**Entrega final de proyecto**

Ariel Eduardo Bedoya Marín.

Identificación: 71275506

Mayo 2023.

Universidad de Antioquia.

Ingeniería de Sistemas.

Introducción a la inteligencia artificial para las ciencias e ingenierías.

**Resultados alcanzados**

Con los análisis iniciales del dataset, se llega a entender que este posee información muy básica para realizar predicciones a cerca de qué géneros de videojuegos podrían ser potencialmente deseables para que las comercializadoras y/o editoras quieran iniciar un desarrollo. Sin embargo, se hace interesante la idea de abordar experimentos que permitan buscar relaciones o patrones de los géneros de videojuegos más vendidos en determinados años, para cuáles plataformas fueron lanzados y cuáles has sido sus empresas publicadoras.

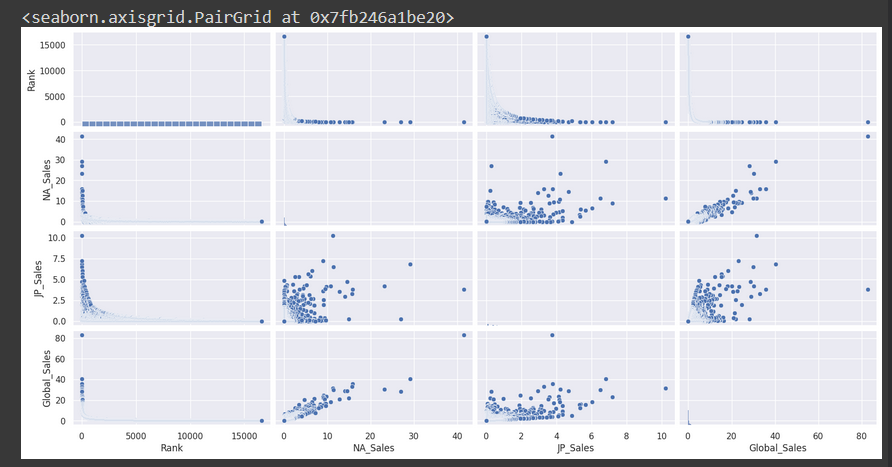
De momento, uno de los mayores obstáculos para manipular el dataset es, irónicamente, su simpleza, debido a que contiene pocas columnas, es decir, hay pocas variables con las cuales buscar correlaciones que lleguen a generar un proyecto complejo. Lo anterior no quiere decir que no sea posible implementar modelos predictivos interesantes y complejos que permitan ver algún tipo de resultado sea positivo o no y así llegar a conclusiones y análisis.

Ya dicho lo anterior, se puede decir que, al momento de hacer este informe, se ha realizado una exploración de los datos, incluyendo procesos para llamar el dataset desde kaggle (por medio de su API). Dentro de esta exploración se ha notado que hay años de publicación faltantes para algunos videojuegos y así mismo faltan algunos nombres de sus publicadores.

Texto

Descripción generada automáticamente

Se ha realizado un reemplazo del índice, el cual estaba numerado, por el nombre del videojuego. Esto debido a que el nombre es único y no proveerá mucha información dentro de las predicciones. Se verificó el tipo de datos, dentro de estos se observó que el año era de tipo ‘Float64’ y por lo tanto se convirtió a un tipo entero (‘Int64’). Se verificaron los datos faltantes y también se contrastó a través de gráficas, las relaciones que podrían o no, tener las variables del conjunto de datos.



Los análisis aquí se hacen evidentes y además es el comportamiento natural, dado que en todos los casos el ranking de un juego es directamente proporcional al número de copias vendidas. También es evidente que el número de copias vendidas a nivel global es directamente proporcional al número de copias vendidas a nivel de regiones y/o países. Hay una pequeña excepción con Japón, en el cual se puede notar una mayor cantidad de copias vendidas respecto a otras regiones y a nivel global.

Debido a la naturaleza de los datos de las columnas "Year" y "Publisher" se ha decido borrar las filas que no poseen estos datos ya que no es posible reemplazarlos por valores calculados. A continuación, se puede observar la cantidad de entradas que quedaron respecto a las que habían antes de la limpieza.

Texto

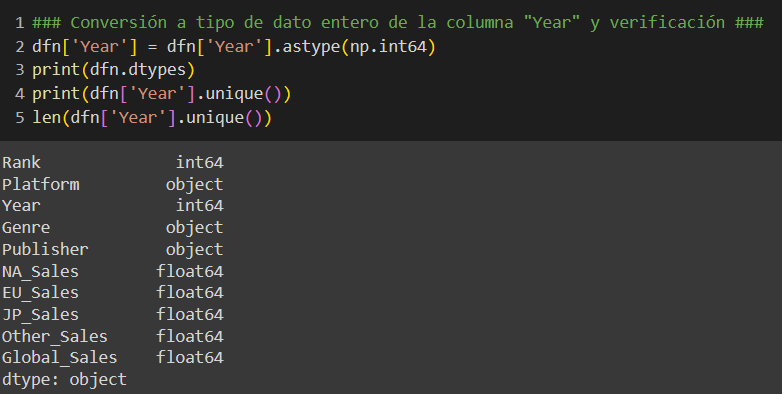
Descripción generada automáticamente

Seguidamente se verifica que ya no existan datos faltantes dentro del dataset.

Tabla

Descripción generada automáticamente con confianza media

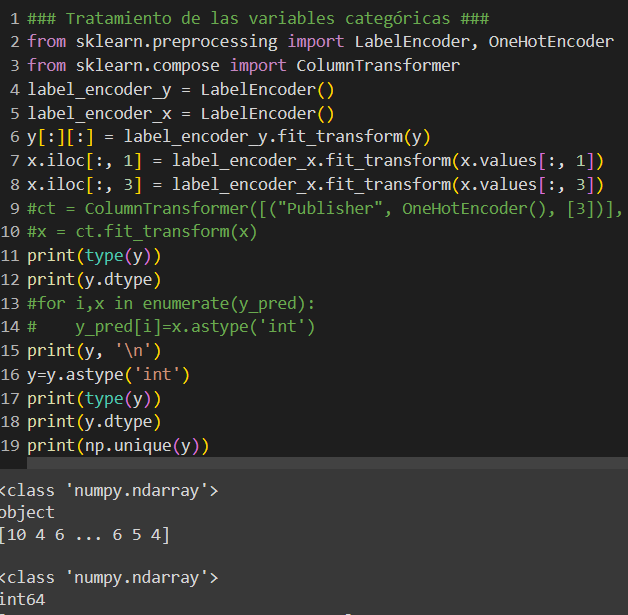
El siguiente paso es convertir el tipo de dato de año en un entero ya que se encuentra en un formato de Float.



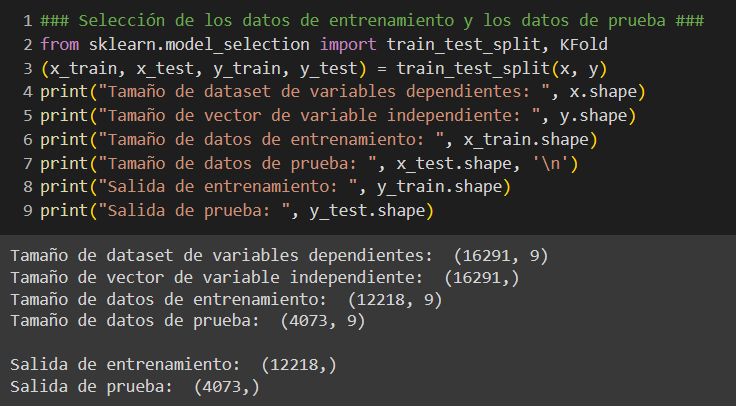
Ya con el dataset preparado se procede a generar un conjunto de datos con las variables dependientes y la variable independiente o de salida.



El siguiente paso ha sido el de codificar las variables categóricas. Es de resaltar que en este punto, no se debe olvidar que al etiquetar las variables categóricas con números, los algoritmos podrían dar un peso numérico, lo cual puede acarrear lecturas incorrectas.



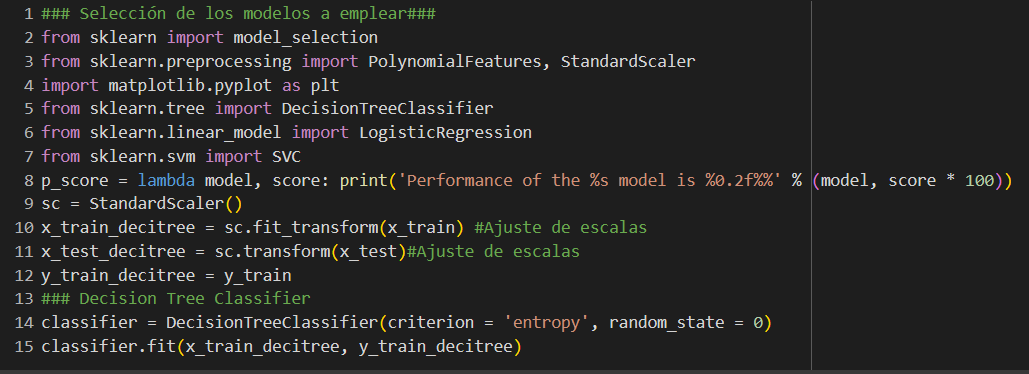
Se procede a asignar los datos que serán para entrenamiento y los que serán para prueba del conjunto de datos



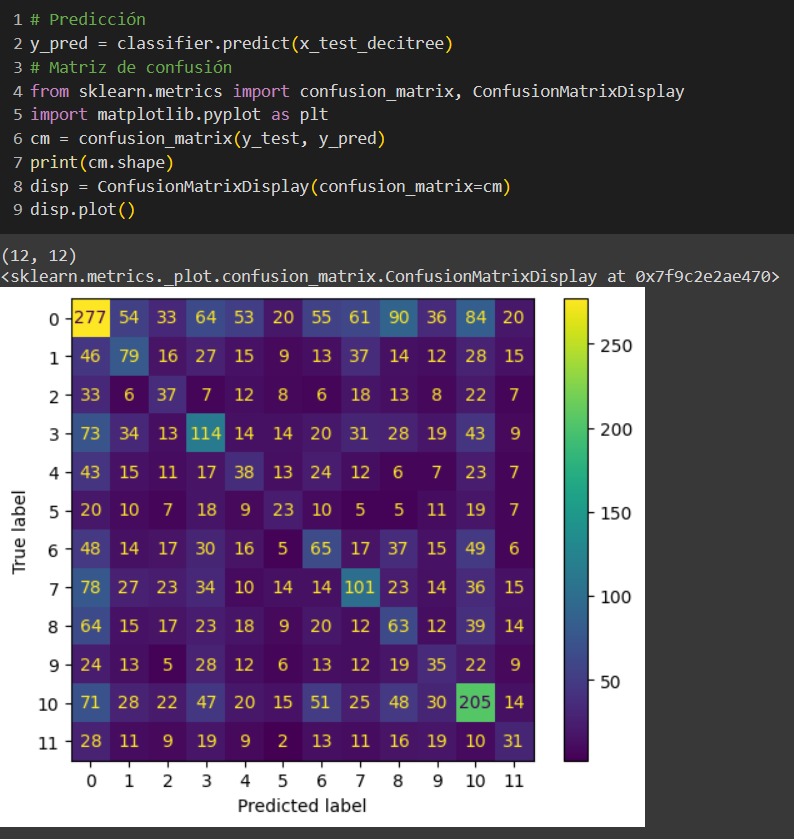
Para los datos de entrenamiento se emplearán 12218 entradas del total de 16291.

Las entradas restantes (4073), serán empleados como datos de test o prueba.

Los modelos con los que se trabajó fueron arboles de decisión de clasificación y regresión logística. Se intentó ejecutar un modelo de máquinas de soporte vectorial, pero esta presentó problemas al momento de realizar los ajustes por lo cual se descartó.



Se realiza el ajuste y la predicción con el árbol de decisión y se visualiza una matriz de confusión como métrica. Los resultados se pueden observar a continuación.

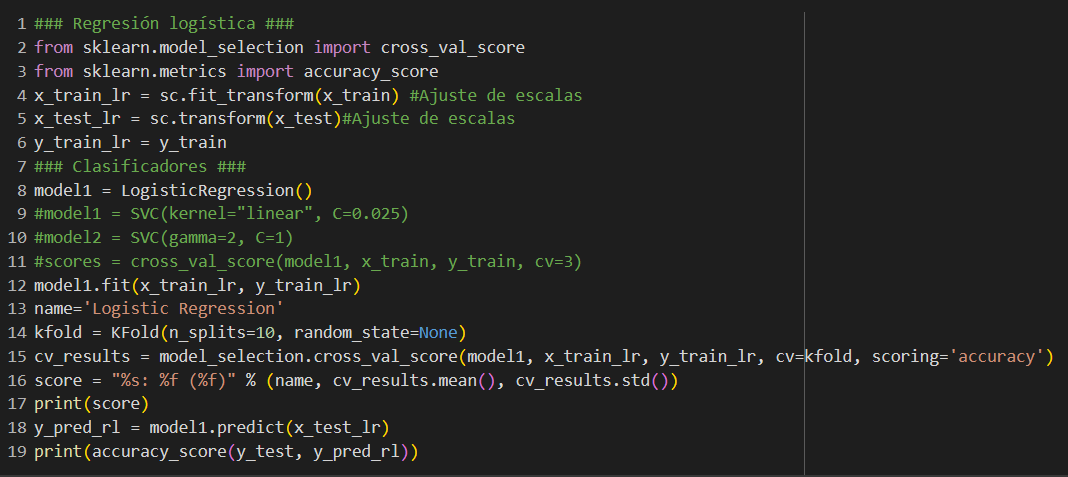


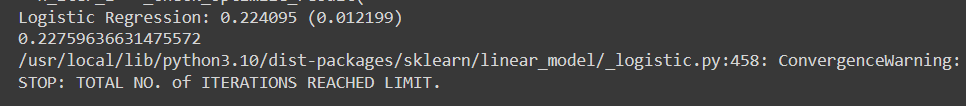
Después hacer el análisis, se concluye que la cantidad de aciertos es baja en este modelo, siendo un total de aciertos de 1068 elementos del total de 4073 del conjunto de datos de salida. Por lo tanto, se observa que tan solo el 26.22% de los datos han sido verdaderos positivos, así que se puede decir que es un porcentaje bastante bajo de acierto. Se puede observar en la matriz que los géneros que mayor cantidad de aciertos tuvieron fueron "Action" y "Sports". El género y su codificación se puede observar a continuación.

Texto

Descripción generada automáticamente

Para el modelo de regresión logística, se realiza el ajuste del modelo con los datos de entrenamiento al igual que con el modelo anterior y se realiza una validación cruzada K-fold. Estos son los resultados a continuación.





Como se puede observar, los resultados de los puntajes tanto de la media y la desviación estándar de la validación cruzada y de la precisión del modelo, no son muy buenos siendo estos de 0.224, 0.012 y 0.227 respectivamente.

Como conclusión se puede decir que se requiere el uso de otros diferentes modelos y de mas pruebas de ensayo y error con diversos parámetros de los modelos usados y otros modelos a implementar.