eliminaron los datos nulos)



El coeficiente de correlación fue de 0.49 y el de determinación de 0.24. Esto también nos indica rendimiento bajo del modelo para predecir la variable objetivo.

Modelo 4 x = mini juego | y = dificultad (Estos datos nulos no fueron eliminados)



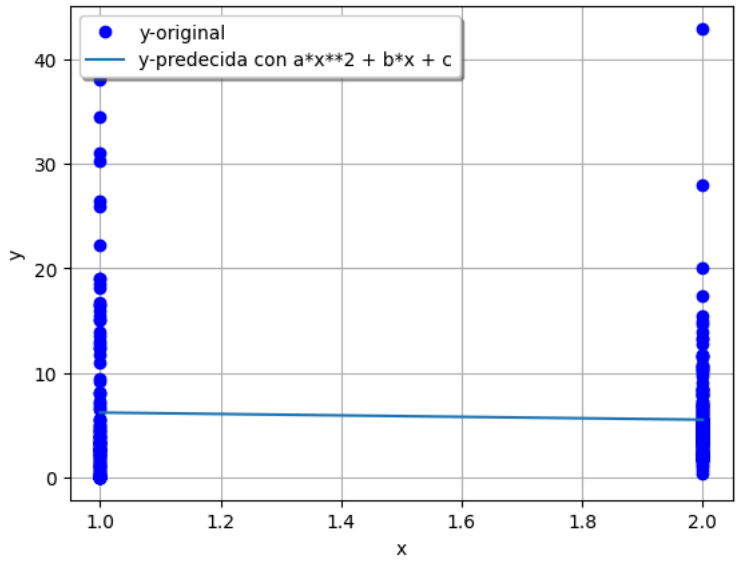
El coeficiente de correlación fue de 0.39 y el de determinación de 0.15. Esto también nos indica rendimiento bajo del modelo para predecir la variable objetivo.

Tecnológico de Monterrey

También realizamos los modelos para Jesus Alejandro, donde volvimos a utilizar la misma función para los 4 modelos:

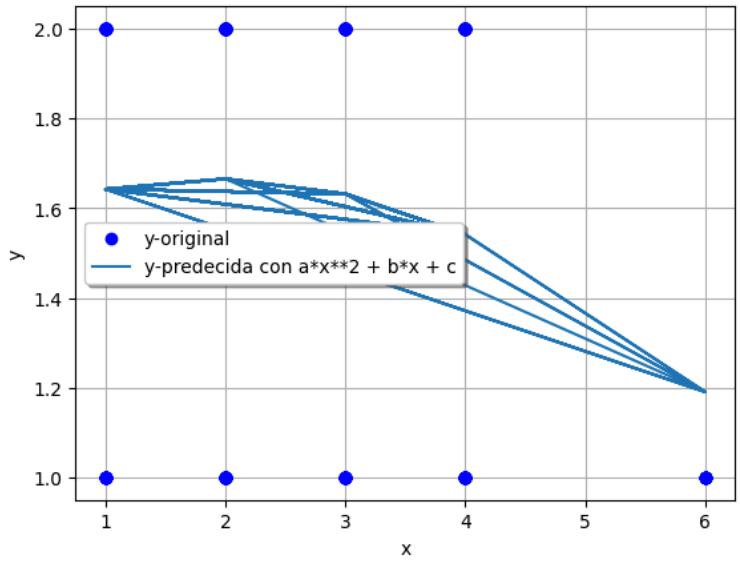
```
y = ax^2 + bx + c    ("Función cuadrática")
```

Modelo 1 x = preionó botón correcto (datos nulos eliminados) | y = tiempo de interacción



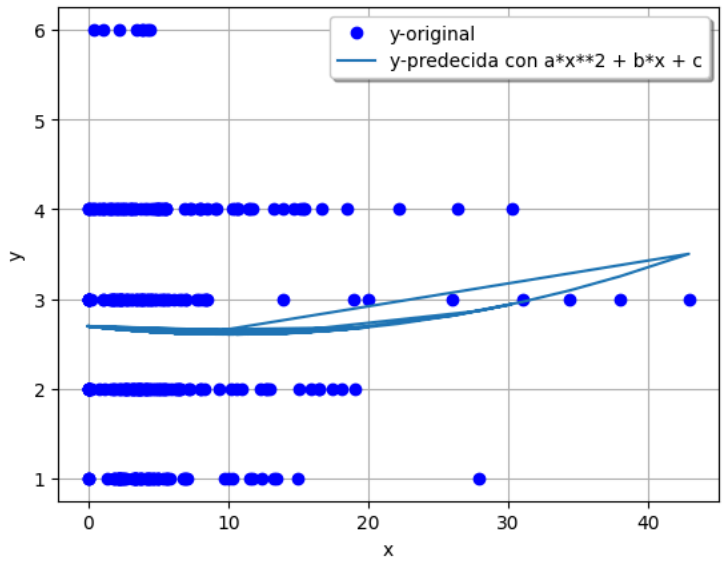
Con un coeficiente de determinación de 0.00295 y coeficiente de correlación con valor de 0.0543 esto nos está indicando un rendimeinto casi nulo para predecir la variable objetivo en el modelo.

Modelo 2 x = color presionado | y = preionó botón correcto (datos nulos eliminados)



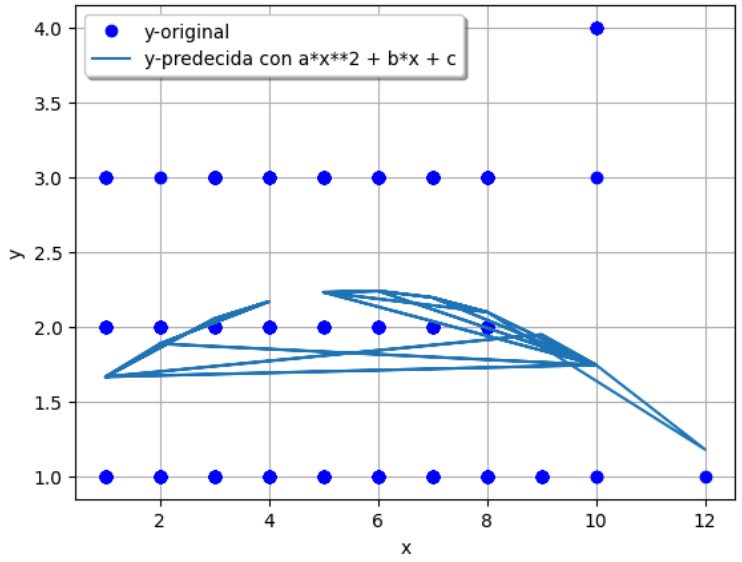
El coeficiente de determinación de 0.0293 y coeficiente de correlación con valor de 0.1713. Esto también nos indica rendimiento bajo del modelo para predecir la variable objetivo.

Modelo 3 x = tiempo de interacción | y = color presionado (en ambos se eliminaron los datos nulos)

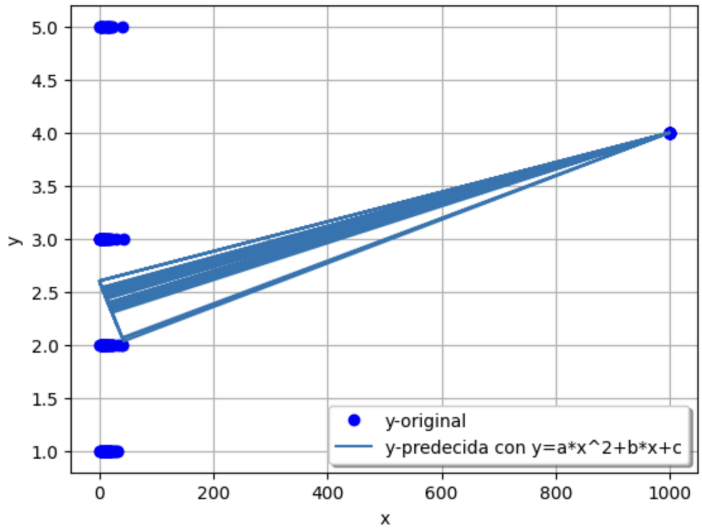


El coeficiente de determinación de 0.0040 y coeficiente de correlación con valor de 0.0635. Esto también nos indica rendimiento demiasiado bajo del modelo para predecir la variable objetivo.

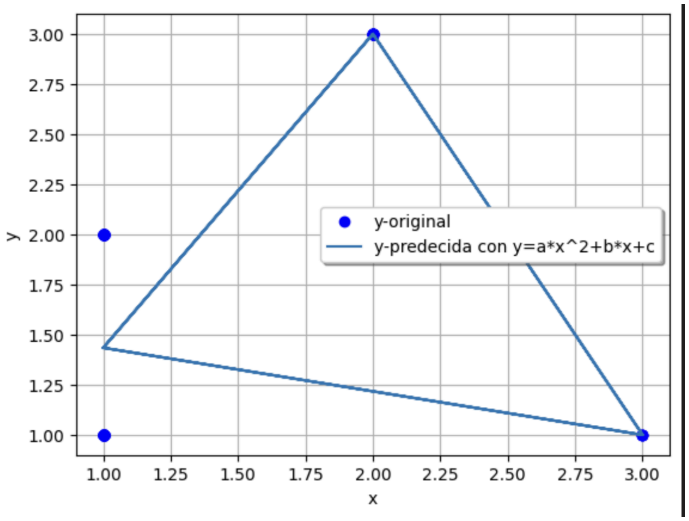
Modelo 4 x = mini juego | y = dificultad (Estos datos nulos no fueron eliminados)



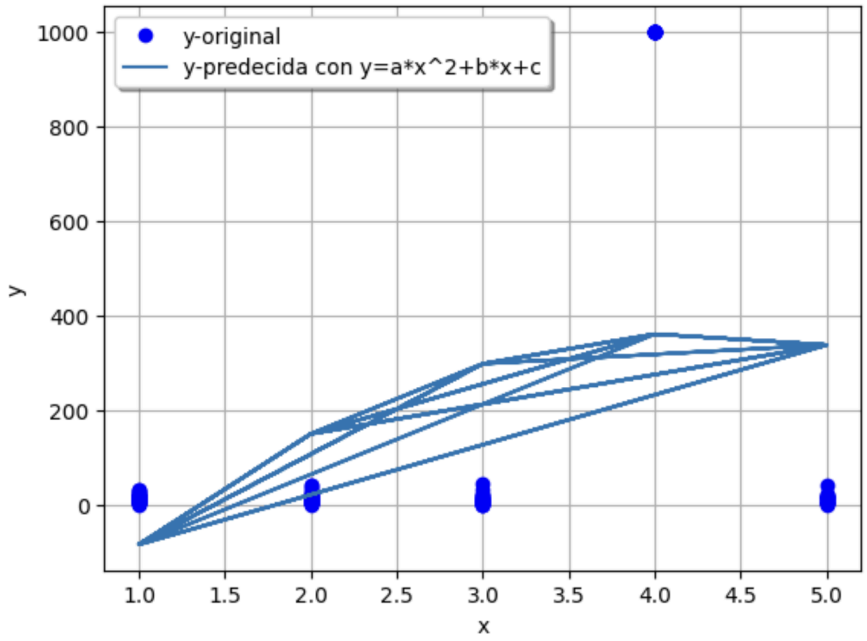
Con un coeficiente de determinación de 0.04486 y coeficiente de correlación con valor de 0.2118 esto nos está indicando un rendimeinto bajo pero mayor a los anteriores para este usuario para predecir la variable objetivo en el modelo.



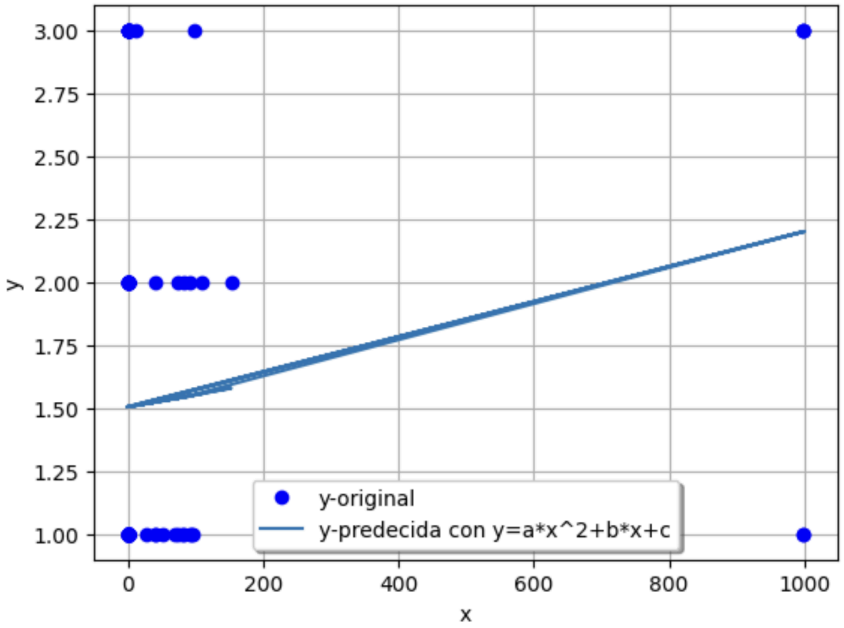
En esta gráfica, algunos valores de x son muy grandes y los demás son pequeños. Eso hace que los puntos se vean todos pegados en un solo lado. La línea azul, que es la predicción del modelo, no pasa bien por los puntos. Esto pasa porque esos números tan grandes hacen que el modelo se confunda y no pueda predecir bien.



Aquí los valores están más parejos, todos están en un rango más corto. Como no hay números raros ni exagerados, el modelo puede dibujar una curva que pasa cerca de los puntos. En este caso, el modelo hace un buen trabajo porque los datos están bien distribuidos.



En esta gráfica, unos pocos puntos tienen valores de y muy altos. Eso hace que el modelo trate de alcanzarlos y la curva se vea rara. No sigue bien el resto de los puntos. Es como si esos datos muy exagerados arruinaran todo el trabajo.



Aquí los puntos están casi alineados, como si siguieran una línea recta. Pero el modelo está usando una curva para predecir. Por eso la línea azul no queda tan bien. En este caso, hubiera sido mejor usar una línea recta sencilla en lugar de una curva.

Cuando los datos nulos son eliminados el coeficiente de correlación incrementa levemente, sin embargo, las columnas usadas para el estudio no están correlacionadas. Por lo que deberían hacerse más filtros por columnas y calcular las correlaciones e identificar con diferentes usuarios para conocer cuál es la correlación más alta. Ya que en la mayoría de los modelos, los coeficientes obtenidos no son lo suficientemente significativos como para poder predecir las variables objetivo propuestas.

Los bajos valores de correlación y de explicación en casi todos los modelos por usuario indican que los modelos lineales o cuadráticos simples no logran representar bien lo que está pasando con cada usuario. Las relaciones entre las variables parecen ser débiles, poco claras y posiblemente afectadas por el comportamiento de cada persona o por factores que no se midieron.

También veo que al quitar los valores nulos, a veces la correlación mejora un poco, lo que sugiere que esos datos faltantes podrían estar escondiendo algunos patrones. Aun así, las correlaciones

siguen siendo débiles, lo que indica que las variables usadas tal vez no son buenas para predecirse entre sí dentro del comportamiento de cada usuario.

En resumen, el análisis muestra que se necesitan modelos más detallados y adaptados a cada persona. Sería útil probar combinaciones distintas de variables, usar modelos no lineales más complejos o incluso agrupar a los usuarios según sus comportamientos, para así encontrar relaciones más claras y útiles dentro de los datos de Wuppi. Este análisis demuestra que un modelo general simple no es suficiente para entender un conjunto de datos tan variado.