Segunda entrega de proyecto

Ariel Eduardo Bedoya Marín.
Identificación: 71275506
Marzo 2023.

Universidad de Antioquia.
Ingeniería de Sistemas.
Introducción a la inteligencia artificial para las ciencias e ingenierías.

Progreso alcanzado

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 y para cuales plataformas fueron lanzados.

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.

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.

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.

Queda pendiente realizar una separación de los datos para el entrenamiento y las pruebas. En este punto corresponderá dejar la variable “Genre” como la variable de salida del modelo a implementar, lo que significa que esta deberá convertirse en un dato discreto y por supuesto, categórico.

Lo anterior lleva a la inquietud a cerca de cómo completar estos años faltantes… Personalmente creo que usar alguna técnica de substitución aprendida durante lo que va del curso, resultaría algo complicado, ya que en la vida real se sabe con certeza en qué año ha sido publicado un determinado juego. De este modo, la forma ideal de completar estos datos faltantes sería basarse en los datos reales. Sin embargo, durante la limpieza de los datos se han completado usando la media y la desviación estándar por el simple hecho de experimentar. Posteriormente se espera completar, de ser posible, estos datos con los valores reales, aunque esto implicaría un trabajo un tanto tedioso aunque no imposible dada la cantidad de datos faltantes respecto al tamaño del dataset. También está en consideración

la búsqueda de un proyecto alternativo a este que permita alcanzar los objetivos requeridos para completar el curso.