

|  |
| --- |
| **Universidad Cenfotec**  **Certificación XTOL Data Analytics & Big Data** |
| Yurguen Peñaranda Thomas  30 de marzo del 2020 |
| Course 5: Data Science with Python  Task 4: Complete a Data Science Capstone Project  “Predicting Bestselling Video Games” |
|  |





Contenido

[Introducción 2](#_Toc36069243)

[1. Planteamiento del problema a resolver 2](#_Toc36069244)

[2. Análisis Inicial de los datos: 2](#_Toc36069245)

[3. Definición de variable a predecir: 2](#_Toc36069246)

[4. Preprocesamiento de los datos: 3](#_Toc36069247)

[4.1. Eliminación de variables 3](#_Toc36069248)

[4.2. Codificación y cambio de tipo de dato de variables categóricas 3](#_Toc36069249)

[4.3. Eliminación de registros incompletos 3](#_Toc36069250)

[4.4. Set de datos final a entrenar: 3](#_Toc36069251)

[5. Labores generales para el Entrenamiento de los Modelos: 5](#_Toc36069252)

[5.1. Valor de aleatorización: 5](#_Toc36069253)

[5.2. Algoritmos a utilizar: 5](#_Toc36069254)

[5.3. Métricas de evaluación de la calidad de predicción de los modelos: 5](#_Toc36069255)

[5.4. Partición de los datos en el Set de Train y Test: 5](#_Toc36069256)

[5.5. Parámetros utilizados en los modelos: 6](#_Toc36069257)

[6. Entrenamiento y optimización de los Modelo de Predicción 6](#_Toc36069258)

[7. Selección del modelo de predicción a utilizar: 6](#_Toc36069259)

[7.1. Afinamiento del modelo (optimización de variables y Cross Validation) 6](#_Toc36069260)

[8. Resultados y análisis de las predicciones 7](#_Toc36069261)

[Bibliografía 8](#_Toc36069262)

# Introducción

*Best Tech* es una cadena de pequeñas tiendas minoristas de accesorios tecnológicos para computadora y celulares que quiere incursionar en la venta de videojuegos para las consolas actuales. *Best Tech* contrató a *Alert Analytics* para que realice un proyecto de Machine Learning que le permita predecir cuales de los videojuegos que serán lanzados próximamente, serán un éxito en ventas y de esta manera *Best Tech* se enfoque inicialmente a comercializar estos videojuegos exitosos.

Este documento es el entregable del proyecto que está desarrollando el *Alert Analytics* para *Best Tech* para generar un modelo que con base en características, permita predecir si un videojuego será un éxito en ventas o no.

En el documento se detallan todos los pasos realizados para generar el modelo de predicción: análisis inicial y preprocesamiento de los datos, definición de parámetros y selección de los algoritmos a utilizar, entrenamiento y validación de los resultados de predicción en el set de test, selección y optimización del mejor algoritmo, y análisis de resultados.

# Planteamiento del problema a resolver

*Best Tech* es una cadena de pequeñas tiendas minoristas de accesorios tecnológicos para computadora y celulares, que incluye: parlantes, cargadores, mouse, estuches, audífonos, baterías portátiles, etc. *Best Tech* quiere incursionar en la venta de videojuegos para las consolas actuales (Nintendo Switch, Playstation 4 y Xbox One). Debido a que la compañía apenas incursionará en la venta de estos productos, inicialmente solo quiere tener a la venta aquellos juegos que sean más exitosos.

Debido a lo anterior, *Best Tech* ha contratado a *Alert Analytics* para que haga un estudio mediante Machine Learning que permita predecir anticipadamente cuales de los videojuegos que serán lanzados próximamente en el mercado serán un éxito en cuanto a ventas, para que *Best Tech* se enfoque a comercializar estos videojuegos.

# Análisis Inicial de los datos:

Para resolver este proyecto se utilizará un dataset extraído del sitio web Kaggle (Kirubi, 2016) que incluye el volumen de unidades de ventas (en millones) de 16717 videojuegos que han sido comercializados desde 1980 hasta 2016.

El dataset incluye 16717 registros y 16 columnas. Las columnas son: nombre del videojuego, consola del videojuego, año de publicación, género, empresa que lo publicó, volumen de unidades vendidas (en millones) a nivel Global y por zonas geográficas (Norteamérica, Europa, Japón y Otras zonas), Clasificación por la Crítica y por los Usuarios, empresa desarrolladora y clasificación del videojuego.

# Definición de variable a predecir:

Como ya se indicó, el dataset incluye de manera cuantitativa el volumen de ventas de los videojuegos a nivel global y por zona, sin embargo, se debe definir un criterio para clasificar si un videojuego es un éxito de ventas o no. Se definirá que el criterio es que si un videojuego tiene un volumen de unidades vendidas mayor o igual a 1 millón será clasificado como un éxito de ventas y de lo contrario no.

Para esto, se creará una nueva columna llamada “Éxito” binaria que tendrá un valor de 1 a los videojuegos con un volumen de ventas global mayor o igual a 1 millón de unidades, o tendrá un valor de 0 si el volumen de ventas global es menor a 1 millón de unidades. Está columna de “Éxito” será la variable a predecir, por lo tanto se requerirá usar algoritmos de Clasificación.

# Preprocesamiento de los datos:

Con el fin disminuir los tiempos de procesamiento de entrenamiento de los modelos y buscar optimizar los resultados de predicción se realizaron diversas tareas de preprocesamiento. A continuación se detallan estas tareas realizadas.

## Eliminación de variables

De acuerdo al criterio de experto del analista, se procederá a eliminar del dataset las variables que detallan a continuación:

* **Name:** El nombre del videojuego no es una variable que se pueda usar para la predicción.
* **Year\_of\_Release:** Se considera que el año de lanzamiento del videojuego no es una variable que pueda usarse para predecir en nuevos videojuegos y que dejarla más bien podría sesgar la predicción.
* **Global\_Sales**, **NA\_Sales**, **EU\_Sales**, **JP\_Sales**, **Other\_Sales:** Debido a que se definió como variable a predecir la clasificación del videojuego como éxito de ventas o no, se considera que las columnas de volumen de ventas por unidades (global y por zonas) no se pueden mantener como predictores, porque para los nuevos videojuegos no se tendrán disponibles estos valores.

## Codificación y cambio de tipo de dato de variables categóricas

Las variables Platform, Genre, Publisher, Developer y Rating son variales categóricas donde se indica la etiqueta de plataforma, género, clasificación, empresa que publicó y desarrolló el videojuego.

Sin embargo, para realizar el entrenamiento de los algortimos de predicción, se requiere que los valores sean númericos, por lo tanto se utiliza una codificación de las etiquetas de estas variables, donde se le asigna un valor numérico a cada una de las posibles etiquetas para cada variable. Posteriormente, a estas variables se les procederá a cambiar el tipo a “**category**”.

## Eliminación de registros incompletos

Realizando un análisis inicial de los datos, se identificó que había registros que tenían vacía la columna de clasificación de la crítica o la clasificación de los usuarios. Para este caso en particular se eliminarán dichos registros que tienen al menos un valor incompleto. Luego de esta depuración quedaron 7017 con todas las columnas completa

## Set de datos final a entrenar:

Con las labores de preprocesamiento realizadas anteriormente, el dataset a entrenar de los clientes está conformado por 7017 registros y 10 variables, que se detallan a continuación.

Figura 1 Detalle del set de datos final a entrenar

| **#** | **Variable** | **Descripción** | **Posible Valor** | **Tipo de dato final** | **Uso** |
| --- | --- | --- | --- | --- | --- |
| 1 | Platform | Consola en la que se puede usar el videojuego: NES, Wii, Wii U, PS2, PS3, PS4, XB, X360, etc. | Etiquetas codificadas con valores de 0 a 30 | category (Categoría) | Variable de entrada (Predictor) |
| 2 | Genre | Género del videojuego. Incluye: Acción, Aventura, Peleas, Plataforma, etc. | Etiquetas codificadas con valores de 0 a 12 | category (Categoría) | Variable de entrada (Predictor) |
| 3 | Publisher | Compañía que publicó el videojuego. Incluye: Activision, Ubisoft, Sega, Electronic Arts, etc. | Etiquetas codificadas con valores de 0 a 581 | category (Categoría) | Variable de entrada (Predictor) |
| 4 | Critic\_Score | Calificación otorgada al videojuego por el personal de Metacritic[[1]](#footnote-1) | Valor numérico de 0 a 100 | int64 (Número entero) | Variable de entrada (Predictor) |
| 5 | Critic\_Count | Cantidad de personal de Metacritic que fueron tomados en cuenta para la calificación de Critic\_Score | Valor numérico | int64 (Número entero) | Variable de entrada (Predictor) |
| 6 | User\_Score | Calificación otorgada al videojuego por los usuarios suscriptores de Metacritic | Valor numérico de 0 a 10 | int64 (Número entero) | Variable de entrada (Predictor) |
| 7 | User\_Count | Cantidad de usuarios de Metacritic que fueron tomados en cuenta para la calificación de User\_Score | Valor numérico | int64 (Número entero) | Variable de entrada (Predictor) |
| 8 | Developer | Compañía que desarrolló el videojuego. Incluye: Nintendo, EA Sport, Game Freak, etc. | Etiquetas codificadas con valores de 0 a 1696 | category (Categoría) | Variable de entrada (Predictor) |
| 9 | Rating | Clasificación del videojuego basado en el contenido y en el público hacia el que va dirigido. Incluye: E (Todo público), E10+ (Mayores de 10 años), T (Mayores de 13 años), M (Mayores de 17 años), etc. | Etiquetas codificadas con valores de 0 a 8 | category (Categoría) | Variable de entrada (Predictor) |
| 10 | Éxito | Indica si el videojuego fue un éxito basado en el volumen de ventas. | * **0 ->** Volumen de ventas menor a 1 millón de unidades * **1 ->** Volumen de ventas mayor o igual a 1 millón de unidades | category (Categoría) | Variable a predecir |

# Labores generales para el Entrenamiento de los Modelos:

## Valor de aleatorización:

Se utiliza un valor de aleatorización (random\_state) de 42. Este valor permite que todas las particiones de datos que se realizan de manera aleatoria, siempre se dividan de la misma manera, lo cual permite la reproducibilidad de los resultados.

## Algoritmos a utilizar:

Para la variable que se requiere predecir (**Exito**) se entrenarán varios algoritmos de clasificación para evaluar cual genera resultados con mejor calidad de predicción. Los 6 algoritmos que se evaluarán en presentan en la sección 4, junto a los resultados de **Accuracy** y **F1** para el set de Test.

## Métricas de evaluación de la calidad de predicción de los modelos:

Para poder cuantificar el desempeño de la calidad de predicción de los modelos se utiliza la siguiente métrica de calidad:

* **Accuracy (Exactitud):** Cantidad de registros correctamente clasificados con respecto al total de registros.

El valor máximo de **Accuracy** es de 1, que quiere decir que la que se predijo la misma cantidad de registros que los registros totales. Por lo tanto, entre más alto es el valor de **Accuracy**, quiere decir que mejor es la calidad de predicción del modelo.

* **F1:** Es el promedio ponderado de Precision (porcentaje de predicciones positivas que fueron correctas) y Recall (el porcentaje de casos positivos detectados). Por lo tanto, este puntaje tiene en cuenta tanto los falsos positivos como los falsos negativos.

|  |  |  |
| --- | --- | --- |
|  |  |  |

* **Donde:**
  + **P ->** Precision
  + **R ->** Recall
  + **VP ->** Verdadero Positivo
  + **FP ->** Falso Positivo
  + **FN ->** Falso Negativo

El valor máximo de **F1** es de 1, que quiere decir que se predijo la totalidad de los registros. Por lo tanto, entre más alto es el valor de **F1**, quiere decir que mejor es la calidad de predicción del modelo.

## Partición de los datos en el Set de Train y Test:

El dataset a entrenar se dividió en un subset de Train y uno de Test. A continuación se detallan los porcentajes y cantidad de registros para cada subset.

Figura 2 Detalle de la partición de los datos en los sets de Training y Testing

|  |  |  |
| --- | --- | --- |
| **Set** | **Porcentaje de datos** | **Cantidad de registros del dataset** |
| **Training** | 80% | 5613 |
| **Testing** | 30% | 1404 |
| **TOTAL** | **100%** | **7017** |

## Parámetros utilizados en los modelos:

Los modelos se entrenan de modo que el algoritmo automáticamente seleccione los valores óptimos para generar el modelo de predicción.

# Entrenamiento y optimización de los Modelo de Predicción

Una vez realizado el preprocesamiento y la partición de los datos según lo indicado anteriormente, se procede a entrenar el set de datos de entrenamiento (**train**) con los 6 algoritmos indicados anteriormente en la variable que se requieren predecir en el dataset**. A continuación el análisis detallado de los resultados de los parámetros utilizados en el entrenamiento y los resultados de predicción en el set de prueba (test) para cada algoritmo.**

Figura 3 Resultados de Accuracy y F1 en el set de Test para los 6 modelos entrenados

|  |  |  |
| --- | --- | --- |
| **Modelo** | **Accuracy** | **F1** |
| LR: Logistic Regression | 0.8198 | 0.7819 |
| RF: Random Forest Classifier | 0.8782 | 0.8683 |
| SVC: Support Vector Classifier | 0.8184 | 0.7692 |
| GBC: Gradient Boosting Classifier | 0.8746 | 0.8629 |
| KNC: K Neighbors Classifier | 0.8205 | 0.8109 |
| GNB: Gaussian Naïve Bayes | 0.8198 | 0.8013 |

# Selección del modelo de predicción a utilizar:

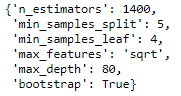
Con base en los resultados presentados anteriormente, se puede ver que el modelo que obtuvo mejores resultados de **Accuracy** y **F1** es el generado con el algoritmo RF (Random Forest Classifier), por lo tanto este será el algoritmo a utilizar para generar las predicciones.

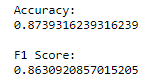
## Afinamiento del modelo (optimización de variables y Cross Validation)

Con el fin de buscar optimizar los niveles de predicción del modelo, se procede a realizar diversas iteraciones de entrenamiento variando los valores de los siguientes 6 parámetros del algoritmo RF: n\_estimators, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf y bootstrap.

También, se utilizó el **Cross Validation**, que es un método de entrenamientoque lo que hace es partir aleatoriamente el set de datos de entrenamiento en k grupos (en este caso 3 grupos), luego realiza el modelo de predicción con k-1 grupos y valida el desempeño de la calidad de predicción con el grupo de datos restantes, al comparar los datos reales contra los datos predichos. Esto se realiza k veces hasta haber validado la calidad del modelo con cada uno de los k grupos. Este método de control lo que busca es tratar de disminuir el overfitting al validar la calidad de predicción con datos que el modelo nunca ha visto. Para este entrenamiento se definió que se realizarían 100 iteraciones, para buscar optimizar los resultados de predicción.

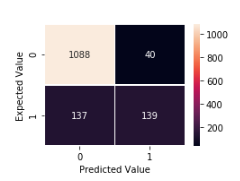
A continuación se presentan los valores de los parámetros optimizados y los resultados de los indicadores de calidad de predicción.





# Resultados y análisis de las predicciones

Figura 4 Gráfico de comparación de valores Reales vs Predichos con el modelo RF optimizado



Con el gráfico anterior y basándose en el resultado de **Accuracy** se podría decir que el modelo tiene un buen nivel de predicción, pues logró predecir correctamente el 87% de los datos en el set de Test.

Sin embargo, para este problema en particular lo realmente no es el nivel de predicción general, sino más bien la capacidad de predicción en cada registro en particular, para que a la hora de que se esté prediciendo si cada videojuego será un éxito de ventas o no y tomar en cuenta este criterio para ver que videojuegos comercializará *Best Tech* en sus tiendas. Para analizar la calidad de predicción individual del modelo es mejor tomar en cuenta el resultado del indicador **F1**.

Con el modelo optimizado, el resultado del indicador **F1** es de 0,86. Con estos resultados se puede concluir que el modelo tiene un excelente nivel de predicción, pues el 86% de las veces el modelo logrará predecir correctamente si un videojuego será un éxito de ventas o no, y por ende *Best Tech* si lo puede usar para tomar sus decisiones comerciales acerca de los videojuegos que tendrá a la venta.

# Bibliografía

Kirubi, R. (2016). *Video Game Sales with Ratings*. Obtenido de Kaggle: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

1. Sitio web que recopila reseñas de álbumes de música, videojuegos, películas, programas de televisión, DVD y libros [↑](#footnote-ref-1)