Predicción de la calificación de una película chilena en el sitio IMDb

Juan Carlos Barría Rival

ARTICLE INFO

Keywords:
Peliculas
Predicción
IMDb
CineChile
Chile
Peliculas Chilenas
Aprendizaje automático
Random Forest

ABSTRACT

Varias producciones chilenas han tenido éxito en otros mercados internacionales, hasta han ganado premios de la academia. En este proyecto se busca obtener la predicción de la nota o rating de una película chilena, tomando datos de sitios especializados y trabajando con algoritmos de aprendizaje automático, con el objetivo de conocer las características que hacen que estas sean populares a través del mundo, con la idea de exportarlas y que sean vistas por otros lugares del mundo.

1. Introducción

La industria del cine en chile se ha expandido también a las plataformas de streaming, por lo que ahora no solo se puede ir a disfrutar de una película chilena al cine, sino que también en la comodidad de la casa a través de estas plataformas. Según datos de JustWatch [3], plataforma que reúne a varios servicios de contenido en demanda de películas y series, el consumo de streaming ha aumentado en un 157% en Chile. Este incremento también se vio reflejado por toda América Latina. Asimismo, esta tendencia está asociada a una reducción en los suscriptores de televisión de pago (-2,8%) durante el primer trimestre de 2020 según los datos entregados por la subsecretaría de telecomunicaciones.

Hoy en día, el streaming se ha convertido en un recurso más para hacernos compañía, durante los días de encierro, ya que como se decía, ha aumentado radicalmente el consumo de plataformas digitales y esto podría provocar un cambio cultural definitivo [5]. Por lo que para este proyecto además de los datos recogidos anteriormente en el sitio CineChile.cl se agregaron nuevos datos pertenecientes al cine chileno en estas plataformas de streaming, específicamente de Netflix y Amazon Prime Video.

Recordemos que IMDb [2] es una base de datos en línea que almacena información relacionada con películas, con sus características, además de la valoración por parte del público con nota de 1 a 10, la cual se utilizará para el entrenamiento del modelo. Ya que en este proyecto se busca predecir las calificaciones de las producciones chilenas, con el objetivo de conocer las características que hacen que estas sean populares, con la idea de exportarlas a otros mercados y sean vistas en otros lugares del mundo, ya sea a través del cine o de plataformas de streaming.

2. Conjunto de datos

Como se comentaba en la introducción, además de la lista que entregaba CineChile, se decidió incluir producciones chilenas que están actualmente en las plataformas de *streaming Netflix* y *Amazon Prime Video*, guiándose por el artículo de MundoPelículas [6]. Por lo tanto, el conjunto de datos se actualizó y se agregaron un total de 27 nuevas películas a este *dataset*, entre las que se encuentran 10 de la plataforma de Netflix y 17 de Amazon Prime Video. Entre las que destacan "Una Mujer Fantástica" ganadora de un Oscar en 2018 a "Mejor Película Extranjera", y "Mi Amigo Alexis" película que cuenta la historia de este destacado deportista de nuestro país (Las películas *No*, *Machuca* y *Tony Manero* que se encuentran en *Netflix*, y *Johnny Cien Pesos* que está en *Prime Video*, ya se encontraban dentro del conjunto de datos).

Por lo que ahora el conjunto de datos cuenta con un total 87 producciones nacionales.

Recordando, el conjunto de datos creado tiene las siguientes características:

1. ID: Número identificador.
2. Title: Título de la película.
3. Year: Año en que se estrenó.
4. Age: Edad a la que esta dirigida.
5. IMDb: Calificación en el sitio IMDb.
6. CineChile: Puntaje asignado por CineChile.
7. Directors: Director/es de la película.
8. Genres: Género/s de la película.
9. Language: Lenguaje/s en el que está disponible.
10. Runtime: Duración en minutos.

2.1. Limpieza y preprocesamiento

Recordemos que como base el proyecto utilizado anteriormente llamado *IMDb Rating Prediction from a data set of Movies* [1], hecho por el usuario "diptaraj23" subido a la plataforma *Kaggle*. Aquí se cargó el nuevo conjunto de datos creado de producciones chilenas con sus nuevas modificaciones. Se hizo limpieza de los datos nulos que se ven en la Figura 1.

Se eliminaron las columnas que no se toman en cuenta para la predicción Title' y 'Cine Chile', y también para este avance durante esta etapa, también se eliminaron todas las películas menores a 60 minutos de duración denominadas cortometrajes y mediometrajes, según Alfred Lopéz en su artículo [4] basándose en la RAE, para finalmente ya quedar listo los datos como se muestra en la Figura 3.

Se copian estos datos de entrenamiento y testeo para evitar cambios del *dataset* original y se aplica *LabelEncoder* para codificar etiquetas de una característica categórica en valores numéricos entre 0 y el número de clases menos 1. El

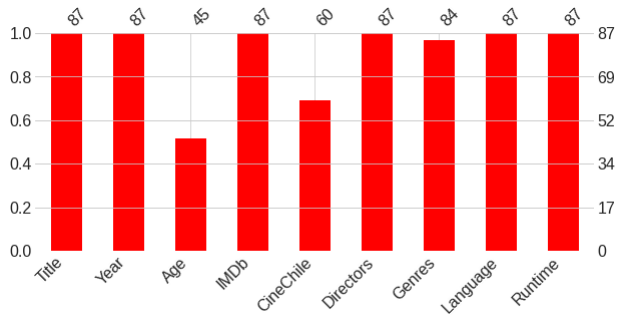

Figure 1: Visualización de los datos faltantes

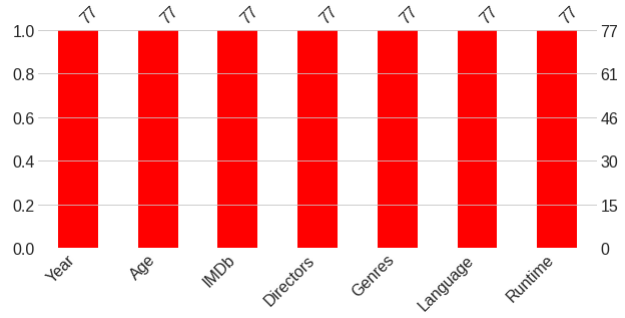

Figure 2: Visualización de los datos sin los datos nulos y las características que no se toman en cuenta

método *fit* lo entrena y el método transform transforma las etiquetas que se incluyan como argumento en los números correspondientes. Posteriormente de las columnas que se utilizarán (*Year, Runtime, Age, Directors, Genre, Language*), se dividirán los datos que se usarán para entrenamiento y testeo.

3. Modelos

Para este caso, se definieron un total de 10 modelos, como se muestra a continuación, (5 más de los utilizados en el proyecto original). usando el algoritmo de aprendizaje automático de Bosques Aleatorios, específicamente el de regresión, **RandomForestRegressor**, que consiste en una técnica de aprendizaje supervisado. Esta es en un conjuntos de árboles de decisión combinados, cada árbol se entrena con distintas muestras de datos, tomando su media aritmética. Luego se compara los rendimientos de cada uno.

```
#Modelos Definidos
model_1 = RandomForestRegressor(n_estimators=50,
                                random_state=1)
model_2 = RandomForestRegressor(n_estimators=100,
                                random_state=1)
model_3 = RandomForestRegressor(n_estimators=100,
                                criterion='mae',
                                random_state=1)
model_4 = RandomForestRegressor(n_estimators=200,
                                min_samples_split=20,
                                random_state=1)
model_5 = RandomForestRegressor(n_estimators=100,
                                max_depth=7,
                                random_state=1)
model_6 = RandomForestRegressor(n_estimators=50,
                                max_depth=7,
                                random_state=1)
model_7 = RandomForestRegressor(n_estimators=300,
                                max_depth=9,
                                random_state=1)
model_8 = RandomForestRegressor(n_estimators=400,
                                max_depth=8,
                                random_state=1)
model_9 = RandomForestRegressor(n_estimators=500,
                                max_depth=10,
                                random_state=1)
model_10 = RandomForestRegressor(n_estimators=1000,
                                max_depth=10,
                                random_state=1)

#Lista de modelos
models = [model_1, model_2, model_3, model_4, model_5,
        model_6, model_7, model_8, model_9, model_10]
```

Luego se compararon estos modelos, para obtener cual de todos tenia mejor puntaje (ó menor Error medio absoluto), los resultados se ven a continuación:

```
Modelo 1 MAE(Error medio absoluto): 0.489400
Modelo 2 MAE(Error medio absoluto): 0.479850
Modelo 3 MAE(Error medio absoluto): 0.466150
Modelo 4 MAE(Error medio absoluto): 0.450187
Modelo 5 MAE(Error medio absoluto): 0.481516
Modelo 6 MAE(Error medio absoluto): 0.492447
Modelo 7 MAE(Error medio absoluto): 0.459241
Modelo 8 MAE(Error medio absoluto): 0.454124
Modelo 9 MAE(Error medio absoluto): 0.453152
Modelo 10 MAE(Error medio absoluto): 0.458659
```

Comparándolo con el conjunto de datos entrenado para el avance 2 cuando se ejecuta el código para comparar los modelos, se obtuvieron los resultados que se ven en la figura, luego de una ejecución de 3.828 segundos:

```
Modelo 1 MAE(Error medio absoluto): 0.536000
Modelo 2 MAE(Error medio absoluto): 0.546133
Modelo 3 MAE(Error medio absoluto): 0.524333
Modelo 4 MAE(Error medio absoluto): 0.579585
Modelo 5 MAE(Error medio absoluto): 0.548294
Modelo 6 MAE(Error medio absoluto): 0.538313
Modelo 7 MAE(Error medio absoluto): 0.531671
Modelo 8 MAE(Error medio absoluto): 0.531943
Modelo 9 MAE(Error medio absoluto): 0.532261
Modelo 10 MAE(Error medio absoluto): 0.533513
```

Figure 3: Resultados con el conjunto de datos anterior (Avance 2)

El mejor puntaje en el Avance 2 vemos que correspondía al modelo 3, el cuál es: **0.5243333333333343**, por lo que ese era el mejor modelo hasta ese momento.

Ya con el nuevo *dataset*, entrenados y testeados nuevamente los modelos, el mejor luego de mostrar los resultados salió el número 4, con un error de **0.450187** o **45%**. Por lo que así se decidió el modelo a utilizar para este avance final.

4. Implementación

Para su implementación, se exportó el modelo con extensión *.joblib* para ser utilizado dentro de una aplicación web programada con *Python* y lenguajes de programación web. Dentro del código de *Python* se cargó el conjunto de datos y el nuevo modelo, además se instalaron las librerías necesarias para correr el programa de forma local, estas son *pandas, sckitlearn, joblib* y Flask que es un microframework para Python, que permite crear aplicaciones web de todo tipo de forma rápida. Esta aplicación se ejecutó con la versión de Python 3.9.6 64-bit. Como base para el CSS se utilizó *Bootstrap*, un framework de desarrollo front-end, junto al archivo HTML, se creó el diseño del entorno de aplicación web para introducir las características y mostrar los valores resultantes según el modelo cargado.

4.1. Resultados y Conclusión

Para mostrar un ejemplo de una predicción, vamos a tomar las características de la película "Mi Amigo Alexis" que se encuentra en la plataforma *Netflix*. En la Figura 4, se ve que se ingresaron las características según los datos obtenidos del conjunto de datos, que recordemos son extraídos de la página web IMDb. Y en la Figura 5, se puede apreciar el resultado de la predicción, luego de dar click en el botón "Predecir". La cual resultó una nota de **5.6** aproximado. El Rating de esta producción en el sitio IMDb es de **4.9**, por lo que existe una diferencia de 7 décimas con lo predecido por la aplicación creada.

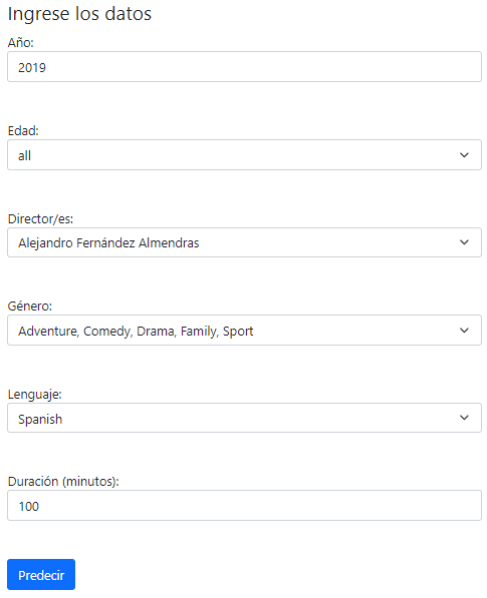

Figure 4: Datos de la película *Mi Amigo Alexis*

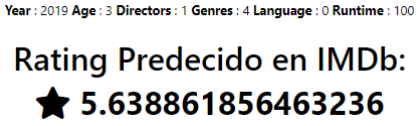

Figure 5: Resultados de predicción de la nota de la película *Mi Amigo Alexis* en IMDb

Según otras pruebas que se hicieron, se obtuvieron resultados algunos más cerca que otros a las notas originales que tienen estas películas en IMDb, aunque hay que decir que la gran mayoría de notas predecidas son distintas a los original. Como conclusión de este proyecto y experimento, se puede inferir que tiene sentido el error que tiene el modelo de **0.45**, con las notas obtenidas de la predicción en comparación de la original. Por lo que, el modelo aún se podría ir mejorando y perfeccionando para así obtener resultados más precisos. La idea de este proyecto era tener una manera de poder predecir el rating/nota de una película chilena, para así encontrar cuales son las características que hacen que la producción sea popular en nuestro país y otros países para así exportarla, mostrarla en salas de cine internacionales, que compre sus derechos una plataforma de Streaming para que pueda ser vista en muchos lugares del mundo o hasta que pueda ser ganadora o nominada a algún premio importante de la academia. Aún así, el proyecto en sí fue muy enriquecedor en conocimiento en técnicas de aprendizaje automático, modelos, conjuntos de datos y se aprendió a aplicar algoritmos que sirven para obtener predicciones respecto a alguna temática o problemática.

References

[1] diptaraj23, 2021. IMDb Rating Prediction from a data set of Movies. URL: https://www.kaggle.com/diptaraj23/imdb-rating-prediction-from-a-data-set-of-movies.

[2] IMDb, . Ratings, reviews, and where to watch the best movies & tv shows. https://www.imdb.com/. Accessed: 2021-8-3.

[3] JustWatch, . Justwatch - ver películas y series online. https://www.justwatch.com/cl. Accessed: 2021-8-3.

[4] López, A., 2016. ¿qué duración debe tener una película para ser considerada como 'largometraje'? https://blogs.20minutos.es/yaestaellistoquetodolosabe/que-duracion-debe-tener-una-pelicula-para-ser-considerada-como-largometraje/. Accessed: 2021-8-3.

[5] Molina, A., Nador, A., 2020. El streaming, un acompañante infaltable en la cuarentena - diplomado periodismo digital UC. http://dpd.comunicaciones.uc.cl/2020/el-streaming-un-acompanante-infaltable-en-la-cuarentena/. Accessed: 2021-8-3.

[6] MundoPelículas, 2021. El cine chileno en la cartelera de netflix y amazon prime. https://www.mundopeliculas.tv/2021/04/05/cine-chileno-online/. Accessed: 2021-8-3.