Modelo para identificar Rising Stars

Carlos Alfonso Alberto Salazar[1 [A01026175]], *Alejandro Méndez Godoy*[1 [A01783325]], *Diego Soria Bandín*[1 [A01747923]], *Carlos Eduardo Medina Flores*[1 [A01753742]], *Paulina Martínez García*[1 [A01748511]]

[1] Tecnológico de Monterrey, Campus Estado de México, México

Abstract. En este artículo se presenta un modelo de predicción para identificar a los *Rising Stars* en el ámbito académico. El modelo se construye utilizando un conjunto de datos que incluye diversas métricas por autor y publicación, y se basa en un enfoque de aprendizaje automático. Los resultados muestran que el modelo es capaz de identificar con un buen nivel de precisión a los autores que tienen el potencial de convertirse en líderes en su campo de estudio. Este enfoque tiene aplicaciones en la identificación temprana de talentos y en la asignación eficiente de recursos en instituciones académicas y de investigación.

Keywords: Rising Stars, Base de Datos, Métricas, Regresión, Predicción, Correlación, Precisión

1. INTRODUCCIÓN

La identificación de *Rising Stars* en el ámbito de la investigación es de gran importancia dentro de las universidades. Identificar a autores con potencial a ser líderes en su campo de trabajo o investigación, en un futuro es de gran utilidad para la asignación de recursos y oportunidades, además de impulsar el desarrollo tecnológico, nuevas ideas o ramas de investigación.

La principal problemática para considerar un autor como una *Rising Star* es la cantidad de publicaciones que hacen en un cierto periodo de tiempo. Leer cada uno de ellos tomaría demasiado tiempo para evaluarlos. Para esto, actualmente se implementan herramientas de la ciencia de datos para crear modelos de recomendación. Esta forma parte esencial para la solución de la problemática. Hemos recibido dos bases de datos. La primera contiene autores anónimos con un SCORPUS ID y una serie de datos para analizar, la segunda muestra las diferentes publicaciones hechas por los investigadores y sus características más útiles.

2. OBJETIVOS

Generar pronóstico del desempeño del investigador utilizando modelos de ciencia de datos.

- Por medio de la base de datos proporcionada de los autores se busca generar los mejores pronósticos con el uso de algunos de estos modelos de ciencia de datos, como lo son: modelos de clasificación, modelos de agrupación en clústeres, modelos de pronóstico, modelos de valores atípicos, modelo de serie temporal entre otros. Los cuales según el método y lo que más no convenga se elige uno o unos de ellos para sacar los tipos de análisis que correspondan para entender el desempeño de los investigadores.

Caracterizar el perfil de un investigador con alto potencial de impacto en años futuros (*Researcher rising stars*).

- Al obtener los resultados de los investigadores automáticamente se creará un perfil con ciertas características cada uno tendrá sus distintivas, unas no resaltan o serán descartadas, algunas coinciden y otras sobresaldrá. Dentro de estos perfiles se buscará cierto grupo el cual en un futuro generará un impacto enorme y de un crecimiento radical.

3. BUSINESS UNDERSTANDING

Para realizar un análisis y emitir una respuesta, principalmente se tomaron en cuenta atributos como la frecuencia con la que un autor publica, la relevancia de sus publicaciones y las citas que este obtiene, entre otras cosas. El primer contratiempo que se identificó, después de entender las métricas y su relevancia, fue que dependiendo del área de publicación del autor, los atributos anteriormente mencionados tienden a tener una variación significativa. Por ejemplo, una publicación de bioquímica suele tener muchas más citas que uno de humanidades o un autor de medicina suele publicar más que uno de ingeniería, por lo que en principio no se podrán comparar estos rubros entre estás áreas, mismamente un factor a tomar en cuenta es que hay varios autores que trabajan en diversos campos. [2]

Cada base de datos tiene datos que pueden ayudar a descartar y tomar en cuenta más a cierto autor, por ejemplo para poder tomar en cuenta la frecuencia de publicaciones variables como ;“scholarly output” la cual mide la cantidad de publicaciones del autor, “most recent

publication" con la publicación más reciente, "oldest publication" con su primera publicación registrada, esto para los casos en los cuales el autor sea muy nuevo o que no haya publicado desde hace mucho tiempo, etc. Estos puntos se verán más a fondo en el Data Understanding.

El estudio de las distintas habilidades y características de los estudiantes ha ido variando mucho durante los años debido a los distintos departamentos que existen al igual que el mercado al que le venden va cambiando. Dentro de estos análisis mostrados usaron cosas como un KPI, conocido también como indicador clave o medidor de desempeño o indicador clave de rendimiento. Algoritmos de agrupamiento utilizado en el aprendizaje no supervisado, los cuales sirven para categorizar datos no etiquetados, es decir, datos sin categorías o grupos definidos. Weight Cumulative Impact una fórmula que suma la relación entre sus citas con los años y su último año durante cierto periodo de tiempo

Sin embargo uno de los datos a analizar más interesantes son la importancia de la publicación, ya que para esto es necesario tomar en cuenta factores como el "h-index", este, como ya se mencionó anteriormente, mide la calidad de las publicaciones de un autor basado en la relevancia de sus publicaciones más exitosas y la cantidad de citas que ha obtenido. Cabe destacar que este es uno de los indicadores que más se toma en cuenta al momento de evaluar a un autor considerado *Rising Star.*

> "Asimismo, el conocido índice h recibió críticas ya que en algunos casos puede proporcionar información engañosa sobre la producción de un científico o puede estar sesgado de varias maneras, como el uso de autocitas, publicar en diferentes dominios de investigación, acceso abierto, la ventaja acumulada de investigadores más experimentados, etc." (Panagopoulos G., Tsatsaronis G. 2014)

Otra observación que se considera no éticamente correcta es puntuar con el mismo valor los trabajos individuales vs los trabajos en equipo al igual que trabajos que compiten y derivan de distintas áreas ya que existe la posibilidad que en unos tome menos tiempo por lo que puede tener un menor valor de aportación. Es por eso que se debe realizar la separación correctamente para así darle los valores exactamente a cada uno para hacer una comparación lo más justa posible.

Otro de los trabajos ya realizados más relevantes es de investigadores chinos apoyados por “National Natural Science Foundation of China”, donde explican que crearon un algoritmo de aprendizaje de clasificación de incremento de impacto o impact increment ranking learning (IIRL). El método usó más de 1.7 millones de autores con los cuales se demostró la eficiencia de este teniendo en promedio una mejora del 8% sobre los métodos de referencia. Aunque no indica precisamente el aumento de 8% sobre que porcentaje se aplicó, se habla de un incremento significativo. Al final nosotros no estaremos implementando las mismas variables, sin embargo, es una buena guía para identificar el tipo de variables más relevantes y cómo segmentar nuestros datos.[1]

Las métricas más reconocidas y de nuestro interés, son las siguientes:

- SJR: el Scimago Journal Rank establece la calidad de las publicaciones en relación de la cantidad de citas que se hace en un documento y toma en cuenta el prestigio de las revistas que lo citan, a mayor prestigio mejor SJR tendrán.
- SNIP: el SNIP compara las publicaciones con otras de su mismo campo, contabilizando la frecuencia con la que los autores citan otros documentos.
- CiteScore: esta métrica toma en cuenta el número de citas recibidas en 4 años dividido entre la cantidad de artículos publicados en ese mismo periodo.
- FWVI: el Field-Weight Views Impact es una métrica que compara la cantidad de vistas que un artículo tuvo en relación al promedio de los demás artículos similares.
- H-Index: el h-index es una relación entre la productividad de publicaciones y el número de citas recibidas
- FWCI: Field-Weighted Citation Impact es la relación entre las citas realmente recibidas y el promedio de citas esperadas en un campo determinado.

[2]

4. DATA UNDERSTANDING

Ya conociendo el objetivo y él área de estudio se analizaron los datos dados. Dos archivos de tipo csv. El primero con los datos de todos los autores, de dimensiones 9 x 84,000 con las métricas anteriormente mencionadas, en estas columnas destacan datos como el h-index, Field-Weighted Citation Impact y citations, sin embargo analizando un poco la base de datos se

pudo apreciar que solo había registro de sus publicaciones en la plataforma “Scholarly” lo cual hacía que datos como el h-index no cuadraran. Por ejemplo, habían personas con una supuesta publicación pero un h-index de 15, lo cual no tiene sentido. Por consiguiente, para arreglar este tipo de inconvenientes se analizó el segundo archivo csv, que de igual manera contiene una base de datos de publicaciones. Este último, sería el que se tomaría en cuenta para los datos de cada una de las publicaciones ya que al contar con todas las publicaciones y sus datos separados individualmente, sería más sencillo de manipular. La base de datos de autores se seguiría tomando en cuenta las variables que son más generales sin importar la plataforma como el h-index y Field Weighted Citation Impact.

Analizando los datos de las publicaciones, se encontró un *data frame* de 58 columnas por 84,000 filas, habían demasiadas variables por lo que lo primero que se hizo fue descartar las variables. Primeramente se descartaron las variables que tuvieran un exceso de datos faltantes (NAN), al tener tantos renglones en blanco, los datos no pueden ser manipulados para crear un modelo. Posteriormente se hizo un breve análisis de correlación para identificar variables que estuvieran altamente relacionadas y el por que de estas.

Correlations

	Number of Authors	Year	Volume	SNIP (publication year)	SNIP percentile (publication ye
Year	-0.023				
Volume	-0.004	-0.033			
SNIP (publication year)	0.061	0.059	-0.047		
SNIP percentile (publication ye	-0.100	-0.000	0.098	-0.590	
CiteScore (publication year)	0.112	0.077	-0.048	0.858	-0.537
CiteScore percentile (publicati	-0.126	-0.034	0.101	-0.484	0.845
SJR (publication year)	0.060	-0.029	-0.039	0.791	-0.496
SJR percentile (publication yea	-0.094	-0.026	0.054	-0.485	0.862
Field-Weighted View Impact	0.405	-0.014	-0.004	0.290	-0.094
Views	0.422	-0.094	-0.025	0.265	-0.140
Citations	0.099	-0.088	-0.016	0.244	-0.134
Field-Weighted Citation Impact	0.097	-0.005	-0.001	0.274	-0.121
Field-Citation Average	0.034	-0.409	-0.122	0.125	-0.241
Outputs in Top Citation Percent	-0.149	-0.088	0.099	-0.349	0.539
Field-Weighted Outputs in Top C	-0.157	-0.074	0.019	-0.323	0.445
Patent citations	-0.002	-0.036	-0.001	0.060	-0.028
Number of Institutions	0.924	-0.012	-0.009	0.175	-0.131
Topic Cluster number	-0.110	0.009	-0.014	-0.038	0.084
Topic number	-0.011	0.084	0.002	0.067	0.012
Topic Cluster Prominence Percen	0.095	0.049	0.006	0.103	-0.158
Topic Prominence Percentile	0.066	0.132	0.007	0.121	-0.181

Figura 1. Primera tabla de análisis de correlación de las variables de la base de datos de publicaciones.

	CiteScore (publication year)	CiteScore percentile (publicati	SJR (publication year)	SJR percentile (publication yea
Year				
Volume				
SNIP (publication year)				
SNIP percentile (publication ye				
CiteScore (publication year)				
CiteScore percentile (publicati	-0.545			
SJR (publication year)	0.891	-0.453		
SJR percentile (publication yea	-0.512	0.908	-0.479	
Field-Weighted View Impact	0.243	-0.099	0.175	-0.079
Views	0.241	-0.140	0.182	-0.110
Citations	0.248	-0.124	0.259	-0.121
Field-Weighted Citation Impact	0.244	-0.116	0.237	-0.115
Field-Citation Average	0.221	-0.226	0.212	-0.197
Outputs in Top Citation Percent	-0.432	0.552	-0.364	0.532
Field-Weighted Outputs in Top C	-0.363	0.455	-0.331	0.453
Patent citations	0.072	-0.025	0.088	-0.026
Number of Institutions	0.209	-0.153	0.143	-0.123
Topic Cluster number	-0.102	0.119	-0.087	0.105
Topic number	0.042	0.022	0.022	0.030
Topic Cluster Prominence Percen	0.199	-0.204	0.147	-0.170
Topic Prominence Percentile	0.188	-0.212	0.138	-0.193

Figura 2. Segunda tabla de análisis de correlación de las variables de la base de datos de publicaciones.

	Field-Weighted View Impact	Views	Citations	Field-Weighted Citation Impact	Field-Citation Average
Year					
Volume					
SNIP (publication year)					
SNIP percentile (publication ye					
CiteScore (publication year)					
CiteScore percentile (publicati					
SJR (publication year)					
SJR percentile (publication yea					
Field-Weighted View Impact					
Views	0.924				
Citations	0.454	0.501			
Field-Weighted Citation Impact	0.467	0.455	0.828		
Field-Citation Average	0.017	0.156	0.165	0.015	
Outputs in Top Citation Percent	-0.150	-0.222	-0.227	-0.232	-0.343
Field-Weighted Outputs in Top C	-0.167	-0.205	-0.225	-0.275	-0.068
Patent citations	0.031	0.044	0.104	0.056	0.048
Number of Institutions	0.625	0.613	0.219	0.222	0.046
Topic Cluster number	-0.034	-0.047	-0.031	-0.016	-0.084
Topic number	0.075	0.049	0.049	0.078	-0.053
Topic Cluster Prominence Percen	0.068	0.098	0.062	0.050	0.164
Topic Prominence Percentile	0.064	0.087	0.064	0.062	0.120

Figura 3. Tercera tabla de análisis de correlación de las variables de la base de datos de publicaciones.

	Outputs in Top Citation Percent	Field-Weighted Outputs in Top C	Patent citations	Number of Institutions	Topic Cluster number
Year					
Volume					
SNIP (publication year)					
SNIP percentile (publication ye					
CiteScore (publication year)					
CiteScore percentile (publicati					
SJR (publication year)					
SJR percentile (publication yea					
Field-Weighted View Impact					
Views					
Citations					
Field-Weighted Citation Impact					
Field-Citation Average					
Outputs in Top Citation Percent					
Field-Weighted Outputs in Top C	0.884				
Patent citations	-0.042	-0.038			
Number of Institutions	-0.187	-0.193	0.004		
Topic Cluster number	0.103	0.074	0.001	-0.115	
Topic number	-0.039	-0.051	-0.001	0.014	0.283
Topic Cluster Prominence Percen	-0.242	-0.169	0.010	0.106	-0.742
Topic Prominence Percentile	-0.275	-0.213	0.014	0.078	-0.264

Figura 4. Cuarta tabla de análisis de correlación de las variables de la base de datos de publicaciones.

	Topic number	Topic Cluster Prominence Percen
Year		
Volume		
SNIP (publication year)		
SNIP percentile (publication ye		
CiteScore (publication year)		
CiteScore percentile (publicati		
SJR (publication year)		
SJR percentile (publication yea		
Field-Weighted View Impact		
Views		
Citations		
Field-Weighted Citation Impact		
Field-Citation Average		
Outputs in Top Citation Percent		
Field-Weighted Outputs in Top C		
Patent citations		
Number of Institutions		
Topic Cluster number		
Topic number		
Topic Cluster Prominence Percen	0.040	
Topic Prominence Percentile	-0.035	0.501

Figura 5. Quinta tabla de análisis de correlación de las variables de la base de datos de publicaciones.

Correlations

	Scholarly Output	Most recent publication	Citations	Citations per Publication	Field-Weighted Citation Impact
Most recent publication	0.179				
Citations	0.834	0.051			
Citations per Publication	0.039	-0.095	0.171		
Field-Weighted Citation Impact	0.047	0.007	0.148	0.788	
h-index	0.438	0.032	0.366	0.157	0.160
Output in Top 10% Citation Perc	0.864	0.056	0.925	0.066	0.075
Oldest publication (since 1996)	-0.257	0.283	-0.107	-0.077	-0.007

	h-index	Output in Top 10% Citation Perc
Most recent publication		
Citations		
Citations per Publication		
Field-Weighted Citation Impact		
h-index		
Output in Top 10% Citation Perc	0.371	
Oldest publication (since 1996)	-0.338	-0.114

Figura 6. Tabla de correlación de las variables de la base de datos de autores.

Después de realizar el análisis de correlación se encontraron datos que podían ser omitidos como por ejemplo la fecha y el año, de las cuales se decidió descartar la fecha porque el análisis se decidió hacer de manera global. De las demás variables había relaciones entre las variables de SNIP, citas y SJR, cabe mencionar que estas variables tienen varias "sub-variables" esto quiere decir que por ejemplo SNIP tiene "SNIP (publication year)", "SNIP (percentile)" y de ese tipo. Analizando un poco más la base de datos hay varias variables del tema o área de publicación, unas especificando más que las otras.

Para hacer un modelo simple había que destacar algunas variables sobre otras y descartar algunas, un ejemplo de esto es que no se tomaría en cuenta el tema de la publicación si no el área para que no se dividiera tanto la base de datos para hacer algunos análisis.

Como se mencionó anteriormente cada área tenía su impacto, media de citas, cantidad de publicaciones, entre otras. Por lo que se separó cada área de publicación y se hizo un análisis de media por cada área de interés.

Tabla I. Medias de la métricas por área de estudio.

.	Number of Authors	SNIP percentile (publication year) *	Citations	FWCI
Engineering & Technology	5.01517413	27.58507463	6.83737562	0.73774689
Arts & Humanities	1.83293173	51.42971888	1.05220884	0.45791165
Life Sciences & Medicine	6.09207082	32.4638953	5.73448807	0.67183449
Natural Sciences	5.69878429	28.73410547	6.83505832	0.76060375
Social Sciences	3.46881886	32.31974438	4.22730278	0.64181578

5. DATA PREPARATION

Para preparar los datos los pasos a seguir fueron los siguientes: seleccionar las columnas de interés, omitir las columnas con una cantidad considerable de valores faltantes (NAN). En este caso se consideró una muestra con 500 observaciones. El primer análisis aplicado a estos datos nos ayudó a darnos cuenta que el comportamiento de libros y journals no era el mismo., Esto se tomó en cuenta para el momento de hacer un modelo, este sea congruente con ambos, en otras palabras, que el modelo incluya datos tanto de libros como de journals por lo que se le aplicó un *drop* a las columnas con las variables que no tuvieran atributos para ambos casos.

Después de realizar esta limpieza, se terminó con un *data frame* con menos de 20 variables para trabajar y hacer más fácil de entender la base de datos. En esta se encontró que por cada publicación generalmente habían varios autores y estaban identificados por el mismo ID que el de la base de datos de autores. Dada esta información, se hizo una separación por cada autor con todos los IDs de sus publicaciones para así obtener un nuevo *data frame* que posteriormente fuera posible unir por medio de las variables de interés de los libros y de los autores.

Seguido de este paso, se realizó una gráfica del comportamiento de todas las medias por año para tener una idea de cómo se comportan ciertas variables con respecto a otras. Con este objetivo en mente, se graficaron las métricas más importantes anteriormente decididas con los

siguientes resultados en la Figura 7.

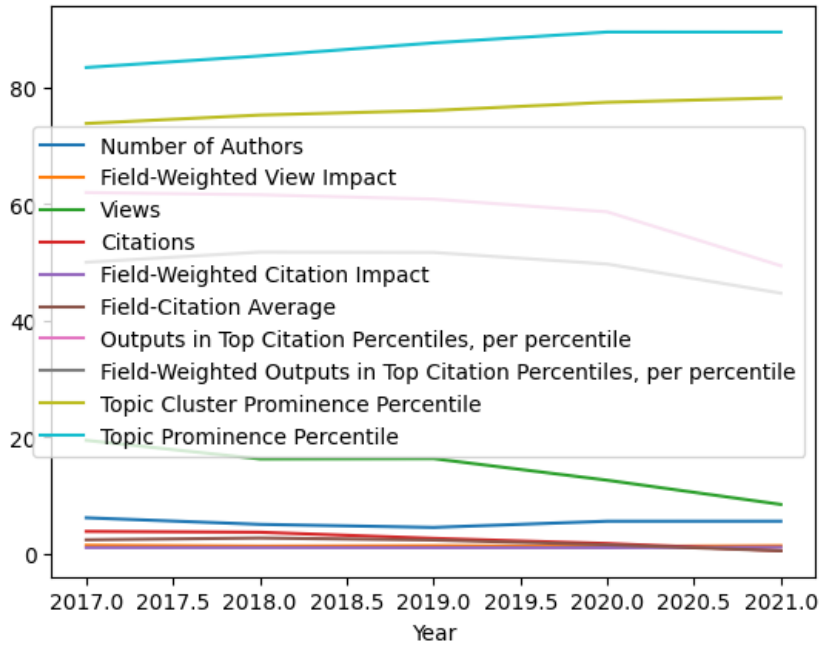

Figura 7. Gráfico del comportamiento de las diferentes métricas a lo largo del tiempo.

En la Figura 7 se puede apreciar que hay ciertas tendencias entre el número de autores y el crecimiento de ciertas variables, además se puede apreciar que hay variables que son considerablemente más altas que otras así que para compararlas habría que hacer un tipo de normalización.

6. CONSTRUCCIÓN DEL MODELO

Tomando en cuenta las características para considerar a un autor una *Rising Star*, como lo es ser una persona recurrente en sus publicaciones ya que puede haber autores que hayan tenido un primer gran éxito y no tener mucho más o que dejen de publicar conforme el tiempo, por lo que evaluamos el tiempo como un factor importante y será otra variable para tomarse en cuenta, graficando la cantidad de publicaciones conforme el tiempo se puede apreciar cómo a lo largo del tiempo se han ido publicando menos artículos:

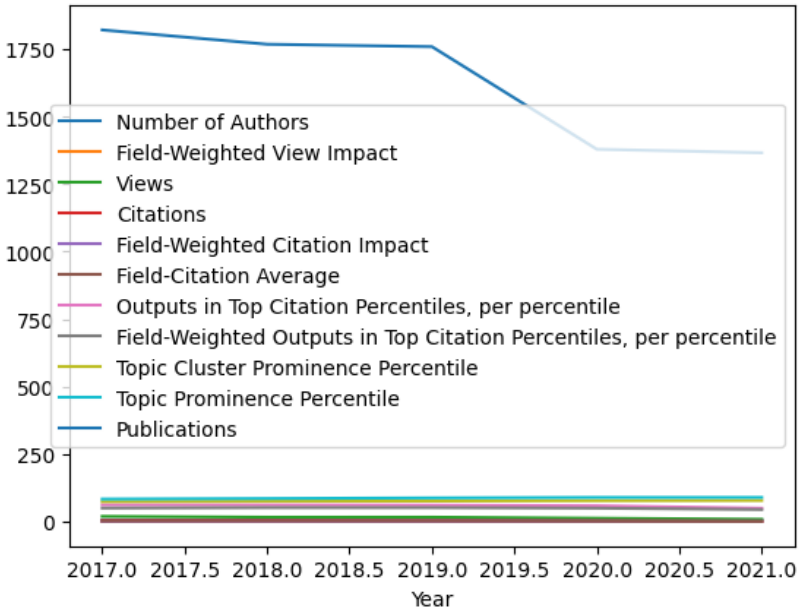

Figura 8. Gráfico de la métricas de publicaciones a lo largo de los años.

Apreciado esto se crearán dos variables y una columna, una columna que diga el área de especialidad del autor que se calculará por medio de la moda, en cuanto las variables será una que ayuda a saber cómo va publicando con respecto al tiempo, si más o menos y otra que indique su rendimiento independientemente de con cuantos autores trabaje.

Para la primer variable se utilizará un modelo de regresión lineal para calcular la pendiente que tiene el autor conforme al tiempo y artículos escritos, para este modelo se utilizará el método “OLS” de la librería de “statsmodel”, se utilizará este método ya que no tiene que ser tan preciso pero sí debe de ser muy rápido y que no gaste muchos recursos para poder aplicarlo a todos los autores.

En cuanto a la variable que diga cómo se comporta el autor publicando sin importar con cuantos autores publique y sin importar el área de estudio, para esto se tomarán todas las variables de cada publicación, se separará y se hará una media por variable de cada año, posteriormente como se mencionó anteriormente se debe normalizar la variable y hacer que el área deje de ser un factor de desigualdad. Para lograr por medio de las medias que habían sido obtenidas anteriormente de cada área (medias de columnas seleccionadas por área) se escogerá

un campo en específico y se normalizará todo con base en él. El área elegida fue medicina ya que es la que más datos tenía, para normalizar se dividirá el promedio de medicina de cada variable entre el promedio del área de la que es el autor, todo sobre el número de autores que a su vez estará dividido entre el número de autores del área del autor dando como resultado una fórmula como la siguiente:

$$y = \frac{\frac{\bar{FWCI}[Med]}{\bar{FWCI}[Área]}FWCI + \frac{\bar{SNIP}[Med]}{\bar{SNIP}[Área]}SNIP + \frac{\bar{VIEWS}[Med]}{\bar{VIEWS}[Área]}VIEWS + \ldots}{\frac{no.\ Autor}{\bar{No.\ Author}[Área]}} \quad (1)$$

Ya con las dos nuevas variables creadas y la columna estas fueron enviadas y juntadas con la base de datos del autor, en esta base de datos habían las siguientes columnas: El ID del autor, FWCI (general del autor), h-index, número de publicaciones, pendiente de tendencia de publicación, área de especialización y el nuevo parámetro. De estás variables solo 3 serán para la evaluación del autor (FWCI (general), h-index y la nueva variable) las demás serán utilizadas para analizar un poco los resultados e identificar los *Rising Star* resultantes.

El proceso de evaluación de estás 3 variables fue hacer un promedio de las 3 variables elegidas para obtener una variable con un autor que mida su rendimiento general e individual. Ya con esta variable se hizo un filtro de que los *Rising Star* deberían estar arriba del promedio en cada variable de importancia (h-index, FWCI y nueva variable) esto para que si en caso de que una persona tenga un gran h-index y FWCI porque ha trabajado en grandes proyectos con muchos colaboradores pero con pocos artículos propios (nueva variable) no le afecte si es que destaca de gran manera en los demás campos.

Además se elaboró un modelo para poder predecir la nueva variable “New Param” por medio de los datos generales; 'h-index','Field-Weighted Citation Impact', 'no.Pubs', 'Pend Pubs Rate’. Para esto se hizo uso de un método de regresión con el método de “Random Forest Regressor” de la librería “sklearn” con un tamaño de test del 30%, con una profundidad máxima de 10 y un máximo de nodos de 50, esto porque se utilizaron diversos métodos y diversas medidas y modelo y esos parámetros son los que regresaron valores más altos.

Finalmente el análisis de regresión arrojó un train score: 0.84 y un test score: 0.65. El train/test es un método el cual nos ayuda a medir la precisión de nuestro modelo. Este divide los datos en distintos porcentajes, un porcentaje para el training y lo que sobre para el test. En este modelo se usó el 70% para el training y el 30% restante para el test. El training es conocido como la parte de la creación y el testing más para la precisión de nuestro modelo. En cuanto a la interpretación de estos datos, en el entrenamiento de datos tenemos un puntaje bastante alto que nos indica que nuestro modelo está bien entrenado y sirve para lo que necesitamos. En el caso del test score no es excelente pero si es bueno, eso debido a que faltaría ajustar los datos o en todo caso agregar más métricas lo que nos indica que existe una buena área de oportunidad. Una posible solución sería hacer de nuevo todo el proceso desde otro punto de vista.

7. RESULTADOS

Tabla II. Top 10 autores candidatos a *Rising Star* con nuestro modelo.

Scopus Author Ids	Field-Weighted Citation Impact	h-index	no.Pubs	Pend PubsRate	Area	New Param (mean)	Predicciones	Mean
20836017400	14.11	87	9	-0.5	Natural Sciences	16.6976	23.437	39.269
24482701000	8.31	35	5	0.07	Engineering & Technology	60.3888	52.229	34.566
55252500400	5.35	147	8	1.5	Natural Sciences	50.9697	49.646	67.773
55316592700	37.89	48	9	0.0	Engineering & Technology	292.0223	217.484	125.971
55316592700	37.89	48	6	0.0	Engineering & Technology	292.0223	217.619	125.971
55623093900	7.52	56	19	0.3	Engineering & Technology	100.7152	86.397	54.745
56208983400	9.46	37	8	1.0	Natural Sciences	93.3623	72.318	46.607
7003572421	96.39	85	6	0.43	Life Sciences & Medicine	4.2169	102.636	61.869
7202378672	7.38	88	7	0.5	Life Sciences & Medicine	26.9613	37.898	40.780
55212570900	18.86	31	6	0.5	Natural Sciences	229.1549	190.051	93.005

En la Tabla II se capturaron los resultados de nuestros 10 candidatos a *Rising Stars*. Elegimos las métricas como el FWCI, el h-index, el número de publicaciones en los últimos años, y nuestro nuevo parámetro. Como se puede ver en la Tabla II, hicimos que nuestro nuevo parámetro tuviera una relación proporcional a el h-index y el FWCI para poder ponderar de manera equitativa las características reales que un *Rising Star* debe de tener para considerarse uno. De esta manera nuestros resultados tienen el mismo peso en un investigador de un campo menos citado como artes y un investigador de campos con muchas citas y visualizaciones como lo es ingeniería y medicina, así podemos obtener resultados más óptimos de las características reales que un *Rising Star* debe tener. Algo que destacar de nuestra tabla es el new param comparado con las predicciones de este mismo, son números muy parecidos, esto demuestra la

eficacia del modelo con base en los resultados. Por último mencionar que el que tiene menor puntuación en nuestro parámetro, el de 4.21 es *Rising Star* gracias a que tiene un mayor FWCI y h-index muy alto. Esto se debe a que colabora bastante y dentro de estas colaboraciones su impacto es importante además de que el promedio de sus citas es elevado.

Referencias

[1] Zhang, C., Liu, C., Yu, L., Zhang, Z.-K., & Zhou, T. (2016). Identifying the academic rising stars. Nature, 535(7610), 298–301. https://doi.org/10.1038/nature18667

[2] Elsevier. (2021). Research Metrics Guidebook. Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0020/1083948/Research-Metrics-Guidebook.pdf

[3] Daud, A., Abbas, F., Amjad, T., Alshdadi, A. A., & Alowibdi, J. S. (2018). Finding rising stars through hot topics detection. Scientometrics, 116(1), 153-176. https://doi.org/10.1007/s11192-018-2677-1

[4] Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2014). Detecting rising stars in dynamic collaborative networks. Journal of Informetrics, 8(1), 117-133. https://doi.org/10.1016/j.joi.2013.10.003