

Análisis de datos

~~Datos interesantes detrás~~ de un concurso de cine.

L.A. Gutierrez-Rodriguez

Posgrado de Ingeniería en Sistemas,
Facultad de Ingeniería Mecánica y Eléctrica,
Universidad Autónoma de Nuevo León

Abstract

Resumen

Contamos con los registros de un concurso de cine, cuyos filmes han sido grabados con dispositivos smartphone. Se busca identificar el alcance del certamen, ayudar a definir las mejores técnicas de difusión, caracterizar las categorías para equilibrar la cantidad de participantes, mejorar la forma en que se registraron los participantes y pronosticar la participación de los próximos años.

La información se encuentra almacenada en cinco documentos de hoja de cálculo. Cada documento cuenta con los registros del concurso de cada año, del 2015 al 2018. Cuatro de ellos son del festival con sede en Colombia y el último del primer festival en México.

Se utiliza el lenguaje de programación Python y las librerías Numpy, Pandas, Scipy y Matplotlib para realizar análisis estadístico sobre los datos y poder graficar los resultados. Se aplican modelos de regresión lineal y múltiple. Se analiza la varianza y las componentes principales de los datos. Se pronostica la cantidad de participantes de futuros años. Se implementa la librería SKLearn para clasificar y agrupar los datos. Y se realiza análisis de texto de las sinopsis e imágenes de los vídeos ganadores. Al aplicar técnicas de estadística descriptiva se puede determinar que la mayoría de extranjeros que participan en el concurso son de México. Por este motivo se decidió en 2018 abrir un festival en ese país. ?

Introducción

En la actualidad contamos con una cantidad inmensa de datos debido a que almacenamos toda clase de información. Se busca darle un sentido útil a estos datos para poder comprenderlos. La ciencia de datos es el estudio de la extracción generalizable de conocimiento a partir de datos [6]. En este trabajo se busca aplicar la ciencia de datos ayudar en la toma de decisiones de dónde y cuándo se deben organizar los eventos, la mejor forma de organizar las categorías y en que otros países se puede expandir el festival.

Se sabe que la mayoría de los concursantes son de regiones cercanas a la capital de Colombia, que estos concursantes usan smartphones para grabar sus vídeos, que proporcionan una sinopsis, y cada participante ubica su filme en un género. Además, no se tiene identificación de los participantes, y por esto no se puede trazar su participación a lo largo de los festivales.

En antecedentes hablaremos de todo lo relacionado con el festival de cine. En la literatura relacionada, veremos que problemas ya han sido tratados y como la obtención de información ayudó a la toma de decisiones. Además, se tratarán temas que están relacionadas con el tipo de información que buscamos. En la metodología aplicada se mencionan las hipótesis que probamos sobre los datos y las herramientas utilizadas. En la sección de los resultados podremos observar cómo se comportaron los

datos con los análisis que probamos. Y en conclusión determinamos el sentido de nuestras hipótesis.

Antecedentes

SmartFilms se ha convertido en el festival de cine más importante del momento en Colombia, gracias a los diferentes escenarios que les permiten a los participantes exhibir todo tipo de contenidos, usando los valores y conceptos de las marcas patrocinadoras de las diferentes categorías por medio del storytelling, product placement y branded content, de esta manera, marcar un nuevo camino hacia el crecimiento de las industrias, utilizando las empresas privadas como aliado en la producción cinematográfica.

Los participantes deben realizar un cortometraje de máximo 5 minutos, incluyendo créditos y este debe ser grabado en su totalidad con un celular o dispositivo móvil, adjuntar un making of (detrás de cámaras), adjuntar cartel de propaganda y enviar el cortometraje antes del cierre de la convocatoria.

El festival es anual. En cada proceso se realizan actividades pedagógicas y de activación para incentivar a las personas a que participen y fomenten la industria cinematográfica

Literatura relacionada

A través del tiempo, el cine ha sido considerado el séptimo arte, la octava maravilla y por ello se han creado certámenes para apreciar el trabajo que conlleva realizar un film. El certamen más importante en el mundo del cine son los premios

Email address: luis.gutierrezrd@uanl.edu.mx (L.A. Gutierrez-Rodriguez)

Oscar. Depende de la academia determinar cual film es ganador de ciertas categorías como iluminación, guión, director, actores, entre otros. Las tecnologías de la información son un rubro tan amplio que se extiende al cine. Se puede analizar la composición de imagen de un film no solo con el ojo de un experto, si no también con visión computacional. En el 2013 se publicó un artículo [7] donde se hizo un análisis predictivo y se buscaba determinar quiénes serían los ganadores de las diversas categorías. Esta aplicación también ilustró cómo Data Science podría implementarse en las industrias de medios y entretenimiento.

En el libro *Movie analytics: a hollywood introduction to big data* [8] se aplicaron la minería de datos, minería de textos y análisis de redes sociales para aprender a analizar los datos de las películas. Se buscó una relación entre la información obtenida del análisis y la cantidad de estrellas otorgadas por el publico en la IMDB. Además compararon lo aprendido con datos reales obtenidos de una película Francesa. En el capítulo *Oscar Prediction and Prediction Markets* [9] se examinó el papel de los mercados de predicción en la evaluación de la probabilidad de que una película nominada reciba un premio de la Academia.

Metodología

En esta sección veremos la información particular de los datos, y que herramientas utilizamos y con que fin. Se aplicó un proceso de limpieza de datos ya que estos presentaban espacios vacíos, errores de tipo de dato o se encontraban formas en las que se llenó un campo, siendo varias de ellas diferentes representaciones de la misma información. Finalmente, de acuerdo a que hipótesis se estaba probando, se seleccionaba solo algunos campos de los registros.

Datos

Contamos con cinco documentos en formato de hoja de cálculo, los cuales son los registros al festival de cine. Estos datos representan los registros desde el año 2015 al 2018.

Herramientas

Utilizamos las librerías Numpy[2] y Pandas[3] para la captura y almacenamiento de los datos limpios y poder trabajar con ellos.

Importamos la librería Scipy[4] para poder realizar análisis estadísticos y Matplotlib[1] para graficar los resultados.

Se implementó SKLearn[5] para realizar clasificaciones y agrupamientos de los datos en base a su comportamiento.

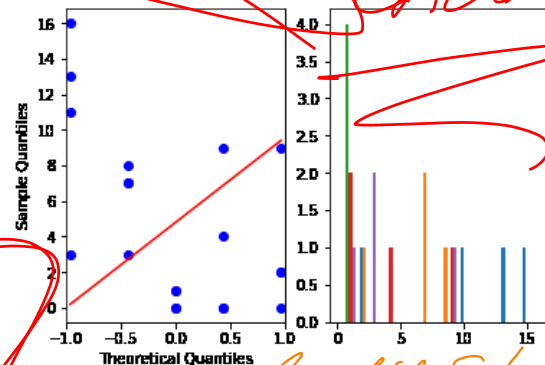
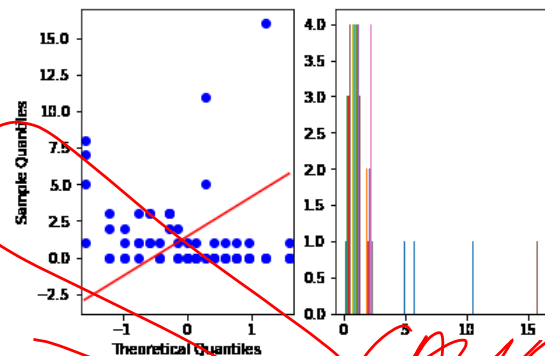
Resultados

Se graficaron los datos para ver si se podía concluir algo sobre estos. Después se seleccionaron los participantes extranjeros de todos los cuatro años, y se ordenaron por el país de procedencia.

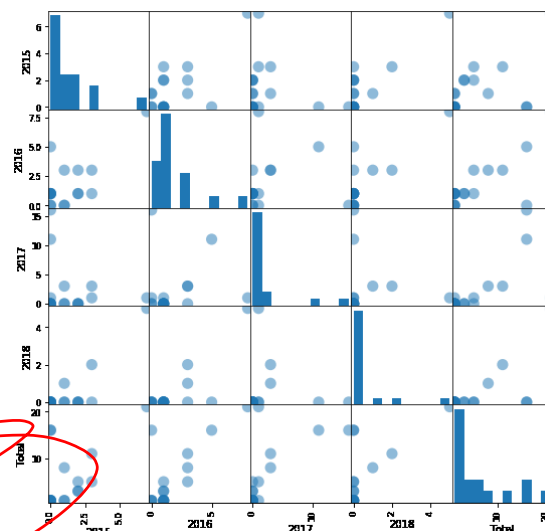
Categorizaciones

Como tenemos muchas cadenas de texto en nuestros datos, es importante hacer categorizaciones ya que hacemos conteos de la información que tenemos disponible y así poder hacer cálculos estadísticos. Aplicamos una categorización por país, pero se removieron los datos que involucran a México y Colombia que son los países anfitriones del concurso y por ende son los que nos sesgan la información, entonces esta información será relevante a los países extranjeros que participan.

Con estas categorías pudimos buscar Modelos lineales y regresión múltiple.



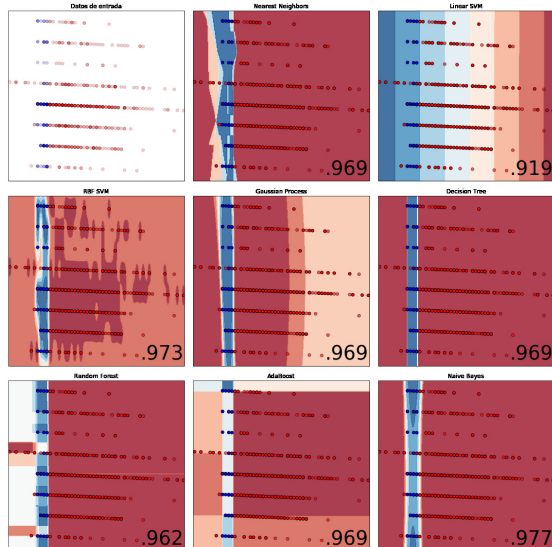
Primero probamos la normalidad de los datos por país de procedencia y por categoría de participación.



Se hizo una regresión lineal múltiple entre los años de participación, para saber si existía correlación.

Machine Learning

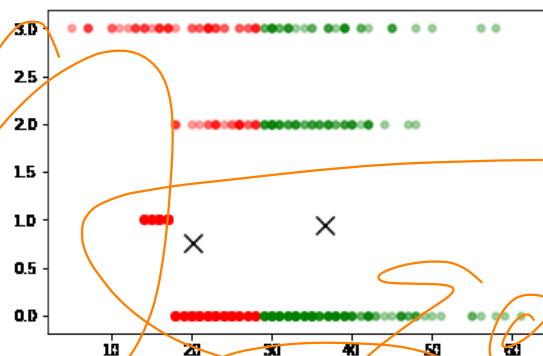
También se usó la librería *sklearn* para ver si las clasificaciones de las categorías ayudaban a entrenar a una red neuronal para que cuando se le entregaran otro conjunto de datos los clasificara correctamente.



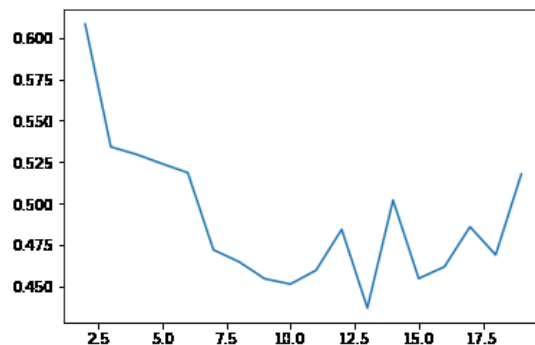
Posteriormente, seguimos utilizando la librería para aplicar algoritmos de agrupamiento, para que los datos se ordenaran de acuerdo a sus propias características.

Algoritmo K-Means

Este algoritmo agrupa los datos al tratar de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como la inercia o la suma de cuadrados dentro del grupo. Este algoritmo requiere que se especifique la cantidad de grupos. Se adapta bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.

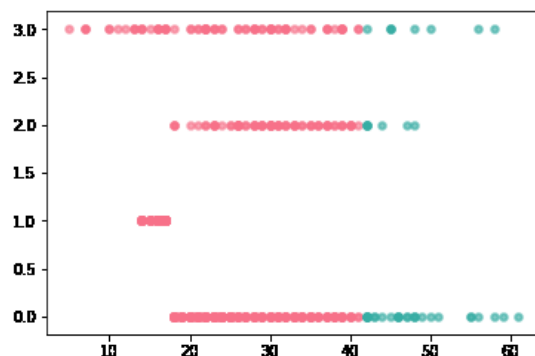


Al aplicar el algoritmo de K-medias, el valor de K usado fue 2. También evaluó cambiar el valor de K, y obtuvimos la siguiente gráfica.



Algoritmo Affinity Propagation

Este algoritmo crea clústeres enviando mensajes entre pares de muestras hasta la convergencia. Luego se describe un conjunto de datos utilizando un pequeño número de ejemplares, que se identifican como los más representativos de otras muestras. Los mensajes enviados entre pares representan la idoneidad para que una muestra sea el ejemplar de la otra, que se actualiza en respuesta a los valores de otros pares. Esta actualización ocurre de manera iterativa hasta la convergencia, momento en el que se eligen los ejemplares finales y, por lo tanto, se proporciona la agrupación final.



Se aplicó el algoritmo Affinity Propagation sobre los mismos datos que el K-medias, donde también se dividió en 2 grupos, pero la división de esos grupos se movió a mayor edad.

Conclusiones

La información proporcionada que se procesó fue insuficiente. Contábamos con aproximadamente 4600 registros de los cuales solo el 10% tenía la información completa.

Debido a que fue el primer año en registrarse el año 2015 fue el cuello de botella para procesar el resto de la información.

El año 2016, se tuvo mayor participación en las categorías infantil y juvenil.

En el año 2017, la cantidad de participantes extranjeros aumentó.

En 2018, debido a que se contemplaron los dos festivales, el Colombiano y el Mexicano, la cantidad de extranjeros que participaron disminuyó, porque se decidió iniciar un festival nuevo en el país que más extranjeros aportaba al festival.

Se deberían agregar limitaciones a los campos de categoría de este festival, como los géneros de los filmes, las Marcas de

celulares y limitar la sinopsis a una cantidad precisa de caracteres.

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por haberme financiado una beca sin la cual no podría haber analizado estos datos.

Agradezco al Sr. Juan Beltrán, Director creativo de Valencia Producciones y al festival SmartFilm, por haberme proporcionado los datos para su análisis.

Referencias

- [1] Matplotlib. URL: <https://matplotlib.org/>.
- [2] Numpy. URL: <https://www.numpy.org/>.
- [3] Python data analysis library. URL: <https://pandas.pydata.org/>.
- [4] Scipy.org. URL: <https://www.scipy.org/>.
- [5] Sklearn. URL: <https://scikit-learn.org/stable/>.
- [6] Vasant Dhar. Data science and prediction. 2012.
- [7] Michael Gold, Ryan McClarren, and Conor Gaughan. The lessons oscar taught us: Data science and media & entertainment. *Big Data*, 1(2):105–109, 2013.
- [8] Dominique Haughton, Mark-David McLaughlin, Kevin Mentzer, and Changan Zhang. *Movie analytics: a hollywood introduction to big data*. Springer, 2015.
- [9] Dominique Haughton, Mark-David McLaughlin, Kevin Mentzer, and Changan Zhang. *Oscar Prediction and Prediction Markets*, pages 37–39. 01 2015.

@ i