Universidad Rey Juan Carlos

Trabajo Fin de Máster

Trabajo Fin de Máster

Máster Universitario En Formación del Profesorado de Ed.Secundaria, Bachillerato, FP e Idiomas

Especialidad en Informática y Tecnología

Curso 2018-2019

AUTOR: Abel de Andrés Gómez
CENTRO: Escuela Politécnica Giner

AGRADECIMIENTOS

Agrademos a...

RESUMEN

Extensión máxima de una página

SUMMARY

Extensión máxima de una página

Índice

Índice de figuras

Índice de cuadros

1. INTRODUCCIÓN

En los últimos años, gracias al gran desarrollo tecnológico que se ha vivido tanto a nivel de computo (mejorando la eficiencia y el uso de los recursos disponibles) como a nivel de transmisión de datos (mejorando las comunicaciones), ha permitido a las organizaciones el almacenamiento de una gran cantidad de información.

Para comprender mejor este gran volumen de información, es necesario utilizar métodos, técnicas, herramientas además de personas con conocimientos (formando todas esta un vínculo estrecho) que permita y ayude a explotar, investigar, predecir y obtener información relevante para tomar decisiones de forma adecuada.

La organización educativa no ha quedado ajena a estas necesidades de una mejor comprensión de los datos. En este sentido, una unidad de Educación Secundaria Obligatoria de la Consejería de Educación de la Comunidad de Madrid ha planteado un problema.

El problema con el que se enfrentan cada día es la planificación de grupos para el siguiente curso. Esta planificación es la base para poder decidir donde se escolariza cada alumno y como se va a repartir la plantilla del profesorado según sus especialidades. Conocer el número de grupos permite, por tanto, un óptimo reparto de la plantilla de docentes y de recursos. De esta forma, además, se evita la existencia de grupos sobrepoblados.

Desde esta unidad, han informado sobre aspectos sobre los que trabajan para poder realizar una predicción acerca del número de grupos para curso venidero.

Estos aspectos son:

1. Escolaridad del curso actual.
    - Número de alumnos y grupos de un determinado centro.
    - El número de alumnos por aula (también conocido como ratio).
    - Matriculación de nuevos alumnos.
        - Principalmente alumnos que superan el nivel de 6º de primaria y pasan a 1º de ESO.
2. Bilingüismo del centro. Muchos alumnos optan por centros bilingües para su mejor formación, por lo que estos centros suelen tener más demanda de alumnos.
3. Posibilidad de creación de nuevas zonas urbanas cerca del centro.

4. Posibilidad de apertura o cierre de centros educativos. El cierre de por ejemplo, de un centro privado, provocara una mayor tasa de matriculación de los centros contiguos.

5. Porcentaje de aprobados. Los alumnos que están ya matriculados tienen prioridad sobre los nuevos alumnos, por lo tanto, si existe una alta tasa de suspensos, quedan pocas plazas de admisión de nueva matricula.

6. El número y la aparición de nuevas enseñanzas. La oferta de nuevas enseñanzas atraerá a nuevos alumnos al centro, incrementando así el número de matriculaciones.

La unidad actualmente utiliza herramientas manuales para conseguir conocer el número de grupos, indicando que es un trabajo mecánico y con herramientas obsoletas, evitando la posibilidad de inclusión de nuevas variables o factores que impliquen nuevos resultados.

Propone dar una solución al problema actual mediante el uso de herramientas y métodos que automaticen dichas tareas y proponga, además, nuevas variables o factores que puedan influir en la toma de decisión.

2. JUSTIFICACIÓN TEÓRICA

En primer lugar, esta investigación se realiza con el propósito de aportar conocimiento existente sobre la importancia de determinadas variables educativas y su relevancia en la predicción en la planificación y la gestión educativa.

En la actualidad existen numerosos informes acerca del uso de la ciencia de datos y sus técnicas en el ámbito educativo. Para la realización de este TFM se han analizado distintas publicaciones de la base de datos científica de ScienceDirect. Para realizar la búsqueda se han utilizado las siguientes palabra claves: educational, data y mining. Se debe recordar que el éxito de la búsqueda depende de estas palabras claves.

De la búsqueda anterior se han obtenido 160 artículos. Posteriormente se han seleccionado aquellos de los últimos 4 años (2016,2017,2018 y 2019). De esta forma obtenemos resultados actuales. Filtrando por fechas, hemos conseguido reducir los resultados a 73 artículos. Se ha realizado una observación sobre los artículos obtenidos y se ha comprobado la existencia de artículos que no resultaban útiles en esta investigación. Por tanto, se ha realizado otra búsqueda utilizando las claves anteriores y añadiendo la clave "prediction". Esta vez, se han obtenido 26 resultados. De todos los resultados obtenido, se han seleccionado 15 artículos que se consideran útiles y que servirán de ayuda.

Debemos destacar que la mayoría de los resultados obtenidos tratan de artículos centrados en la predicción de los resultados académicos del alumnos teniendo en cuenta ciertos factores internos (como las propias calificaciones a lo largo del curso) y externos (como factores etnograficos, edad, situación económica familiar, etc).

En este sentido es interesante realizar un análisis de dichos artículos, puesto que en primer lugar se deberá tener en cuenta cuales son las metodologías de la ciencia de datos que se están utilizando. Ademas, en segundo lugar, se debe tener en cuenta los modelos que se utilizan para predecir variables de carácter educativo.

Una vez que se han analizado los artículos, se ha decido realizar otra búsqueda en ScienceDirect, teniendo en cuenta las palabras claves: "gis", "data", "miningz .[ed]ucation". De esta búsqueda se han obtenido 22 resultados. De estos resultados se ha analizado un único articulo que se considera importante. El motivo de esta búsqueda es intentar encontrar artículos centrados en GIS (Sistemas de Información Geográfica). Los sistemas de información geográfica, como su propio nombre indica, se utilizan para referenciar datos en el espacio.

En el artículo de prensa de Fernandes y cols. (2019), se muestra el uso de técni-

cas como los métodos de clasificación y el algoritmo predictivo de GBM (Gradient Boosting Model) con el objetivo de obtener aquellas variables en el entorno del alumno, que hace que este obtenga mejores o peores resultados escolares. Este estudio, además, tiene el objetivo de aportar información útil para los representantes políticos en el ámbito educativo, el consejo escolar y los profesores con el objetivo de que estos puedan realizar políticas públicas, materiales didácticos y trabajo social para beneficiar a los estudiantes.

Los datos escolares a estudiar proceden de alumnos de colegios de un Distrito Federal de Brasil durante el 2015 y el 2016. Estos datos se han obtenido a partir de la base de datos de iEducar que contiene atributos relacionados con cada alumno.

Algunas de las variables que se estudian en el articulo anterior pertenecen concretamente al ámbito personal, social y geográfico del alumno. Estas variables son:

1. El barrio del alumno.
2. El centro educativo.
3. La edad del alumno.
4. Los ingresos del alumno.
5. Los alumnos con necesidades especiales.
6. El genero.
7. El entorno en el aula.

En esta investigación se utiliza la metodología CRISP-DM (del inglés Cross Industry Standard Process for Data Mining) que es una metodología frecuente en el desarrollo de proyectos de Data Mining. Esta metodología indica cómo debe realizarse el proceso de "data mining". Esta metodología se ha utilizado en otros artículos como Delen (2010) o (Şen, Uçar, y Delen, 2012).

Según este mismo artículo de Şen y cols. (2012), esta metodología contiene las siguientes fases en el ciclo:

1. Entendimiento del negocio. Debe comprenderse los objetivos del negocio. Se debe realizar una descripción del problema. Por ultimo debe hacerse un plan de proyecto para alcanzar los objetivos deseados.

2. Entendimiento de los datos. Debe identificarse las fuentes de los datos y obtener aquellos datos relevantes para la consecución de los objetivos.

3. Preparación de los datos. Conlleva el pre-procesado, la limpieza y la transformación de los datos relevantes con el objetivo de usar algoritmos de minería de datos.

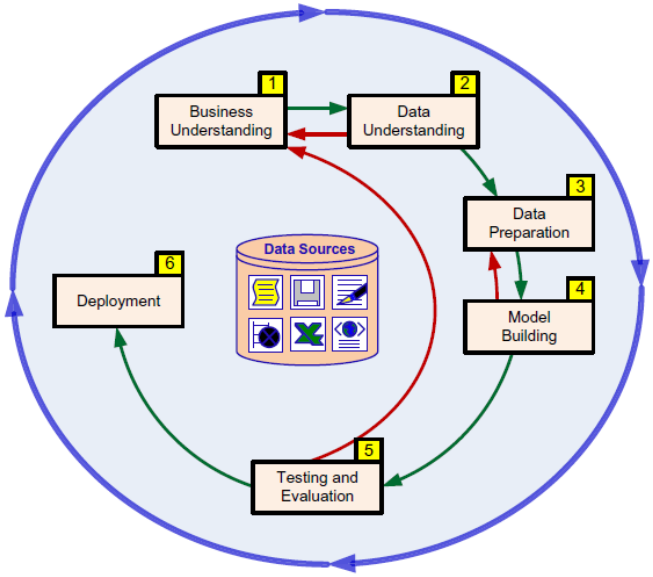
4. Construcción del modelo. Se debe desplegar un gran número de modelos y quedarse con aquellos que devuelvan valores óptimos para los datos utilizados.

5. Evaluación y Test. Debe evaluarse y probarse los modelos. Deben compararse entre sí y comprobar que son útiles para los datos expuestos.

6. Puesta en marcha. Realizar actividades usando los modelos seleccionados en el proceso de la toma de decisión.

En la figura 1, obtenida del artículo de Şen y cols. (2012) se muestra el ciclo de CRISP.

Fig. 1: Ciclo de la metodología CRISP



Para realizar la parte de predicción, utiliza las variables anteriormente comentadas, incluidas en dos conjuntos de datos. En el primer conjunto de datos (DS-I), se almacenan los datos obtenidos antes de comenzar el comienzo del año escolar. El segundo conjunto de datos incluye las variables del primer conjunto de datos y alguna nueva que se ha obtenido después del segundo mes del año escolar. Siendo algunas de estas variables nuevas las asignaturas, las notas y las ausencias. Estas dos últimas variables son las que mayor importancia tienen en la revelación de los resultados académicos finales.

El primer conjunto de datos (DS-I) se usa para entrenar el modelo de clasificación I (CM-I), que identifica la probabilidad que tiene un alumno de suspender teniendo en cuenta los datos del comienzo de curso. El segundo conjunto de datos (DS-II) se usa para entrenar el modelo de clasificación II, que también identifica la probabilidad

que tiene un alumno de suspender teniendo en cuenta los datos del comienzo de curso e incluyendo las nuevas variables. Una vez que se han entrenado los modelos, se ha utilizado la matriz de confusión para obtener la bondad o efectividad del modelo respecto al conjunto de datos. Los datos obtenidos han mostrado que las variables de "vecindario", "colegio", "ciudad" y "edad"son factores relevantes que afectan a los resultados académicos de los alumnos.

Como conclusión, se indica en esta investigación que el entorno social y sus variables tienen una influencia directa en el proceso de enseñar-aprender. Esta investigación puede aportar información a los profesionales que busquen herramientas o métodos para mejorar los resultados escolares de los alumnos.

Por otro lado, en el artículo de prensa de Asif, Merceron, Ali, y Haider (2017) se citan otras investigaciones realizadas, donde también se utilizan variables sociales como la edad, sexo, nacionalidad, estado civil, desplazamiento (si el alumno vive fuera del distrito), necesidades especiales, tipo de admisión, situación laboral, situación económica, etc.

En el artículo de Asif y cols. (2017), se analiza el rendimiento de los alumnos matriculados en el 4 año del grado universitario de Tecnología Informática. El objetivo es, nuevamente, obtener información sobre el rendimiento de estudiantes para que las personas interesadas (directores y docentes) puedan mejorar el programa educativo. Los enfoques para lograr este objetivo son los siguientes:

1. En primer lugar se generan clasificadores para predecir el rendimiento de los estudiantes al final del curso académico tan pronto como sea posible. Estos clasificadores toman las calificaciones de admisión y las calificaciones finales del primer y segundo año. No se consideran características socio-económicas o demográficas.

2. En segundo lugar, utilizando estos clasificadores, el objetivo es utilizar cursos que puedan servir como indicadores efectivos del desempeño de los estudiantes. De esta forma se puede ayudar o estimular a los alumnos en riesgo.

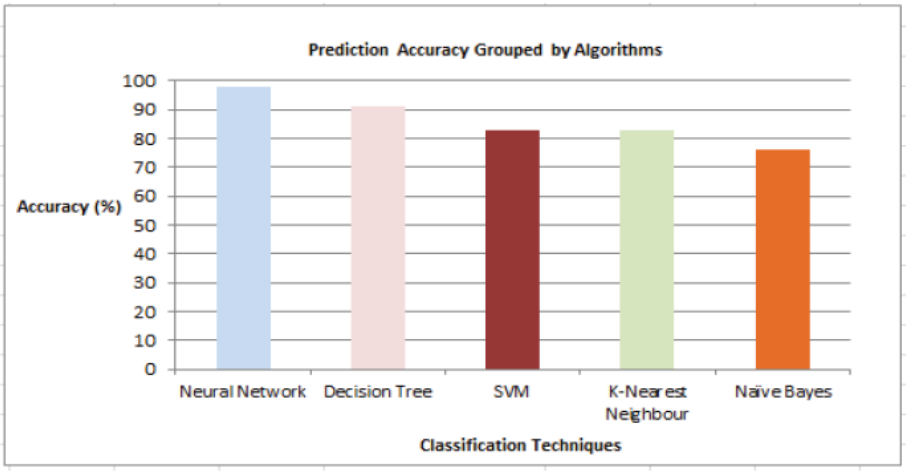
3. Por último, se va a investigar como el rendimiento académico progresa sobre el cuarto año del grado. Para ello, se va a utilizar técnicas de *clustering* y se van a dividir a los alumnos en grupos, donde los alumnos de un mismo grupo van a tener la misma progresión en el rendimiento. De esta forma, se van a agrupar los alumnos que hayan tenido bajas calificaciones a lo largo de sus estudios y aquellos que han tenido altas calificaciones a lo largo de sus estudios. La clave es obtener y comprender los indicadores propuestos en el segundo paso.

Los datos utilizados en este articulo proceden de las calificaciones del cuarto año del grado de ingeniería de Tecnología Informática de una universidad de Pakistán. Se van a tomar 210 alumnos que se han matriculado en los cursos de 2007-2008 y 2008-2009. Los datos contienen variables relacionadas con las calificaciones de pre-admisión de los alumnos y de las calificaciones de estos en los siguientes 4 años del programa de grado. Por tanto, para lograr los objetivos establecidos, Asif y cols. (2017), va a utilizar los arboles de decisión y clúster como técnicas de minería de datos.

Otro de los artículos que se ha utilizado como referencia ha sido el de Shahiri, Husain, y Rashid (2015). En este artículo, nuevamente se han utilizado técnicas predictivas para la mejora del rendimiento académico de los alumnos. En este caso, los datos utilizados proceden de instituciones malayas. De nuevo se han tenido en cuenta los resultados académicos internos como las calificaciones de prácticas o tareas, exámenes, actividades en el laboratorio, test de clase y atención. También se ha tenido en cuenta factores externos como el género, la edad, el entorno familiar y la discapacidad. Este articulo sirve de referencia para obtener y acotar los modelos a utilizar en este TFM.

En este artículo se indica que "a priori", sin tener en cuenta la experiencia, es necesario realizar un proyecto piloto, que responda a dos preguntas en concreto. La primera pregunta que se plantea son los atributos o variables a utilizar en la investigación. La segunda pregunta planteada es sobre los métodos predictivos a utilizar. La siguiente figura 2 obtenida del artículo, muestra la precisión en la predicción de los algoritmos entre los años 2002 y 2015.

Fig. 2: Predicción en la precisión agrupada por algoritmos desde 2002 a 2015



Teniendo en cuenta dicha figura, vemos que las redes neuronales son las que obtienen mejores resultados junto con los arboles de decisiones, lo que significa que se ajustan más a los datos.

Los resultados obtenidos en otro artículo, concretamente el de Ashraf, Zaman, y Ahmed (2018) indican que el mejor modelo para los datos propuestos ha sido obtenido utilizando el algoritmo de bosques aleatorios. Este algoritmo ha obtenido mejores resultados que otros algoritmos como los arboles de decisión o árbol aleatorio. Este articulo utiliza también datos académicos de alumnos, en este caso, pertenecientes a la Universidad Kashmir.

Una vez que se ha realizado un análisis sobre la metodología utilizada y los algoritmos predictivos, además de tener en cuenta las variables utilizadas (relacionadas con el entorno del alumno), se va a realizar una investigación sobre nuevas variables que podrían incluirse en este TFM.

En el libro de Panahi y cols. (2019), se ha realizado una serie de investigaciones cuyo objetivo ha sido determinar la idoneidad de construir o emplazar centros educativos según pesos dados a factores. Estos factores son los siguientes:

- **Facilidades Urbanas:** En este punto se incluyen las gasolineras, las tuberías de gas de alta presión y las líneas de alta tensión. Cuanto más cerca estén los centros de estas zonas, más riesgo existe para los alumnos. Se tiende por tanto a alejar los centros de estos puntos.

- **Densidad de población y áreas residenciales:** La proximidad de los colegios a zonas residenciales con una gran población de estudiantes es importante, puesto que, a menor distancia entre los estudiantes, los colegios y sus casas menor es el gasto de las familias y menor es la probabilidad de que los alumnos sean secuestrados.

- **Accesibilidad a red de carreteras urbanas:** La distancia de las calles y las autovías es otro factor importante para situar los colegios. Cuanto más cerca estén los colegios a estas vías, más facilidades tendrán los alumnos, y por lo tanto más ahorro de tiempo y costes. Sin embargo, la cercanía de los colegios a las autovías o autopistas, puede implicar mayor riesgo de accidentes. Sin embargo, si las autovías o autopistas se encuentran lejos, se reduce la accesibilidad a los colegios. Es necesario situar los centros en puntos intermedios (100-200m).

- **Servicios Urbanos:** Las distancias a los hospitales, a las estaciones de bomberos y de policía tienen mayor influencia. Sin embargo, estos deben situarse a distancias prudenciales de los centros (100-200m).

- **Centros culturales:** La proximidad de los centros culturales incrementa la salud espiritual y psicológica del alumno, incrementando así sus conocimientos. Curiosamente, si existen estos tipos de centros cercanos al colegio, entonces no es necesario que dichos colegios dispongan de estos servicios (pudiéndose ampliar las aulas, el comedor, etc)

La investigación se ha llevado a cabo en la ciudad de Tehran. Se han tomado para el estudio dos distritos. Uno de ellos contiene 106 colegios y el otro 137. A partir de la geolocalización de dichos colegios y de los subfactores comentados, se ha realizado un estudio sobre la relación existente entre los factores y subfactores y los colegios.

Los pesos dados a cada factor y subfactor se han determinado utilizando un algoritmo llamado SWARA. Los resultados finales obtenidos indican que existen subfactores que influyen más o menos en la posición del centro.

Por tanto, una vez que se ha estudiado una metodología de trabajo, y se han observado un gran número de variables relacionadas con los alumnos y con los centros, además de haber agrupado ciertos algoritmos que pueden aportar mayor información sobre los datos educativos, se va a realizar la investigación propia para resolver el problema propio de este TFM.

3. BIBLIOGRAFÍA

Ashraf, M., Zaman, M., y Ahmed, M. (2018). Using ensemble stackingc method and base classifiers to ameliorate prediction accuracy of pedagogical data. *Procedia Computer Science*, *132*, 1021 - 1040. Descargado de http://www.sciencedirect.com/science/article/pii/S1877050918307506 (International Conference on Computational Intelligence and Data Science) doi: https://doi.org/10.1016/j.procs.2018.05.018

Asif, R., Merceron, A., Ali, S. A., y Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177 - 194. Descargado de http://www.sciencedirect.com/science/article/pii/S0360131517301124 doi: https://doi.org/10.1016/j.compedu.2017.05.007

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498 - 506. Descargado de http://www.sciencedirect.com/science/article/pii/S0167923610001041 doi: https://doi.org/10.1016/j.dss.2010.06.003

Şen, B., Uçar, E., y Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, *39*(10), 9468 - 9476. Descargado de http://www.sciencedirect.com/science/article/pii/S0957417412003752 doi: https://doi.org/10.1016/j.eswa.2012.02.112

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., y Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research*, *94*, 335 - 343. Descargado de http://www.sciencedirect.com/science/article/pii/S0148296318300870 doi: https://doi.org/10.1016/j.jbusres.2018.02.012

Panahi, M., Yekrangnia, M., Bagheri, Z., Pourghasemi, H. R., Rezaie, F., Aghdam, I. N., y Damavandi, A. A. (2019). 7 - gis-based swara and its ensemble by rbf and ica data-mining techniques for determining suitability of existing schools and site selection of new school buildings. En H. R. Pourghasemi y C. Gokceoglu (Eds.), *Spatial modeling in gis and r*

for earth and environmental sciences (p. 161 - 188). Elsevier. Descargado de http://www.sciencedirect.com/science/article/pii/B9780128152263000077 doi: https://doi.org/10.1016/B978-0-12-815226-3.00007-7

Shahiri, A. M., Husain, W., y Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414 - 422. Descargado de http://www.sciencedirect.com/science/article/pii/S1877050915036182 (The Third Information Systems International Conference 2015) doi: https://doi.org/10.1016/j.procs.2015.12.157