Memoria de trabajo

Tratamiento Inteligente de Datos

Proyecto final de la asignatura Tratamiento Inteligente de Datos de 3º Ingeniería Informática, Itinerario de Computación, Escuela Superior de Ingeniería Técnica.

Autores		
Nombre	**Apellidos**	**Correo**
Erik Andreas	Barreto de Vera	alu0100774054@ull.edu.es
Jorge	Alonso Hernández	alu0100767803@ull.edu.es

Índice

1. Introducción

Este documento contiene el trabajo del proyecto final de la asignatura Tratamiento Inteligente de Datos de 3º Ingeniería Informática, Itinerario de Computación, Escuela Superior de Ingeniería y Tecnología. En él se explica cómo ha sido todo el proceso para analizar los datos, el entorno de trabajo, la fuentes de los datos, etc.

2. Objetivo

El objetivo del proyecto es intentar conseguir un modelo capaz de predecir cuando una película va a tener éxito o no lo va a tener. Se va a analizar toda la información relativa a cada película estrenada en el año 2015 en España.

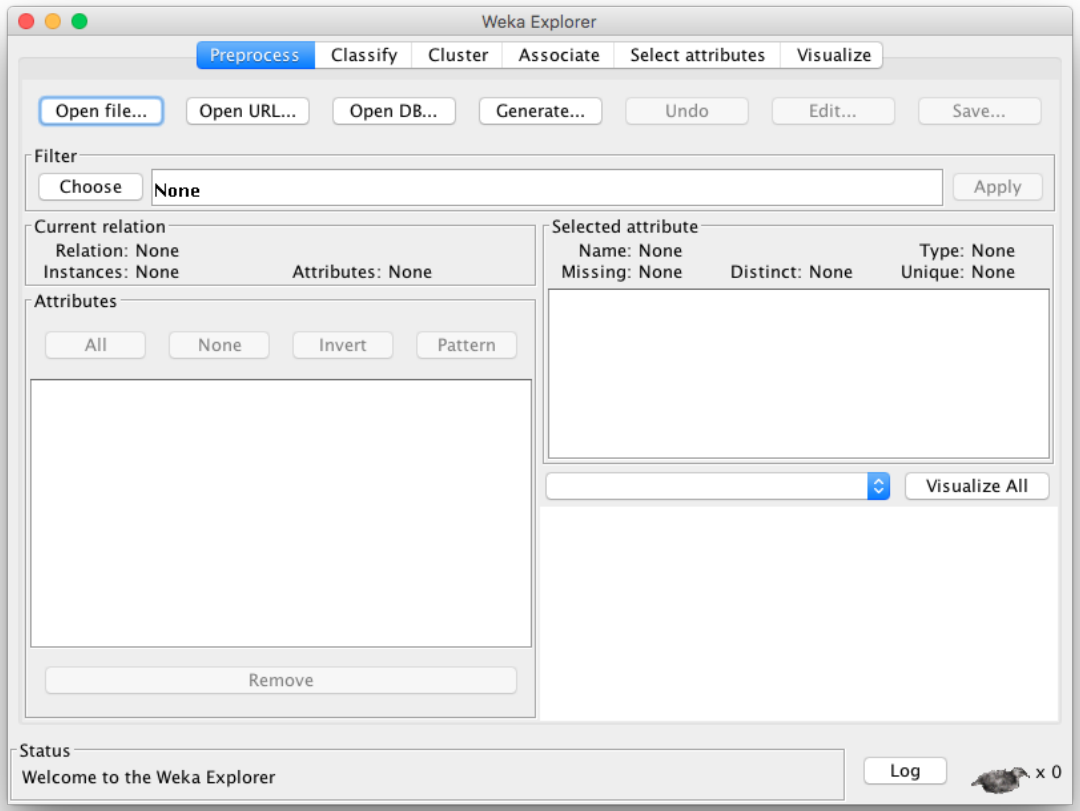
En este proceso se van a seguir una serie de directrices y pasos que veremos más adelante pero que vamos a resumir en lo siguiente; En primer lugar, se va a proceder a recolectar datos sobre las películas estrenadas en España en el año 2015, partiendo de toda la información que provee la base de datos se hace un proceso de definir reglas para decidir la influencia de qué datos son importantes y cuáles no, para empezar a estructurar la forma que va a tener nuestro problema. Después de haber descrito el problema, Con el entorno de trabajo Weka se trabajará toda la parte del análisis de los datos. Con este entorno se recogen todos los valores estadísticos que nos explican cómo funciona el problema en la realidad, Si el modelo funciona nos va a dar un modelo que es capaz de acertar cuanto de buena va a ser una película.

Por último, el modelo que se obtiene podría actuar como un recomendador para las compañías de la industria que les va a permitir definirse estrategias para decidir preguntas como ¿Cuándo estrenar una película?, ¿Qué género puede ser más rentable?, etc.

3. Entorno de trabajo

3.1. Weka

Weka es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL.



Razones a favor de Weka

- Está disponible libremente bajo la licencia pública general de GNU.

- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.

- Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado.

- Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

Weka soporta varias tareas estándar de minería de datos, especialmente, preprocesamiento de datos, clustering, clasificación, regresión, visualización, y selección.

Herramientas del programa

La interfaz Explorer (Explorador) dispone de varios paneles que dan acceso a los componentes principales del banco de trabajo:

- El panel "Preprocess" dispone de opciones para importar datos de una base de datos, de un fichero CSV, etc., y para pre-procesar estos datos utilizando los denominados algoritmos de filtrado. Estos filtros se pueden utilizar para transformar los datos (por ejemplo convirtiendo datos numéricos en valores discretos) y para eliminar registros o atributos según ciertos criterios previamente especificados.

- El panel "Classify" permite al usuario aplicar algoritmos de clasificación estadística y análisis de regresión a los conjuntos de datos resultantes. También permite estimar la exactitud del modelo predictivo resultante, mediante curvas ROC, etc. Finalmente, tiene utilidades para visualizar el propio modelo, en aquellos casos en que esto sea posible, como por ejemplo un árbol de decisión.

- El panel "Associate" proporciona acceso a las reglas de asociación aprendidas que intentan identificar todas las interrelaciones importantes entre los atributos de los datos.

- El panel "Cluster" da acceso a las técnicas de clustering o agrupamiento de Weka como por ejemplo el algoritmo K-means. Este es sólo una implementación del algoritmo expectación-maximización para aprender una mezcla de distribuciones normales.

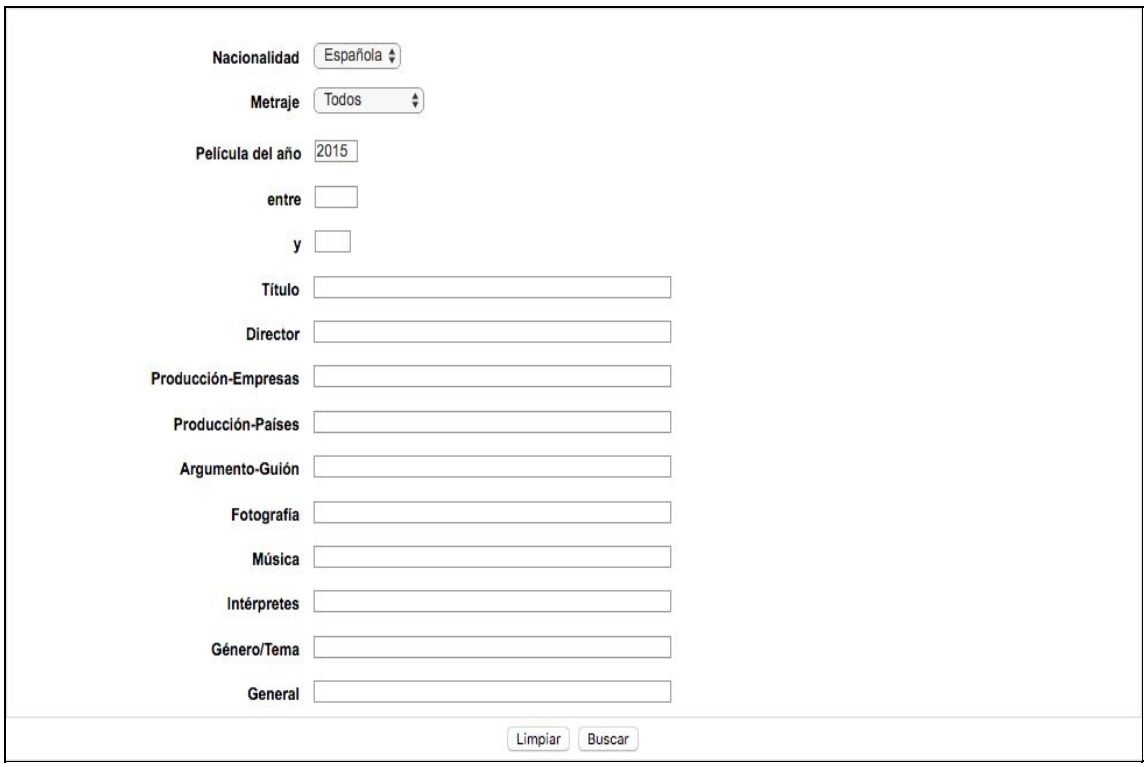
- El panel "Selected attributes" proporciona algoritmos para identificar los atributos más predictivos en un conjunto de datos.

- El panel "Visualize" muestra una matriz de puntos dispersos (scatterplot) donde cada punto individual puede seleccionarse y agrandarse para ser analizados en detalle usando varios operadores de selección.

4. Descripción de los datos a analizar

La muestra de datos que vamos a utilizar es de un total de 198 películas estrenadas en los cines de España durante el año 2015. La fuente de los datos ha sido la base de datos del Ministerio de educación, cultura y deporte, . El buscador tiene un aspecto similar a la imagen siguiente.

Página del ministerio (buscador)

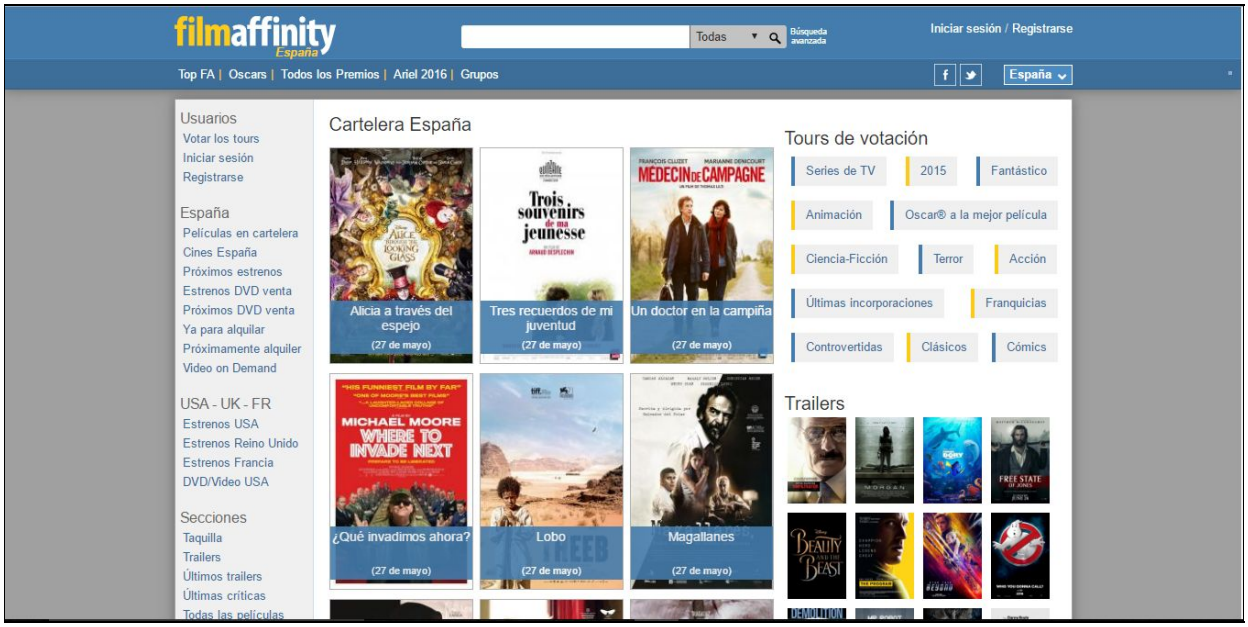


Para saber todas las películas que fueron estrenadas en el año 2015 iremos a la página de labutaca la cual tiene un listado con todas las películas estrenadas en el año 2015 ordenadas por la semana de estreno, la vista del listado de películas de la web es la siguiente:

Labutaca

También nos apoyaremos en dos webs de cines bastantes conocidas, para obtener otros datos que no encontramos en la página del ministerio. Estas webs son sensacine y filmaffinity:

filmaffinity



Sensacine

Para cada película se han establecido una serie de atributos que hemos definido en base a unos criterios que se explicarán en el siguiente apartado en detalle. Se han buscado las películas una a una en base al año 2015 y se han pasado los atributos que queremos a un fichero con extensión *".arff"* que es el formato estándar con el que trabaja el programa Weka. Veámos el fichero en este punto.

FicheroWeka.arff

@relation espectadores

@attribute genero { Accion, Aventura, Animacion, Comedia, Drama, Terrorl, Sci-fi, Thriller, Otros }

@attribute edad { Todos, >7, >12, >16, >18 }

@attribute mesEstreno { Enero, Febrero, Marzo, Abril, Mayo, Junio, Julio, Agosto, Septiembre, Octubre, Noviembre, Diciembre }

@attribute duracion numeric

@attribute productora { WaltDisneyCompany, TwentiethCenturyFox, UniversalPicturesInternational, WarnerBrosEntertainment, SonyPicturesReleasing, AurumProducciones, Paramount, HispanoFoxfilm, AContracorrienteFilms, CastelaoPictures, DeaPlaneta, Tripictures, Otra }

@attribute pais { ESP, EEUU }

@attribute acotrRenombre {Si, No}

@attribute basadaEn { Libro, Comic, HechosReales, Original }

@data

...

5. Reglas definidas

Aquí se definen los criterios que se han establecido para analizar las películas. Para cualquier otro atributo posible para una película que no se encuentre entre los siguientes es porque ha sido descartado por no ser relevante en el análisis.

- **Género:** Es un atributo de tipo nominal. En la muestra del proyecto se han agrupado películas para cada género. Entre los valores definidos tenemos aquellos cuyo número de películas es lo suficientemente grande como para que podamos incluirlo en un solo grupo, por el contrario, existen algunos género que no han aportado un gran número de películas a lo largo del año, para todos estos se ha definido un grupo a parte donde se engloban llamado *otros*.

- **Edad:** Es un atributo de tipo nominal. Hemos escogido la edad como un atributo importante por el hecho de que el tipo de público es diferente para cada rango de edades, lo que espera una persona sobre una película de un cierto género de una determinada edad es diferente al de otra persona con una edad mucho mayor, por ejemplo una película infantil y una de terror.

- **Mes de estreno:** Es un atributo de tipo nominal. Vamos a intentar ver cuánto de influyente es el estrenar una película en un determinado momento del año, si realmente una película determinada puede generar un número mucho mayor que otra por el hecho de estrenarse en época de vacaciones donde la gente tiene tiempo libre para poder ir al cine o por ejemplo el caso de meses de muchos gastos como es Enero debido al gasto de las festividades frente a otros.

- **Duración:** Es un atributo de tipo numérico. Se quiere comprobar si la gente decide ver una película por el tiempo que dure. Veamos el caso donde una persona no quiera ver una película muy larga por razones de tiempo invertido, o el caso que la gente prefiera una película larga porque considera que estará mucho más trabajada y será mejor(Son algunos ejemplos, más tarde se harán las conclusiones.

- **Productora:** Es un atributo de tipo nominal. Se pretende hacer hincapié en sí las grandes productoras acaparan todo el público por las facilidades que tienen para generar publicidad de sus películas, presupuesto para contratar o producir, etc. O si por el contrario la gente se fija más en la idea o el guión de una película.

- **País:** Es un atributo de tipo nominal. Se quiere ver si en España tienen más éxito las películas producidas en España o en Estados Unidos.

- **Basado en:** Es un atributo de tipo nominal. Se quiere comprobar si la gente prefiere una película que esté basada en algún libro, por ejemplo novelas de éxito que son pasadas al mundo del cine, o

por ejemplo basadas en hechos reales, es decir, cuanto de influyente es esta afirmación sobre el éxito.

- **Espectadores:** Se ha incluido la variable espectadores que cuenta el número de espectadores en que han visto la película en las salas de cine durante el mismo año. Se pretende comprobar si el porcentaje de éxito es determinado por el número de espectadores, por la cantidad de dinero recaudado (siguiente variable) o ambos.

Para la variable de espectadores hemos definido unos rangos de lo que se considera una buena, mala, regular, etc película. Los intervalos son:

- Muy Pocos [0 - 23206]
- Pocos [23206 - 124531]
- Medio [124531 - 490350]
- Bastantes [490350 - 1360220]
- Muchos [> 1360220]

- **Recaudación:** Esta variable se comporta del mismo modo que la anterior y cuenta la cantidad de dinero recaudado de la película en euros por las salas de cine españolas.

Para la variable de recaudación hemos definido unos rangos de lo que se considera una buena, mala, regular, etc., recaudación. Los intervalos son:

- Muy Poca [0 - 133281,25]
- Poca [133281,25 - 716432,37]
- Media [716432,37 - 2916532,35]
- Bastante [2916532,35 - 11431071,11]
- Mucha [> 11431071,11]

Las variables espectadores y recaudación se van a analizar independientemente una de la otra, con el fin de saber cuál es más determinante en el resultado.

- **Actores de éxito:** Por último hemos dejado esta variable con el fin de comentarla al final. Es una variable de tipo nominal que toma los valores de si la película tiene actores de renombre o no. Se

pretende comprobar si los espectadores se decantan por películas con actores de renombre entre sus integrantes o por el contrario no, esto podría suponer que el dinero invertido en grandes actores podría reducirse. Esta variable va a ser especial porque se van a realizar diferentes análisis para ver cuánto influye en el resultado.

6. Análisis de los datos

Variables escogidas para cada uno de los aná:

Análisis 1	Análisis 2	Análisis 3	Análisis 4
Género Calificación Mes de estreno Duración Productora País Basada en **Número de espectadores**	Género Calificación Mes de estreno Duración Productora País Basada en **Actores de éxito** **Número de espectadores**	Género Calificación Mes de estreno Duración Productora País Basada en **Recaudación (Euros)**	Género Calificación Mes de estreno Duración Productora País Basada en **Actores de éxito** **Recaudación (Euros)**

NOTA: Se resaltan en negrita las variables que cambian en cada uno de los análisis.

1. Análisis 1.

Árbol de decisión de weka:

```
productora = WaltDisneyCompany
|   edad = Todos
|   |   duracion <= 99: Medio (3.0/1.0)
|   |   duracion > 99: Muchos (2.0)
|   edad = >7
|   |   duracion <= 132: Bastantes (2.0)
|   |   duracion > 132: Muchos (2.0)
|   edad = >12: Muchos (0.0)
|   edad = >16: Muchos (0.0)
|   edad = >18: Muchos (0.0)
productora = TwentiethCenturyFox: Bastantes (5.0/2.0)
productora = UniversalPicturesInternational: Medio (18.0/11.0)
productora = WarnerBrosEntertainment: Medio (22.0/13.0)
productora = SonyPicturesReleasing: Medio (19.0/12.0)
productora = AurumProducciones: Medio (18.0/9.0)
productora = Paramount
|   mesEstreno = Enero: Pocos (2.0)
|   mesEstreno = Febrero: Bastantes (1.0)
```

```
|   mesEstreno = Marzo: Pocos (0.0)
|   mesEstreno = Abril: Pocos (0.0)
|   mesEstreno = Mayo: Pocos (0.0)
|   mesEstreno = Junio: Pocos (0.0)
|   mesEstreno = Julio: Bastantes (1.0)
|   mesEstreno = Agosto: Muchos (2.0)
|   mesEstreno = Septiembre: Pocos (0.0)
|   mesEstreno = Octubre: Medio (1.0)
|   mesEstreno = Noviembre: Pocos (0.0)
|   mesEstreno = Diciembre: Pocos (0.0)
productora = HispanoFoxfilm
|   edad = Todos: Bastantes (2.0)
|   edad = >7: Medio (3.0)
|   edad = >12
|   |   genero = Accion: MuyPocos (0.0)
|   |   genero = Aventura: MuyPocos (0.0)
|   |   genero = Animacion: MuyPocos (0.0)
|   |   genero = Comedia: MuyPocos (3.0/1.0)
|   |   genero = Drama: MuyPocos (0.0)
|   |   genero = Terror: Medio (1.0)
|   |   genero = Sci-fi: MuyPocos (0.0)
|   |   genero = Thriller: Pocos (2.0/1.0)
|   |   genero = Otros: MuyPocos (0.0)
|   edad = >16: Pocos (7.0/4.0)
|   edad = >18: Medio (0.0)
productora = AContracorrienteFilms: MuyPocos (15.0/6.0)
productora = CastelaoPictures: Pocos (7.0/3.0)
productora = DeaPlaneta: Pocos (6.0/2.0)
productora = Tripictures
|   duracion <= 111: Pocos (3.0)
|   duracion > 111: Medio (2.0/1.0)
productora = Otra: MuyPocos (49.0/22.0)
```

Matriz de confusión del Árbol:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 35  5 10  0  0 |  a = MuyPocos
 24 10 11  4  0 |  b = Pocos
 12  8 25  3  2 |  c = Medio
  0  8 16  2  3 |  d = Bastantes
  2  2 11  5  0 |  e = Muchos
```

Clasificador Naive Bayes de weka:

Attribute	Class MuyPocos (0.25)	Pocos (0.25)	Medio (0.25)	Bastantes (0.15)	Muchos (0.1)
genero					
Accion	1.0	7.0	6.0	7.0	4.0
Aventura	1.0	2.0	2.0	4.0	3.0
Animacion	3.0	1.0	5.0	4.0	5.0
Comedia	17.0	14.0	15.0	7.0	3.0
Drama	14.0	16.0	11.0	3.0	4.0
Terror	6.0	5.0	5.0	3.0	1.0
Sci-fi	2.0	3.0	2.0	4.0	4.0
Thriller	8.0	7.0	6.0	2.0	3.0
Otros	7.0	3.0	7.0	4.0	2.0
[total]	59.0	58.0	59.0	38.0	29.0
edad					
Todos	9.0	6.0	5.0	7.0	6.0
>7	9.0	7.0	9.0	7.0	6.0
>12	14.0	20.0	20.0	7.0	7.0
>16	15.0	17.0	19.0	11.0	4.0
>18	8.0	4.0	2.0	2.0	2.0
[total]	55.0	54.0	55.0	34.0	25.0
mesEstreno					
Enero	4.0	8.0	2.0	7.0	1.0
Febrero	5.0	3.0	4.0	3.0	4.0

Marzo	3.0	6.0	3.0	3.0	3.0
Abril	9.0	4.0	6.0	2.0	3.0
Mayo	4.0	5.0	5.0	3.0	1.0
Junio	6.0	1.0	8.0	2.0	3.0
Julio	6.0	3.0	6.0	6.0	1.0
Agosto	2.0	6.0	8.0	1.0	3.0
Septiembre	3.0	5.0	6.0	3.0	2.0
Octubre	7.0	8.0	6.0	6.0	4.0
Noviembre	8.0	6.0	3.0	2.0	4.0
Diciembre	5.0	6.0	5.0	3.0	3.0
[total]	62.0	61.0	62.0	41.0	32.0

duracion					
mean	95.0907	106.4453	105.7334	113.8914	119.994
std. dev.	13.0091	14.2712	13.4866	17.9922	22.2032
weight sum	50	49	50	29	20
precision	1.5075	1.5075	1.5075	1.5075	1.5075

productora					
WaltDisneyCompany	1.0	1.0	3.0	3.0	6.0
TwentiethCenturyFox	1.0	2.0	1.0	4.0	2.0
UniversalPicturesInternational	2.0	2.0	8.0	5.0	6.0
WarnerBrosEntertainment	3.0	4.0	10.0	7.0	3.0
SonyPicturesReleasing	6.0	3.0	8.0	4.0	3.0
AurumProducciones	3.0	4.0	10.0	4.0	2.0
Paramount	1.0	3.0	2.0	3.0	3.0
HispanoFoxfilm	3.0	6.0	7.0	5.0	2.0
AContracorrienteFilms	10.0	6.0	2.0	1.0	1.0
CastelaoPictures	3.0	5.0	1.0	2.0	1.0
DeaPlaneta	1.0	5.0	3.0	1.0	1.0
Tripictures	1.0	4.0	2.0	2.0	1.0
Otra	28.0	17.0	6.0	1.0	2.0
[total]	63.0	62.0	63.0	42.0	33.0

pais					
ESP	22.0	15.0	12.0	2.0	7.0
EEUU	30.0	36.0	40.0	29.0	15.0

[total]	52.0	51.0	52.0	31.0	22.0
basadaEn					
Libro	3.0	12.0	7.0	4.0	6.0
Comic	1.0	1.0	2.0	5.0	2.0
HechosReales	6.0	5.0	4.0	2.0	3.0
Original	44.0	35.0	41.0	22.0	13.0
[total]	54.0	53.0	54.0	33.0	24.0

Matriz de confusión del clasificador:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 31  9 10  0  0 |  a = MuyPocos
 17 16 10  5  1 |  b = Pocos
 10 13 17  6  4 |  c = Medio
  2  3 10  7  7 |  d = Bastantes
  1  3  5  8  3 |  e = Muchos
```

2. Análisis 2.

Árbol de decisión de weka:

productora = WaltDisneyCompany
| edad = Todos
| | duracion <= 99: Medio (3.0/1.0)
| | duracion > 99: Muchos (2.0)
| edad = >7
| | duracion <= 132: Bastantes (2.0)
| | duracion > 132: Muchos (2.0)
| edad = >12: Muchos (0.0)
| edad = >16: Muchos (0.0)
| edad = >18: Muchos (0.0)
productora = TwentiethCenturyFox: Bastantes (5.0/2.0)
productora = UniversalPicturesInternational: Medio (18.0/11.0)
productora = WarnerBrosEntertainment: Medio (22.0/13.0)
productora = SonyPicturesReleasing: Medio (19.0/12.0)
productora = AurumProducciones: Medio (18.0/9.0)
productora = Paramount
| mesEstreno = Enero: Pocos (2.0)

| mesEstreno = Febrero: Bastantes (1.0)
| mesEstreno = Marzo: Pocos (0.0)
| mesEstreno = Abril: Pocos (0.0)
| mesEstreno = Mayo: Pocos (0.0)
| mesEstreno = Junio: Pocos (0.0)
| mesEstreno = Julio: Bastantes (1.0)
| mesEstreno = Agosto: Muchos (2.0)
| mesEstreno = Septiembre: Pocos (0.0)
| mesEstreno = Octubre: Medio (1.0)
| mesEstreno = Noviembre: Pocos (0.0)
| mesEstreno = Diciembre: Pocos (0.0)
productora = HispanoFoxfilm
| edad = Todos: Bastantes (2.0)
| edad = >7: Medio (3.0)
| edad = >12
| | genero = Accion: MuyPocos (0.0)
| | genero = Aventura: MuyPocos (0.0)
| | genero = Animacion: MuyPocos (0.0)
| | genero = Comedia: MuyPocos (3.0/1.0)
| | genero = Drama: MuyPocos (0.0)
| | genero = Terror: Medio (1.0)
| | genero = Sci-fi: MuyPocos (0.0)
| | genero = Thriller: Pocos (2.0/1.0)
| | genero = Otros: MuyPocos (0.0)
| edad = >16: Pocos (7.0/4.0)
| edad = >18: Medio (0.0)
productora = AContracorrienteFilms: MuyPocos (15.0/6.0)
productora = CastelaoPictures: Pocos (7.0/3.0)
productora = DeaPlaneta: Pocos (6.0/2.0)
productora = Tripictures
| duracion <= 111: Pocos (3.0)
| duracion > 111: Medio (2.0/1.0)
productora = BettaPictures: MuyPocos (3.0/1.0)
productora = InopiaFilms: Pocos (3.0/1.0)
productora = GolemDistribuciones
| duracion <= 98: Pocos (2.0)
| duracion > 98: MuyPocos (2.0/1.0)

productora = VertigoFilms: Pocos (3.0)

productora = Otra

| basadaEn = Libro: Pocos (4.0/1.0)

| basadaEn = Comic: MuyPocos (0.0)

| basadaEn = HechosReales: MuyPocos (3.0/1.0)

| basadaEn = Original: MuyPocos (29.0/8.0)

Matriz de confusión del árbol

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 34  5 11  0  0 |  a = MuyPocos
 19 15 12  3  0 |  b = Pocos
  9  9 25  5  2 |  c = Medio
  0  8 16  2  3 |  d = Bastantes
  1  3 11  5  0 |  e = Muchos
```

Clasificador de Naive Bayes de weka:

	Class				
Attribute	MuyPocos (0.25)	Pocos (0.25)	Medio (0.25)	Bastantes (0.15)	Muchos (0.1)
genero					
Accion	1.0	7.0	6.0	7.0	4.0
Aventura	1.0	2.0	2.0	4.0	3.0
Animacion	3.0	1.0	5.0	4.0	5.0
Comedia	17.0	14.0	15.0	7.0	3.0
Drama	14.0	16.0	11.0	3.0	4.0
Terror	6.0	5.0	5.0	3.0	1.0
Sci-fi	2.0	3.0	2.0	4.0	4.0
Thriller	8.0	7.0	6.0	2.0	3.0
Otros	7.0	3.0	7.0	4.0	2.0
[total]	59.0	58.0	59.0	38.0	29.0
edad					
Todos	9.0	6.0	5.0	7.0	6.0
>7	9.0	7.0	9.0	7.0	6.0

>12	14.0	20.0	20.0	7.0	7.0
>16	15.0	17.0	19.0	11.0	4.0
>18	8.0	4.0	2.0	2.0	2.0
[total]	55.0	54.0	55.0	34.0	25.0

mesEstreno

Enero	4.0	8.0	2.0	7.0	1.0
Febrero	5.0	3.0	4.0	3.0	4.0
Marzo	3.0	6.0	3.0	3.0	3.0
Abril	9.0	4.0	6.0	2.0	3.0
Mayo	4.0	5.0	5.0	3.0	1.0
Junio	6.0	1.0	8.0	2.0	3.0
Julio	6.0	3.0	6.0	6.0	1.0
Agosto	2.0	6.0	8.0	1.0	3.0
Septiembre	3.0	5.0	6.0	3.0	2.0
Octubre	7.0	8.0	6.0	6.0	4.0
Noviembre	8.0	6.0	3.0	2.0	4.0
Diciembre	5.0	6.0	5.0	3.0	3.0
[total]	62.0	61.0	62.0	41.0	32.0

duracion

mean	95.0907	106.4453	105.7334	113.8914	119.994
std. dev.	13.0091	14.2712	13.4866	17.9922	22.2032
weight sum	50	49	50	29	20
precision	1.5075	1.5075	1.5075	1.5075	1.5075

productora

WaltDisneyCompany	1.0	1.0	3.0	3.0	6.0
TwentiethCenturyFox	1.0	2.0	1.0	4.0	2.0
UniversalPicturesInternational	2.0	2.0	8.0	5.0	6.0
WarnerBrosEntertainment	3.0	4.0	10.0	7.0	3.0
SonyPicturesReleasing	6.0	3.0	8.0	4.0	3.0
AurumProducciones	3.0	4.0	10.0	4.0	2.0
Paramount	1.0	3.0	2.0	3.0	3.0
HispanoFoxfilm	3.0	6.0	7.0	5.0	2.0
AContracorrienteFilms	10.0	6.0	2.0	1.0	1.0
CastelaoPictures	3.0	5.0	1.0	2.0	1.0

DeaPlaneta	1.0	5.0	3.0	1.0	1.0
Tripictures	1.0	4.0	2.0	2.0	1.0
BettaPictures	3.0	1.0	2.0	1.0	1.0
InopiaFilms	2.0	3.0	1.0	1.0	1.0
GolemDistribuciones	2.0	3.0	2.0	1.0	1.0
VertigoFilms	1.0	4.0	1.0	1.0	1.0
Otra	24.0	10.0	4.0	1.0	2.0
[total]	67.0	66.0	67.0	46.0	37.0
pais					
ESP	22.0	15.0	12.0	2.0	7.0
EEUU	30.0	36.0	40.0	29.0	15.0
[total]	52.0	51.0	52.0	31.0	22.0
basadaEn					
Libro	3.0	12.0	7.0	4.0	6.0
Comic	1.0	1.0	2.0	5.0	2.0
HechosReales	6.0	5.0	4.0	2.0	3.0
Original	44.0	35.0	41.0	22.0	13.0
[total]	54.0	53.0	54.0	33.0	24.0
actorRenombre					
Si	10.0	18.0	21.0	19.0	15.0
No	42.0	33.0	31.0	12.0	7.0
[total]	52.0	51.0	52.0	31.0	22.0

Matriz de confusión del clasificador:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 32  7 10  0  1 |  a = MuyPocos
 14 16 10  8  1 |  b = Pocos
 11 12 15  8  4 |  c = Medio
  2  2 10  9  6 |  d = Bastantes
  2  0  6 11  1 |  e = Muchos
```

3. Análisis 3.

Árbol de decisión de weka:

productora = WaltDisneyCompany: Bastante (9.0/4.0)

productora = TwentiethCenturyFox: Bastante (5.0)

productora = UniversalPicturesInternational: Media (18.0/11.0)

productora = WarnerBrosEntertainment: Media (22.0/14.0)

productora = SonyPicturesReleasing

| mesEstreno = Enero: Poca (2.0/1.0)

| mesEstreno = Febrero: MuyPoca (2.0/1.0)

| mesEstreno = Marzo: Mucha (1.0)

| mesEstreno = Abril: Media (2.0)

| mesEstreno = Mayo: Poca (1.0)

| mesEstreno = Junio: Bastante (1.0)

| mesEstreno = Julio: MuyPoca (2.0/1.0)

| mesEstreno = Agosto: Media (2.0)

| mesEstreno = Septiembre: Media (0.0)

| mesEstreno = Octubre: Bastante (1.0)

| mesEstreno = Noviembre: MuyPoca (2.0/1.0)

| mesEstreno = Diciembre: MuyPoca (3.0/1.0)

productora = AurumProducciones: Media (18.0/7.0)

productora = Paramount

| mesEstreno = Enero: Poca (2.0)

| mesEstreno = Febrero: Bastante (1.0)

| mesEstreno = Marzo: Poca (0.0)

| mesEstreno = Abril: Poca (0.0)

| mesEstreno = Mayo: Poca (0.0)

| mesEstreno = Junio: Poca (0.0)

| mesEstreno = Julio: Poca (1.0)

| mesEstreno = Agosto: Bastante (2.0)

| mesEstreno = Septiembre: Poca (0.0)

| mesEstreno = Octubre: Media (1.0)

| mesEstreno = Noviembre: Poca (0.0)

| mesEstreno = Diciembre: Poca (0.0)

productora = HispanoFoxfilm: Media (18.0/12.0)

productora = AContracorrienteFilms: MuyPoca (15.0/6.0)

productora = CastelaoPictures: MuyPoca (7.0/4.0)

productora = DeaPlaneta: Poca (6.0/2.0)

productora = Tripictures: Poca (5.0/1.0)

productora = BettaPictures: MuyPoca (3.0/1.0)

productora = InopiaFilms: Poca (3.0/1.0)

productora = GolemDistribuciones

| duracion <= 98: Poca (2.0)

| duracion > 98: MuyPoca (2.0/1.0)

productora = VertigoFilms: Poca (3.0)

productora = Otra

| basadaEn = Libro: Poca (4.0/1.0)

| basadaEn = Comic: MuyPoca (0.0)

| basadaEn = HechosReales: MuyPoca (3.0/1.0)

| basadaEn = Original: MuyPoca (29.0/8.0)

Matriz de confusión del árbol:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 33  7  8  2  1 |  a = MuyPoca
 19 12 10  8  0 |  b = Poca
 10  6 26  6  0 |  c = Media
  4  7 16  9  2 |  d = Bastante
  3  1  4  4  0 |  e = Mucha
```

Clasificador de Naive Bayes de weka:

	Class				
Attribute	MuyPoca	Poca	Media	Bastante	Mucha
	(0.26)	(0.25)	(0.24)	(0.19)	(0.06)
genero					
Accion	1.0	5.0	6.0	9.0	4.0
Aventura	1.0	2.0	2.0	5.0	2.0
Animacion	3.0	1.0	3.0	7.0	4.0
Comedia	17.0	13.0	16.0	8.0	2.0
Drama	15.0	16.0	10.0	4.0	3.0
Terror	6.0	5.0	7.0	1.0	1.0
Sci-fi	2.0	4.0	2.0	4.0	3.0
Thriller	8.0	7.0	5.0	5.0	1.0

Otros	7.0	5.0	6.0	4.0	1.0
[total]	60.0	58.0	57.0	47.0	21.0

edad

Todos	9.0	5.0	4.0	11.0	4.0
>7	9.0	7.0	9.0	9.0	4.0
>12	14.0	22.0	17.0	12.0	3.0
>16	15.0	17.0	21.0	9.0	4.0
>18	9.0	3.0	2.0	2.0	2.0
[total]	56.0	54.0	53.0	43.0	17.0

mesEstreno

Enero	4.0	7.0	3.0	7.0	1.0
Febrero	5.0	3.0	4.0	5.0	2.0
Marzo	3.0	6.0	2.0	4.0	3.0
Abril	9.0	4.0	6.0	2.0	3.0
Mayo	4.0	6.0	4.0	3.0	1.0
Junio	6.0	1.0	7.0	3.0	3.0
Julio	6.0	4.0	6.0	4.0	2.0
Agosto	2.0	5.0	9.0	3.0	1.0
Septiembre	3.0	4.0	5.0	5.0	2.0
Octubre	8.0	9.0	6.0	7.0	1.0
Noviembre	8.0	6.0	3.0	3.0	3.0
Diciembre	5.0	6.0	5.0	4.0	2.0
[total]	63.0	61.0	60.0	50.0	24.0

duracion

mean	95.65	106.2915	105.5224	114.4085	121.7276
std. dev.	13.4743	13.7391	13.9525	18.9447	22.0705
weight sum	51	49	48	38	12
precision	1.5075	1.5075	1.5075	1.5075	1.5075

productora

WaltDisneyCompany	1.0	1.0	2.0	6.0	4.0
TwentiethCenturyFox	1.0	1.0	1.0	6.0	1.0
UniversalPicturesInternational	2.0	2.0	8.0	5.0	6.0
WarnerBrosEntertainment	3.0	5.0	9.0	9.0	1.0

SonyPicturesReleasing	6.0	3.0	7.0	6.0	2.0
AurumProducciones	3.0	3.0	12.0	3.0	2.0
Paramount	1.0	4.0	2.0	4.0	1.0
HispanoFoxfilm	3.0	6.0	7.0	5.0	2.0
AContracorrienteFilms	10.0	5.0	3.0	1.0	1.0
CastelaoPictures	4.0	4.0	1.0	2.0	1.0
DeaPlaneta	1.0	5.0	3.0	1.0	1.0
Tripictures	1.0	5.0	1.0	2.0	1.0
BettaPictures	3.0	1.0	2.0	1.0	1.0
InopiaFilms	2.0	3.0	1.0	1.0	1.0
GolemDistribuciones	2.0	3.0	2.0	1.0	1.0
VertigoFilms	1.0	4.0	1.0	1.0	1.0
Otra	24.0	11.0	3.0	1.0	2.0
[total]	68.0	66.0	65.0	55.0	29.0

pais

ESP	23.0	16.0	9.0	7.0	3.0
EEUU	30.0	35.0	41.0	33.0	11.0
[total]	53.0	51.0	50.0	40.0	14.0

basadaEn

Libro	3.0	10.0	8.0	7.0	4.0
Comic	1.0	1.0	2.0	5.0	2.0
HechosReales	6.0	5.0	4.0	4.0	1.0
Original	45.0	37.0	38.0	26.0	9.0
[total]	55.0	53.0	52.0	42.0	16.0

Matriz de confusión del clasificador:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 30 11  9  0  1 |  a = MuyPoca
 17 15 10  7  0 |  b = Poca
  9 10 20  8  1 |  c = Media
  3  6 10 16  3 |  d = Bastante
  1  1  1  9  0 |  e = Mucha
```

4. Análisis 4.

Árbol de decisión de weka:

productora = WaltDisneyCompany: Bastante (9.0/4.0)

productora = TwentiethCenturyFox: Bastante (5.0)

productora = UniversalPicturesInternational: Media (18.0/11.0)

productora = WarnerBrosEntertainment: Media (22.0/14.0)

productora = SonyPicturesReleasing

| mesEstreno = Enero: Poca (2.0/1.0)

| mesEstreno = Febrero: MuyPoca (2.0/1.0)

| mesEstreno = Marzo: Mucha (1.0)

| mesEstreno = Abril: Media (2.0)

| mesEstreno = Mayo: Poca (1.0)

| mesEstreno = Junio: Bastante (1.0)

| mesEstreno = Julio: MuyPoca (2.0/1.0)

| mesEstreno = Agosto: Media (2.0)

| mesEstreno = Septiembre: Media (0.0)

| mesEstreno = Octubre: Bastante (1.0)

| mesEstreno = Noviembre: MuyPoca (2.0/1.0)

| mesEstreno = Diciembre: MuyPoca (3.0/1.0)

productora = AurumProducciones: Media (18.0/7.0)

productora = Paramount

| mesEstreno = Enero: Poca (2.0)

| mesEstreno = Febrero: Bastante (1.0)

| mesEstreno = Marzo: Poca (0.0)

| mesEstreno = Abril: Poca (0.0)

| mesEstreno = Mayo: Poca (0.0)

| mesEstreno = Junio: Poca (0.0)

| mesEstreno = Julio: Poca (1.0)

| mesEstreno = Agosto: Bastante (2.0)

| mesEstreno = Septiembre: Poca (0.0)

| mesEstreno = Octubre: Media (1.0)

| mesEstreno = Noviembre: Poca (0.0)

| mesEstreno = Diciembre: Poca (0.0)

productora = HispanoFoxfilm: Media (18.0/12.0)

productora = AContracorrienteFilms: MuyPoca (15.0/6.0)

productora = CastelaoPictures: MuyPoca (7.0/4.0)

productora = DeaPlaneta: Poca (6.0/2.0)

productora = Tripictures: Poca (5.0/1.0)

productora = BettaPictures: MuyPoca (3.0/1.0)

productora = InopiaFilms: Poca (3.0/1.0)

productora = GolemDistribuciones

| duracion <= 98: Poca (2.0)

| duracion > 98: MuyPoca (2.0/1.0)

productora = VertigoFilms: Poca (3.0)

productora = Otra

| basadaEn = Libro: Poca (4.0/1.0)

| basadaEn = Comic: MuyPoca (0.0)

| basadaEn = HechosReales: MuyPoca (3.0/1.0)

| basadaEn = Original: MuyPoca (29.0/8.0)

Matriz de confusión del árbol:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 33  7  8  2  1 |  a = MuyPoca
 18 13 10  8  0 |  b = Poca
 10  6 26  6  0 |  c = Media
  4  7 15  9  3 |  d = Bastante
  3  1  4  4  0 |  e = Mucha
```

Clasificador de Naive Bayes de weka:

	Class				
Attribute	MuyPoca	Poca	Media	Bastante	Mucha
	(0.26)	(0.25)	(0.24)	(0.19)	(0.06)
genero					
Accion	1.0	5.0	6.0	9.0	4.0
Aventura	1.0	2.0	2.0	5.0	2.0
Animacion	3.0	1.0	3.0	7.0	4.0
Comedia	17.0	13.0	16.0	8.0	2.0
Drama	15.0	16.0	10.0	4.0	3.0
Terror	6.0	5.0	7.0	1.0	1.0
Sci-fi	2.0	4.0	2.0	4.0	3.0
Thriller	8.0	7.0	5.0	5.0	1.0
Otros	7.0	5.0	6.0	4.0	1.0

[total]	60.0	58.0	57.0	47.0	21.0
edad					
Todos	9.0	5.0	4.0	11.0	4.0
>7	9.0	7.0	9.0	9.0	4.0
>12	14.0	22.0	17.0	12.0	3.0
>16	15.0	17.0	21.0	9.0	4.0
>18	9.0	3.0	2.0	2.0	2.0
[total]	56.0	54.0	53.0	43.0	17.0
mesEstreno					
Enero	4.0	7.0	3.0	7.0	1.0
Febrero	5.0	3.0	4.0	5.0	2.0
Marzo	3.0	6.0	2.0	4.0	3.0
Abril	9.0	4.0	6.0	2.0	3.0
Mayo	4.0	6.0	4.0	3.0	1.0
Junio	6.0	1.0	7.0	3.0	3.0
Julio	6.0	4.0	6.0	4.0	2.0
Agosto	2.0	5.0	9.0	3.0	1.0
Septiembre	3.0	4.0	5.0	5.0	2.0
Octubre	8.0	9.0	6.0	7.0	1.0
Noviembre	8.0	6.0	3.0	3.0	3.0
Diciembre	5.0	6.0	5.0	4.0	2.0
[total]	63.0	61.0	60.0	50.0	24.0
duracion					
mean	95.65	106.2915	105.5224	114.4085	121.7276
std. dev.	13.4743	13.7391	13.9525	18.9447	22.0705
weight sum	51	49	48	38	12
precision	1.5075	1.5075	1.5075	1.5075	1.5075
productora					
WaltDisneyCompany	1.0	1.0	2.0	6.0	4.0
TwentiethCenturyFox	1.0	1.0	1.0	6.0	1.0
UniversalPicturesInternational	2.0	2.0	8.0	5.0	6.0
WarnerBrosEntertainment	3.0	5.0	9.0	9.0	1.0
SonyPicturesReleasing	6.0	3.0	7.0	6.0	2.0

AurumProducciones	3.0	3.0	12.0	3.0	2.0
Paramount	1.0	4.0	2.0	4.0	1.0
HispanoFoxfilm	3.0	6.0	7.0	5.0	2.0
AContracorrienteFilms	10.0	5.0	3.0	1.0	1.0
CastelaoPictures	4.0	4.0	1.0	2.0	1.0
DeaPlaneta	1.0	5.0	3.0	1.0	1.0
Tripictures	1.0	5.0	1.0	2.0	1.0
BettaPictures	3.0	1.0	2.0	1.0	1.0
InopiaFilms	2.0	3.0	1.0	1.0	1.0
GolemDistribuciones	2.0	3.0	2.0	1.0	1.0
VertigoFilms	1.0	4.0	1.0	1.0	1.0
Otra	24.0	11.0	3.0	1.0	2.0
[total]	68.0	66.0	65.0	55.0	29.0

pais					
ESP	23.0	16.0	9.0	7.0	3.0
EEUU	30.0	35.0	41.0	33.0	11.0
[total]	53.0	51.0	50.0	40.0	14.0

basadaEn					
Libro	3.0	10.0	8.0	7.0	4.0
Comic	1.0	1.0	2.0	5.0	2.0
HechosReales	6.0	5.0	4.0	4.0	1.0
Original	45.0	37.0	38.0	26.0	9.0
[total]	55.0	53.0	52.0	42.0	16.0

actorRenombre					
Si	10.0	16.0	24.0	25.0	8.0
No	43.0	35.0	26.0	15.0	6.0
[total]	53.0	51.0	50.0	40.0	14.0

Matriz de confusión del clasificador:

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 31  9 10  0  1 |  a = MuyPoca
 16 16 10  7  0 |  b = Poca
  9  8 20 10  1 |  c = Media
  4  3 11 16  4 |  d = Bastante
  1  0  2  9  0 |  e = Mucha
```

7. Conclusiones

- Análisis 1.

    - Árbol de decisión (J48): En el árbol resultante destacamos que la variable productora ha sido muy importante, se ha escogido como el nodo raíz. A parte de la productora, encontramos la variable edad para productoras como Walt Disney, la cual es una productora de películas sobretodo infantiles. También podemos apreciar que para la compañía de Paramount el mes de estreno ha influido en el éxito de sus películas, esta es una productora que estrena películas durante todo el año, sin embargo, sólo en los meses de verano (Julio y Agosto) es dónde registra un mayor éxito.

    - Matriz de confusión (J48): La matriz resultante tiene un porcentaje de acierto del 38%. Es un porcentaje bastante bajo, aunque no se equivoca para rangos muy lejanos, es decir, una película determinada como muy poco éxito no la va a clasificar como bastante exitosa pero si como poco exitosa. Otra cosa que cabe destacar es que el porcentaje de acierto para una película que pueda ser medianamente, poco o muy poco exitosa es mayor que para las películas más exitosas.

    - Naive Bayes: Utilizando Naive Bayes no hemos conseguido mejorar el porcentaje de acierto, por lo que las conclusiones son bastante parecidas al algoritmo anterior.

- Análisis 2.

    - Árbol de decisión (J48): En el árbol resultante obtenemos prácticamente los mismos resultados que para el análisis 1.

- Matriz de confusión (J48): La matriz resultante tiene un porcentaje de acierto del 38%. Los valores para la matriz de confusión son idénticos.

- Naive Bayes: Utilizando Naive Bayes no hemos conseguido mejorar el porcentaje de acierto sino empeorarlo un poco, 36%, por lo que las conclusiones son bastante parecidas al algoritmo anterior.

- Análisis 3.

  - Árbol de decisión (J48): Ahora la productora SonyPicturesReleasing también vemos que depende bastante su éxito de las fechas de estreno de sus películas en donde los meses de verano (Junio, Julio, Agosto, Septiembre) son los ganadores.

  - Matriz de confusión (J48): En la matriz resultante obtenemos una pequeña mejora, el porcentaje de acierto ahora es del 40% con la variable de recaudación, pero aún sigue siendo bastante bajo. El árbol sigue generando las mismas conclusiones, es capaz de distinguir películas malas pero le es difícil acertar las buenas.

- Análisis 4.

  - Árbol de decisión (J48): No se obtienen conclusiones nuevas.

  - Matriz de confusión (J48): En la matriz resultante obtenemos una pequeña mejora, el porcentaje de acierto ahora es del 41% con la variable de actores de éxito, pero aún sigue siendo bastante bajo. El árbol sigue generando las mismas conclusiones, es capaz de distinguir películas malas pero le es difícil acertar las buenas.

8. Conclusiones del proyecto

Una vez finalizado el análisis de los datos de las películas estrenadas en el año 2015 en España es que resulta difícil predecir cuáles van a ser las "*películas del año",* lo que sí podemos predecir es aquellas características que hacen a un película, una película sin éxito.

Como hemos podido comprobar entre las variables más importante de la industria son las productoras, las buenas productoras tienen mejores condiciones para producir grandes películas (actores de éxito, presupuesto, cantidad de películas, etc) por lo que las estadísticas las convierte en candidatos favorables para obtener el éxito.

Otras variables a destacar han sido las fechas de estreno, para algunas productoras los meses de verano como son Junio, Julio, Agosto y Septiembre han determinan el éxito de sus películas, esto confirmaría que en verano la gente va más al cine pues están de vacaciones y tienen más tiempo libre.

Para productoras muy pequeñas hemos visto que el éxito ha recaído en la originalidad de la película, su variable más importante ha sido en el hecho en que están basadas y hemos visto que cuando tienen mayor éxito es para películas basadas en libros, las basadas en comics por ejemplo no resultan nada exitosas para estas productoras, esto puede deberse a que el presupuesto para vestuario y efectos especiales son menores y las escenas no quedan tan elaboradas.

9. Bibliografía

Labutaca:

http://www.labutaca.net/guiacine/estrenos-de-cine-peliculas-2015.htm

Buscador del ministerio de educación cultura y deporte:

http://www.mecd.gob.es/bbddpeliculas/cargarFiltro.do?layout=bbddpeliculas&cache=init&language=es

Sensacine:

http://www.sensacine.com/

Filmaffinity:

http://www.filmaffinity.com/es/main.html

web de Weka:

http://www.cs.waikato.ac.nz/ml/weka/index.html

Repositorio donde pueden encontrarse los ficheros empleados para los análisis así como todos los datos del proyecto:

https://github.com/alu0100774054/ProyectoWeka