Identificación de jóvenes talento utilizando algoritmos de machine learning.

Carlos Daniel Ruvalcaba Serrano.
18/08/2019

Introducción.

Las Olimpiadas del conocimiento son eventos que se llevan año con año, las cuales tratan de distintas disciplinas cada una: Biología, Química, Informática, Matemáticas, Física, entre otras disciplinas. Los estudiantes más sobresalientes son aquellos que a temprana edad comenzaron a incursionar en estas disciplinas y se prepararon hasta realizarse. Más de la mitad de los estudiantes que participan en este tipo de competencias realizan estudios postdoctorales en áreas técnicas de STEM una vez que terminan su carrera, por lo que su participación y preparación son importantes para el desarrollo de Zacatecas y del país. La identificación temprana de estudiantes con aptitudes necesarias para competir otorga una ventaja competitiva en las competencias. Fomentar y cultivar competidores desde temprana edad es una tarea difícil en cualquier disciplina. El resultado del proyecto es una herramienta en forma de prueba que ayuda en la identificación de jóvenes talento a temprana edad, que, dadas las características del estudiante, otorga una probabilidad de ser un joven talento.

Metodología.

El problema de la predicción de talento en jóvenes a temprana edad se atacó utilizando análisis de datos y machine learning. Se realizó una encuesta (la cual se encuentra anexa a este documento) a 79 estudiantes considerados *no talentos en matemáticas* y a 20 estudiantes, competidores estatales de la Olimpiada Mexicana de Matemáticas, con el fin de estudiar las características que definen a los jóvenes talentosos en matemáticas más allá del coeficiente intelectual.

En resumen, el proceso para estudiar a los estudiantes consistió en las siguientes fases.

1- Investigación de campo: en esta primera fase se investigó acerca de los estudios y soluciones alternas que otras personas han propuesto para resolver la problemática. El objetivo de esta fase fue reconocer los atributos que se han utilizado en estudios previos y seleccionar aquellos factibles para la población zacatecana. Finalmente, se obtuvo un cuestionario con el que fue posible recolectar información.
2- Recolección de la materia prima de estudio: esta fase consistió en la recolección de información de parte de los estudiantes talentos y no talentos mediante la aplicación del cuestionario. Se visitaron distintas escuelas de nivel preparatoria, la población objetivo fueron estudiantes en primer y segundo año de preparatoria.
3- Limpieza de los datos: en esta fase se formalizó el resultado de la recolección de datos, creando un dataset que expresara cada uno de los atributos o características estudiados de forma que una computadora fuera capaz de interpretarlos.
4- Análisis estadístico de los datos: en esta fase se realizó un análisis estadístico de los datos utilizando métodos computarizados. Se calcularon las medidas de tendencia central de los datos, además de que se obtuvieron las correlaciones importantes entre los distintos atributos de los datos. El resultado fue una serie de observaciones de las que se obtuvieron recomendaciones las cuales se expresan más adelante en este documento.
5- Generación del modelo clasificatorio: en esta fase se generaron distintos modelos para encontrar el que mejor describiera los datos. Se crearon árboles de decisión utilizando el algoritmo C 5.0 además de modelos de regresión logística. Esta decisión se llevó a cabo dado que estos modelos permiten saber la razón del resultado o cómo se llega a este. Para discriminar las características que modelan mejor a un joven talento se utilizaron árboles de decisión, para después crear un modelo de regresión logística con los atributos obtenidos y de esta forma, formar un cuestionario reducido que contiene únicamente las preguntas necesarias de los atributos seleccionados.
6- Exposición de los resultados: finalmente, una vez que el modelo seleccionado fue entrenado, se creó una página web la cual contiene un cuestionario con el que es posible crear una predicción en base a lo que se responda en este. Sin embargo, por cuestiones de tiempo, no fue posible montar la página en un servidor, pero se puede encontrar el código fuente y los archivos del estudio aquí https://github.com/Charly52830/JovenesTalento

Resultados.

Se entrenó un modelo de machine learning que dada una serie de atributos seleccionados, predice la probabilidad de que un estudiante sea joven talento en

matemáticas. De esta forma, se pretende demostrar que a través de esta metodología es posible contribuir a la identificación de talento a temprana edad.

Los atributos que se utilizaron en este estudio fueron:

1- **Género:** atributo que describe si el estudiante es hombre o mujer.
2- **Creencia religiosa:** atributo que describe si el estudiante practica o no alguna religión.
3- **Actividad lúdica:** atributo que describe si el estudiante practica o se desenvuelve en actividades lúdicas como deportes o artes.
4- **Número de libros:** atributo que describe la cantidad de libros que el estudiante tiene en su casa.
5- **Número de hijo:** atributo que describe el número de hijo que es de sus padres.
6- **Número de hermanos:** atributo que describe la cantidad de hermanos que tiene.
7- **Trabaja:** Atributo que describe si el estudiante trabaja o no.
8- **Tipo localidad:** atributo que describe si el estudiante radica en una localidad urbana o rural.
9- **Población:** atributo que describe la cantidad de personas que radican en la localidad.
10- **Nivel de marginación de la localidad.**
11- **Tipo de escuela:** atributo que describe si el estudiante estudia en una escuela pública o privada.
12- **Turno:** atributo que describe si el estudiante estudia en la tarde o en la mañana.
13- **Promedio:** atributo que describe el promedio general actual del estudiante.
14- **Grado académico.**
15- **Unión de padres:** atributo que describe si los padres biológicos del estudiante se encuentran unidos o separados.
16- **Nivel de estudio del padre.**
17- **Nivel de estudio de la madre.**
18- **Edad del padre.**
19- **Edad de la madre.**
20- **Profesión del padre:** atributo que describe el tipo de actividad económica que realiza el padre.
21- **Profesión de la madre:** atributo que describe el tipo de actividad económica que realiza la madre.
22- **Ingreso mensual:** atributo que describe el ingreso económico mensual que percibe la familia del estudiante.
23- **Apoyo emocional:** atributo que describe de forma cuantificable el apoyo emocional que recibe el estudiante por parte de su familia.
24- **Apoyo académico:** atributo que describe de forma cuantificable el apoyo que percibe por parte de sus profesores en la escuela.

25- **Presión familiar:** atributo que describe de forma cuantificable la presión que percibe por parte de sus padres para terminar la escuela.
26- **Ambiente escolar:** Atributo que describe de forma cuantificable la disposición de los compañeros del estudiante por el aprendizaje escolar.
27- **Atención de los padres en la escuela:** atributo que describe de forma cuantificable la atención que los padres del estudiante prestan en las actividades escolares de sus hijos.
28- **Ambiente hostil:** atributo que describe de forma cuantificable la hostilidad que percibe el estudiante en su escuela.
29- **Relación esfuerzo-éxito:** atributo que describe en una escala del 1 al 5 cuánto cree el estudiante que el esfuerzo define al éxito.
30- **Grado de felicidad:** atributo que describe en una escala del 1 al 5 qué tan feliz es el estudiante.

Los atributos seleccionados, resultado de la discriminación realizada por los árboles de decisión, son:

1- Unión de padres.
2- Presión familiar.
3- Ambiente escolar.
4- Promedio.
5- Género.
6- Libros.
7- Apoyo académico.
8- Edad padre.
9- Relación esfuerzo-éxito.

Utilizando estos atributos se crea un modelo predictivo de regresión logística, ignorando por completo el resto de los atributos que no se encuentran en esta lista.

El mejor modelo que se encontró fue utilizando estos atributos, no obstante, el desempeño de este no es el óptimo. A pesar de que tiene medidas estadísticas que demuestran el valor que otorga el modelo, tiene una sensitividad de 0.667 y una especificidad de 0.95, es decir, que del 100% de los estudiantes que predice como talento, solo el 66% de ellos son predichos correctamente. Esto hace que en la práctica el modelo sea muy ineficiente y por lo tanto, no se debe poner en producción.

Un análisis de curva de aprendizaje reveló que el síntoma que sufre el modelo es un desbalance del número de jóvenes talento con respecto a los que no lo son, además de un dataset con muy pocos datos que no permiten una fase de entrenamiento óptima.

¿Cómo puede mejorar?

La aplicación de machine learning para predecir potenciales competidores puede ser utilizada en cualquier disciplina, siempre y cuando las preguntas del cuestionario o las formas de obtener los datos sean las correctas. Para este estudio se tiene una cantidad de 99 observaciones entre los jóvenes talento y los no talento, es una cantidad muy pequeña por lo que es necesario recaudar más información de ambas partes. Para aumentar la credibilidad del resultado y ampliar la posibilidad de nuevos modelos, es necesario tener una mayor diversidad de los estudiantes respecto a aspectos rurales y escuelas privadas.

Las siguientes observaciones se basan en el resultado de la aplicación de medidas estadísticas tales como medidas de tendencia central y correlación de los atributos del dataset.

El análisis sugiere que entre la población de estudiantes talento, aquellos que perciben una mayor presión familiar tienden a obtener calificaciones menores a aquellos que reciben poca presión familiar, situación contraria, pero explicada con menor fuerza, en los jóvenes normales, quienes obtienen calificaciones más altas cuanto más se les presione. Asumiendo que los jóvenes talento son más autodidactas que el resto de la población juvenil, podemos concluir que entre más se presione a una mente autodidacta, peores resultados se obtendrán, pero si aquella mente no es autodidacta entonces es mejor que reciba algo de presión.

En la población en general, los estudiantes que perciben una mayor atención por parte de sus profesores tienden a obtener mejores calificaciones. Este es un aspecto muy interesante, ya que involucra una de las principales diferencias entre la educación pública y la educación privada. En una escuela privada, en la que la cantidad de estudiantes es considerablemente menor respecto a los que hay en una escuela pública, los estudiantes recibirán más apoyo por parte de sus profesores debido a que es más fácil tratar con grupos pequeños. Esto puede explicar por ejemplo, la razón de porqué las escuelas privadas tienden a tener un promedio más alto que las escuelas públicas cuando son evaluadas por la SEP.

En los jóvenes talento, las mujeres perciben un ambiente más hostil que los hombres. En la población general, las mujeres obtienen calificaciones más altas que los hombres, aquellos que perciben mayor apoyo emocional por parte de sus padres, se consideran ellos mismos más felices. Quienes trabajan, obtienen calificaciones más bajas.

La correlación del promedio con la clase, demuestra que este atributo siempre será un factor determinante en la detección de un joven talento (en lo que a olimpiadas del conocimiento se refiere), además, una observación importante es que de los 20 jóvenes talento, solo uno tiene a sus padres separados, es decir, 95% de los jóvenes talento tiene a sus padres unidos, mientras que en el resto de los jóvenes solo el 72% de ellos tiene a sus padres unidos. La mayoría de los atributos relacionados

con los padres, tales como profesión, nivel de estudios e ingreso económico demostraron no tener importancia en la decisión de ser o no un joven talento. Un joven talento puede venir de cualquier familia.

Conclusión.

La detección de jóvenes talento a temprana edad puede ser abordada utilizando algoritmos de machine learning, sin embargo, una deficiente etapa de recolección de datos puede conducirnos a modelos no óptimos. Existen patrones entre los jóvenes talento que nos permiten entender cómo son, cómo actúan y cómo perciben el ambiente. La aplicación de prácticas innovadoras en la resolución de problemas educativos supone un avance en el progreso social, una ventaja competitiva en las olimpiadas y nuevas oportunidades para la población joven estudiantil.

Reconocimientos.

Expreso mi agradecimiento al equipo de la Secretaría del Zacatecano Migrante y al equipo del Instituto de la Juventud por haberme dado la oportunidad de participar en este enriquecedor proyecto.

Agradezco especialmente al Dr. José Ramón Pasillas por haberme guiado en el proceso de la aplicación del machine learning, así como al Dr. Carlos Erick Galván por haberme guiado en la evaluación de los resultados obtenidos del modelo.