Instituto Tecnológico de Costa Rica

Programa de Ciencias de Datos

Propuesta de Proyecto:

Sistema de Predicción Binaria para la clasificación de un conjunto de datos de video juegos

Curso: Big Data

Profesor: Dr. Juan Manuel Esquivel Rodríguez

Estudiante: Fabián Morera Gutiérrez

Abril, 2021

Investigación Preliminar

Aspectos generales

Se propone la realización de un sistema que permita la predicción binaria de una variable para un conjunto de datos de ventas de videojuegos a nivel global, el cual es recopilado por el sitio web especializado vgchartz.com. El dataset en cuestión describe datos puntuales sobre cada video juego, tales como el nombre del registro (videojuego), las plataformas en las que está disponible, el género, la clasificación por edad (Apto para todos, solo audiencias mayores de edad, etc.), la cantidad total de copias enviadas, las ventas totales a nivel global, así como por regiones geográficas (Europa, Norteamérica, Japón y otros) y el año de publicación.

Como set de datos secundario se propone utilizar los registros depurados y almacenados por la popular página web especializada Metacritics. Dicho conjunto de datos registra información puntual acerca de las críticas recibidas por cada videojuego, es decir, cada uno de esos representa una fila en la tabla. Entre los datos que se disponen en este dataset, se encuentran la cantidad de críticas positivas, negativas y neutrales aportadas por especialistas (revistas u otros sitios web), el puntaje final generado por Metacritic, la cantidad total de críticas de usuarios, tanto positivas, negativas y neutrales y un puntaje final de los usuarios.

La unión de los datos se realizará por medio del nombre estandarizado del registro (Nombre del videojuego sin espacios, mayúsculas o caracteres especiales).

Variable a predecir

La propuesta plantea realizar la predicción binaria de si el puntaje final de la crítica especializada (Es decir, no usuarios corrientes), supera el umbral de los 70 puntos. Dicha puntuación está basada en una escala de 100 y representa un valor cercano al promedio, lo cual permite una distribución bastante equitativa de los valores.

Detalles de los datasets

A continuación, se describen en mayor detalle los campos a utilizar en ambos datasets.

1. Ventas de videojuegos durante el 2019

- Nombre(String)
- Plataforma (String): Nombre de la plataforma en la que corre el videojuego.
- Género(String): Género del videojuego.
- Clasificación ESRB (String): Clasificación de audiencia según ESRB.

- Total de copias enviadas (Integer): Total de copias del videojuego enviadas por la casa de publicación.
- Ventas globales (Integer): Total de ventas en millones.
- Ventas en NA (Integer): Total de ventas en Norteamérica en millones.
- Ventas en PAL (Integer): Total de ventas en Europa en millones.
- Ventas en JP (Integer): Total de ventas en Japón en millones.
- Ventas en Otros (Integer): Total de ventas en otros países en millones.
- Año (Integer): Año de publicación del videojuego.

Link del recurso:
https://www.kaggle.com/ashaheedq/video-games-sales-2019?select=vgsales-12-4-2019-short.csv

2. Estadísticas de Videojuegos de Metacritic:

- Nombre(String)
- Crítico Positivo (Integer): Número de críticas positivas de medios especializados.
- Crítico Negativo (Integer): Número de críticas negativas de medios especializados.
- Crítico Neutral (Integer): Número de críticas neutrales de medios especializados, en base 100.
- Puntaje general de Críticos (Integer): Promedio la crítica especializada.
- Usuario Positivo (Integer): Número de críticas positivas de usuarios.
- Usuario Negativo (Integer): Número de críticas negativas de usuarios.
- Usuario Neutral (Integer): Número de críticas neutrales de usuarios.
- Puntaje general de Usuario (Integer): Promedio la crítica de usuarios regulares en base 10.

Link del recurso:
https://www.kaggle.com/skateddu/metacritic-all-time-games-stats