Universidad Rey Juan Carlos

Trabajo Fin de Máster

Explotación y modelos para predicción de datos en el Sistema Educativo de la Comunidad de Madrid

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
DIRECTOR: Aurelio Berges García

AGRADECIMIENTOS

Agrademos a...

RESUMEN

Extensión máxima de una página

SUMMARY

Extensión máxima de una página

Índice

Índice de figuras VII

Índice de cuadros VIII

1 INTRODUCCIÓN 1

1.1 Introducción . . . 1
1.2 Contexto . . . 2
1.3 Objetivos . . . 3
1.4 Metodología . . . 4
1.5 Organización del TFM . . . 4

2 JUSTIFICACIÓN TEÓRICA 7

2.1 Introducción . . . 7
2.2 Análisis trabajos previos relevantes . . . 8
2.3 Estudios más relacionados con la minería de datos en educación . . . 8
2.4 Metodología de trabajo en el desarrollo de proyectos de minería de datos . . . 11
2.5 Modelos utilizados en el desarrollo de proyectos de minería de datos en el entorno educativo . . . 13
2.6 Herramientas analizadas para la minería de datos . . . 15
2.7 Conclusiones . . . 17

3 PROPUESTA DE INTERVENCIÓN 19

3.1 Justificación . . . 19
3.2 Minería de datos . . . 20
3.3 Lenguaje R y RStudio . . . 21
3.4 Modelos seleccionados . . . 21
3.4.1 Criterio de selección . . . 21
3.4.2 Métricas de precisión . . . 22

4 BIBLIOGRAFÍA 25

Índice de figuras

1 Velocidad Procesador (MIPS) a lo largo del tiempo. (Fuente: Kurzweil http://www.kurzweilai.net) 1
2 Velocidad de transferencia a lo largo del tiempo. (Fuente: Nielsen Norman Group) 1
3 Artículos aceptados y publicados desde 2011. Recuperado de Sin y Muthu (2015) 3
4 Comparación metodologías de Minería de Datos. Recuperado de Moine, Gordillo, y Haedo (2011) 12
5 Ciclo de la metodología CRISP. Recuperado de Şen, Uçar, y Delen (2012) 13
6 Predicción en la precisión agrupada por algoritmos desde 2002 a 2015. Recuperado de Shahiri, Husain, y Rashid (2015) 14
7 Comparación de los resultados obtenidos. Recuperado de Adekitan y Salau (2019) 15
8 Herramientas más usadas. Recuperado de (Piatetsky, 2013) 16

Índice de cuadros

1. INTRODUCCIÓN

1.1. 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.

Esto se debe a que, como se pueden observar en las siguientes figuras (1 y 2), las millones de instrucciones por segundo (MIPS) que realiza un procesador (relacionado con el tiempo de cómputo) y la velocidad de transmisión de datos en bits por segundo (BPS) han crecido a lo largo de los últimos años (Nielsen, 2018).

Fig. 1: Velocidad Procesador (MIPS) a lo largo del tiempo. (Fuente: Kurzweil http://www.kurzweilai.net)

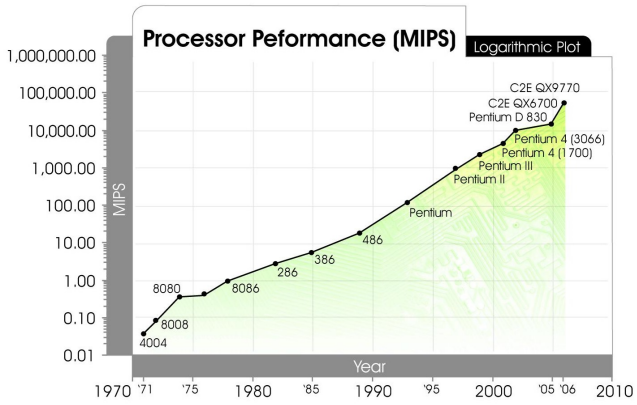
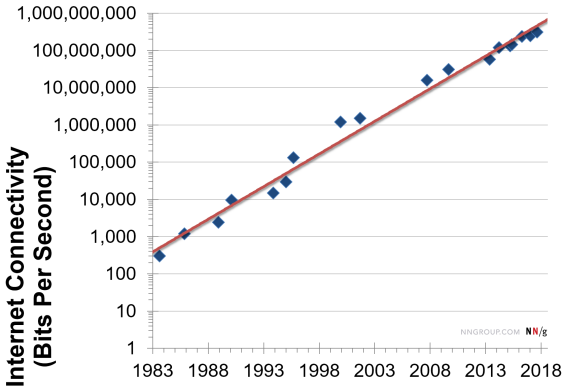


Fig. 2: Velocidad de transferencia a lo largo del tiempo. (Fuente: Nielsen Norman Group)



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.

Para descubrir la información en estos grandes volúmenes de datos, es necesario abordar el concepto de minería de datos. Según Martínez (2016), la minería de datos es el "proceso que permite transformar información en conocimiento útil para el negocio, a través del descubrimiento y cuantificación de relaciones en una gran base de datos". La minería de datos aplica técnicas estadísticas y matemáticas para poder obtener esta información implícita en los datos.

Algunas de las aplicaciones de la minería de datos según (Riquelme Santos, Ruiz, y Gilbert, 2006) son: comercio y banca, medicina y farmacia, seguridad y detección de fraude, astronomía, geología, minería, agricultura, pesca, ciencias ambientales y ciencias sociales.

1.2. Contexto

La organización educativa no ha quedado ajena a estas necesidades de una mejor comprensión de los datos. Según Romero y Ventura (2010) la minería de datos educativa (EDM) se encarga del desarrollo de métodos para explotar los datos del entorno educativo y entender mejor a los estudiantes y las herramientas que se utilizan para el aprendizaje de estos.

Por un lado, tanto el software educativo como las bases de datos institucionales, han generado una gran cantidad de datos acerca de alumnos, reflejando el aprendizaje de estos a lo largo del tiempo. (Romero y Ventura, 2010)

Por otro lado, el uso de pedagógico de Internet (eLearning), ha generado también grandes cantidades de datos acerca de la enseñanza-aprendizaje (técnicas, herramientas, etc). (Romero y Ventura, 2010).

"Toda esta información es una mina de oro, en el contexto educativo". (Romero y Ventura, 2010).

El proceso de EDM convierte los datos en bruto,obtenidos de sistemas educativos en información útil que puede tener un gran impacto en las investigaciones y practicas educativas. (Romero y Ventura, 2010)

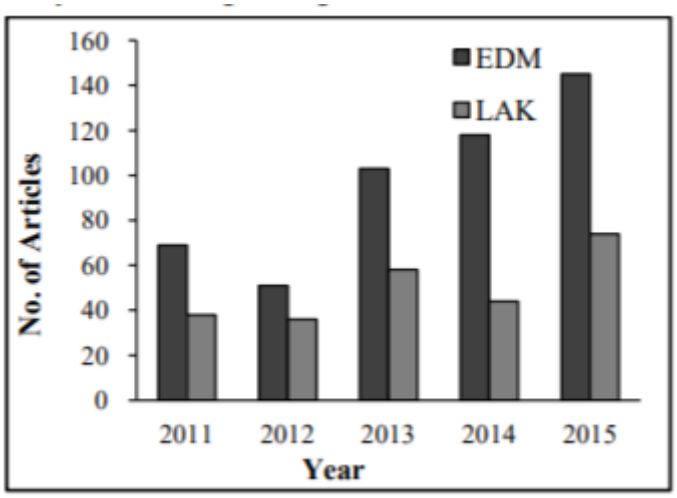
No obstante, como se indica en el libro de (Silva y Fonseca, 2017) la EDM se ha desarrollado mas lentamente que en el resto de ámbitos.

Como se puede observar en el articulo de Sin y Muthu (2015) y en la figura 3, el numero de artículos publicados en la conferencia internacional sobre la minería de

datos ha crecido desde 2011.

Este aumento de artículos publicados implica un aumento en el uso de la minería de datos en la educación.

Fig. 3: Artículos aceptados y publicados desde 2011. Recuperado de Sin y Muthu (2015)



A su vez, Sin y Muthu (2015), ha categorizado los artículos publicados según su contenido en categorías que definen las distintas aplicaciones de la minería de datos en la educación. Algunas de estas categorías son: detección de comportamiento, estimación de habilidades, predicción de mejora académica, etc

1.3. Objetivos

Con este TFM se propone dar una solución a un problema actual de una unidad (de Secundaria) de la Consejería de Madrid, mediante el uso de herramientas y métodos flexibles que automaticen dichas tareas y proponga, además, nuevas variables o factores que puedan influir en la toma de decisión.

Los objetivos que se quiere cumplir con este TFM son los siguientes:

- Seleccionar variables que interesen estudiar y que aporten valor en el desarrollo de este TFM.
- Obtener modelos que se ajusten correctamente a los datos.
- Probar distintos modelos y seleccionar aquellos que aporten mayor precisión en la predicción.
- Realizar predicciones con datos existentes.

1.4. Metodología

El proceso o metodologia llevado a cabo en este TFM ha seguido las siguientes fases:

1. En primer lugar, se ha detectado una determinada necesidad en una unidad de la Consejería de Educación.
2. Una vez detectada la necesidad, se han realizado reuniones con dicha unidad para obtener la mayor información posible acerca de sus necesidades y la forma en la que satisfacerlas. Antes de comenzar la investigación, se debe tener un claro conocimiento sobre las necesidades existentes y establecer un plan de acción.
3. Una vez se tiene claro cual es la necesidad, se va a realizar una propuesta para poder satisfacer las necesidades de la unidad.
4. La propuesta establecida debe ser validada por la propia unidad.
5. Una vez validada la propuesta, se deben estudiar distintos modelos. Se debe analizar cual de los modelos es el que mayor precisión obtiene.
6. Por ultimo, se debe validar el modelo seleccionado y realizar las predicciones correspondientes con los datos de la unidad.

1.5. Organización del TFM

La estructura que se va a seguir en el TFM va a ser la siguiente:

- **Capitulo 1. Introducción:** En el primer capítulo se van a definir las necesidades existentes que justifican el desarrollo de este trabajo. También se va a definir los objetivos que se persiguen con la realización de este. Por último, se presenta la estructura que tendrá el presente documento.
- **Capitulo 2. Justificación teórica:** En este segundo capítulo se va a realizar una investigación sobre el estado de la cuestión. Se va a realizar un estudio sobre los métodos, modelos y usos de la minería de datos en el ámbito educativo.
- **Capitulo 3. Propuesta de intervención:** En este tercer capítulo se va a plantear una solución al problema existente.
- **Capitulo 4. Diseño de la investigación:** Este capítulo va a definir los pasos que se seguirán en la realización de un proyecto de minería de datos. Se van a detallar también las tareas que se van a desempeñar en cada uno de los pasos.

- **Capitulo 5. Conclusiones:** En este capítulo se van a detallar las conclusiones obtenidas a partir de los resultados alcanzados.

2. JUSTIFICACIÓN TEÓRICA

2.1. Introducción

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, la planificación y la gestión educativa.

En segundo lugar, y, teniendo en cuenta los propósitos de esta investigación, se debe establecer el objeto de búsqueda. Por tanto, esta labor de búsqueda se va a centrar en obtener documentación científica acerca de la minería de datos en el ámbito educativo, más concretamente, en la educación secundaria.

A partir del objeto de búsqueda, se debe establecer las fuentes que se van a utilizar para obtener resultados fiables, ya que en la actualidad existen numerosos artículos acerca del uso de la ciencia de datos, pero es necesario acotar la búsqueda a lo relativo a educación.

Para la realización de este TFM se han analizado distintas publicaciones de la base de datos científica de Web of Science.

Para realizar la búsqueda se han utilizado las siguientes palabras clave: "educational", "data" y "mining". Se debe recordar que el éxito de la búsqueda depende de estas palabras claves. También se ha realizado una búsqueda utilizando estas claves en Teseo, ScienceDirect y Google Academics.

De la búsqueda en "Web of Science" con las claves comentadas se han obtenido un gran numero de publicaciones. Por tanto, se ha tenido que acotar la búsqueda incluyendo nuevas claves ("models" y "predictions"). De esta nueva búsqueda se han obtenido 47 artículos. Posteriormente se ha realizado una observación sobre los artículos obtenidos y se ha comprobado la existencia de artículos que no resultan útiles en esta investigación, por lo que se han descartado.

Además, de la primera búsqueda, donde se obtenían gran cantidad de resultados, se han revisado, entre otros, los artículos más populares (respecto a su número de citas) y actuales, y se han seleccionado 15 que se ajustan a las necesidades de este estudio.

Complementariamente, también se ha realizado búsquedas incluyendo la clave de "gis". 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.

Debemos destacar que la mayoría de los resultados obtenidos tratan de artículos

centrados en la predicción de los resultados académicos de los alumnos teniendo en cuenta ciertos factores internos (como las propias calificaciones a lo largo del curso) y externos (como factores etnográficos, edad, situación económica familiar, etc.).

2.2. Análisis trabajos previos relevantes

En este apartado se va a tener en cuenta los artículos más representativos, que son aquellos que realizan un estado de la cuestión. Estos artículos investigan acerca de las publicaciones sobre la minería de datos en el entorno educativo.

En este aspecto se pueden destacar los artículos de Silva y Fonseca (2017), Romero y Ventura (2010) y Peña-Ayala (2014).

Silva y Fonseca (2017) en su artículo, realiza una revisión sobre las publicaciones realizadas, citando diversos artículos y resumiendo brevemente el estudio y las técnicas y algoritmos utilizadas en este. Además, de forma genérica agrupa los algoritmos más utilizados en las técnicas de clasificación, "clustering" y regresión.

En el artículo de Peña-Ayala (2014) se muestra el número de publicaciones existentes hasta el momento que utilizan ciertos algoritmos predictivos como el K-Means, J-48, Naives Bayes, etc. En este mismo artículo, se clasifican las publicaciones en seis categorías. Podemos destacar que la categoría mayoritaria (con un 21 %) es el modelado del comportamiento del alumno seguida del rendimiento académico del alumno (con un 20 %). Romero y Ventura (2010) utiliza también categorías para clasificar las publicaciones.

Estos artículos sirven como referencia y ayuda en la localización de artículos de una determinada temática según las clasificaciones dadas.

2.3. Estudios más relacionados con la minería de datos en educación

Como ya se ha comentado con anterioridad, existen una serie de clasificaciones sobre las publicaciones realizadas en la minería de datos en educación. La mayoría de los artículos se centran en el rendimiento y en las calificaciones de los alumnos y, cómo teniendo en cuenta estas investigaciones, se puede mejorar la calidad educativa.

En el artículo de Fernandes et al. (2019), 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 este artículo pertenecen concretamente al ámbito personal, social y geográfico del alumno. Estas variables son: el barrio del alumno, el centro educativo, la edad del alumno, los ingresos del alumno, los alumnos con necesidades especiales, el genero y el entorno en el aula.

Como conclusiones, se indica en este artículo 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 Asif, Merceron, Ali, y Haider (2017), se realizan otras investigaciones relativas al rendimiento académico, 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.

El objetivo es, nuevamente, obtener información sobre el rendimiento de estudiantes para que las personas interesadas (directores y docentes) puedan mejorar el programa educativo.

Otro de los artículos que se ha utilizado como referencia ha sido el de Shahiri et al. (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.

Relacionado con el rendimiento académico, existe también un artículo en el que se realiza labores de predicción para evitar el fracaso escolar. En este artículo, (Vera, Morales, y Soto, 2012), se han seleccionado variables en el que se incluyen si el alumno fuma, bebe, si tiene alguna discapacidad física, la edad, el nivel económico entre otras muchas. Los datos de este artículo se han obtenido a partir de encuestas realizadas a alumnos, del Centro Nacional de Evaluación y del Departamento de Servicios Escolares. Relacionado también con el rendimiento académico, es el artículo de Kaur, Singh, y Josan (2015) donde se utilizan variables como el uso del móvil por parte del alumno, el tipo del colegio, la localización de este (áreas urbanas o rurales), el acceso a Internet del alumno, etc. Siendo las variables de la existencia de Internet y ordenador en casa las que más afectan en la predicción. En este estudio, la variable a predecir en esta investigación es si el alumno se gradúa o no.

Se ha encontrado un artículo referente a la educación en España, este artículo es el

de José (2016), en el que se analiza las calificaciones y las tareas para cada trimestre de estudiantes de Bachillerato y ESA (Educación Secundaria para Adultos). José en este artículo, utiliza alumnos de un determinado centro público de Andalucía. Los cursos de alumnos que evalúa son 1º y 2º de Bachillerato y de ESA.

También se ha revisado un artículo relacionado con la mejora académica de alumnos de ingeniería en los primeros 3 años de titulación. Este artículo de Adekitan y Salau (2019) utiliza datos de una universidad de Nigeria. Inicialmente se consideraron 18 variables, sin embargo, solo se utilizaron 6 variables que son las siguientes: matriculación, genero, especialidad de los estudiantes, ciudad del estudiante, calificaciones y tipo de educación secundaria recibida previamente.

En el artículo de Álvarez García et al. (2010) se analiza la relación entre la violencia y la repetición de curso. En la investigación se han realizado un cuestionario a 1742 estudiantes de 7 centros. Según el artículo, los resultados obtenidos han indicado que la violencia es mayor cuando los alumnos han repetido de curso. Alguna de las variables que se estudian son: Violencia de profesorado hacia alumnado, violencia física indirecta por parte del alumnado, Violencia verbal de alumnado hacia alumnado, violencia física directa entre alumnado y violencia verbal de alumnado hacia profesorado.

En el libro de Panahi et al. (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 resultados finales obtenidos indican que existen subfactores que influyen más o menos en la posición del centro.

2.4. Metodología de trabajo en el desarrollo de proyectos de minería de datos

En primer lugar, y antes de realizar cualquier trabajo, es necesario tener en cuenta una metodología valida a seguir, es decir, se debe seguir una serie de pasos para conseguir los objetivos determinados. Existen distintas metodologías de trabajo para realizar un proyecto de minería de datos. Sin embargo, según el artículo de Moine et al. (2011), las metodologías que abarcan todas las posibles etapas de un proyecto serían las metodologías CRISP-DM y Catalyst. El resto de metodologías no completan todas las fases que se debiera o simplemente establecen los pasos a seguir, pero no las tareas. La comparativa se muestra en la figura 4.

Fig. 4: Comparación metodologías de Minería de Datos. Recuperado de Moine et al. (2011)

Fases	KDD	CRISP – DM	SEMMA	CATALYST
Análisis y comprensión del negocio	Comprensión del dominio de aplicación	Comprensión del negocio		Modelado del negocio
Selección y preparación de los datos	Crear el conjunto de datos Limpieza y pre-procesamiento de los datos Reducción y proyección de los datos	Entendimiento de los datos Preparación de los datos	Muestreo Comprensión Modificación	Preparación de los datos
Modelado	Determinar la tarea de minería Determinar el algoritmo de minería Minería de datos	Modelado	Modelado	Selección de herramientas y modelado inicial
Evaluación	Interpretación	Evaluación	Valoración	Refinamiento del modelo
Implementación	Utilización del nuevo conocimiento	Despliegue		Comunicación

La metodología de trabajo predominante en los artículos observados de carácter educativo ha sido la metodología CRISP-DM (del inglés Cross Industry Standard Process for Data Mining) que es una metodología frecuente en el desarrollo de proyectos de minería de datos. Esta metodología indica cómo debe realizarse, mediante tareas, dichos proyectos. Esta metodología se ha utilizado en artículos como Fernandes et al. (2019), Delen (2010), Şen et al. (2012), Jaramillo y Arias (2015) y Chalaris, Gritzalis, Maragoudakis, Sgouropoulou, y Tsolakidis (2014).

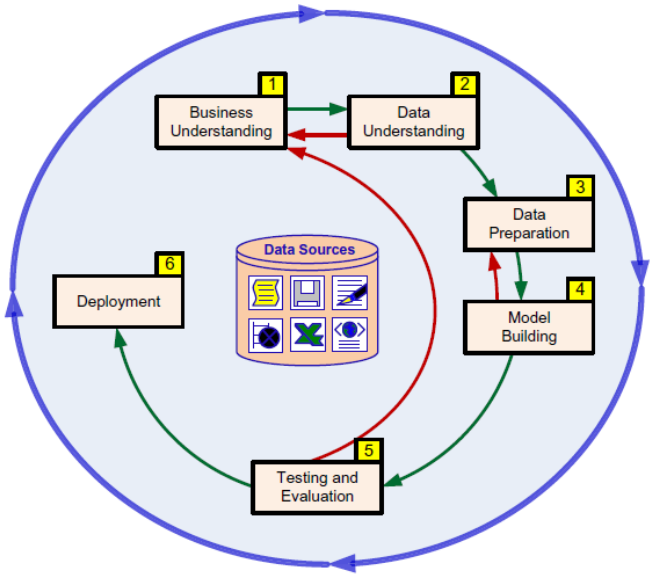
Según este mismo artículo de Şen et al. (2012), esta metodología contiene las siguientes fases en el ciclo:

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 5, obtenida del artículo de Şen et al. (2012) se muestra el ciclo de CRISP.

Fig. 5: Ciclo de la metodología CRISP. Recuperado de Şen et al. (2012)



Según Jaramillo y Arias (2015), en su investigación se ha seleccionado Crisp-DM por las siguientes razones: "La metodología a utilizar es Crisp-DM ya que cada una de sus fases se encuentra claramente estructurada definiendo de tal forma las actividades y tareas que se requieren para lograr el objetivo planteado, es decir, la más completa entre las metodologías comparadas, es flexible por ende se puede hacer usos de cualquier herramienta de minería de datos", idea ya presentada en el artículo de (Moine et al., 2011).

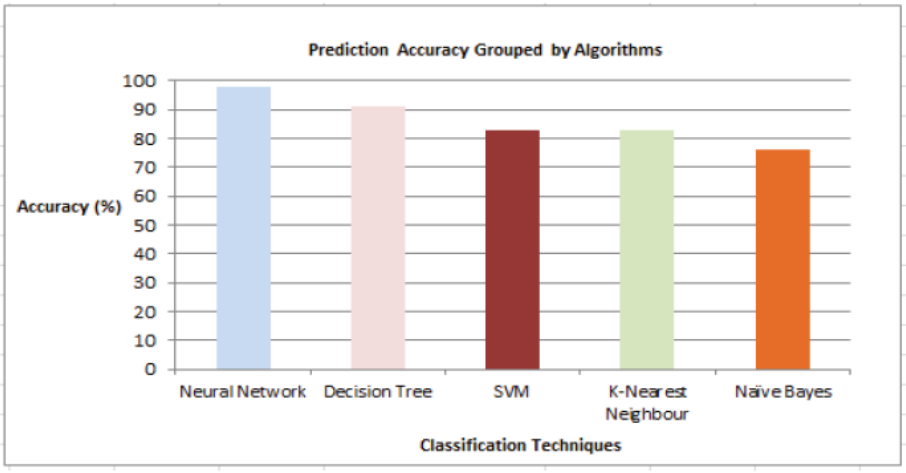
2.5. Modelos utilizados en el desarrollo de proyectos de minería de datos en el entorno educativo

Una vez que se ha revisado los artículos existentes, se debe abstraer la información relativa a los modelos utilizados con el objetivo de obtener un estado de la situación de dichos modelos.

En el artículo de prensa de Fernandes et al. (2019), se muestra el uso de técnicas 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.

En el artículo de Shahiri et al. (2015) 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 6 obtenida del artículo, muestra la precisión en la predicción de los algoritmos entre los años 2002 y 2015.

Fig. 6: Predicción en la precisión agrupada por algoritmos desde 2002 a 2015. Recuperado de Shahiri et al. (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.

Para lograr los objetivos establecidos en el análisis del rendimiento académico, Asif et al. (2017), va a utilizar los arboles de decisión, Naïves Bayes, Redes Neuronales, 1-Vecino-Cercano y Bosques Aleatorios. Los mejores resultados se han obtenido utilizando el algoritmo de Naïves Bayes, obteniendo un 85 % de precisión.

En cuanto al artículo de Adekitan y Salau (2019), nuevamente se han utilizado algoritmos como redes neuronales, bosques aleatorios, arboles de decisión, Naïve Bayes, combinación de árboles y regresión logística. En la figura 7 se puede observar la comparación entre los modelos.

Fig. 7: Comparación de los resultados obtenidos. Recuperado de Adekitan y Salau (2019)

	PNN	Random Forest	Decision Tree	Naive Bayes	Tree Ensemble	Logistic Regression
Correct Classified	475	485	477	478	486	493
Accuracy	85.895%	87.70%	87.85%	86.438%	87.884%	89.15%
Cohen's Kappa (k)	0.767	0.799	0.803	0.782	0.803	0.823
Wrong Classified	78	68	66	75	67	60
Error	14.105%	12.297%	12.155%	13.562%	12.116%	10.85%

En este artículo Adekitan y Salau (2019), se puede observar como la regresión logística obtiene la mayor precisión en los resultados. Por tanto, es capaz de clasificar correctamente los datos y, en consiguiente, obtener mejores predicciones. Otro artículo donde la regresión logística es la que mayor precisión da es el de Lehr et al. (2016), donde además se utilizan los algoritmos de Naïves Bayes, Bosques aleatorios, arboles de decisión y K-vecinos-cercanos.

2.6. Herramientas analizadas para la minería de datos

Existen diversas herramientas para realizar minería de datos, por lo tanto, se deberá analizar cuales se están utilizando en los artículos estudiados e incluso se debe revisar no solo en el ámbito educativo, sino de forma general. De esta forma obtendremos las herramientas más utilizadas.

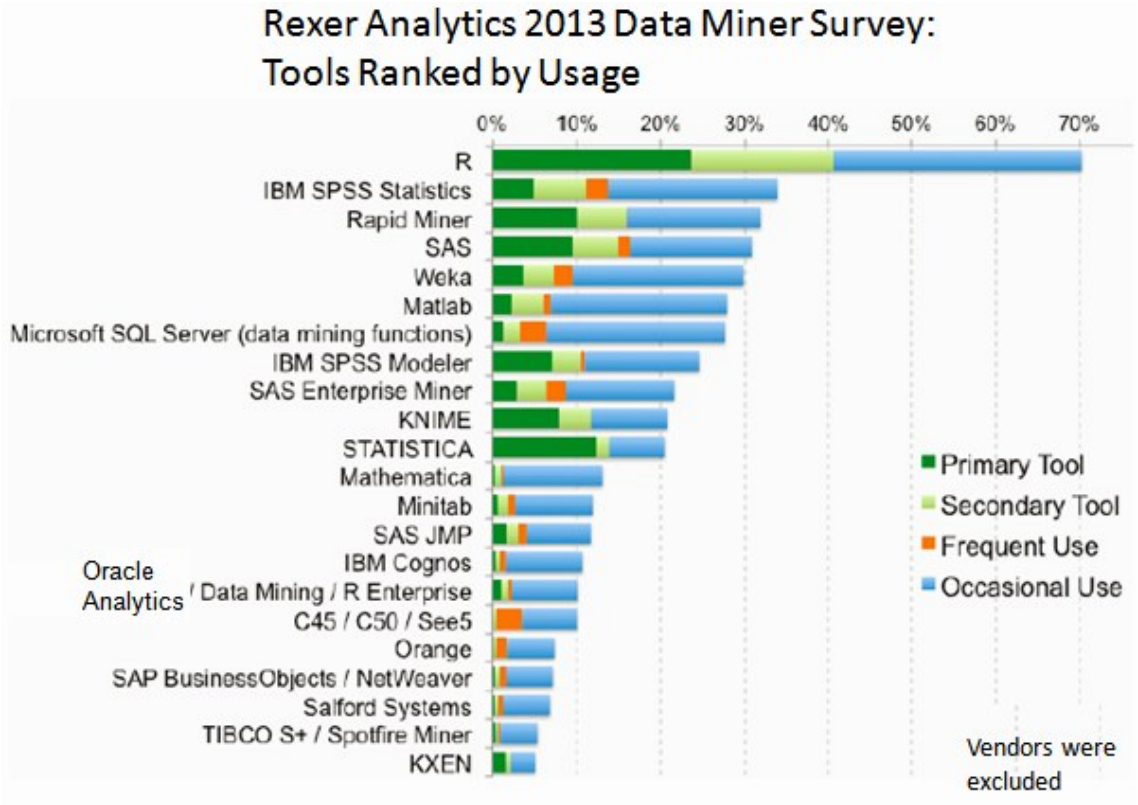
En el artículo de Rodríguez Suárez y Díaz Amador (2009) se recogen algunas de las más utilizadas. Entre ellas se puede destacar SPSS Clementine, WEKA y Oracle

Data Miner. Además, artículos como el de José (2016) han utilizado R, que es un lenguaje estadístico.

En el artículo de Jaramillo y Arias (2015), se ha realizado una breve comparación nuevamente entre las herramientas de WEKA, RapidMiner y Knime. De esta comparación, los autores han seleccionado la herramienta de RapidMiner para realizar las investigaciones por las siguientes características: "posee una licencia libre, combinación de modelos, interfaz amiga, multiplataforma, empleo de técnicas, además permite aplicar varios algoritmos de minería de datos..." (Jaramillo y Arias, 2015)

En la figura 8 se pueden observar las 10 herramientas más utilizadas en 2013 según Rexer Analytics. (Piatetsky, 2013)

Fig. 8: Herramientas más usadas. Recuperado de (Piatetsky, 2013)



Como se puede observar en la figura anterior, las herramientas más utilizadas son R, IBM SPSS Statistics, RapidMiner, y también en puestos superiores se encuentra Weka.

Teniendo conocimiento sobre las herramientas más utilizadas, se debe elegir una de ellas para realizar la investigación de este TFM.

2.7. Conclusiones

Teniendo en cuenta los artículos anteriores, vemos que existen metodologías, técnicas y herramientas comunes, sin embargo, dependiendo del artículo, unas técnicas obtienen mejores resultados que otras. Esto se debe al carácter de los datos.

En cuanto a las metodologías, existen muchas investigaciones que utilizan sus propias metodologías en vez de utilizar aquellas de uso frecuente. No obstante, es importante tener claro el procedimiento a seguir.

Por otro lado, también se ha comprobado que existen un gran número de variables comunes de estudio en la mayoría de los artículos. Esto se debe a que la mayoría de los artículos tienen una gran relación, y es investigar acerca de factores que impliquen un mejor rendimiento en los alumnos. Algunas de estas variables se pueden considerar en la investigación de este TFM.

Respecto a los modelos utilizados, se han utilizado algoritmos comunes en varios artículos, y como ya se ha comentado, en ciertas ocasiones han sido más precisos que en otras. Esto se debe a los propios datos. Por tanto, en este TFM se van a realizar pruebas con distintos modelos y se va a seleccionar el que mejor resultados obtenga.

Por último, en muchos artículos no se ha referenciado las herramientas utilizadas para llevar a cabo la investigación, sin embargo, se ha acudido a otras fuentes para obtener las herramientas con mayor uso. El uso de unas herramientas u otras, no es relevante, puesto que, a día de hoy, la mayoría de las herramientas satisfacen las necesidades de los investigadores, sin embargo, se debe tener en cuenta la licencia comercial, las posibles librerías, etc. que nos pueda o no proporcionar la herramienta.

3. PROPUESTA DE INTERVENCIÓN

3.1. Justificación

Desde la Consejería de Educación de Madrid, también se ha querido obtener información intrínseca de los datos que poseen. De esta forma, 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 enfrenta esta unidad 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 recursos. De esta forma, además, se evita la existencia de grupos sobrepoblados.

Desde esta unidad, se ha informado sobre aspectos con 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, 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.

3.2. Minería de datos

Con el objetivo de resolver el problema comentado anteriormente y satisfacer los objetivos, se plantea el uso de la ciencia de datos como proceso para descubrir relaciones entre los datos, que sean significativas. Además, se van a buscar patrones y tendencias en los datos que ayuden a la toma de decisiones.

En primer lugar, se debe tener en cuenta que la ciencia de datos aúna métodos y tecnologías que provienen del campo de las matemáticas, la estadística y la informática entre las que se pueden encontrar el análisis descriptivo o exploratorio, el aprendizaje automático ("machine learning"), el aprendizaje profundo ("Deep learning"), etc. (Marín, 2018). En esta propuesta de intervención, se va a centrar en el **análisis descriptivo** y el **aprendizaje automático**.

El análisis descriptivo, como ya se ha comentado, va a ser útil para observar características de los propios datos. Entre estas características se va a poder observar cuales son las variables que más convienen al estudio por su importancia, utilizando técnicas como el análisis principal de componentes. El artículo de Costa, Fonseca, Santana, de Araújo, y Rego (2017) incluye el apartado de pre-procesado, en el que realiza un estudio para reducir la dimensionalidad de las variables, puesto que están trabajando con un gran número de ellas.

El aprendizaje automático, se divide en dos áreas: el aprendizaje supervisado y el no supervisado.

- El aprendizaje supervisado (o predictivos): se basa en algoritmos que intentan encontrar una función, que, dadas las entradas, asigne unas salidas adecuadas. Estos algoritmos se entrenan mediante datos históricos y de esta forma aprende a asignar salidas adecuadas en función de dichas entradas, dicho de otra forma, predice el valor de salida. A su vez, el aprendizaje supervisado se divide en regresión (si la salida es de tipo numérico) y clasificación (si la salida es del tipo categórico). (Recuero, 2017)
- El aprendizaje no supervisado: se utiliza en datos en los que existen variables de entrada, pero no existen variables de salida para dichas variables de entrada. Por consiguiente, solo se puede describir la estructura de los datos,

para intentar conseguir algún tipo de tendencias y patrones que simplifiquen el análisis.(Recuero, 2017) (Rodríguez Suárez y Díaz Amador, 2009)

3.3. Lenguaje R y RStudio

Una vez que se tienen claros los conceptos y las técnicas, se deberá elegir la herramienta de trabajo. En esta línea de investigación se va a utilizar R como lenguaje de programación y RStudio como entorno de desarrollo para R.

Como ya se ha comentado, R es un lenguaje de programación para el análisis estadístico. Al estar orientado a la estadística, proporciona un gran número de bibliotecas y herramientas. Destaca también por la generación de gráficos estadísticos de gran calidad. Posee muchos paquetes dedicados a la graficacion. Además, es una herramienta que facilita el cálculo numérico y el uso en la minería de datos. (Goette, 2014)

Su potencia reside fundamentalmente en que es un software gratuito y de código abierto. Como ya se ha comentado, posee un gran número de herramientas que pueden ampliarse mediante paquetes, librerías o definiendo funciones propias.

Por otro lado, RStudio es el entorno de desarrollo para R. Es también software libre y tiene la ventaja que se puede ejecutar sobre distintas plataformas (Windows, Mac y Linux).

3.4. Modelos seleccionados

Los métodos de predicción que se van a utilizar para resolver el problema en cuestión van a ser aquellos que mejores resultados han obtenido utilizando los datos de varias líneas de investigación estudiadas. Estos métodos son los siguientes: árboles de decisión, redes neuronales, k-vecinos cercanos, bosques aleatorios y regresión logística. Obviamente se debe destacar que, aunque se van a utilizar dichos métodos, pueden existir otros que se ajusten mejor a los datos.

3.4.1. Criterio de selección

Una vez que se han seleccionado las variables y los algoritmos a estudiar, es hora de realizar el propio modelado. Al realizar el modelado, deberemos tener en cuenta que variables son mejores para este modelado. Es posible que existan variables que únicamente empeoren los resultados del modelado, por lo tanto, se deberán desestimar. Para ello se va a utilizar el criterio de Akaike (AIC).

Este criterio indica el ajuste que tienen los datos experimentales con el modelo utilizado. Obviamente, el criterio de AIC solo tiene sentido cuando se realizan comparaciones con otros modelos (utilizando el mismo conjunto de datos). (Martınez et al., 2009)

Cuanto menor sea el valor de este criterio, mejor se ajustan los datos al modelo. Por tanto, se deberá seleccionar el modelo que menor AIC tenga. (Martınez et al., 2009)

3.4.2. Métricas de precisión

Las métricas que se va a utilizar para obtener la precisión de los modelos van a ser aquellas descritas en el artículo de Costa et al. (2017), Helal et al. (2018) y Ashraf et al. (2018). Estas métricas son frecuentes en ámbitos como la obtención de información, aprendizaje automático y otros dominios como la clasificación binaria. Dichas métricas van a ser las siguientes:

- FMeasure: es la media armónica de la precisión y recuperación de un clasificador; es decir, FMeasure = 2 * Precision * Recall / (Precision + Recall).
- Precision: es la fracción de verdaderos positivos entre todos los ejemplos clasificados como positivos. P= TP/(FP+TP).
- Recall: es la fracción de verdaderos positivos clasificados correctamente. R = TP/(FN+TP).
- AUC: el área bajo la característica de operación del receptor. La curva (ROC) indica la probabilidad de que un clasificador clasifique un positivo seleccionado aleatoriamente sobre un negativo. Un AUC con valor de 1 indica un perfecto clasificador, mientras que 0.5 implica que el clasificador lo hace de forma aleatoria.

Donde:

- TP - Verdadero Positivo: es el número de instancias positivas clasificadas correctamente como positivas.
- FP - Falso Negativo: es el número de instancias positivas clasificadas incorrectamente como negativas.
- FP - Falso Positivo: es el número de instancias negativas clasificadas incorrectamente como positivas.
- TN - Verdadero Negativo: es el número de instancias negativas clasificadas correctamente como negativas.

Estas métricas se van a mostrar utilizando un gráfico conocido como matriz de confusión.

4. BIBLIOGRAFÍA

Adekitan, A. I., y Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250. Descargado de http://www.sciencedirect.com/science/article/pii/S240584401836924X doi: https://doi.org/10.1016/j.heliyon.2019.e01250

Álvarez García, D., Álvarez Pérez, L., Núñez Pérez, J. C., González Castro, M. P., González García, J. A., Rodríguez Pérez, C., y Cerezo Menéndez, R. (2010). Violencia en los centros educativos y fracaso académico. *Revista Iberoamericana de Psicología y salud*.

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

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., y Tsolakidis, A. (2014). Improving quality of educational processes providing new knowledge using data mining techniques. *Procedia-Social and Behavioral Sciences*, *147*, 390–397.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., y Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247 - 256. Descargado de http://www.sciencedirect.com/science/article/pii/S0747563217300596 doi: https://doi.org/10.1016/j.chb.2017.01.047

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

Goette, P. E. (2014). *R, un lenguaje y entorno de programación para análisis estadístico.* Descargado 2019-04-16, de https://www.genbeta.com/desarrollo/r-un-lenguaje-y-entorno-de-programacion-para-analisis-estadistico

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., y Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, *161*, 134 - 146. Descargado de http://www.sciencedirect.com/science/article/pii/S0950705118303939 doi: https://doi.org/10.1016/j.knosys.2018.07.042

Jaramillo, A., y Arias, H. P. P. (2015). Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje. *Revista Tecnológica-ESPOL*, *28*(1).

José, B. G. F. (2016). Explotación y modelos para predicción de datos de un centro educativo.

Kaur, P., Singh, M., y Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, *57*, 500–508.

Lehr, S., Liu, H., Kinglesmith, S., Konyha, A., Robaszewska, N., y Medinilla, J. (2016). Use educational data mining to predict undergraduate retention. En *2016 ieee 16th international conference on advanced learning technologies (icalt)* (pp. 428–430).

Marín, J. L. (2018). *Ciencia de datos, machine learning y deep learning.* Descargado de https://datos.gob.es/es/noticia/ciencia-de-datos-machine-learning-y-deep-learning (Recuperado 17 enero, 2019)

Martınez, D. R., Julio, L. A., Cabaleiro, J. C., Pena, T. F., Rivera, F. F., y Blanco, V. (2009). El criterio de información de akaike en la obtención de modelos estadısticos de rendimiento. *XX Jornadas de Paralelismo*.

Martínez, M. (2016). Minería de datos. *Universidad Nacional del Noroeste Facultad de Ciencias Exactas, Naturales y Agrimensura, Argentina*.

Moine, J. M., Gordillo, S. E., y Haedo, A. S. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. En *Congreso argentino de ciencias de la computación* (Vol. 17).

Nielsen, J. (2018). *Nielsen's law of internet bandwidth.* Descargado de https://www.nngroup.com/articles/law-of-bandwidth/

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

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4, Part 1), 1432 - 1462. Descargado de http://www.sciencedirect.com/science/article/pii/S0957417413006635 doi: https://doi.org/10.1016/j.eswa.2013.08.042

Piatetsky, G. (2013). *Rexer analytics 2013 data miner survey highlights.* Descargado de https://www.predictiveanalyticsworld.com/patimes/rexer-analytics-2013-data-miner-survey-highlights/2777/

Recuero, P. (2017). *Los 2 tipos de aprendizaje en machine learning: supervisado y no supervisado.* Descargado 2019-01-17, de https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html

Riquelme Santos, J. C., Ruiz, R., y Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.*.

Rodríguez Suárez, Y., y Díaz Amador, A. (2009). Herramientas de minería de datos. *Revista Cubana de Ciencias Informáticas*, *3*(3-4).

Romero, C., y Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

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

Silva, C., y Fonseca, J. (2017, 09). Educational data mining: A literature review. En (p. 87-94). doi: 10.1007/978-3-319-46568-5_9

Sin, K., y Muthu, L. (2015). Application of big data in education data mining and learning analytics–a literature review. *ICTACT journal on soft computing*, *5*(4).

Vera, C. M., Morales, C. R., y Soto, S. V. (2012). Predicción del fracaso escolar mediante técnicas de minería de datos. *Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem*, *109*.