PREMIOS AL CINE

INDICE.-

- FEATURE BINNING

REDUCCION DE DIMENSIONALIDAD

- PCA
- MCA

MODELOS MACHINE LEARNING

- ALGORITMOS DE CLASIFICACION
  - REGRESIÓN LOGÍSTICA
  - KNN
  - ARBOLES DE DECISIÓN
  - RANDOM FOREST
  - SVM
- ANÁLISIS DE LAS MÉTRICAS MODELO DE CLASIFICACIÓN
  - BALANCEO DE CLASES
    - SMOTE
- MÉTODOS DE ENSAMBLE
  - BOOSTING
    - XG BOOST
    - LIGHT GBM

ALTERNATIVA VARIABLES DEFINIDAS POR EL MODELO

- RANDOM FOREST
- XG BOOST

CLUSTERING

- K MEANS Y MÉTRICAS

CONCLUSIONES

INTRODUCCIÓN.-

El objetivo de este análisis es estudiar el comportamiento de la cantidad de premios recibidos por los distintos géneros de la industria cinematográfica e identificar si hay posibilidades de que una película reciba o no un premio, analizando si existen relaciones entre la recepción de premios y otras variables como el género de la misma.

CONTEXTO COMERCIAL

La premiación y reconocimiento dentro de la industria cinematográfica es un misterio. Hemos escuchado cientos de casos de películas muy taquilleras o actuaciones fantásticas que nunca fueron reconocidas con un premio a pesar de recibir excelentes críticas por parte de la audiencia.

DEFINICION DEL OBJETIVO

He sido contratada por una empresa de la industria cinematográfica que produce películas de diversos géneros, quien me solicita estimar la posibilidad de que una película del género "Fantasía y/o animación" reciba un premio.

PROBLEMA COMERCIAL

Existe la creencia de que siempre películas del género dramático, por el simple hecho de ser de este género serán mayormente reconocidas que otras, lo cual lleva a productores y directores de otros géneros a la preocupación. En particular, en este caso, se analiza el género de la "Fantasía y animación", ya que cada vez se logran producciones más asombrosas gracias a los avances tecnológicos, pero los productores no ven el

reconocimiento reflejado del trabajo y el costo que representa producir una película de esta índole.

Se responderán las siguientes preguntas:

- ¿A medida que pasan los años, más premios se entregan?
- ¿Las películas más taquilleras son las más premiadas?
- ¿Existe relación entre la popularidad, las críticas obtenidas y la recepción de premios?
- ¿El género dramático es el género más premiado?
- ¿Cómo ha sido la premiación para el género fantasía/animación en relación al género dramático a lo largo de la historia?
- ¿El género de la fantasía/animación ha aumentado sus niveles de popularidad a lo largo de la historia?
- ¿Hay algún director de la firma que haya ganado más premios en este género que en otros?

CONTEXTO ANALÍTICO

Para este proyecto se analizará un datasets que contiene información de 2000 películas, para cada título, podemos conocer: su género, año, director, duración, calificación, ingresos, presupuesto y recaudación en DVD en millones de dólares, el país de procedencia, la cantidad de premios, espectadores y criticas (positivas y negativas) recibidas y la popularidad de las mismas.

EDA Y DATA WRANGLING.-

El **Análisis Exploratorio de Datos o Exploratory Data Analysis**, tiene como finalidad examinar los datos previamente a la aplicación de cualquier técnica estadística, para conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas.

Data wrangling también se conoce como organización de datos. Es un término general que describe varios procesos, todos diseñados para tomar datos sin procesar y transformar los conjuntos de datos complejos y desordenados en formatos más fáciles de usar.

ANÁLISIS PREMILINAR DE DATOS

DICCIONARIO DE VARIABLES

El dataset original contiene 2000 filas y 15 columnas. Las columnas originales del dataset son:

- Título: es el nombre de la película. Tipo de variable *object*.
- Género: según las características y el formato la película puede ser del género comedia, musical, western, fantasía y animación, terror, thriller, ciencia ficción, drama, épica o romántica. Tipo de variable *object*.
- Año: es el año de la producción de la película, desde 1980 a 2023. Tipo de variable *category*.
- Director: refiere a quien dirigió la producción de la misma. Tipo de variable *object*.
- Duración: en minutos. Tipo de variable *object*.
- Calificación: refiere a la descripción de la madurez del contenido del 0 (cero) al 10 (diez). Se establecen en función de varios factores, como el contenido sexual, el nivel de violencia, el consumo de drogas y el uso de lenguaje profano. Tipo de variable *float*.
- Ingresos (millones): se refiere al dinero recibido por la reproducción en cines y el merchandaising. Tipo de variable *float*.
- Presupuesto (millones): gastos incurridos para su producción. Tipo de variable *float*.
- País: procedencia de la película Tipo de variable *object*.
- Premios: cantidad de premios nacionales/internacionales recibidos. Tipo de variable *int*.
- Espectadores: cantidad de personas que vieron la película en cines. Tipo de variable *int*.
- Críticas Positivas: cantidad de reseñas positivas recibidas por la audiencia. Tipo de variable *int*.
- Críticas Negativas: cantidad de reseñas negativas recibidas por la audiencia. Tipo de variable *int*.
- Popularidad: refiere al interés que genera la película entre los espectadores, del 0 (cero) al 10 (diez). Tipo de variable *float*.
- Recaudación en DVD (millones): dinero recibido por la venta de DVD. Tipo de variable *float*.

A lo largo del trabajo se agregan nuevas columnas a los fines de obtener otro tipo de información para el análisis:

- Resultado (millones): resultado neto por película. Tipo de variable *float*.

Ingresos (millones) + Recaudación en DVD (millones) – Presupuesto (millones)

- Criticas: criticas netas por película. Tipo de variable *int*.

Críticas Positivas – Críticas Negativas

- Premios_binario: se redefine la variable “premios” como binaria, para utilizarla en modelos de machine learning. Tipo de variable *int*.

Encodeo manual de la variable. Definiendo como 0 (cero) NO RECIBIO PREMIOS y 1 (uno) RECIBIO PREMIO.

- genero_binario: se redefine la variable “género” como binaria, para utilizarla en modelos de machine learning. Tipo de variable *int*.

Encodeo manual de la variable. Definiendo como 0 (cero) GENERO FANTASÍA/ANIMACIÓN y 1 (uno) RESTO DE LOS GENEROS.

- genero_label: se redefine la variable “género” como binaria, para utilizarla en modelos de machine learning. Tipo de variable *int*.

Encodeo con label encode de la variable. Definiendo como 0 (cero) GENERO FANTASÍA/ANIMACIÓN y 1 (uno) RESTO DE LOS GENEROS.

- Categoría: creación de una columna para convertir la variable “Premios” a variable categórica para ser utiliza en modelos de machine learning. Tipo de variable *category*.

Categorización de la variable según bins definidos en función de la cantidad de premios recibidos. En la columna se encuentran las categorías MUY GANADORAS para películas que han recibido 9 o 10 premios, MEDIANAMENTE GANADORAS para películas que han recibido entre 4 y 8 premios inclusive y como POCO GANADORAS aquellas que han recibido 3 premios o menos.

DETECCIÓN DE DUPLICADOS, NULOS Y ERRÓNEOS

A través de la aplicación de métodos, se identifica que no existen en el dataset valores duplicados, nulos ni erróneos.

```
# Detección y tratamiento de valores nulos

valores_nulos = df_peliculas.isnull().sum()
print(valores_nulos)
```

```
Título                          0
Género                          0
Año                             0
Director                        0
Duración                        0
Calificación                    0
Ingresos (millones)             0
Presupuesto (millones)          0
País                            0
Premios                         0
Espectadores                    0
Críticas Positivas              0
Críticas Negativas              0
Popularidad                     0
Recaudación en DVD (millones)   0
dtype: int64
```

```
# Detección y tratamiento de valores NaN

valores_NaN = df_peliculas.isna().sum()
print(valores_NaN)
```

```
Título                          0
Género                          0
Año                             0
Director                        0
Duración                        0
Calificación                    0
Ingresos (millones)             0
Presupuesto (millones)          0
País                            0
Premios                         0
Espectadores                    0
Críticas Positivas              0
Críticas Negativas              0
Popularidad                     0
Recaudación en DVD (millones)   0
dtype: int64
```

```
# Análisis datos duplicados

valores_duplicados = df_peliculas.duplicated().sum()
print (valores_duplicados)
```

```
0
```

REEMPLAZO DE DATOS Y CATEGORIZACIÓN DE LAS VARIABLES

Se modifica el tipo de la variable año para que sea mejor su utilización y se asigna nombre a cada uno de los géneros.

También se crean las columnas detalladas en el apartado anterior.

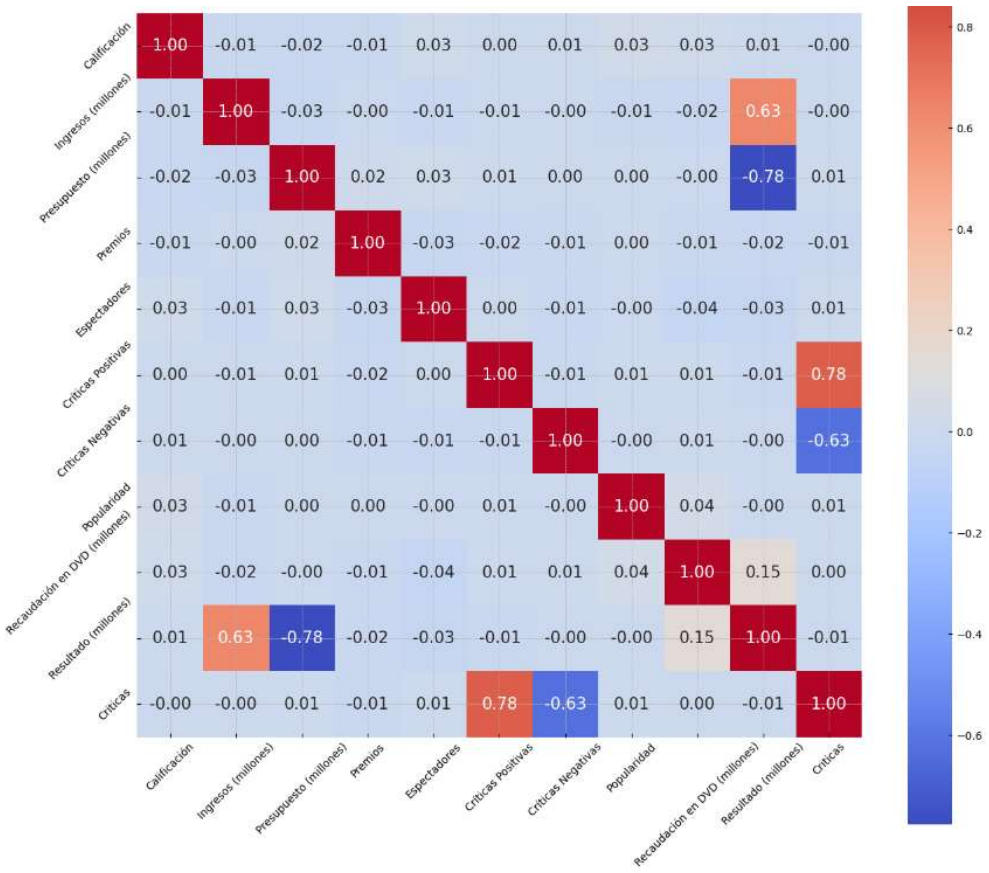
ANÁLISIS ESTADÍSTICO PRELIMINAR

A través del método *.describe* se visualiza un análisis estadístico preliminar de las variables numéricas.

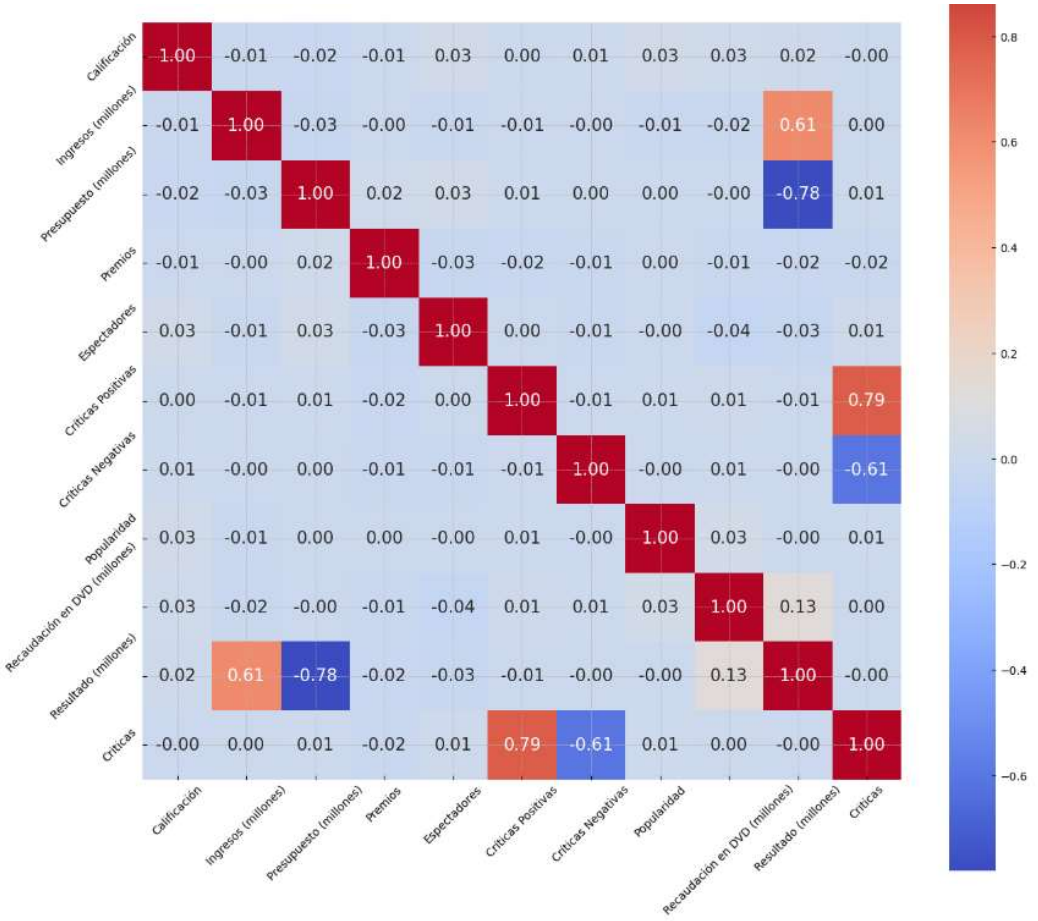
	Calificación	Ingresos (millones)	Presupuesto (millones)	Premios	Espectadores	Críticas Positivas	Críticas Negativas	Popularidad	Recaudación en DVD (millones)	Resultado (millones)	Criticas
count	2000.000000	2000.00000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	4.899800	9.96965	17.357900	5.019000	511094.616500	75.359500	29.643500	4.980850	2.497050	-4.891200	45.716000
std	2.900113	5.86335	7.267939	3.159688	286494.162803	14.639447	11.659648	2.941201	1.460021	9.564332	18.781739
min	0.000000	0.00000	5.000000	0.000000	1123.000000	50.000000	10.000000	0.000000	0.000000	-28.000000	0.000000
25%	2.400000	4.90000	10.900000	2.000000	274333.500000	63.000000	19.000000	2.400000	1.200000	-11.800000	32.000000
50%	5.000000	9.90000	17.300000	5.000000	508741.500000	75.500000	29.000000	4.900000	2.500000	-5.100000	46.000000
75%	7.400000	15.10000	23.800000	8.000000	758798.250000	88.000000	40.000000	7.600000	3.800000	1.900000	60.000000
max	10.000000	20.00000	30.000000	10.000000	999530.000000	100.000000	50.000000	10.000000	5.000000	19.100000	89.000000

ANÁLISIS DE CORRELACIÓN ENTRE VARIABLES

Se elabora un heatmap para analizar si existe correlación entre las variables. Primero se elabora con el método Pearson.



Como se visualiza no existe una correlación lineal alta entre las variables. Por este motivo, se analiza aplicando el método Spearman.



Tampoco se observa niveles de correlación altos entre las variables.

ANÁLISIS DE LA VARIABLE TARGET "PREMIOS"

Se define a la variable PREMIOS como la variable target a analizar.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central se representan a través de un número que permite entender la región central de un conjunto de valores de datos. Las tres medidas de tendencia central más utilizadas son la media, la mediana y la moda. En el caso de análisis estos valores son:

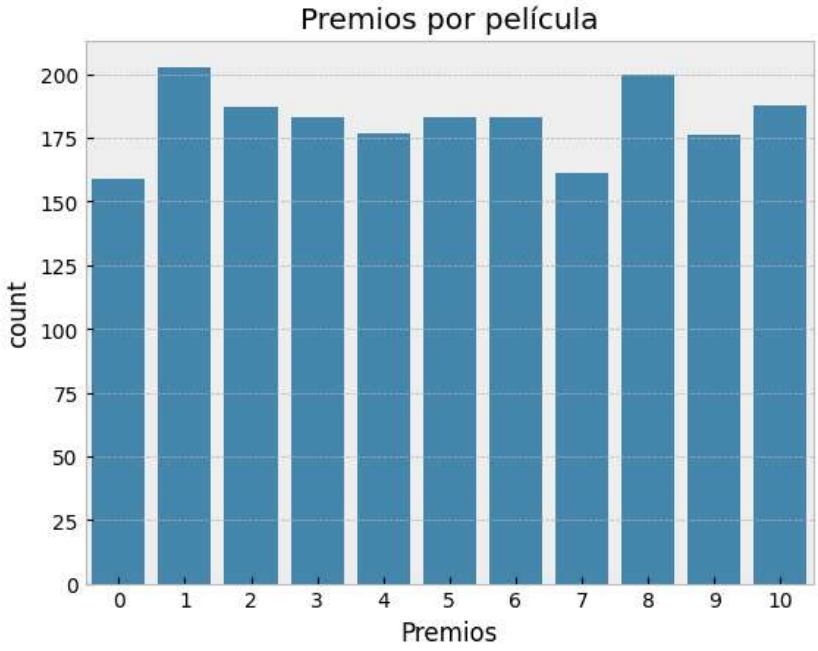
- Media: 5.019
  Es el promedio de la cantidad de premios recibidos
- Mediana: 5.0
  Es el valor medio del conjunto de datos cuando los valores se ordenan de forma ascendente o descendente
- Moda: 1
  Representa el valor más común dentro del conjunto de datos.

Con estos valores, podría anticiparse que la variable no tiene una distribución normal, lo confirmaremos en el próximo apartado.

ANÁLISIS DE LA DISTRIBUCIÓN

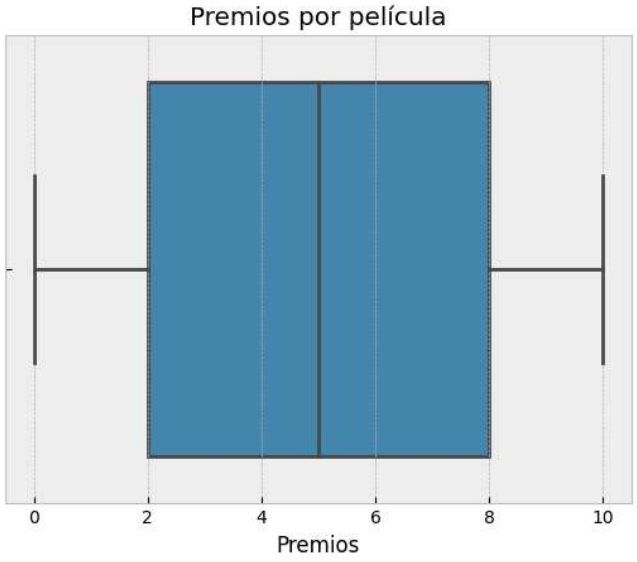
Mediante la aplicación del "Test de Shapiro" se analiza la distribución de la vaiable. Si el valor p es menor que el nivel de significancia predefinido (0.05), se rechaza la hipótesis nula y se concluye que los datos no siguen una distribución normal. Si el valor p es mayor que el nivel de significancia, no hay suficiente evidencia para rechazar la hipótesis nula y se puede considerar que los datos siguen una distribución normal.

El p valor es menor a 0.05, por lo tanto la variable Premios **no sigue una distribución normal.**



IDENTIFICACIÓN DE OUTLIERS

Aplicando MAD arroja que no existen outliers, se visualiza en el gráfico.



ANALISIS DE LAS HIPÓTESIS

Con el objetivo de responder las preguntas de interés del proyecto, se plantean hipótesis que serán aceptadas o rechazadas, a través de visualizaciones gráficas y mediante el análisis de coeficientes de correlación y el p-valor.

Coeficiente de Correlación: El coeficiente de correlación mide la fuerza y la dirección de una relación lineal entre dos variables. El valor del coeficiente de correlación varía entre -1 y 1. Un valor de 1 indica una correlación positiva perfecta. Un valor de -1 indica una correlación negativa perfecta. Un valor cercano a 0 indica una correlación débil o inexistente.

Dos tipos comunes de coeficientes de correlación son el coeficiente de correlación de Pearson y el coeficiente de correlación de Spearman. El primero se utiliza cuando las variables son continuas y tienen una relación lineal, mientras que el segundo se utiliza cuando las variables pueden tener una relación monotónica pero no necesariamente lineal.

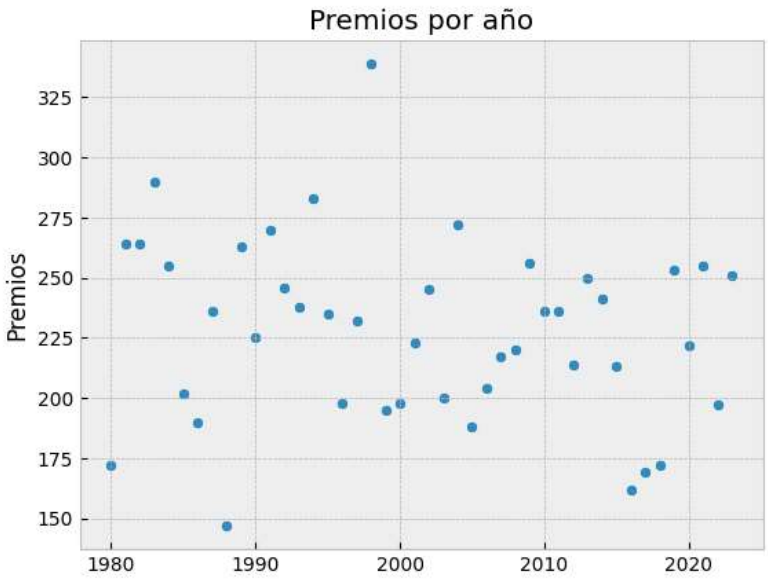
Valor p (p-value): El valor p es una medida de la evidencia en contra de una hipótesis nula en estadística. En el contexto de la correlación, el valor p se utiliza para evaluar la significancia estadística de la relación observada. Un valor p pequeño (generalmente menor que un umbral, como 0.05) sugiere que la relación observada no es probable que sea el resultado del azar y que la correlación es estadísticamente significativa.

En resumen, el coeficiente de correlación cuantifica la fuerza y la dirección de la relación entre dos variables, mientras que el valor p evalúa la significancia estadística de esa correlación. Ambos son útiles para interpretar y entender las relaciones entre variables en un conjunto de datos.

PREMIOS POR AÑO

Pregunta hipótesis: ¿A medida que pasan los años, más premios se entregan?

Gráficamente:



Hipótesis nula: No existe relación entre las variables año y premios

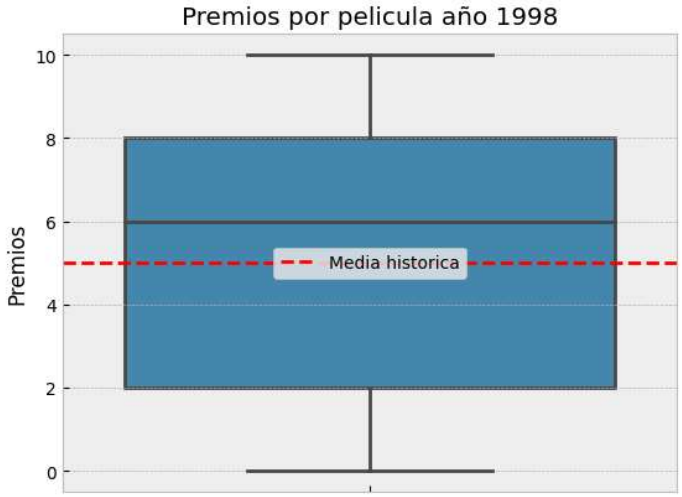
Hipótesis alternativa: Existe relación entre las variables año y premios

Coeficiente de Spearman: -0.18948259600000783. Podría significar una correlación negativa entre las variables, pero se trataría de una relación débil, ya que el coeficiente es bajo.

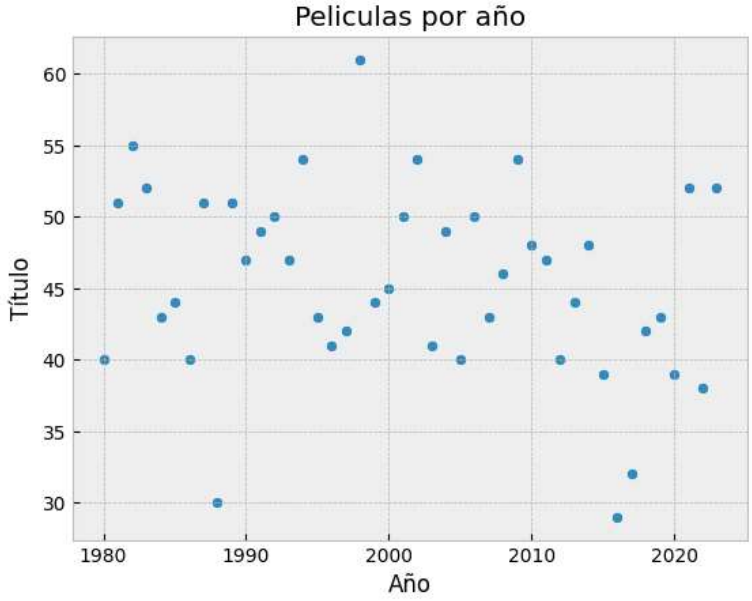
P-valor: 0.21798703796320967. En este caso no es menor al 0.05, lo cual sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Interpretación: Estadísticamente no se puede rechazar la hipótesis nula. Gráficamente se observa que no hay una relación lineal entre las variables, y que la entrega de premios no aumenta linealmente a medida que pasan los años. Se visualiza en el grafico que hubo un año, 1998, en el que más premios se entregaron. Esto se debe a dos factores:

1- En promedio cada película históricamente recibe 5 premios, para ese año en particular hubieron 13 películas que recibieron más premios de lo normal.



2- Por otro lado fue el año en que se produjeron la mayor cantidad de películas.

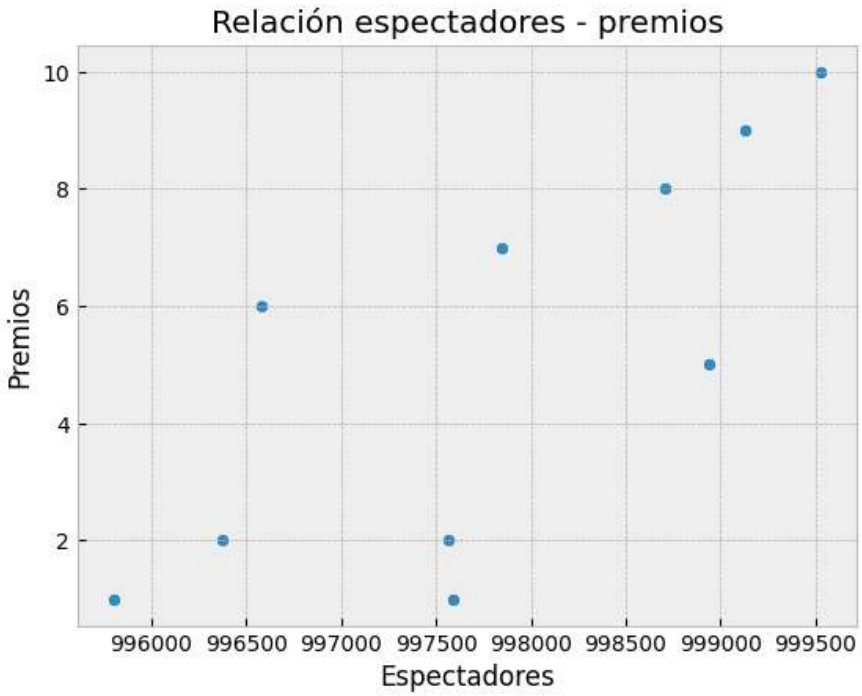


Esto demuestra, no solo que la cantidad de premios no aumenta año tras año, si no que existen años donde por la cantidad o calidad solamente de las películas producidas se entregaron más premios.

TOP 10 PELICULAS MAS TAQUILLERAS

Pregunta hipótesis: ¿Las películas más taquilleras son las más premiadas?

Gráficamente:



Hipótesis nula: No existe relación entre las variables premios y espectadores

Hipótesis alternativa: Existe relación entre las variables premios y espectadores

Coeficiente de Pearson entre las variables espectadores y premios: 0.75. Habría una correlación fuerte entre las variables.

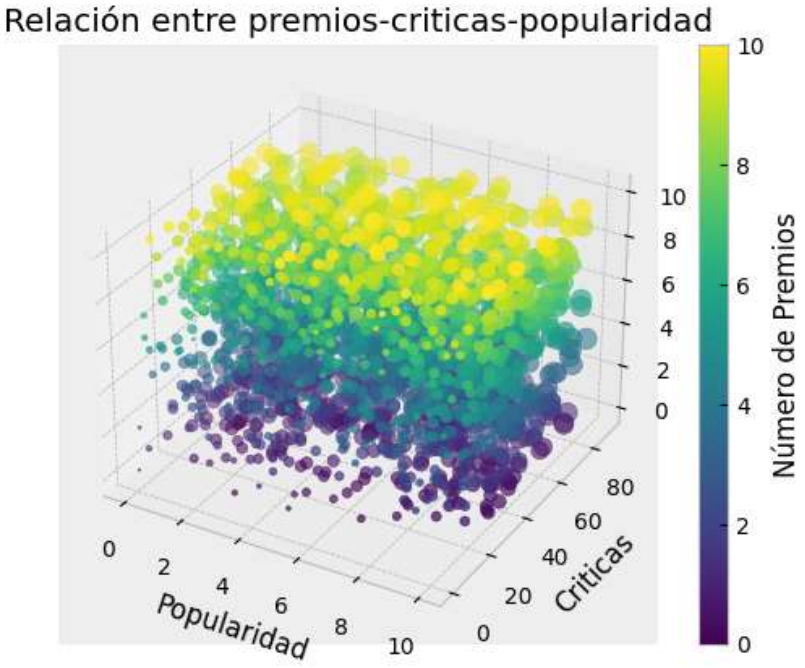
P-valor entre las variables espectadores y premios: 0.0119 En este caso es menor al 0.05, lo cual sugiere que hay suficiente evidencia para rechazar la hipótesis nula.

Interpretación: Estadísticamente se podría rechazar la hipótesis nula. Gráficamente e observa que existe relación entre la cantidad de premios recibidos por las 10 películas más taquilleras y la cantidad de espectadores.

ANÁLISIS PREMIOS-CRÍTICAS-POPULARIDAD

Pregunta hipótesis: ¿Existe relación entre la popularidad, las críticas obtenidas y la recepción de premios?

Gráficamente:



Hipótesis nula: No existe relación entre las variables premios, popularidad y críticas

Hipótesis alternativa: Existe relación entre las variables premios, popularidad y críticas

Coeficiente de Pearson entre las variables premios y críticas: -0.007065761096766035. No habría una correlación fuerte entre las variables.

P-valor entre las variables premios y críticas: 0.7521567217775721. En este caso no es menor al 0.05, lo cual sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Coeficiente de Pearson entre las variables premios y popularidad: 0.0012503294066048198. No habría una correlación fuerte entre las variables.

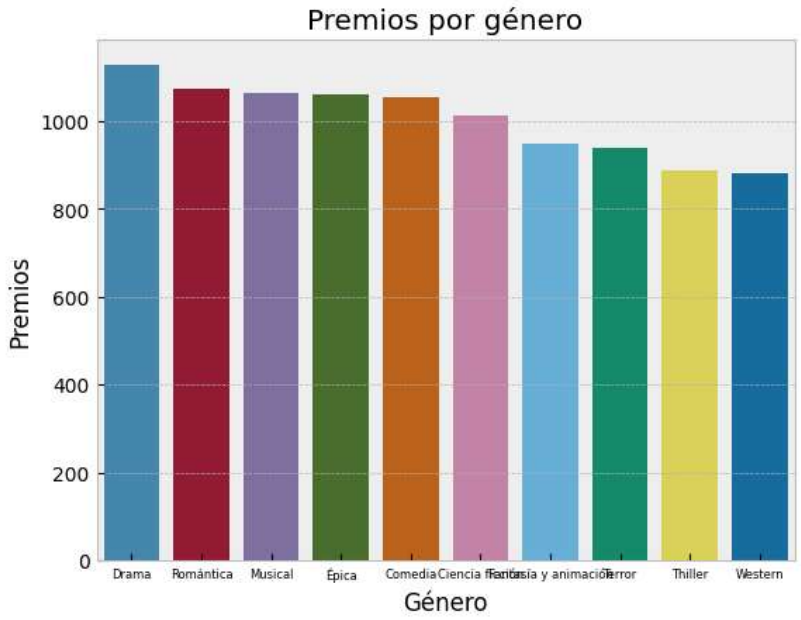
P-valor entre las variables premios y popularidad: 0.9554362126679201. En este caso no es menor al 0.05, lo cual sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

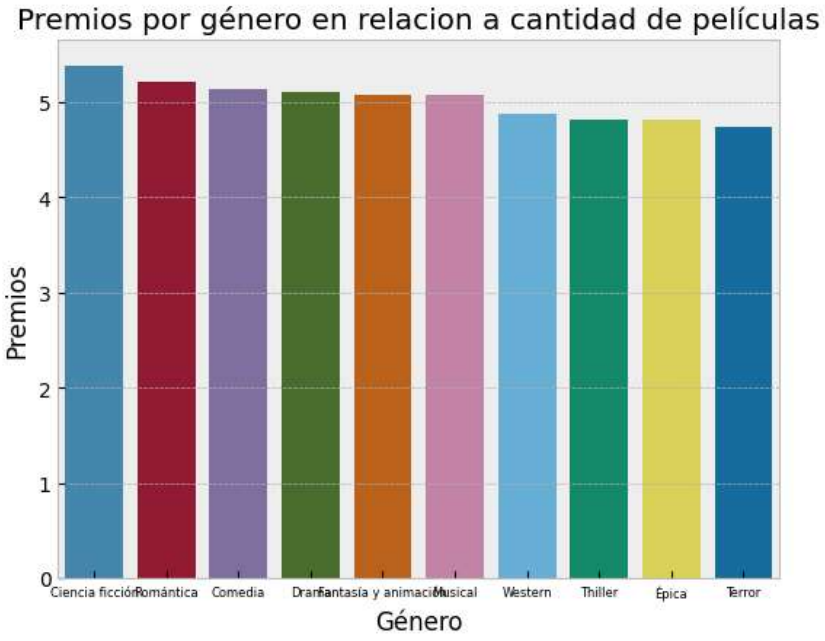
Interpretación: Estadísticamente no se puede rechazar la hipótesis nula. Gráficamente se visualiza que no existe relación lineal entre la popularidad, las críticas y la recepción de premios.

ANÁLISIS POR GÉNERO

Pregunta hipótesis: El género drama, ¿es el género más premiado?

Gráficamente:





Hipótesis nula: No existe relación entre las variables premios y género

Hipótesis alternativa: Existe relación entre las variables premios y género

Coeficiente de Spearman entre las variables premios y género: -0.7939393939393938. Habría una correlación inversa fuerte entre las variables.

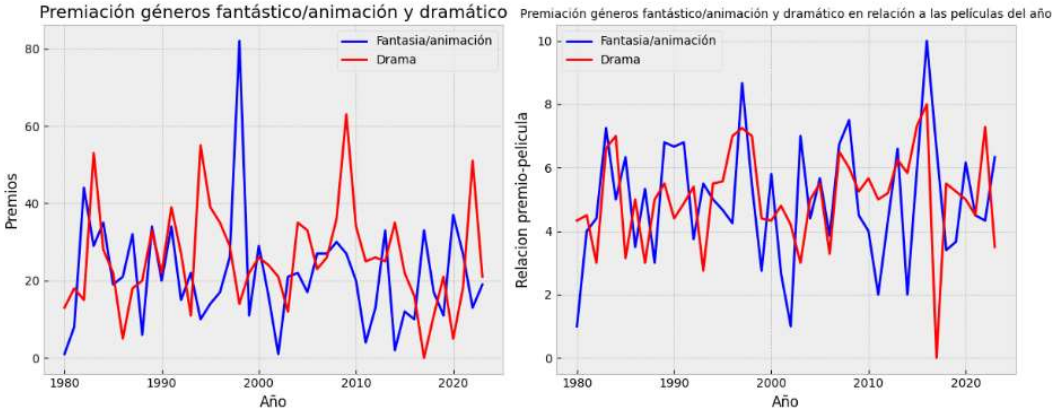
P-valor entre las variables premios y género: 0.0060999233136969115. En este caso es menor al 0.05, lo cual sugiere que hay suficiente evidencia para rechazar la hipótesis nula.

Interpretación: Estadísticamente podría rechazarse la hipótesis nula. Analizando gráficamente, en una primera instancia podría observarse que efectivamente es el género dramático el que más premios ha recibido. Pero si se analiza la cantidad de premios en relación a la cantidad de películas de cada género producidas, no sería este género el más premiado, si no el género de la Ciencia Ficción.

ANÁLISIS PREMIACIÓN GÉNERO FANTASIA/ANIMACIÓN

Pregunta hipótesis: ¿Existe relación entre los premios recibidos por el género fantasía/animación en relación al género dramático a lo largo de la historia?

Gráficamente:



Hipótesis nula: No existe relación entre las variables premios y género, para los géneros específicos drama y fantasía/animación

Hipótesis alternativa: Existe relación entre las variables premios y género, para los géneros específicos drama y fantasía/animación

Coeficiente de Spearman entre las premios y género, para los géneros específicos drama y fantasía/animación: 0.1793279878979584. Habría una correlación débil entre las variables.

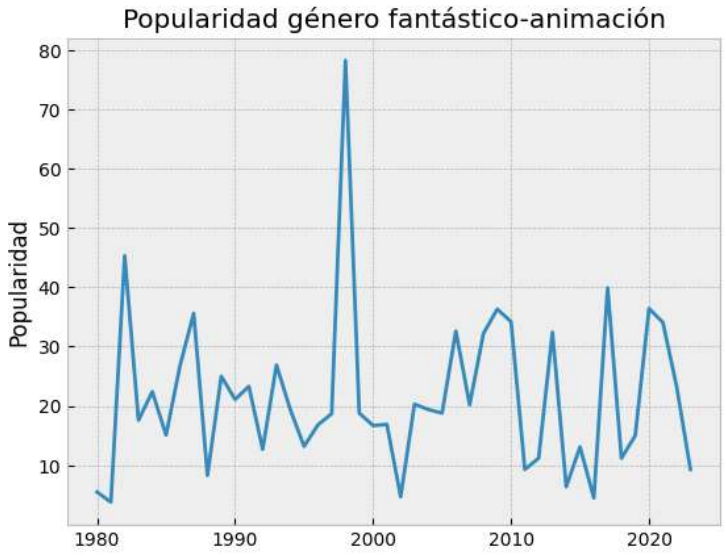
P-valor entre las variables premios y género, para los géneros específicos drama y fantasía/animación: 0.2441185101894924. En este caso no es menor al 0.05, lo cual sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Interpretación: Estadísticamente no se puede rechazar la hipótesis nula. Gráficamente se visualiza que no existe ninguna relación entre los premios recibidos por cada género a lo largo de los años ni en término de total de premios ni en promedio en base a la cantidad de películas producidas para cada año.

ANÁLISIS POPULARIDAD GÉNERO FANTASIA/ANIMACIÓN

Pregunta hipótesis: ¿El género de la fantasía/animación ha aumentado sus niveles de popularidad a lo largo de la historia?

Gráficamente:



Hipótesis nula: No existe relación entre las variables premios y género, para los géneros específicos drama y fantasía/animación

Hipótesis alternativa: Existe relación entre las variables premios y género, para los géneros específicos drama y fantasía/animación

Coeficiente de Spearman entre las variables popularidad y año, para el género fantasía/animación: 0.02452777019012809. No habría correlación entre las variables, o sería prácticamente nula.

P-valor entre las variables entre popularidad y año, para el género fantasía/animación: 0.8744263353507186. En este caso no es menor al 0.05, lo cual sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Interpretación: Estadísticamente no se puede rechazar la hipótesis nula. Gráficamente se visualiza que el género de la fantasía/animación no ha aumentado sus niveles de popularidad a lo largo de la historia, hay un pico de popularidad en el año 1998, pero luego los niveles varian año a año, aumentan y disminuyen constantemente.

ANÁLISIS DIRECTORES

Pregunta hipótesis: ¿Hay algún director de la firma que haya ganado más premios en este género que en otros?

Gráficamente:

Director	Ciencia ficción	Comedia	Drama	Fantasía y animación	Musical	Romántica	Terror	Thiller	Western	Épica	Total
Director 81	10	15	0	39	12	30	16	10	10	0	142
Director 40	0	0	33	36	6	5	0	4	6	11	101
Director 85	13	27	0	34	17	14	14	1	9	3	132
Director 100	1	5	22	30	10	15	0	11	3	25	122
Director 80	16	0	9	29	18	12	16	3	0	7	110

Hipótesis nula: No existe relación entre las variables premios y directores, para cada género.

Hipótesis alternativa: Existe relación entre las variables premios y directores, para cada género.

Coeficiente de Spearman entre las variables premios y directores, para cada género: -0.05422012768786661. No habría correlación entre las variables, o sería prácticamente nula.

P-valor entre las variables premios y directores, para cada género: 0.11355739662209467. En este caso no es menor al 0.05, lo cual sugiere que no hay suficiente evidencia para rechazar la hipótesis nula.

Interpretación: Estadísticamente, no se puede rechazar la hipótesis nula. Gráficamente hay directores que han ganado más premios en el género fantasía/animación que en otros, siendo el que más premios ha ganado en este género el director 81 con un total de 39 premios.

CONCLUSIONES SOBRE LAS HIPÓTESIS

En principio, y según lo analizado hasta el momento, existiría una falsa creencia de que las películas que reciben mejores críticas y/o son más populares son las que más cantidad de premios reciben o las que mayores ingresos generan.

Se visualiza cómo cambia el panorama al tener en cuenta solo la cantidad total de premios recibidos para cada género, dando por válida la hipótesis de que el género más premiado es el género dramático. Pero al analizar la cantidad de premios recibidos en relación a la cantidad de películas producidas esta hipótesis queda descartada.

Es real que el género fantástico/animado no es de los más premiados, pero esto puede deberse a otros factores, no al hecho de que no se trate de películas buenas o populares, por ejemplo este género no recibiría nunca premios en categorías como “Mejor actor/actriz”, “Mejor protagónico” u otras que hagan referencia a la actuación de una persona en particular, ya que este género no está interpretado por personas si no por personajes justamente fantásticos, y si bien pueden recibir premiación en categorías como “Mejor producción animada” también existen categorías específicas para los otros géneros.

Lo que si se observa es que existen directores que reciben mejores premiaciones según el género de la película que se dirige, por lo que podría recomendarse hacer alguna reunión con estos directores y los que no reciben o reciben pocos premios para hacer algún intercambio de ideas, o bien, directamente enfocar la dirección del género fantástico a los directores que más se destacan en este género.

FEATURE ENGINEERING.-

ENCODEO DE LAS VARIABLES

Por medio de un encodeo manual, mediante la definición de diccionarios, se crean las columnas "Premios_binario" y "genero_binario", con las características definidas en el apartado de "Diccionario de variables". De esta manera se transforman variables categóricas en numéricas para poder utilizarla en los modelos de Machine Learning.

También se crea la columna "genero_label", a través de label encode.

FEATURE BINNING

Mediante de la definición de bins para categorizar la variable "Premios", se crea la columna "Categoria" con las características definidas en el apartado de "Diccionario de variables". De esta manera se transforma la variable numérica en variable categórica para poder utilizarla en los modelos de Machine Learning.

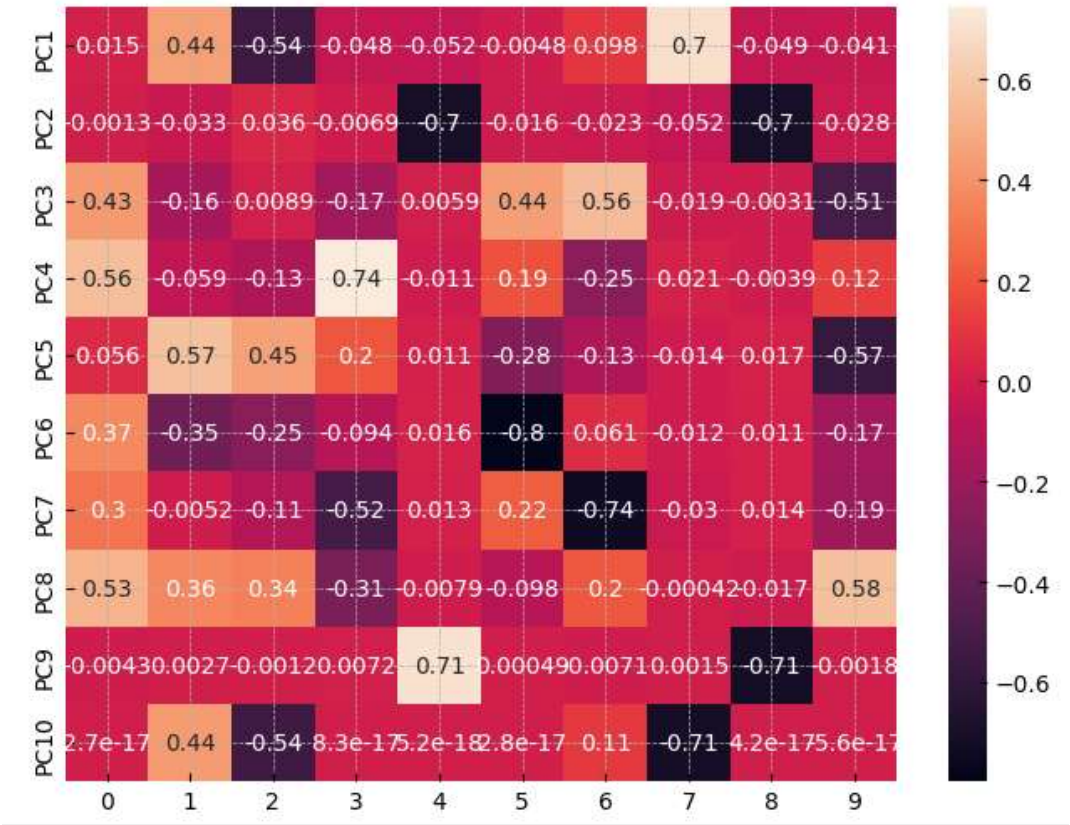
REDUCCIÓN DE DIMENSIONALIDAD.-

Es una técnica de aprendizaje no supervisado que consiste en la transformación de datos desde un espacio de alta dimensión a uno de baja dimensión para que la representación de baja dimensión retenga algunas propiedades significativas de los datos originales.

PCA (Principal Component Analysis)

Es una técnica utilizada para la identificación de un número más pequeño de variables no correlacionadas conocidas como componentes principales de un conjunto más grande de datos. Permite enfatizar la variación y capturar patrones fuertes en un conjunto de datos. Se utiliza principalmente para analizar variables continuas.

Luego del escalado de las variables numéricas, se aplica PCA y se obtiene la siguiente lista de componentes:



En función de las varianzas las listas de componentes aportan las siguientes varianzas:

```
['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9', 'PC10']
[0.2  0.18 0.11 0.1  0.1  0.1  0.09 0.09 0.02 0.  ]
```

Como se identifica en el análisis de las varianzas, la selección de los componentes que más varianza acumulan no permite reducir significativamente la dimensionalidad, esto podría deberse a que la mayoría de las variables numéricas analizadas, en realidad tienen un comportamiento de tipo categórico.

MCA (Análisis de Correspondencia Múltiple)

MCA se utiliza cuando se trabaja con variables categóricas. Su objetivo es analizar la relación entre las categorías de varias variables categóricas.

Se analizan las variables propiamente categóricas y con el método .get_dummies se encodean las variables. Al analizar las varianzas acumuladas de los 6 componentes se observa que tampoco aportan varianzas acumuladas significativas, por lo que tampoco puede aplicarse en este caso la reducción de dimensionalidad.

component	eigenvalue	% of variance	% of variance (cumulative)
0	0.007	0.65%	0.65%
1	0.006	0.58%	1.23%
2	0.006	0.58%	1.81%
3	0.006	0.57%	2.38%
4	0.006	0.56%	2.94%
5	0.006	0.55%	3.49%

MODELOS DE MACHINE LEARNING.-

ALGORITMOS DE CLASIFICACIÓN

La elección del modelo de ML y del tipo de algoritmo a utilizar depende de varios factores, entre ellos la naturaleza del problema y el tipo de resultado que se desea obtener.

En este análisis en particular, lo que se desea analizar es la posibilidad de que una película reciba o no un premio, es decir la respuesta objetivo al problema planteado es una respuesta categórica binaria, SI o NO. Por este motivo, es que se aplicará el modelo de CLASIFICACION, definiendo como variables las siguientes:

- VARIABLE TARGET .> PREMIOS
- VARIABLES INDEPENDIENTES -> GÉNERO, POPULARIDAD, CRTÍCAS, CALIFICACIÓN Y ESPECTADORES.

Dentro del modelo de clasificación, se analizarán y compararan los resultados obtenidos al aplicar los algoritmos de **regresión logística, KNN, árboles de decisión, Random Forest y SVM**. Se elabora la matriz de confusión y se analizan las métricas para validar el modelo.

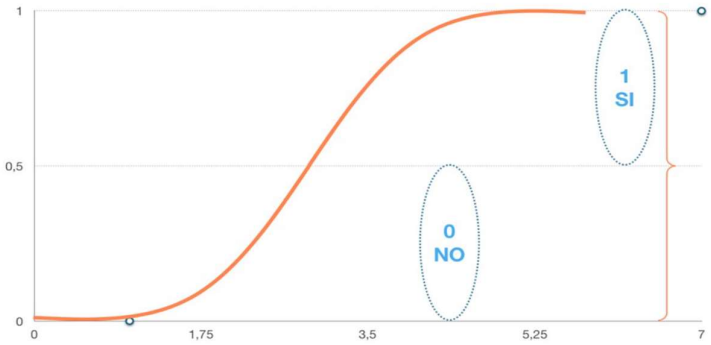
La **matriz de confusión** es una herramienta que permite visualizar el desempeño del algoritmo. Nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo.

En primera instancia se desarrollan los algoritmos con **validación simple**, luego se aplica la validación cruzada y mejora de hiperparámetros, en este sentido se desarrolla **Grid Search CV y Randomized Search** CV. Se fija como métrica para obtener la mejora el F1-score, ya que en el problema que analizamos, se observa que existe un desbalanceo de clases por lo que no sería bueno tomar como métrica el Acurracy, también se aplica el método **Statified K-Fold**, para lograr mantener equilibradas las clases al momento de hacer las divisiones del conjunto de datos para entrenamiento.

En segunda instancia, con los mejores parámetros detectados se aplica **SMOTE** para lograr el balanceo de clases.

REGRESION LOGISTICA

Es un modelo estadístico que se utiliza para determinar si una variable independiente tiene un efecto sobre una variable dependiente binaria. A partir del resultado obtenido al aplicar la función Sigmoide, se puede clasificar el resultado como SI (mayor al 0.5) o NO (menor al 0.5)

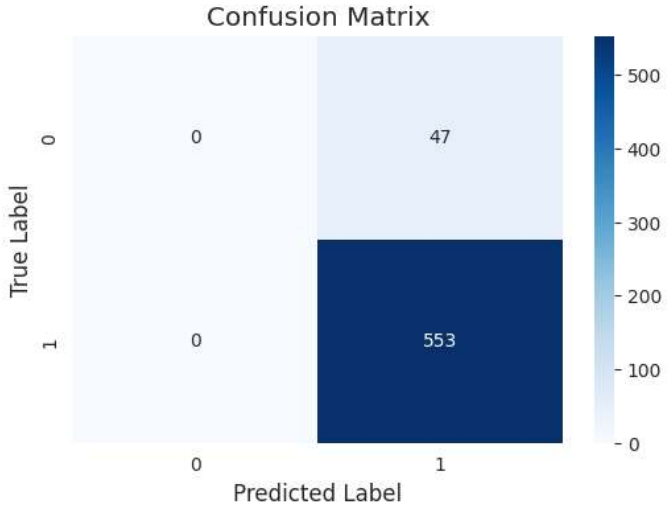


APLICACIÓN

El algoritmo requiere que las variables sean numéricas, por este motivo se utilizan para el análisis las variables previamente encodeadas.

A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje de 0.92 con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo.

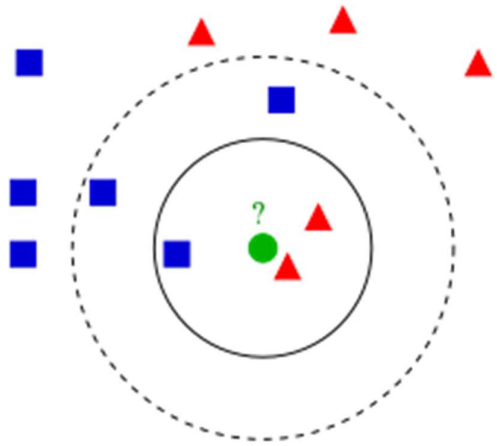
MÉTRICAS



	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

KNN

Este algoritmo utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Dada una nueva instancia, de la cual no sabemos cuál es su clase, vamos a recurrir a sus vecinos cercanos para clasificarla.

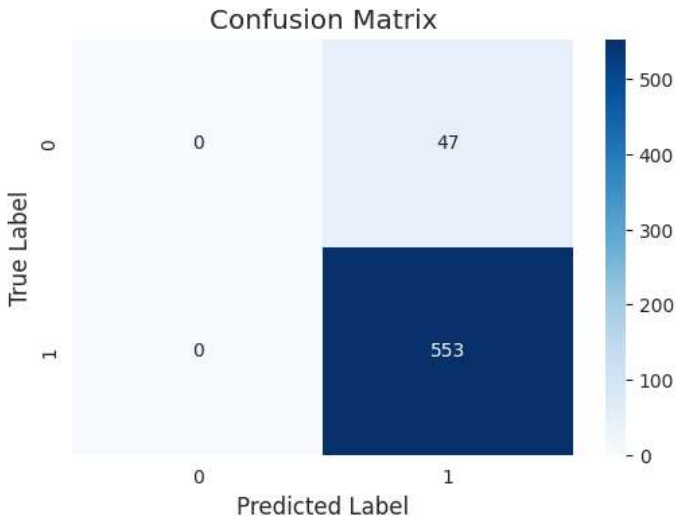


APLICACIÓN

El algoritmo requiere que las variables sean numéricas, por este motivo se utilizan para el análisis las variables previamente encodeadas.

A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje máximo de 0.958 de F1-Score con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo, tomando como mejor numero K=9 en Grid Search CV y K=11 en Randomized Search CV

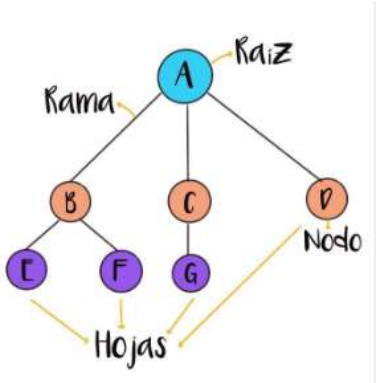
MÉTRICAS



	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

ÁRBOLES DE DECISIÓN

Este modelo de machine learning representa una estructura jerárquica en forma de árbol, divide los datos en diferentes ramas, basándose en características particulares, hasta llegar a las hojas del árbol, donde se asignan las etiquetas. En el proceso de entrenamiento, el árbol de decisión aprende a tomar decisiones al seleccionar la mejor característica para dividir los datos en cada nodo interno del árbol. Se utilizan métricas como la ganancia de información o la reducción de la impureza Gini para evaluar la calidad de las divisiones. A lo largo del entrenamiento, el árbol se ajusta a los datos de entrenamiento, y luego puede realizar predicciones sobre nuevos datos durante la fase de prueba.



APLICACIÓN

Para mantener la uniformidad con los algoritmos anteriores, se utilizan para el análisis las variables previamente encodeadas.

En este algoritmop la validación simple arroja resultados diferentes a los demás:

	precision	recall	f1-score	support
0	0.06	0.09	0.07	47
1	0.92	0.89	0.91	553
accuracy			0.83	600
macro avg	0.49	0.49	0.49	600
weighted avg	0.85	0.83	0.84	600

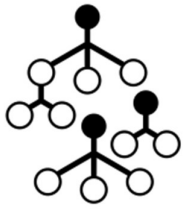
A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje máximo de 0.958 de F1-Score con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo, tomando como mejores parámetros en Grid Search CV max_depth: 6, max_leaf_nodes: 5, min_samples_leaf: 20, random_state=42 y en Randomized Search CV max_depth: 8, max_leaf_nodes: 5, min_samples_leaf: 20, random_state=42

MÉTRICAS

	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

RANDOM FOREST

Se trata de un tipo de ensamble en ML en el cual se combinan distintos tipos de árboles de decisión.

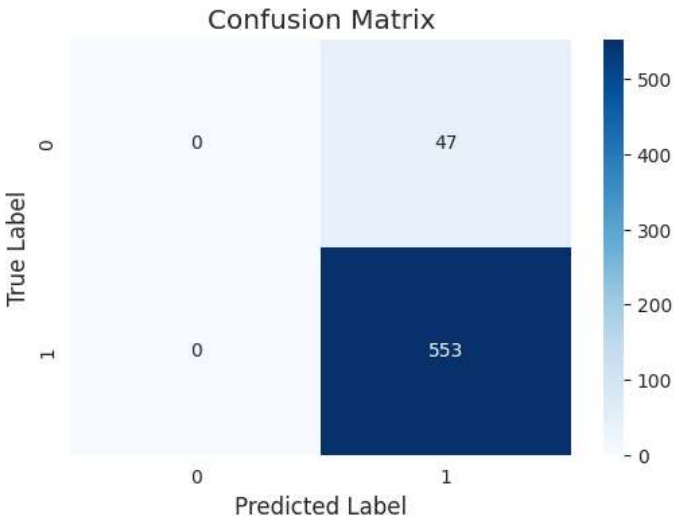


APLICACIÓN

Para mantener la uniformidad con los algoritmos anteriores, se utilizan para el análisis las variables previamente encodeadas.

A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje máximo de 0.958 de F1-Score con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo, tomando como parámetros max_depth: 6, max_leaf_nodes: 5, min_samples_leaf: 20, n_estimators: 150 en Grid Search CV. En Randomized Search CV max_depth: 6, max_leaf_nodes: 10, min_samples_leaf: 30, n_estimators: 150, random_state=42

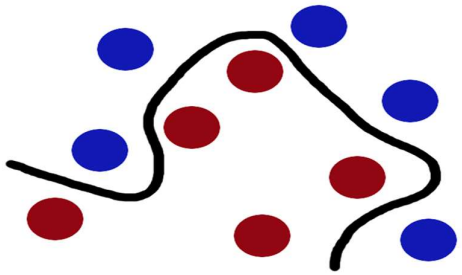
MÉTRICAS



	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

SVM – SUPPORT VECTOR MACHINES

Es un algoritmo que se fundamenta en la construcción de hiperplanos de segmentación. La idea principal detrás de las SVM es encontrar el hiperplano óptimo que separe de manera más clara las diferentes clases en un conjunto de datos. Busca un hiperplano que maximice la separación entre las clases, de manera que los puntos de una clase estén en un lado del hiperplano y los de la otra clase estén en el otro lado. Los puntos que están más cerca del hiperplano y que son cruciales para definir su posición se llaman vectores de soporte. El objetivo es maximizar la margen, que es la distancia entre el hiperplano y los vectores de soporte.

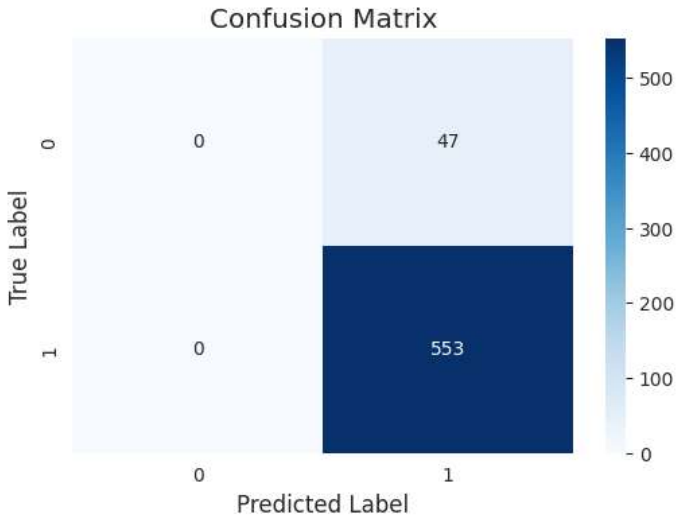


APLICACIÓN

Para mantener la uniformidad con los algoritmos anteriores, se utilizan para el análisis las variables previamente encodeadas.

A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje máximo de 0.958 de F1-Score con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo, tomando como parámetros modelo__C: 1, modelo__kernel: linear en Grid Search CV, mismos parámetros arroja Randomized Search.

MÉTRICAS



	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

ANÁLISIS MÉTRICAS DE CLASIFICACIÓN

Analizando las métricas dentro de la opción RECIBE PREMIO (1) se identifica que el modelo sería bueno, tiene buenos valores para precisión, recall y F1-Score. El F1-SCORE tanto en entrenamiento como en testeo no tiene diferencia marcada, lo que significa que **no hay sobreajuste ni subajust**e en lo que respecta a la opción de RECIBIR PREMIO(1). En caso que existiesen habría que ajustar los parámetros del modelo.

Por el lado de la opción NO RECIBE PREMIO (0) **hay que realizar un ajuste en las clases** ya que como se visualiza en la matriz de confusión, la predicción de no recibir premio es 0 y el modelo no puede realizar predicciones válidas para estos valores.

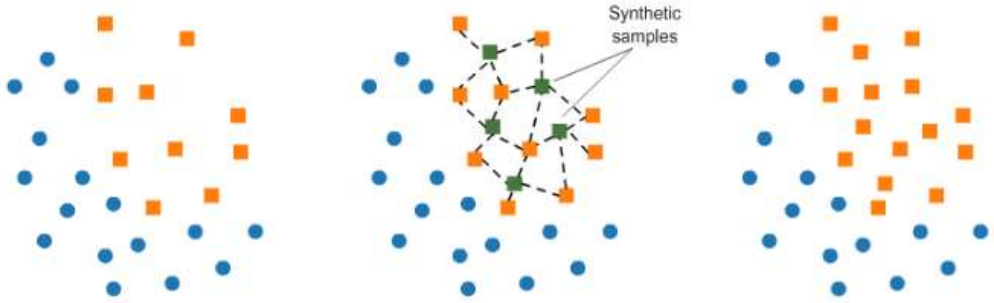
BALANCEO DE CLASES

El desbalanceo de clases significa que el modelo puede volverse sesgado hacia la clase dominante (recibir premio es el 92.05%). Por ello es que el modelo puede tener dificultades para generalizar y hacer predicciones precisas para las clases minoritarias (no recibir premio es el 7.95%).

Para ajustar esto existen técnicas que pueden basarse en aumentar la representación de la clase minoritaria (SMOTE y oversampling), reducir la representación de la clase mayoritaria (undersampling) o una combinación de ambas. La elección y combinación de la técnica dependerá del conjunto de datos y el problema que se está abordando.

SMOTE

Esta técnica sintetiza elementos para la clase minoritaria. El algoritmo funciona seleccionando ejemplos que están cerca en el espacio de características, trazando una línea entre los ejemplos en el espacio de características y dibujando una nueva muestra en un punto a lo largo de esa línea.



Aplicando SMOTE a cada uno de los modelos, se observan las siguientes métricas en cuanto a su rendimiento:

Regresión Logística

	precision	recall	f1-score	support
0	0.08	0.38	0.13	47
1	0.92	0.61	0.73	553
accuracy			0.59	600
macro avg	0.50	0.50	0.43	600
weighted avg	0.85	0.59	0.69	600

KNN

	precision	recall	f1-score	support
0	0.09	0.45	0.14	47
1	0.93	0.59	0.72	553
accuracy			0.58	600
macro avg	0.51	0.52	0.43	600
weighted avg	0.86	0.58	0.68	600

Árbol de decisión

	precision	recall	f1-score	support
0	0.10	0.62	0.17	47
1	0.94	0.51	0.66	553
accuracy			0.52	600
macro avg	0.52	0.56	0.41	600
weighted avg	0.87	0.52	0.62	600

Random Forest

	precision	recall	f1-score	support
0	0.09	0.45	0.15	47
1	0.93	0.62	0.74	553
accuracy			0.61	600
macro avg	0.51	0.53	0.45	600
weighted avg	0.86	0.61	0.70	600

SVM

	precision	recall	f1-score	support
0	0.09	0.60	0.15	47
1	0.93	0.46	0.61	553
accuracy			0.47	600
macro avg	0.51	0.53	0.38	600
weighted avg	0.86	0.47	0.58	600

MÉTODOS DE ENSAMBLE

También llamados "métodos combinados", intentan ayudar a mejorar el rendimiento de los modelos de Machine Learning al mejorar su precisión.

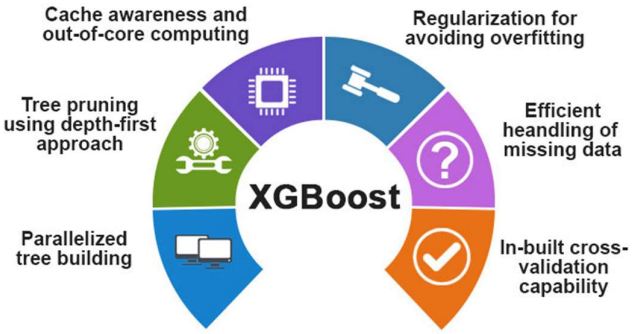
Este es un proceso mediante el cual se construyen estratégicamente varios modelos de Machine Learning para resolver un problema particular.

Dentro de estos métodos existen diversas técnicas, una es el **bagging** donde los algoritmos simples son usados en paralelo con el principal objetivo de aprovecharse de la independencia que hay entre los algoritmos simples, ya que el error se puede reducir bastante al promediar las salidas de los modelos simples. Uno de los algoritmos que utiliza esta técnica es Random Forest, ya analizando anteriormente. La otra técnica es el boosting, que se desarrolla más en profundidad a continuación

Boosting es un método de aprendizaje conjunto que combina un conjunto de modelos simples para minimizar los errores de entrenamiento. Se selecciona una muestra aleatoria de datos y se ajusta con un modelo y luego se entrena secuencialmente. Con cada iteración, las reglas débiles de cada clasificador individual se combinan para formar una regla de predicción fuerte.

XG BOOST

Extreme Gradient Boosting, es uno de los algoritmos de machine learning de tipo supervisado más usados en la actualidad. Este algoritmo se caracteriza por obtener buenos resultados de predicción con relativamente poco esfuerzo, en muchos casos equiparables o mejores que los devueltos por modelos más complejos computacionalmente, en particular para problemas con datos heterogéneos. Cada modelo es comparado con el anterior. Si un nuevo modelo tiene mejores resultados, entonces se toma este como base para realizar nuevas modificaciones. Si, por el contrario, tiene peores resultados, se regresa al mejor modelo anterior y se modifica ese de una manera diferente.



APLICACIÓN

Para mantener la uniformidad con los algoritmos anteriores, se utilizan para el análisis las variables previamente encodeadas.

A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje máximo de 0.958 de F1-Score con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo, tomando como parámetros alpha: 0.30000000000000004, gamma: 0.5, max_depth: 10, min_samples_leaf: 10, min_samples_split: 20, n_estimators: 300. En este caso se aplica **Halving Grid Search CV** para la búsqueda de hiperparámteros, ya que van mejor con este tipo de algoritmos.

MÉTRICAS

	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

SMOTE

	precision	recall	f1-score	support
0	0.08	0.19	0.11	47
1	0.92	0.81	0.86	553
accuracy			0.76	600
macro avg	0.50	0.50	0.49	600
weighted avg	0.86	0.76	0.80	600

LIGHT GBM

Utiliza la técnica Gradient Boosting. Con este método los árboles se construyen de manera secuencial y cada uno que se agrega aporta su granito de arena para refinar la predicción anterior. Es decir, se comienza con un valor constante y cada árbol nuevo se entrena para predecir el error en la suma de todas las predicciones de los árboles anteriores. Una vez terminado el proceso, las predicciones se calculan sumando los resultados de todos los árboles que se construyeron. El efecto que tiene esto es que cada vez que se agrega un árbol nuevo se le presta atención a las muestras en las que el modelo está funcionando peor y se trabaja para mejorar ese aspecto.

APLICACIÓN

Para mantener la uniformidad con los algoritmos anteriores, se utilizan para el análisis las variables previamente encodeadas.

A través de la técnica de validación cruzada se identifican los ajustes y selección adecuada de hiperparámetros para lograr un buen rendimiento del modelo. La validación arroja un puntaje máximo de 0.958 de F1-Score con los hiperparámetros ingresados, lo que significa un buen rendimiento del modelo, tomando como parámetros alpha: 0.1, gamma: 0.7, max_depth: 8, min_samples_leaf: 10, min_samples_split: 20, , random_state=42. En este caso se aplica **Halving Grid Search CV** para la búsqueda de hiperparámteros, ya que van mejor con este tipo de algoritmos.

MÉTRICAS

	precision	recall	f1-score	support
0	0.00	0.00	0.00	47
1	0.92	1.00	0.96	553
accuracy			0.92	600
macro avg	0.46	0.50	0.48	600
weighted avg	0.85	0.92	0.88	600

SMOTE

	precision	recall	f1-score	support
0	0.06	0.15	0.09	47
1	0.92	0.81	0.86	553
accuracy			0.76	600
macro avg	0.49	0.48	0.48	600
weighted avg	0.85	0.76	0.80	600

CONCLUSIONES

En resumen las métricas del análisis con las diferentes técnicas y algoritmos quedan:

ALGORITMO / MÉTRICA		REGRESION LOGISTICA		KNN				ÁBOL DECISIÓN			
		VALIDACION SIMPLE	SMOTE	VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH	VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH
					SIN SMOTE	CON SMOTE			SIN SMOTE	CON SMOTE	
PRECISION	0 - No recibe premio	0,00	0,08	0,00	0,00	0,09	0,00	0,05	0,00	0,10	0,00
	1 - Recibe premio	0,92	0,92	0,92	0,92	0,93	0,92	0,92	0,92	0,94	0,92
RECALL	0 - No recibe premio	0,00	0,38	0,00	0,00	0,45	0,00	0,06	0,00	0,62	0,00
	1 - Recibe premio	1,00	0,61	1,00	1,00	0,59	1,00	0,90	1,00	0,51	1,00
F1-SCORE	0 - No recibe premio	0,00	0,13	0,00	0,00	0,14	0,00	0,06	0,00	0,17	0,00
	1 - Recibe premio	0,96	0,73	0,96	0,96	0,72	0,96	0,91	0,96	0,66	0,96
ACCURACY		0,92	0,59	0,92	0,92	0,58	0,92	0,83	0,92	0,52	0,92

ALGORITMO / MÉTRICA		RANDOM FOREST				SVM			
		VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH	VALIDACION SIMPLE	GRID SEARCH		RANDOMIZED SEARCH
			SIN SMOTE	CON SMOTE			SIN SMOTE	CON SMOTE	
PRECISION	0 - No recibe premio	0,00	0,00	**0,09**	0,00	0,00	0,00	0,09	0,00
	1 - Recibe premio	0,92	0,92	**0,93**	0,92	0,92	0,92	0,93	0,92
RECALL	0 - No recibe premio	0,00	0,00	**0,45**	0,00	0,00	0,00	0,60	0,00
	1 - Recibe premio	1,00	1,00	**0,62**	1,00	1,00	1,00	0,46	1,00
F1-SCORE	0 - No recibe premio	0,00	0,00	**0,15**	0,00	0,00	0,00	0,15	0,00
	1 - Recibe premio	0,96	0,96	**0,74**	0,96	0,96	0,96	0,61	0,96
ACCURACY		0,92	0,92	**0,61**	0,92	0,92	0,92	0,47	0,92

ALGORITMO / MÉTRICA		XG BOOST		LIGHT GBM	
		HALVING GRID SEARCH		HALVING GRID SEARCH	
		SIN SMOTE	CON SMOTE	SIN SMOTE	CON SMOTE
PRECISION	0 - No recibe premio	0,00	0,08	0,00	0,06
	1 - Recibe premio	0,92	0,92	0,92	0,92
RECALL	0 - No recibe premio	0,00	0,19	0,00	0,15
	1 - Recibe premio	1,00	0,81	1,00	0,81
F1-SCORE	0 - No recibe premio	0,00	0,11	0,00	0,09
	1 - Recibe premio	0,96	0,86	0,96	0,86
ACCURACY		0,92	0,76	0,92	0,76

Es decir que el mejor modelo seria Random Forest luego de aplicar SMOTE y la mejora de hiperparámetros con validación mediante Grid Search CV. Por otro lado, con método de ensamble la mejor alternativa seria el algoritmo XG Boost.

ALTERNATIVA VARIABLES DEFINIDAS POR EL MODELO

RANDOM FOREST

Es posible que al definir las variables uno mismo, se dejen de lado variables que harían mejor al modelo, es por esto que aplicando Random Forest se analizan las métricas pero obteniendo del mismo modelo cuales serían las variables a incluir según su importancia, esta consulta arroja que los mejores predictores serian:

Importancia de los predictores en el modelo

	predictores	importancia
4	Espectadores	0.104021
9	Resultado (millones)	0.095566
7	Popularidad	0.089314
8	Recaudación en DVD (millones)	0.088267
2	Ingresos (millones)	0.086476
10	Criticas	0.086089
0	Año	0.084102
1	Calificación	0.081218
3	Presupuesto (millones)	0.079629
6	Críticas Negativas	0.072619
5	Críticas Positivas	0.072402
12	genero_label	0.053920
11	genero_binario	0.006378

Tomando las 5 primeras variables y los mejores hiperparámtros para el algoritmo junto con SMOTE, las métricas quedan de la siguiente manera:

	precision	recall	f1-score	support
0	0.12	0.26	0.16	47
1	0.93	0.83	0.88	553
accuracy			0.79	600
macro avg	0.52	0.54	0.52	600
weighted avg	0.87	0.79	0.82	600

Es decir, que se mejora el rendimiento del modelo dejando que el algoritmo defina cuales son los mejores predictores.

XG BOOST

El mismo análisis se realiza con el algoritmo de ensamble, el cual arroja que las variables a incluir según su importancia serian:

Importancia de los predictores en el modelo

	predictores	importancia
11	genero_label	0.11
3	Espectadores	0.10
7	Recaudación en DVD (millones)	0.10
9	Criticas	0.09
5	Críticas Negativas	0.09
4	Críticas Positivas	0.09
0	Calificación	0.08
8	Resultado (millones)	0.08
1	Ingresos (millones)	0.08
2	Presupuesto (millones)	0.08
6	Popularidad	0.07
10	genero_binario	0.03

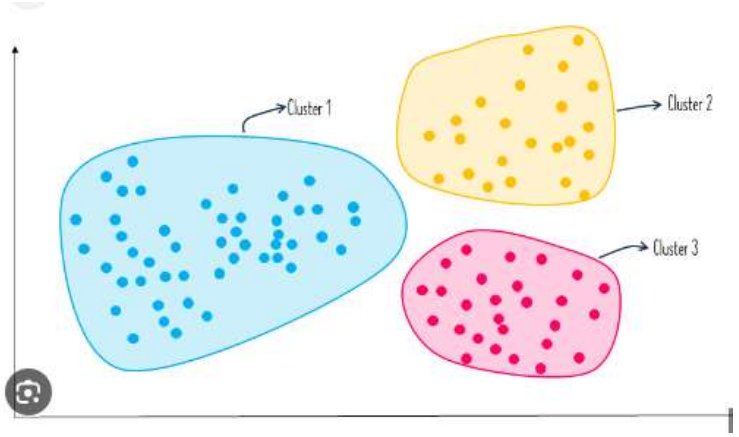
Tomando las 5 primeras variables y los mejores hiperparámtros para el algoritmo junto con SMOTE, las métricas quedan de la siguiente manera:

	precision	recall	f1-score	support
0	0.08	0.13	0.10	47
1	0.92	0.88	0.90	553
accuracy			0.82	600
macro avg	0.50	0.50	0.50	600
weighted avg	0.86	0.82	0.84	600

Es decir, que también se mejora el rendimiento del modelo dejando que el algoritmo defina cuales son los mejores predictores.

CLUSTERING.-

Los algoritmos de agrupamiento son métodos no supervisados, donde la entrada no está etiquetada y la resolución de problemas se basa en la experiencia del algoritmo. Estos algoritmos aprenden de los atributos disponibles en la matriz de diseño X con el fin de generar grupos de compartan características similares en los datos analizados. Se asignan los objetos a grupos homogéneos asegurando la mínima varianza intra-cluster y la máxima varianza inter-cluster.



K-MEANS Y MÉTRICAS

K-Means es uno de los algoritmos que pueden utilizarse para este tipo de aprendizaje, y es el que se utilizará en el análisis. Este algoritmo lo que hace es definir los centroides iniciales y el método comienza las iteraciones.

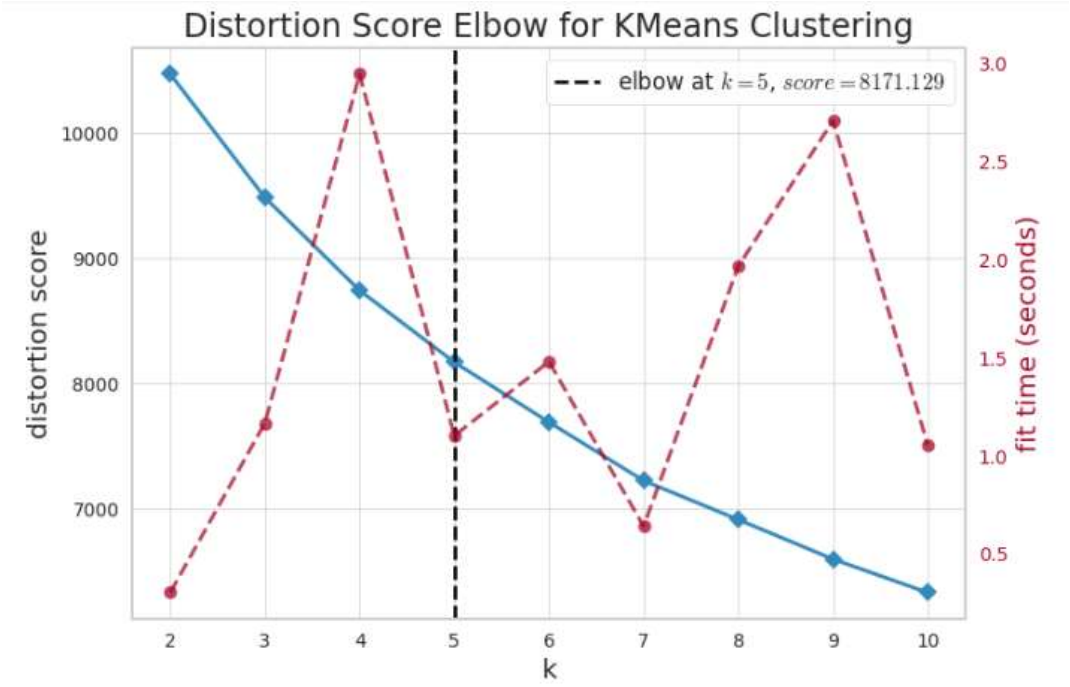
Una vez elegido el algoritmo, hay que definir la cantidad óptima de cluster, no existe una métrica que lo defina por si solo, es criterio del Data Science, en base a la observación, conocimiento y experiencia, quien termina decidiendo cual es la cantidad óptima de cluster. Si existen métricas que ayudan a la definición del número de cluster óptimo. En el análisis se aplicarán el método del codo, el Score de Silhouette, Indice de Kalinski y el Indice de Davies-Bouldin.

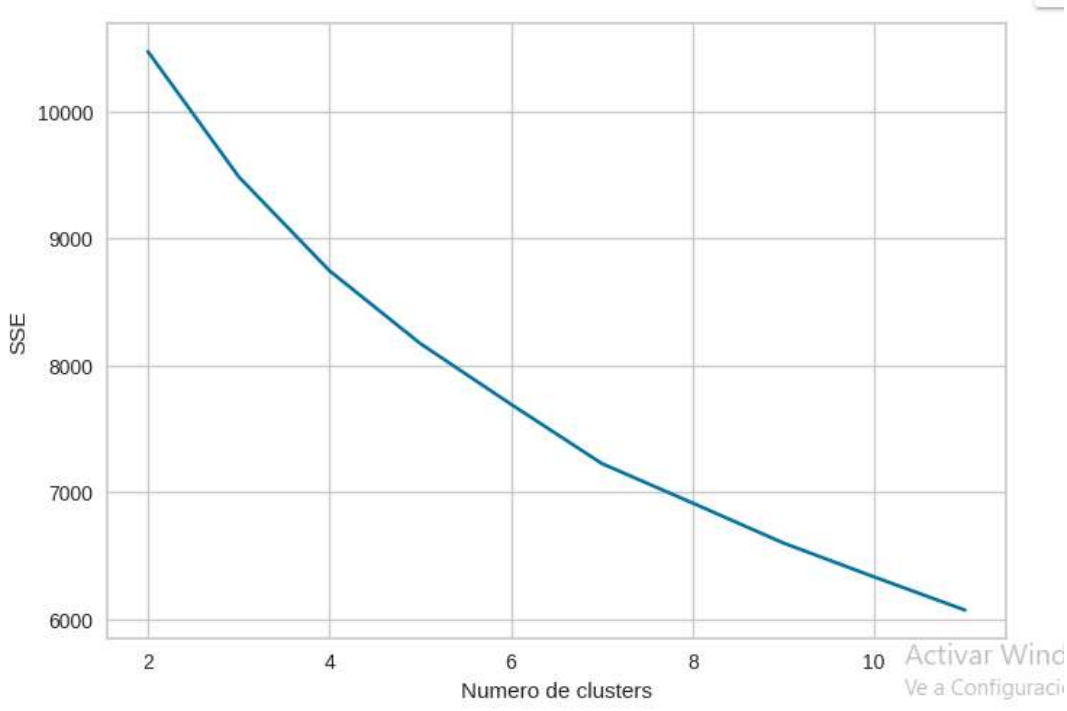
APLICACIÓN

Se definen las variables que deben ser tenidas en cuenta como características para la agrupación, esta son: premios, género_label, calificación, espectadores, críticas y popularidad.

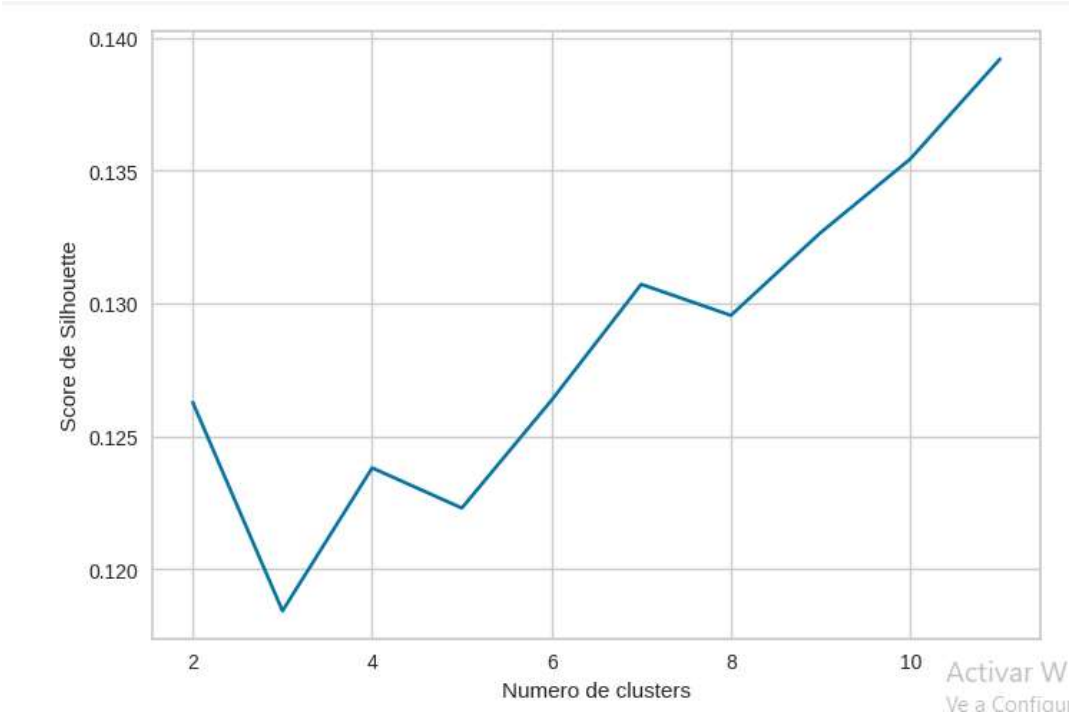
Se aplican los procedimientos para encontrar el k óptimo

Método del codo

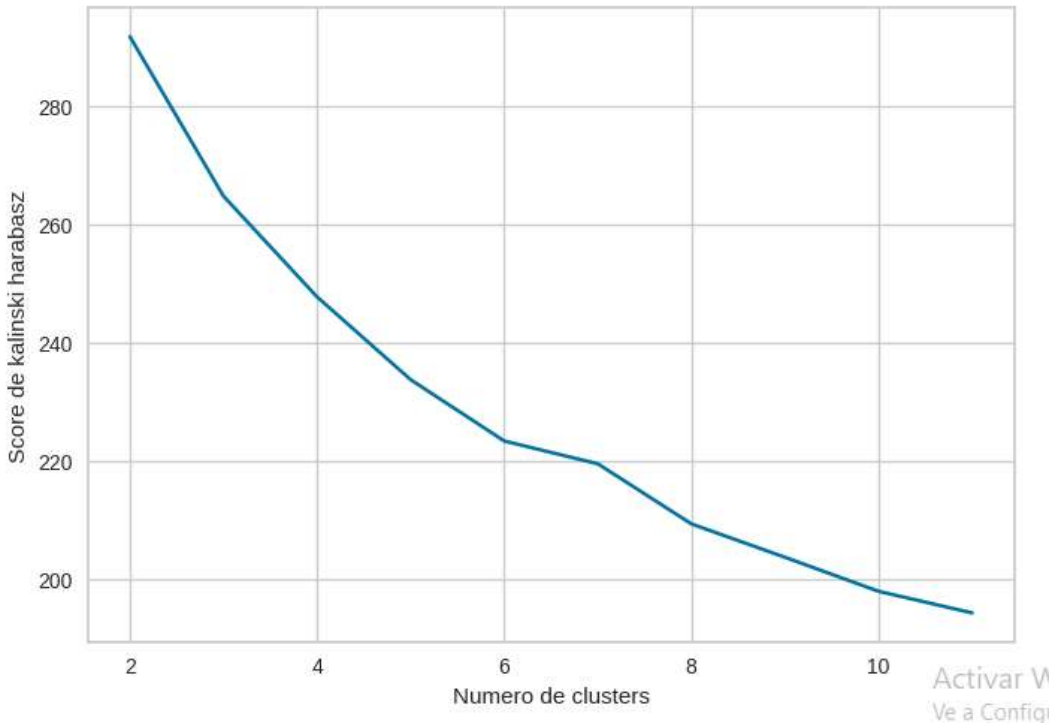




Score de Silhouette



Índice de Kalinski



Índice de Davies-Bouldin

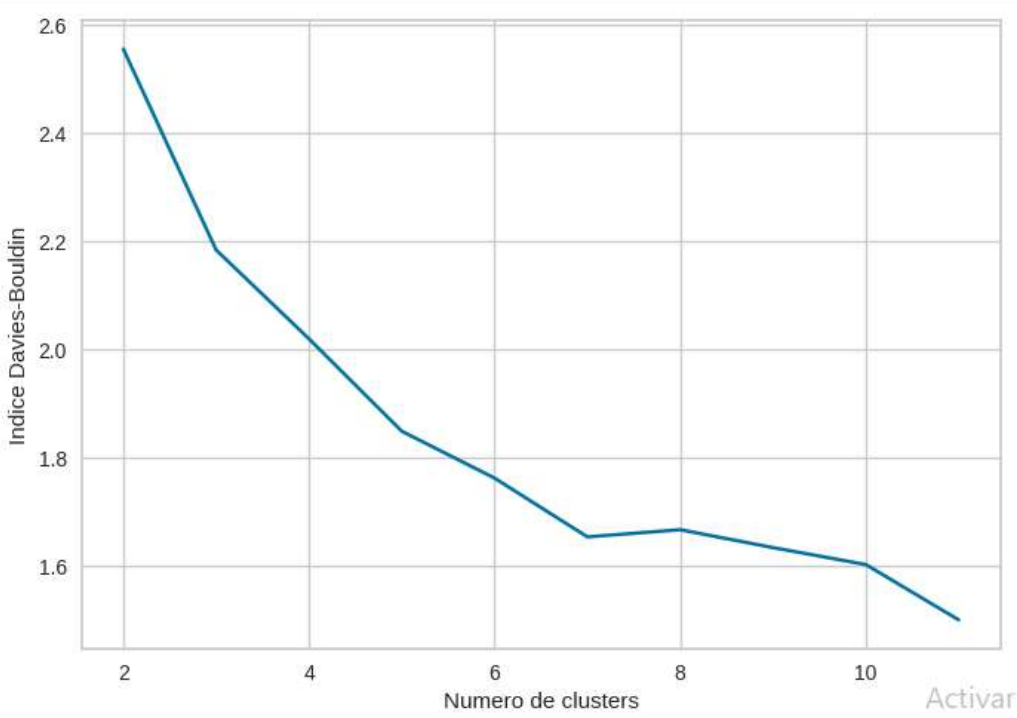
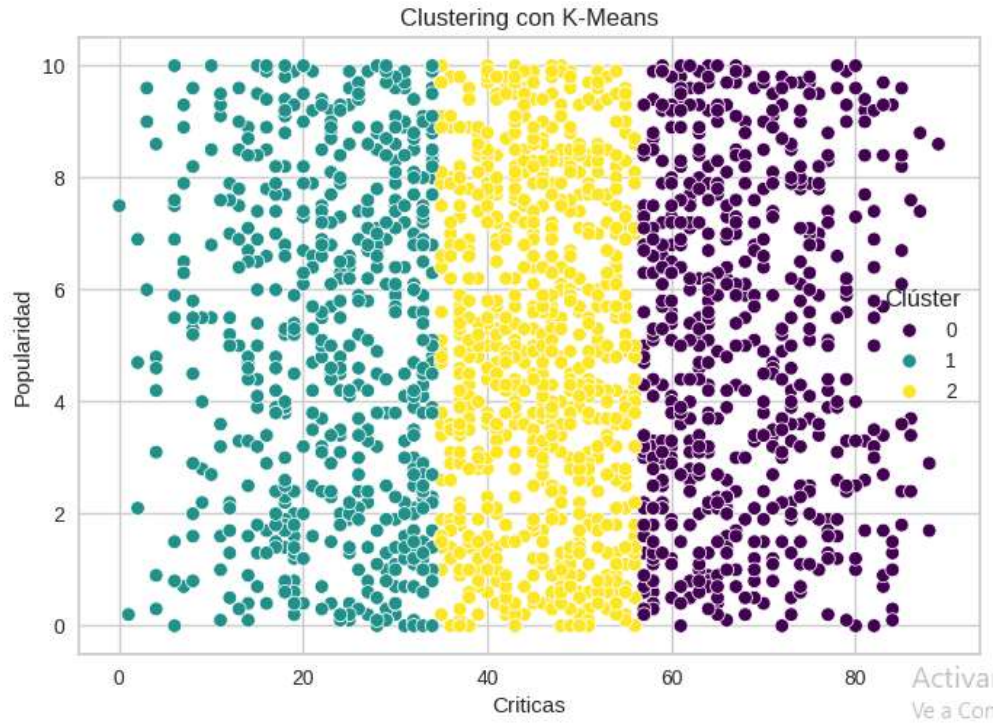


Gráfico con k=3

Tomando solo 2 características como Popularidad y Críticas. Gráficamente los cluster quedan:



CONCLUSIONES.-

El objetivo del presente análisis es poder identificar la posibilidad de que una película del género fantasía/animación reciba un premio, utilizando modelos de predicción de machine learning.

Luego de un exhaustivo análisis de los datos, las variables y las hipótesis planteadas, se identifica la necesidad de realizar algunos cambios en los datos para favorecer el proceso de aplicación de los modelos.

Una vez que se llega al desarrollo de la parte predictiva, se plante el problema como problema de clasificación y se desarrollan diferentes algoritmos a los cuales se les aplican mejoras en hiperparámetros para lograr obtener el mejor modelo. Como se identifica un importante desbalanceo en las clases es necesario aplicar otras técnicas para lograr un mejor análisis.

En una primera instancia de análisis la aplicación de los algoritmos se efectúa sobre variables independientes predefinidas, considerando a criterio del autor cuales serían las mejores, con este criterio se identifica que los algoritmos que mejores métricas arrojan son Random Forest y XG Boost, previa configuración de hiperparámetros y aplicación de las técnicas de balanceo de datos.

En una segunda instancia y se aplican estos mismos algoritmos pero sobre las variables independientes definidas por el mismo algoritmo, y se identifica que las métricas mejoran.

En conclusión el problema planteado como modelo de clasificación, aplicando Random Forest o XG boost dejando al algoritmo definir las variables a analizar funciona correctamente obteniendo buenas métricas y permitiría poder predecir la posibilidad de que una determinada película reciba o no su premio.-