Universidad del Valle de Guatemala
Inteligencia Artificial - CC3045
Catedrático: Luis Alberto Suriano Saravia
Mayo 2020

Integrantes:
JULIO ROBERTO HERRERA SABAN
JUAN PABLO PINEDA MELENDEZ
DIEGO ANDRES CRESPO MEALLA

Proyecto Final
Predicción de éxito de videojuegos usando Redes Neuronales

Link al repositorio:
https://github.com/jurhs2000/ai-proyecto

Problemática

La industria de los videojuegos ha ido creciendo durante los últimos años, en 2020 logró ingresos de $160 millones lo que la pone por encima de la industria de las películas y la música juntas, siendo la segunda industria más grande del entretenimiento solamente siendo drásticamente superada por la industria de la televisión. (Statista, 2022)

Los desarrolladores de videojuegos deben esperar un ingreso mayor a la inversión para el desarrollo del mismo, que puede superar los $100 millones para juegos AAA (alto presupuesto) e incluso llegando a los $500 millones (Baltezarevic, R., et al., 2018). Aunque existe un público de 3200 millones de jugadores (*gamers*) en todo el mundo (Statista, 2022) que se divide en distintas plataformas (PlayStation, XBox, PC, Nintendo, etc) y en distintas regiones del mundo, los *gamers* son un público al que hay que convencer con un buen juego, sin llegar a producir *hype* (expectativas muy altas) que termine los termine decepcionando, para así asegurar las ventas.

Entre una vasta cantidad de videojuegos, algunas similares propuestas de distintas compañías, otros juegos que proponen ser únicos, existen muchos factores que contribuyen a la decisión de compra de un videojuego, desde la reputación de la compañía que lo desarrolla, entregas anteriores de la misma saga, campaña de *marketing*, críticas prelanzamiento y/o comparaciones con otros juegos, entre otras variables. En cuanto a la influencia de las críticas, se consideran las críticas de “expertos” que vendrían a calificar aspectos más técnicos del producto y a representar a una entidad como portal de noticias o blog y que en nuestro *dataset* basado en las *reviews* de metacritic.com vendría a ser el “metascore”; por otro lado la crítica de “usuarios” a la que puede contribuir cualquier usuario registrado en la plataforma es representada por el “userscore” en el *dataset* basado en metacritic.com. Ambos tipos de *reviews* son subjetivos y aunque algunos puedan tener mayor influencia a corto o largo plazo, existen estudios como el de EEDAR que muestran que un mayor porcentaje de personas recomienda un videojuego tras ser influenciado por *reviews* positivas, otro porcentaje menor lo hace sin influencia de *reviews* y el menor porcentaje lo hace tras ser influenciado por *reviews* negativas (Meer, A., 2010). Otros estudios indican que las *reviews* de estos sitios de reseñas como metacritic.com, ign.com,

etc, tienen correlación positiva en las ventas de los videojuegos, así como otro tipo de *reviews* en formato de videos de YouTube o podrían considerarse también las publicaciones a través de redes sociales de usuarios más independientes pero con una audiencia parte del público objetivo. (Adigüzel, F., 2021)

Preprocesamiento del *dataset*

Los datos utilizados provienen de dos *datasets*, ambos obtenidos de la plataforma Kaggle de Google. El primero llamado ***Video Game Sales*** contiene 16598 videojuegos y contiene las siguientes columnas:

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC,PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

De estas se descartó Rank ya que no nos interesa saber cuál tuvo más ventas, así como Publisher y Genre ya que se comparó con las columnas que representan lo mismo en el otro *dataset* y se consideró que las otras representan mejor el dato.

El segundo *dataset* llamado ***Games of All Time from Metacritic*** contiene 8831 videojuegos y contiene las siguientes columnas:

- game name
- user score
- meta score
- platform information
- description
- developer
- url to game
- genre
- rating (E, M, T etc)
- type (multi/single player).

De estas se descartaron las columnas description, url to game, rating y type ya que no son variables representativas que queremos tomar en cuenta para nuestra variable dependiente (ventas del primer *dataset*). Este primer *dataset* tiene la particularidad que la columna de genre y platform information contiene distintas plataformas y géneros en un mismo registro, por lo que se separaron y así poder cuadrar con las distintas ventas para distintas plataformas en el primer *dataset* y en el caso de géneros para que el modelo pueda saber la influencia de cada género en las ventas y no de un conjunto de géneros.

Ambos *dataset* se unieron utilizando el nombre del videojuego y la plataforma, quedando así un total de 15759 registros de los cuales muchos nombres de videojuegos están repetidos

pero varía en su plataforma y ventas o en su género. Cabe aclarar que si el *dataset* de ventas no tenía un registro de ventas diferente para cada plataforma igualmente los registros quedaron separados por plataforma pero con el mismo valor de ventas.

También se realizó la limpieza general para quitar datos nulos, infinitos y para categorizar las variables cualitativas con un número para representar cada plataforma, género y desarrollador.

Análisis de variables del *dataset*

Con el objetivo de verificar que los registros de nuestro *dataset* ya procesado cumpla con el comportamiento esperado de las variables según lo investigado en la problemática se realiza un análisis estadístico de las variables, empezando por una matriz de correlaciones que mediante un mapa de calor muestra los coeficientes de correlación entre la intersección de todas las variables.

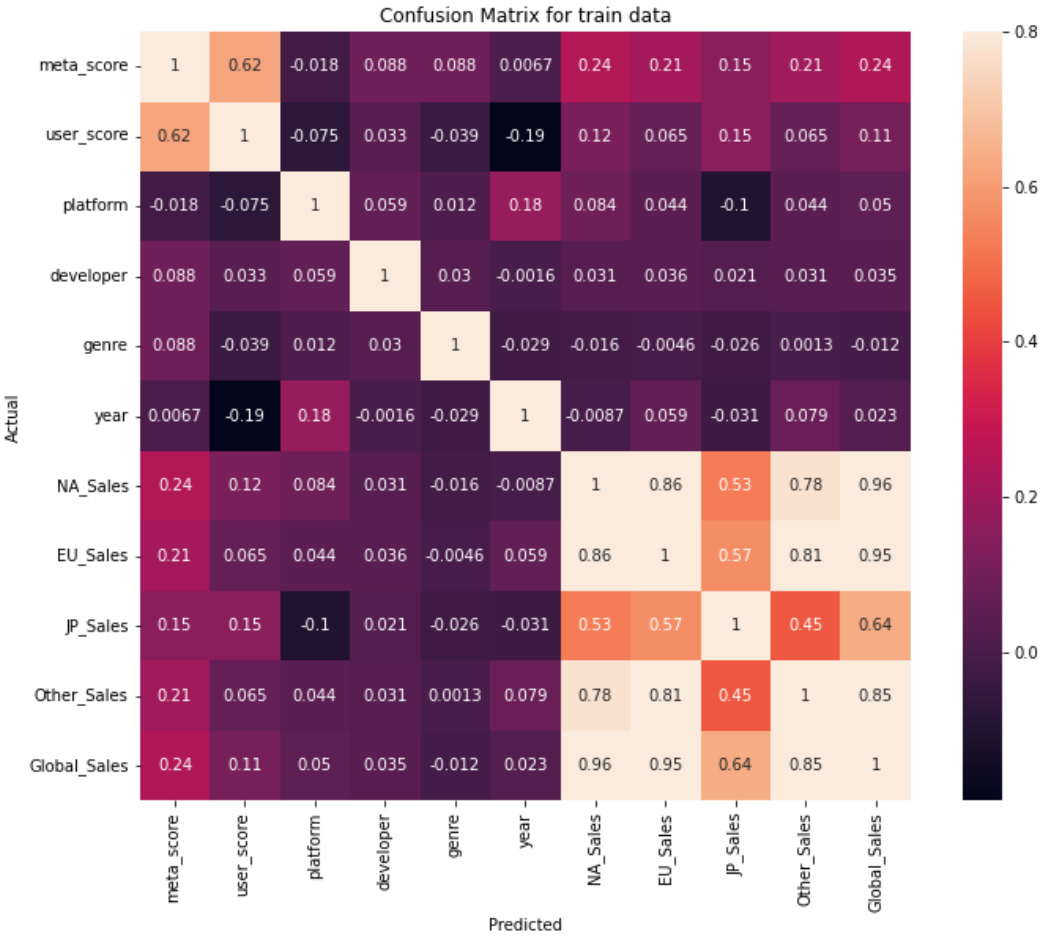


Figura 1: Matriz de correlación de todas las variables.

Esto muestra que el metascore y el userscore sí tienen correlación con las ventas, en algunas regiones más que otras pero sí la hay. Por otro lado nos muestra que la plataforma, el desarrollador, el género y el año no muestran una correlación fuerte respecto a las ventas.

Antes de corroborar estas correlaciones vemos si en las columnas de las ventas existen datos atípicos usando gráficas de caja y bigotes, si existen se removerán. Que existan datos atípicos en este *dataset* no significa que sean datos incorrectos pero pueden llegar a confundir al modelo y que las predicciones de este no sean precisas aunque sí exactas.

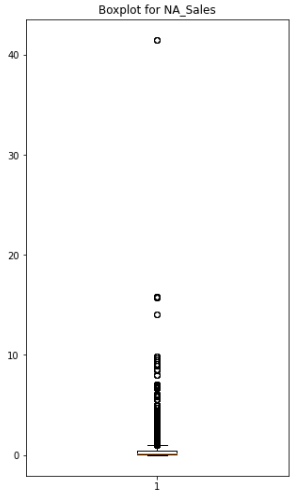


Figura 2: Caja y bigotes de NA_Sales.

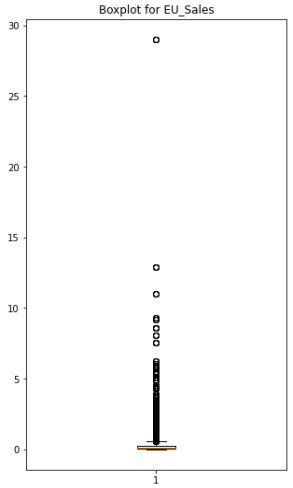


Figura 3: Caja y bigotes de EU_Sales.

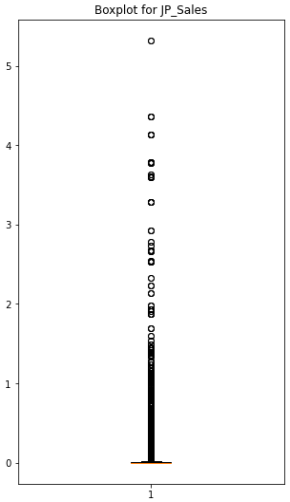


Figura 4: Caja de bigotes de JP_Sales.

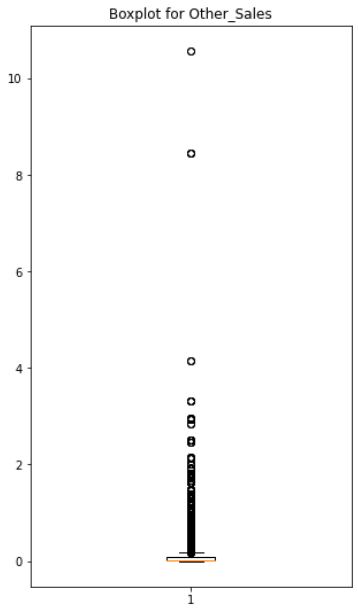


Figura 5: Caja de bigotes de Other_Sales.

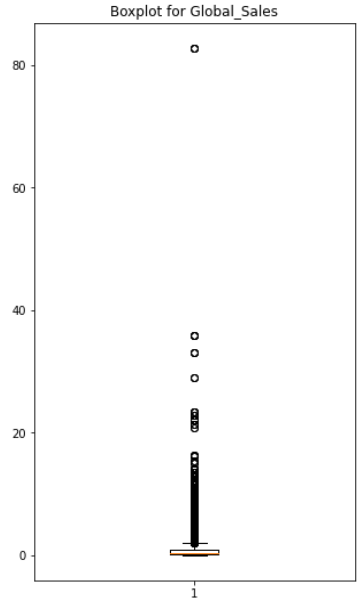
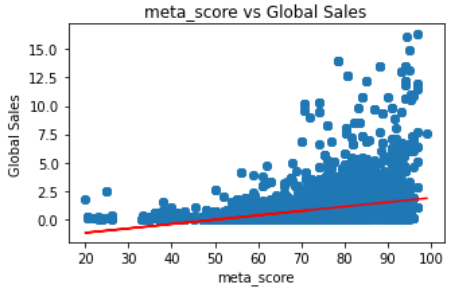


Figura 6: Caja de bigotes de Global_Sales.

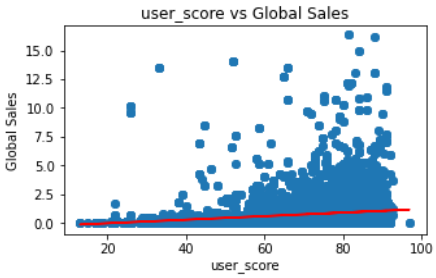
No se removerán todos los datos atípicos ya que como se mencionó, no son datos incorrectos, pero se limitará para evitar que el modelo llegue a considerar estas cifras para más datos solo por uno que existe en el *dataset*. Los límites considerados (en millones) son 11 para NA_Sales, 10 para EU_Sales, 3 para JP_Sales, 3 para Other_Sales y 20 para Global_Sales.

Ahora sí, para corroborar las correlaciones realizamos un gráfica de cada una de las variables contra las ventas globales que en teoría son las ventas generales entre todas las regiones, de estas gráficas también sacamos la correlación de pearson y el P-value, este último al ser menor que 0.05 nos dice que los datos son estadísticamente significativos para rechazar una hipótesis nula. (Beers, B., 2022)



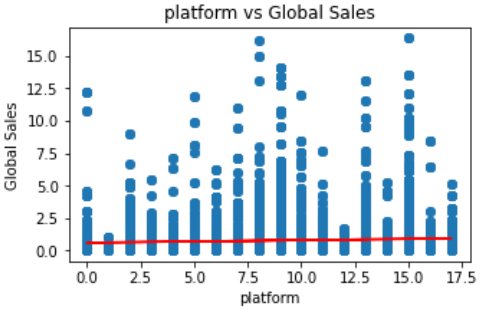

```
meta_score vs Global Sales: Pearson Correlation Coefficient: 0.3431351249617914
meta_score vs Global Sales: p-value: 0.0
```

Figura 7: meta_score vs Global_Sales.



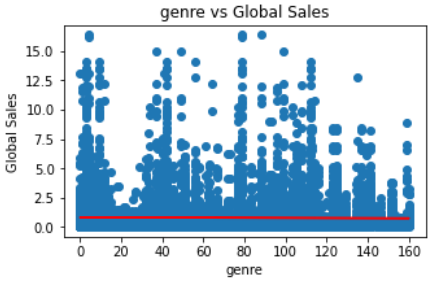

```
user_score vs Global Sales: Pearson Correlation Coefficient: 0.1303219452032923
user_score vs Global Sales: p-value: 1.7990579225153507e-60
```

Figura 8: user_score vs Global_Sales.



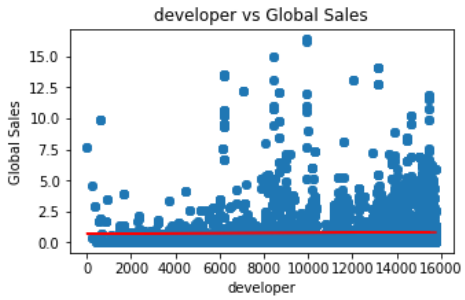

```
platform vs Global Sales: Pearson Correlation Coefficient: 0.06149339406884187
platform vs Global Sales: p-value: 1.2147005444456585e-14
```

Figura 9: platform vs Global_Sales.



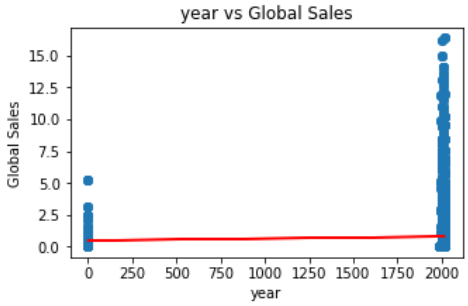

```
genre vs Global Sales: Pearson Correlation Coefficient: -0.01470119646280976
genre vs Global Sales: p-value: 0.06537972203635341
```

Figura 10: genre vs Global_Sales.




```
developer vs Global Sales: Pearson Correlation Coefficient: 0.014905562076697906
developer vs Global Sales: p-value: 0.06172432582642953
```

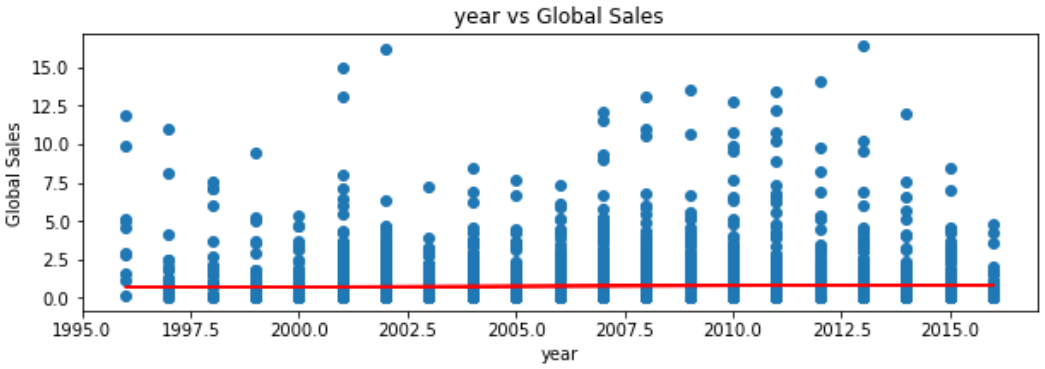
Figura 11: developer vs Global_Sales.




```
year vs Global Sales: Pearson Correlation Coefficient: 0.02882027924964735
year vs Global Sales: p-value: 0.0003027932915892241
```

Figura 12: year vs Global_Sales.

De estas gráficas y valores podemos ver las variables que no están aportando al rechazo de la hipótesis nula con un p-value mayor a 0.05, por lo tanto procedemos a descartar la columna genre y developer; también existían filas que solo difieren en la columna genre, por lo que se deben eliminar estos duplicados que quedan al eliminar dicha columna, pasando así de 15759 filas a 4415 filas, ahora sin tomar en cuenta los distintos géneros de cada juego ni los desarrolladores. También se observa que la columna de year tiene valores en 0 por lo que procedemos a quitar esos valores, dejando los siguientes valores para esa gráfica.

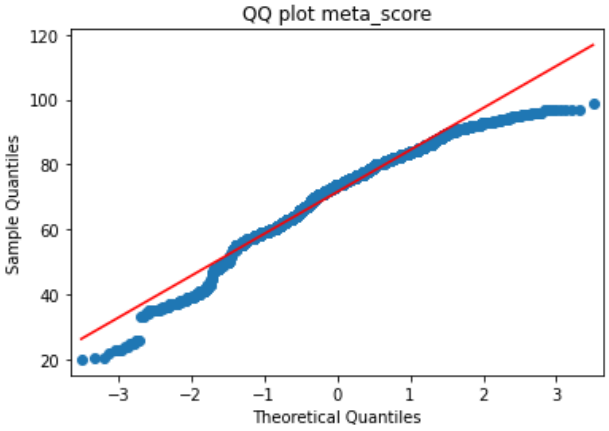



```
year vs Global Sales: Pearson Correlation Coefficient: 0.016672796021478448
year vs Global Sales: p-value: 0.2725852241568882
```

Figura 13: year vs Global_Sales sin datos de year en 0.

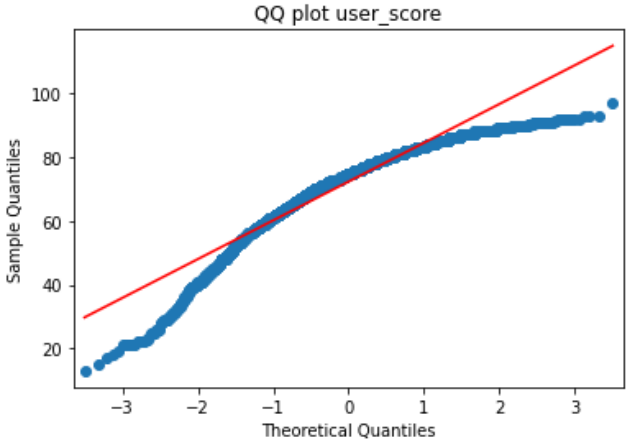
Esto nos confirma que el dato del año tampoco aporta al rechazo de la hipótesis nula, lo cual es curioso, ya que teníamos contemplado agregar una variable computada a partir de esta que sería “time_on_market” para que el modelo supiera el tiempo que el juego lleva en el mercado y asumir que un juego con más años tiene mayores ventas ya que las ha acumulado pero al parecer no es así.

Siguiendo con el análisis estadístico, procedemos a comprobar que los datos de las columnas meta score, user score y platform estén normalizados para comprobar que nuestro modelo no esté sesgado a un conjunto de críticas o plataformas, para ello utilizaremos la gráfica QQ y los valores de sesgo y curtosis.



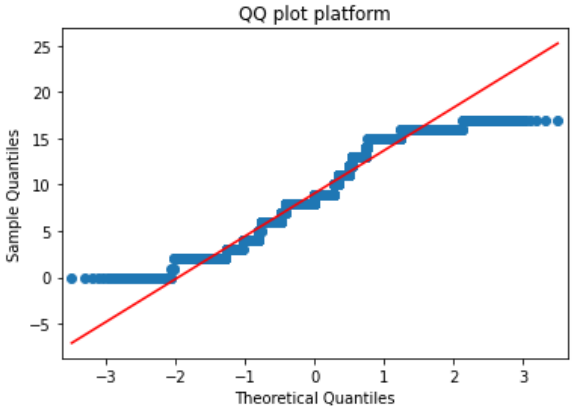

```
skew:  -0.6900625509399992
kurtosis:  0.4576112276926452
```

Figura 14: QQ plot meta_score.




```
skew:  -1.2447611325537924
kurtosis:  2.0379105368234445
```

Figura 15: QQ plot user_score.

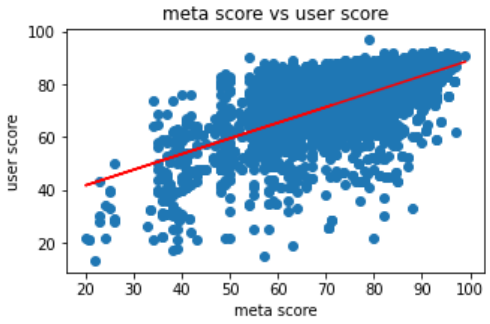



```
skew:  0.015592349635656106
kurtosis:  -1.0370606341283388
```

Figura 16: QQ plot platform.

En cuanto al meta score y al user score vemos que como los sesgos son negativos están un poco sesgados a las críticas con alto puntaje pero al ser la curtosis mayor a 0 se puede decir que existe una concentración de datos a la media. En cuanto a las plataformas, la curtosis es negativa lo que significa que los datos están dispersos lo cual es bueno, aparte están muy poco sesgados.

Por último queremos ver la relación entre el meta score y el user score por lo que graficamos ambos datos en eje X y eje Y y obtendremos el coeficiente de pearson así como el p-value.




```
meta score vs user score: Pearson Correlation Coefficient: 0.6311859505840607
meta score vs user score: p-value: 0.0
```

Figura 17: meta_score vs user_score

Con esta gráfica vemos que ambos tipos de críticas sí están relacionados, lo cual nos confirma el coeficiente de pearson y en cuanto al p-value al ser 0 nos dice que ambas variables son estadísticamente significativas entre sí.

Modelo a utilizar

Claramente al querer calcular los datos de ventas debemos usar un modelo de regresión, si solo dependieramos de las variables de meta score y user score tal vez se podría tomar en cuenta un modelo lineal o logístico que aplique dichas regresiones sin embargo la variable de la plataforma al no tener una alta correlación con las variables respuesta de las ventas pero sí ser estadísticamente significativa y nosotros al depender de ella por tener distintos datos de ventas para distintas plataformas, se optó por implementar una red neuronal de regresión, específicamente un perceptrón multicapa con optimización basada en el error cuadrático. Para ello usamos la librería sklearn de python la cual en su clase `neural_network.MLPRegressor` nos permite especificar el número de capas, la función de activación, el *learning rate*, la tolerancia y el número máximo de iteraciones. Cabe aclarar que esta clase si no encuentra mejora en la pérdida mayor al valor de la tolerancia en las últimas 10 iteraciones (valor que también se puede cambiar), para el entrenamiento.

Antes de crear el modelo se realizó una separación del *dataset* para usar el 70% de los datos para entrenamiento y el 30% restante para *testeo*. Tanto esta separación de datos como el entrenamiento del modelo utilizan un valor de *seed* para poder replicar los resultados en todo momento.

Para recordar, nuestras variables predictoras, la X para el modelo, son:

- meta_score
- user_score
- platform

Y las variables dependientes o variables respuesta, la Y para el modelo, son:

- NA_Sales
- EU_Sales
- JP_Sales
- Other_Sales
- Global_Sales

Luego de realizar varias pruebas (no registradas) utilizando distinta cantidad de *hidden layers* y nodos en ellas, con máximos de 25, 100 o 500, se llegaron a obtener resultados considerablemente buenos comparado con las anteriores pruebas con una estructura que utiliza 5 *hidden layers* y un máximo de 1000 nodos en una de ellas, específicamente (800, 1000, 500, 250, 100). Así mismo se definió que se utiliza la función de activación “ReLu”, un *learning rate* y tolerancia de 0.000001 y un máximo de iteraciones de 100000.

La gráfica de la pérdida durante las iteraciones en el aprendizaje es la siguiente:

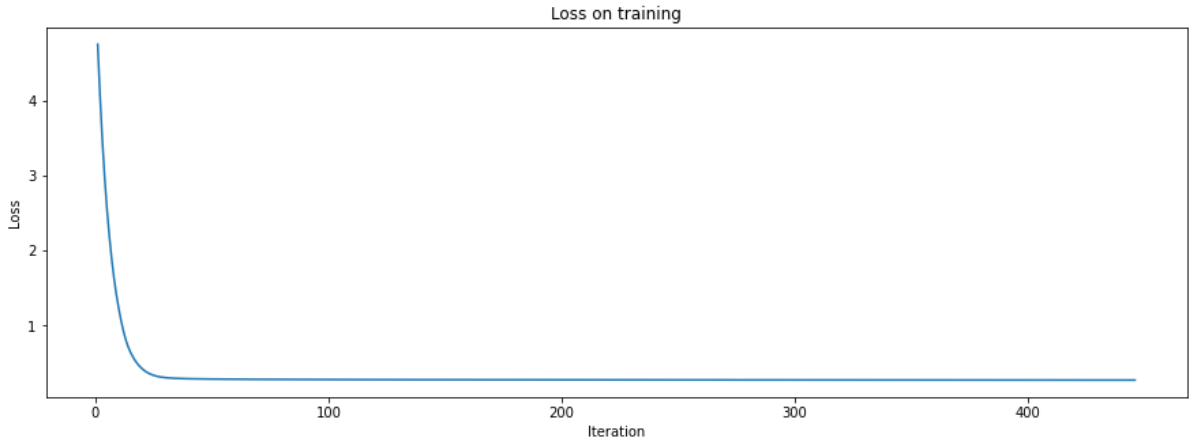


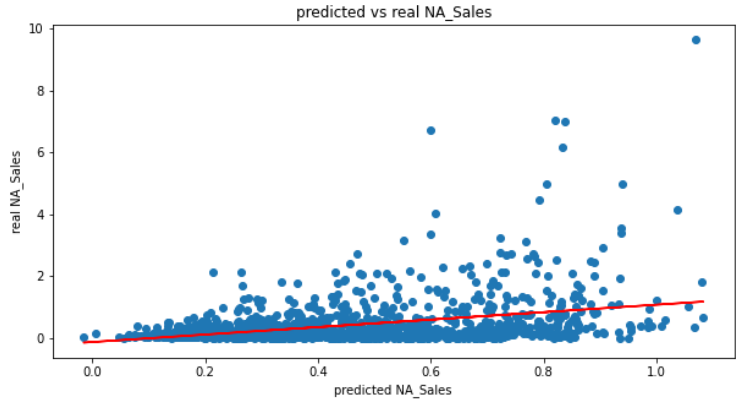
Figura 18: Loss on each iteration on training

Hipótesis

Antes de poner a nuestro modelo a predecir, definiremos la hipótesis a partir de la problemática investigada y de lo que sabemos de nuestros datos. Prácticamente nuestra hipótesis es que la crítica de “expertos” y la crítica de usuarios para cada plataforma, sea la misma o distinta crítica, influye positivamente en las ventas en cada región, es decir que a mayor valor de crítica, mayor número de ventas. La hipótesis nula sería la contradicción de esto, que ambos tipos de crítica no influyen en las ventas de los videojuegos. Por lo tanto esperamos que nuestro modelo pueda predecir las ventas de cada región dadas las críticas de “expertos” y de usuario así como la plataforma o plataformas del videojuego.

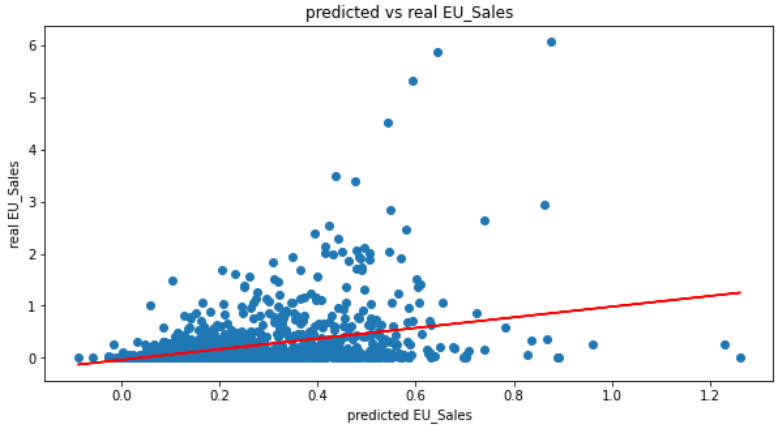
Resultados y Evaluación del modelo

Para observar los resultados de nuestras predicciones de ventas en cada región sobre el 30% del *dataset* que separamos para *testeo*, graficamos en el eje X el valor predicho y en el eje Y el valor real de cada variable. Si tuviéramos un 100% de éxito en las predicciones las gráficas se verían como una recta con pendiente, sin embargo vemos la dispersión a base del error en las predicciones. También se muestra el coeficiente de correlación de pearson y el p-value.



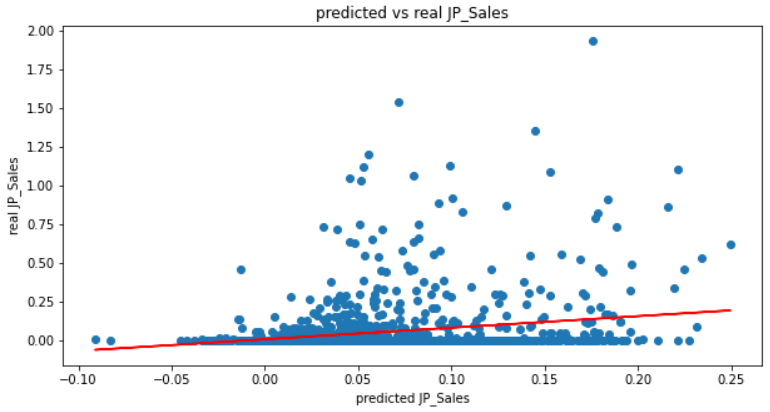

```
NA_Sales vs Global Sales: Pearson Correlation Coefficient: 0.934658458234439
NA_Sales vs Global Sales: p-value: 0.0
```

Figura 19: predicted vs real NA_Sales.



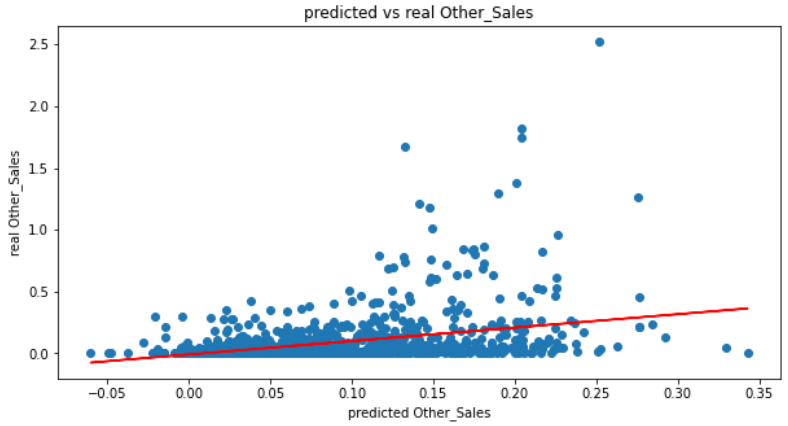

```
EU_Sales vs Global Sales: Pearson Correlation Coefficient: 0.9058185290480484
EU_Sales vs Global Sales: p-value: 0.0
```

Figura 20: predicted vs real EU_Sales



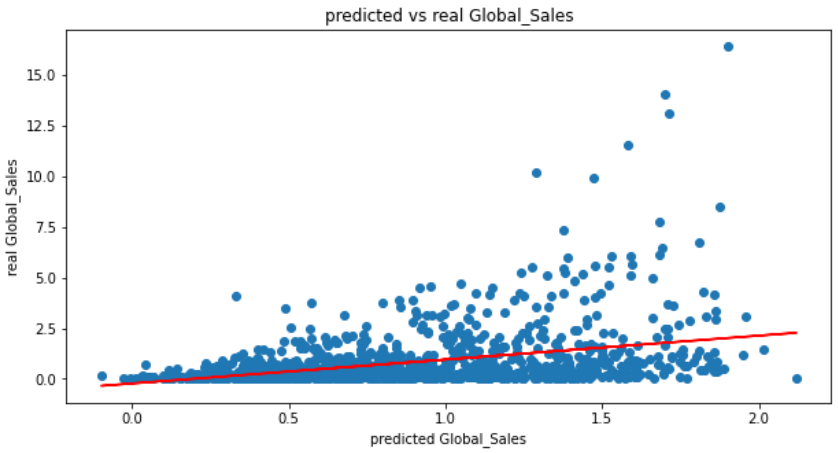

```
JP_Sales vs Global Sales: Pearson Correlation Coefficient: 0.5045458932597402
JP_Sales vs Global Sales: p-value: 1.3985741549379756e-278
```

Figura 21: predicted vs real JP_Sales




```
Other_Sales vs Global Sales: Pearson Correlation Coefficient: 0.8139783295058733
Other_Sales vs Global Sales: p-value: 0.0
```

Figura 22: predicted vs real Other_Sales




```
Global_Sales vs Global Sales: Pearson Correlation Coefficient: 1.0
Global_Sales vs Global Sales: p-value: 0.0
```

Figura 23: predicted vs real Global_Sales

De estos resultados observamos que, aunque se ven datos bastante dispersos, el coeficiente de Pearson muestra una correlación alta entre los resultados predichos y los reales, a excepción de las ventas en Japón. Por otro lado, los p-value indican que entre las predicciones y los valores reales, podemos rechazar la hipótesis nula.

Vemos que las predicciones en realidad fallan con aquellas ventas reales que superan por mucho a las del resto del *dataset* y eso que limpiamos algunos datos atípicos. También cabe mencionar que algunas predicciones llegaron a dar datos negativos aunque en la Y, es decir en los datos de las ventas no existen valores negativos.

Para evaluar las métricas de regresión de nuestro modelo, aplicamos algunas utilizando igualmente las predicciones contra los resultados reales. El *mean squared error* nos indica el promedio de las distancias de nuestros puntos predichos a la línea de regresión, es decir qué tanto nos alejamos en promedio de un valor exacto, aunque nuestro *mse* es de 0.4, recordemos que estos valores están en millones. El *r2 score* puede ir de 0 a 1, indicando la proporción de varianza que explican las variables independientes del modelo, es decir qué tan bien encaja nuestro modelo con los datos reales a través de los *inputs*, aquí es donde ya vemos la debilidad de nuestro modelo con un 0.1104 de r2 lo que nos dice que nuestro modelo no predice muy bien a partir de las variables ingresadas.

```
mse:  0.4576711117993647
r2:  0.11040733598722577
```

Figura 24: Métricas *mse* y r2 del modelo.

Conclusiones y Recomendaciones

El modelo nos permitió rechazar la hipótesis nula, lo cual nos permite concluir que las críticas de “expertos” y de usuarios, representadas en nuestro *dataset* como meta_score y user_score respectivamente, al menos de la plataforma de reseñas obtenidas que es metacritic.com, sí tienen influencia sobre las ventas de los videojuegos en cada región. La

otra principal conclusión es que solo estas críticas no son suficientes para generar un modelo de predicción de ventas, ya que en promedio se tuvo un error de casi medio millón de dólares y en las gráficas de resultados (Figuras 19-23) se observa que existen juegos con ventas relativamente mayores a lo predicho según las críticas, esto quiere decir que pueden haber juegos muy bien calificados que generan buenas ventas pero otros que igualmente tienen calificaciones altas en *reviews* pero que por otros factores no considerados en este proyecto, generan ventas mucho mayores a otros juegos.

Una de las variables que se considera puede influir en estas ventas que superan por mucho a otras en este *dataset* es el presupuesto del juego, por lo que se recomienda poseer estos datos para mejorar esta carencia en nuestro actual modelo.

Por último se puede decir que las ventas de Japón son las que menos influenciadas se ven por las *reviews* de estas plataformas de reseñas, por lo visto en la Figura 21.

Referencias

Adigüzel, F. (2021). The Effect of YouTube Reviews on Video Games Sales. Journal of Business Research - Turk. DOI: http://dx.doi.org/10.20491/isarder.2021.1249. https://www.researchgate.net/publication/353737385_The_Effect_of_YouTube_Reviews_on_Video_Game_Sales

Baltezarevic, R., Baltezarevic, B., Baltezarevic, V. (2018). THE VIDEO GAMING INDUSTRY (from play to revenue). Researchgate. https://www.researchgate.net/publication/330324266_THE_VIDEO_GAMING_INDUSTRY_from_play_to_revenue

Beers, B. (2022). P-Value. Investopedia. https://www.investopedia.com/terms/p/p-value.asp

Meer, A. (2010). EEDAR study: review scores do affect sales. gamesindustry.biz. https://www.gamesindustry.biz/articles/eedar-review-scores-do-affect-sales

Statista. (2022). Estimated media revenue worldwide in 2020, by category. Statista. https://www.statista.com/statistics/1132706/media-revenue-worldwide/

Statista. (2022). Number of video gamers worldwide in 2021, by region. https://www.statista.com/statistics/293304/number-video-gamers/