

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

Desde estas líneas me gustaría expresar mi mas sincero agradecimiento:

A mi tutor Aurelio Berges, por ayudarme y guiarme durante todo el proyecto.

A Felipe Retortillo, por ayudarme con toda la información necesaria para realizar este TFM.

A mis padres, por haberme proporcionado la mejor educación y lecciones de vida.

En especial a mi padre, por haberme enseñado que, con esfuerzo, trabajo y constancia todo se consigue, y que en esta vida nadie regala nada.

En especial a mi madre, por hacerme ver cada dia la vida de una forma diferente y confiar en mis decisiones. Sin olvidar su infinita ayuda durante estos últimos años.

A todos mis familiares por haberme apoyado y animado.

A mis compañeros del máster, con los que he compartido buenos momentos a lo largo de las intensas tardes del máster.

A mis amigos, por estar siempre a mi lado.

A todos aquellos que siguen estando cerca de mi y que le regalan a mi vida algo de ellos.

RESUMEN

La planificación de grupos escolares es una tarea complicada que debe realizar la Consejería de Educación e Investigación de la Comunidad de Madrid. Gestionar los recursos de la mejor forma es vital para mejorar la calidad educativa. Por tanto, se hace necesario el estudio de las variables más importante que permitan un mejor reparto de los medios disponibles.

En los estudios que se realizan se utilizan datos de los grupos actuales de distintos centros, entre otras variables. Con esta información se han construido varios modelos que permitan predecir el número de grupos que deben planificarse para el curso siguiente y poder así repartir los recursos disponibles (profesores, suministros, etc.).

En este estudio se realiza una clasificación sobre la importancia de las variables existentes, además se realizan estudios estadísticos para obtener la relación que existen entre estas variables para posteriormente utilizar distintos modelos con el objeto de conseguir la mayor predicción posible.

Para obtener el mejor modelo se utilizan distintas métricas con el objetivo de poder compararlos entre sí en función de las necesidades que se requieran.

ABSTRACT

ÍNDICE

Índice de tablas	III
Índice de figuras	III
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
4 Diseño de Investigación	23
4.1 Introducción	23
4.2 Arquitectura Inicial	23
4.3 Diseño de la minería de datos	26
4.3.1 Comprensión del negocio	28
4.3.2 Comprensión de los datos	28
4.3.3 Preparación de los datos	29
4.3.4 Modelado	31
4.3.5 Evaluación	35
4.3.6 Distribución	36
4.4 Herramientas utilizadas	36

4.4.1	Suite de Pentaho BI	36
4.4.2	Lenguaje R y RStudio	37
5	ANÁLISIS DE RESULTADOS	39
5.1	Análisis exploratorio de datos	39
5.2	Análisis predictivo	40
6	CONCLUSIONES	43
6.1	Conclusiones	43
6.2	Lineas Futuras	43
7	REFERENCIAS	45
	ANEXOS	50
A	Gráficas del Análisis Exploratorio de datos	53
A.A	Definición de Variables	53
A.B	Análisis exploratorio	54
A.B.1	Estadísticos más relevantes	54
A.B.2	Análisis de normalidad	55
A.B.3	Relaciones entre variables	59
A.C	Selección de Variables	62
A.C.1	Usando Random Forest	62
A.C.2	Regresión Paso a Paso	63
B	Análisis Predictivo	67
B.A	Comparación de Modelos	67
B.B	Predicción utilizando el modelo de árbol de decisión	69
B.C	Sistema de Explotación de la Consejería de Educación e Investigación . .	70

Índice de tablas

A.1	Nomenclatura de las Variables	53
B.1	Tabla comparación modelos (precisión y tiempo)	68

Índice de figuras

1.1	Velocidad Procesador (MIPS) a lo largo del tiempo. (Fuente: Kurzweil http://www.kurzweilai.net)	1
1.2	Velocidad de transferencia a lo largo del tiempo. (Fuente: Nielsen Norman Group)	1
1.3	Artículos aceptados y publicados desde 2011. Recuperado de Sin y Muthu (2015)	3
2.1	Comparación metodologías de Minería de Datos. Recuperado de Moine, Gordillo, y Haedo (2011)	12
2.2	Predicción en la precisión agrupada por algoritmos desde 2002 a 2015. Recuperado de Shahiri, Husain, y Rashid (2015)	13
2.3	Comparación de los resultados obtenidos. Recuperado de Adekitan y Salau (2019)	14
2.4	Herramientas más usadas. Recuperado de (Piatetsky, 2013)	16
2.5	Plataformas de ciencia de datos y aprendizaje de máquina. Recuperado de Gartner (https://www.gartner.com)	17
4.1	Arquitectura lógica del proyecto. Recuperado de la Consejería de Educación e Investigación de la Comunidad de Madrid.	23
4.2	Esquema en estrella. Recuperado de Pesquera (2019)	24
4.3	Cubo de Mondrian. Recuperado de Urquizu (2019)	25
4.4	Fases del ciclo de vida de CRISP-DM. Recuperado de <i>Manual CRISP-DM de IBM SPSS Modeler</i> (2012).	27
4.5	Funcionamiento Árboles Decisión. Recuperado de Sayad (2019)	32

4.6	Red Neuronal. Recuperado de Piqueras (2017)	33
4.7	Ejemplo SVM. Recuperado de Álvarez (2016)	34
4.8	KNN. Recuperado de Klein (2018)	34
4.9	Fórmula de MAE. Recuperado de https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d	35
4.10	Fórmula de RMSE. Recuperado de https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d	36
A.1	Estadísticos de las variables. Elaboración propia	54
A.2	Diagrama de cajas normalizado. Elaboración propia	55
A.3	Resumen de normalidad de variables. Elaboración propia	56
A.4	Resumen de normalidad de variables. Elaboración propia	56
A.5	Distribución 1. Elaboración propia	57
A.6	Distribución 2. Elaboración propia	57
A.7	Distribución 3. Elaboración propia	58
A.8	Diagrama de barras 1. Elaboración propia	59
A.9	Diagrama de barras 2. Elaboración propia	59
A.10	Matriz de correlaciones. Elaboración propia	60
A.11	Variables más correladas. Elaboración propia	60
A.12	Mayor correlación con variable a predecir. Elaboración propia	61
A.13	Correlación entre variables NM_ALUMNOS y NM_GRUPOS. Elaboración propia	62
A.14	Correlación entre variables NM_UNIDADES y GRUPOS_PREDECIR. Elaboración propia	62
A.15	Variables más importantes usando Random Forest. Elaboración propia	63
A.16	Resultado Backward Selection. Elaboración propia	64
A.17	Gráfico Backward Selection. Elaboración propia	64
A.18	Resultado Stepwise Selection. Elaboración propia	64
A.19	Gráfico Stepwise Selection. Elaboración propia	65
A.20	Resultado Forward Selection. Elaboración propia	65
A.21	Gráfico Forward Selection. Elaboración propia	66
B.1	Comparación Modelos y Precisión. Elaboración propia	68
B.2	Comparación Modelos y Tiempo en Entrenar. Elaboración propia	69
B.3	Aumento o disminución de grupos en la predicción. Elaboración propia	69
B.4	Cuadro de Mandos de Centros I. Elaboración propia	70
B.5	Cuadro de Mandos de Centros II. Elaboración propia	71
B.6	Cuadro de Mandos de Ausencias de Alumnado I. Elaboración propia	72
B.7	Cuadro de Mandos de Ausencias de Alumnado II. Elaboración propia	73
B.8	Cubo OLAP de Unidades. Elaboración propia	74

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.

Un ejemplo de esta evolución se puede observar en las figuras (1.1 y 1.2), 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).

Figura 1.1: Velocidad Procesador (MIPS) a lo largo del tiempo. (Fuente: Kurzweil <http://www.kurzweilai.net>)

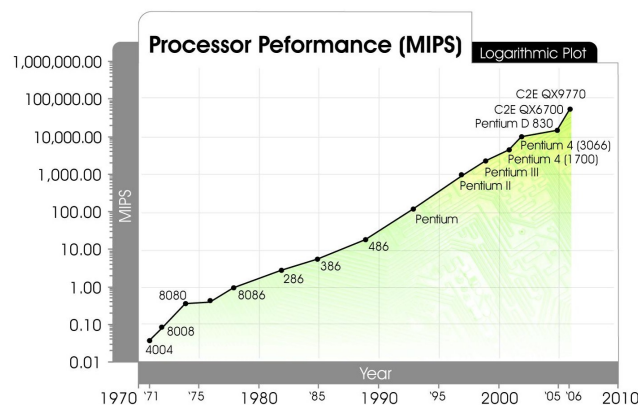
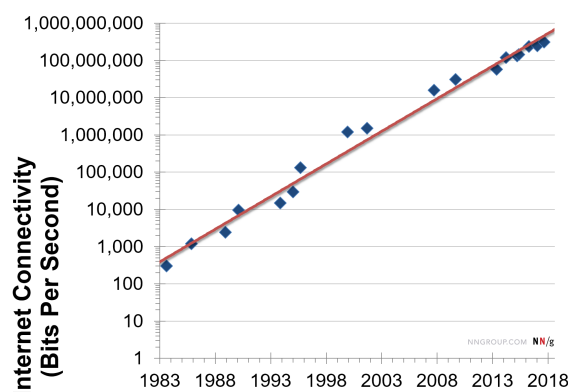


Figura 1.2: Velocidad de transferencia a lo largo del tiempo. (Fuente: Nielsen Norman Group)



Para comprender mejor este gran volumen de información que disponen las organizaciones, 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

Las organizaciones de ámbito educativo no han quedado ajenas 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. Por otro, el uso pedagógico de Internet (eLearning), ha generado también grandes cantidades de datos acerca de la enseñanza-aprendizaje (técnicas, herramientas, etc). "Toda esta información es una mina de oro, en el contexto educativo". (Romero y Ventura, 2010).

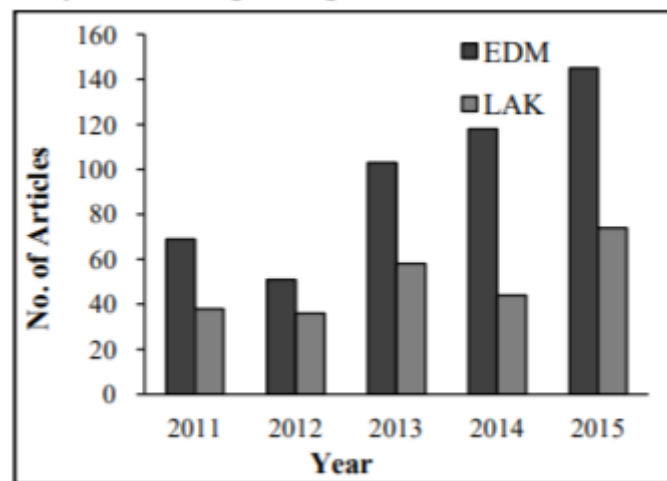
Mohamad y Tasir (2013) define EDM como una disciplina emergente, relacionada con el desarrollo de métodos para la exploración de datos que proceden del entorno educativo, para entender mejor a los estudiantes y las herramientas que estos utilizan para aprender. Coincide con el artículo de Silva y Fonseca (2017) en el que se indica que la EDM se ha desarrollado mas lentamente que en el resto de ámbitos.

Como se puede observar en el artículo de Sin y Muthu (2015) y en la figura 1.3, el numero de artículos publicados en la conferencia internacional sobre la minería de datos

ha crecido desde 2011 casi exponencialmente.

Este aumento de artículos publicados muestra como existe un aumento en el uso de la minería de datos en la educación.

Figura 1.3: 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. Por tanto, como se puede observar, dentro de la minería de datos en el entorno educativo, existen distintos campos estudiados y distintos ámbitos de aplicación.

1.3. Objetivos

En este trabajo fin de máster se plantea una solución a la necesidad que tiene la Consejería de Educación e Investigación de la Comunidad de Madrid (CEI) para dar respuesta a las necesidades de la demanda concreta de plazas escolares del nuevo período escolar. Para ello, se plantea 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 decisiones.

Dicho objetivo global se pretende alcanzar mediante los siguientes sub-objetivos:

- Seleccionar variables de interés, relativas a la resolución de las necesidades anteriormente expuesta, para 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.
- Utilizar los modelos seleccionados para realizar predicciones con los datos existentes.

1.4. Metodología

El proceso o metodología llevado a cabo en este TFM sigue las siguientes fases:

1. En primer lugar, se ha detectado una determinada necesidad en una unidad de la Consejería de Educación e Investigación de la Comunidad de Madrid.¹
2. Una vez detectada la necesidad, se realizan 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. A partir del conocimiento sobre cuales son las necesidades, se realiza una propuesta para poder satisfacer dichas necesidades
4. La propuesta establecida debe ser validada por la propia unidad.
5. Una vez validada la propuesta, se estudian distintos modelos. Se debe analizar cual de estos es el que mayor precisión obtiene.
6. Por ultimo, se valida 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 es la siguiente:

- **Capítulo 1. Introducción:** En el primer capítulo se definen las necesidades existentes que justifican el desarrollo de este trabajo. También se definen los objetivos que se persiguen con la realización de este. Por último, se presenta la estructura que tiene el presente documento.
- **Capítulo 2. Justificación teórica:** En este segundo capítulo se realiza una investigación sobre el estado de la cuestión, estudiando los métodos, modelos y usos de la minería de datos en el ámbito educativo.

¹El autor de este TFM ha colaborado con la Consejería de Educación e Investigación de la Comunidad de Madrid en el desarrollo del proyecto inicial

- **Capítulo 3. Propuesta de intervención:** En este tercer capítulo se define el problema existente.
- **Capítulo 4. Diseño de la investigación:** Este capítulo define los pasos que se siguen en la realización de un proyecto de minería de datos. Se detallan también las tareas que se van a desempeñar en cada uno de los pasos.
- **Capítulo 5. Análisis de los resultados:** Una vez realizado el análisis exploratorio y predictivo, se mostraran los resultados obtenidos en este capítulo.
- **Capítulo 6. Conclusiones:** En este capítulo se detallan las conclusiones obtenidas a partir de los resultados alcanzados.

2 | Justificación Teórica

2.1. Introducción

Uno de los intereses que se tienen a la hora de realizar cualquier proyecto de “Big Data”, “Business Intelligence” o un análisis predictivo son las variables a utilizar, ya que son las que van a permitir obtener la mayor información esperada.

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.

Para ello, y, teniendo en cuenta los propósitos de esta investigación, se debe establecer el objeto de búsqueda de 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 utilizan 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 número de publicaciones. Por tanto, se ha tenido que acotar la búsqueda incluyendo nuevas palabras 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 cen-

trados 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.). Estos factores se han utilizado principalmente para obtener una aproximación sobre las calificaciones, y el fracaso escolar. Mostrando de forma implícita las relaciones de estos aspectos con el resultado (calificación).

2.2. Análisis trabajos previos relevantes

En este apartado se resaltan los artículos más representativos que han realizado un estado del arte sobre la minería de datos en la educación, recogiendo información genérica de otros artículos. Estos artículos más representativos van a servir de referencia no sólo para obtener nuevos artículos sino para ver las metodologías comunes utilizadas y las aplicaciones concretas de la ciencia de datos en el ámbito educativo.

En este sentido 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, realizan 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, Nai- ves 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.

Mediante estos artículos se ha obtenido una vista general de la minería de datos, proporcionando información y realizando comparaciones que acotan la búsqueda de nuevas técnicas, herramientas, algoritmos, etc.

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

Este apartado se centra en las categorías -comentadas anteriormente- que tienen mayor relación con el problema a resolver en este tipo de investigación.

Teniendo en cuenta esta serie de clasificaciones o categorías 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 género 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, género, 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. Algunas 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 lleva 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 sub-factores comentados, se ha realizado un estudio sobre la relación existente entre los factores y sub-factores y los colegios.

El objetivo de esta investigación es determinar la idoneidad para seleccionar los lugares de construcción de los centros, investigando los sub-factores dados.

Los resultados finales que se obtienen indican que los factores por orden de importancia para la construcción son: los posibles daños que puedan amenazar a los alumnos y sus familias, la reducir del coste para las familias y el incremento de la eficiencia escolar.

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. Por tanto, en este apartado se va a estudiar los métodos de trabajo existentes en los artículos estudiados, ver sus ventajas y desventajas para posteriormente seleccionar el que se considere apto para esta investigación.

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 2.1.

Figura 2.1: Comparación metodologías de Minería de Datos. Recuperado de Moine et al. (2011)

Fases	KDD	CRISP – DM	SEMMA	CATALYST
<i>Análisis y comprensión del negocio</i>	Comprensión del dominio de aplicación	Comprensión del negocio		Modelado del negocio
<i>Selección y preparación de los datos</i>	Crear el conjunto de datos	Entendimiento de los datos	Muestreo	
	Limpieza y pre-procesamiento de los datos		Comprensión	
	Reducción y proyección de los datos	Preparación de los datos	Modificación	Preparación de los datos
<i>Modelado</i>	Determinar la tarea de minería			
	Determinar el algoritmo de minería	Modelado	Modelado	Selección de herramientas y modelado inicial
	Minería de datos			
<i>Evaluación</i>	Interpretación	Evaluación	Valoración	Refinamiento del modelo
<i>Implementación</i>	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, Uçar, y Delen (2012), Jaramillo y Arias (2015) y Chalaris, Gritzalis, Maragoudakis, Sgouropoulou, y Tsolakidis (2014).

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).

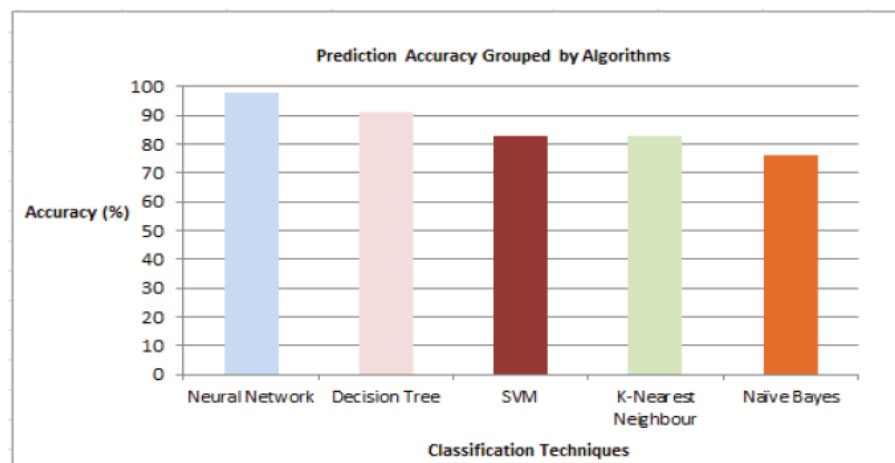
2.5. Modelos utilizados en el desarrollo de proyectos de minería de datos en el entorno educativo

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 2.2 obtenida del artículo, muestra la precisión en la predicción de los algoritmos entre los años 2002 y 2015.

Figura 2.2: 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 árboles 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 artículo 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 2.3 se puede observar la comparación entre los modelos.

Figura 2.3: 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 ello, se debe analizar cuales se utilizan 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 destacan: SPSS Clementine, WEKA y Oracle Data Miner. Además, artículos como el de José (2016) han utilizado R, que es un lenguaje estadístico.

R 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 graficación. Además, es una herramienta que facilita el cálculo numérico y el uso en la minería de datos. (Goette, 2014)

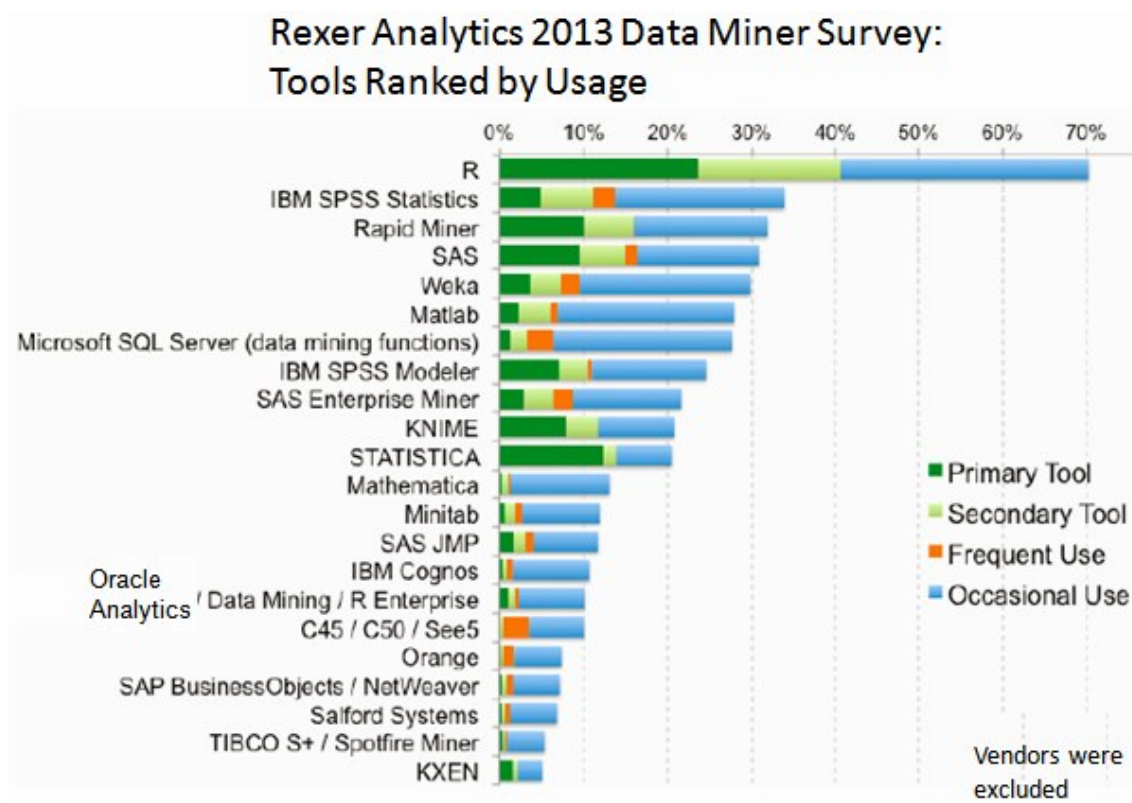
Su potencia reside fundamentalmente en que es un software gratuito y de código abierto. Como ya se ha comentado, posee un gran número de herramientas que pueden ampliarse mediante paquetes, librerías o definiendo funciones propias.

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).

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 amigable, multiplataforma, empleo de técnicas, además permite aplicar varios algoritmos de minería de datos...” (Jaramillo y Arias, 2015)

En la figura 2.4 se pueden observar las 10 herramientas más utilizadas en 2013 según Rexer Analytics. (Piatetsky, 2013)

Figura 2.4: Herramientas más usadas. Recuperado de (Piatetsky, 2013)



Como se puede observar en la figura 2.4, las herramientas más utilizadas son R, IBM SPSS Statistics, RapidMiner, y también en puestos superiores se encuentra Weka.

Por otro lado, también se ha consultado la página de Gartner, que es una empresa consultora y de investigación de las tecnologías de la información y que realiza informes sobre las herramientas existentes. En la siguiente figura 2.5 se muestran las herramientas más usadas divididas en cuadrantes dependiendo de la habilidad necesaria para usarlas y la visión completa de estas.

Figura 2.5: Plataformas de ciencia de datos y aprendizaje de máquina. Recuperado de Gartner (<https://www.gartner.com>)



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

2.7. Conclusiones

A partir de 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 realizan pruebas con distintos modelos y se selecciona el que mejor resultados obtenga. Para esta investigación, se deben seleccionar modelos de carácter regresivos.

Por último, en muchos artículos no se han 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.

Por ello, se considera muy oportuno definir, en base a la literatura existente y a los métodos, técnicas, herramientas, etc. más extendidos entre la comunidad científica, una metodología para diseñar esta investigación y unas herramientas para utilizar en dicha metodología y así satisfacer las necesidades dadas.

3 | Propuesta de Intervención

3.1. Justificación

Desde la Consejería de Educación de Madrid (CEI), también se ha querido obtener información intrínseca de los datos que poseen. Una unidad integrada en la Subdirección General de Centros de Educación Secundaria ha planteado un problema que se describe a continuación.

Cada año la escolarización de alumnos en el Sistema Educativo requiere adoptar una serie de medidas que den respuesta a las necesidades de la **demandas concretas de plazas escolares del nuevo período escolar**. Estas medidas suelen centrarse en materia de nueva construcción de centros, ampliación y adaptación de sus espacios, número y distribución del profesorado, ordenación de nuevas enseñanzas y la determinación del número de unidades de escolarización en los centros.

En consecuencia, para asegurar la adecuación de dicha demanda a la oferta de escolarización del alumnado en cada nuevo curso, es indispensable que las Unidades de Gestión de la Consejería de Educación e Investigación realicen un estudio de las zonas en las que se encuentran ubicados los centros, de la diversidad de su alumnado y de las consiguientes tendencias al aumento o disminución de alumnado y de las unidades. Como resultado del oportuno análisis se ponen en marcha actuaciones para ampliar o reducir el número de grupos y plazas autorizados en cada centro, las enseñanzas a impartir y la plantilla de recursos humanos necesaria. Puede darse incluso la necesidad de agrupamientos de centros dentro de una misma localidad o distrito, o en su caso la supresión de alguno, para atender con mayor racionalidad y eficacia las distintas necesidades.

En el caso particular de la determinación de grupos o unidades de la Enseñanza Secundaria, la Unidad de Planificación de Centros Públicos, los Servicios de Inspección Educativa y la Unidad de Programas Educativos de cada Dirección de Área Territorial (DAT) deben seguir un procedimiento específico para que, dentro de su ámbito respectivo, se alcance el fin anteriormente expuesto.

Se detalla seguidamente dicho procedimiento, contextualizándolo a la planificación para el próximo curso escolar:

1. Las DAT remiten a la Subdirección General de Centros de Educación Secundaria (SGCES) la distribución definitiva de grupos autorizados de cada centro, así como el número de alumnos matriculados, en el presente curso 2018/2019, desglosada por centros, niveles educativos, turnos y cursos de Educación Secundaria. Para los centros bilingües, se desglosa la información del total de grupos autorizados y alumnos

matriculados, en grupos y alumnos de programa y sección bilingüe, y en secciones lingüísticas. Así mismo se indicará el número de grupos mixtos que el centro tenga autorizados. Para ello se utiliza un formulario, en formato Microsoft Access.

Así mismo, en el envío, las DAT remiten, en formato editable (Excel), la distribución definitiva del cupo de profesorado asignado para cada centro, desglosado por centro y por cada uno de los distintos conceptos de cupo.

2. Las DAT realizan la propuesta de oferta educativa de cada centro para el curso 2019/2020, distribuyendo los grupos previstos por centros, niveles, turnos y cursos de Educación Secundaria. Dicha propuesta se envía a la SGCES.

Para facilitar dicha labor, se envía por correo electrónico a las DAT un fichero de datos, en formato Microsoft Access, que contiene un formulario con el listado de centros para su autorización.

3. En la Subdirección General de Centros de Educación Secundaria, se analizan todas las propuestas recibidas. Para obtener dicha distribución de grupos autorizados, el personal debe realizar trabajos manuales de predicción. Los aspectos que se tienen en cuenta para realizar la predicción son los siguientes:

a) **Escolaridad del curso actual:**

- Número de alumnos y grupos de un determinado centro.
- 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.

b) **Bilingüismo** del centro. Muchos alumnos optan por centros bilingües para su mejor formación, por lo que estos centros suelen tener más demanda de alumnos.

c) Posibilidad de creación de **nuevas zonas urbanas** cerca del centro.

d) 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.

e) **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 matrícula.

f) 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.

Para facilitar dicha labor, se envía por correo electrónico a las DAT un fichero de datos, en formato Microsoft Access, que contiene un formulario con el listado de centros para su autorización.

4. Una vez analizadas las propuestas enviadas a la SGCES, esta se encarga de distribuir por centros los grupos de escolarización necesarios para el curso 2019/2020 y se comunicara a las DAT la distribución provisional de grupos por centro.
5. Las DAT pueden enviar las alegaciones oportunas a la propuesta provisional.
6. La Dirección General de Educación Infantil, Primaria y Secundaria autorizará el número de grupos y se lo comunicará a las Direcciones de Área Territorial con el fin de que cada Área Territorial remita a los centros docentes la oferta de grupos para la escolarización del curso 2019/2020 según las fechas establecidas en la planificación del proceso de admisión.

Si se considera necesario, a fin de analizar las propuestas y observaciones remitidas, se podrán mantener reuniones de trabajo conjuntas con las Direcciones de Área Territorial.

7. Una vez resuelto el proceso de admisión, en el plazo de 10 días, los Servicios de Inspección Educativa de las respectivas Direcciones de Área Territorial estudiarán con detalle los distritos y localidades con mayor o menor demanda de plazas de escolarización de la prevista. En función de estos análisis y con el fin de precisar las actuaciones para atender las necesidades del curso escolar 2019/2020, especialmente para su incidencia en 1º ESO, se remite a la Subdirección General de Centros de Educación Secundaria el informe justificativo correspondiente indicando las variaciones producidas de alumnos y grupos en los centros, por distritos o localidad, respecto de la autorización comunicada a la que se hace referencia en el apartado anterior.

Este procedimiento se encuentra de forma detallada en la Instrucción de la Dirección General de Educación Infantil, Primaria y Secundaria con el siguiente título: *“Instrucciones de la dirección general de educación infantil, primaria y secundaria sobre la planificación del próximo curso escolar 2019/2020 en los centros públicos que imparten eso y bachillerato, creación de nuevos centros y modificación de la red, implantación y autorización de enseñanzas y propuesta de grupos”*. (Consejería de Educación e Investigación, 2018)

Actualmente, la Unidad de Planificación de Centros Públicos utiliza herramientas poco automatizadas para conocer el número de alumnos y unidades, y no disponen de algoritmos predictivos que faciliten y mejoren esta labor.

Por ello, con esta investigación, lo que se persigue es diseñar un sistema global y flexible que sea capaz de ayudar en la predicción a la Unidad de Planificación de Centros Públicos (teniendo en cuenta los aspectos dados), otorgando así una mayor garantía en la planificación de grupos. El sistema es flexible ya permite la incorporación de nuevas variable de estudio en la predicción.

A partir de esta investigación no solo se obtienen los mejores modelos que se ajusten a los datos, sino que se va a realizar un “Script” que sirva de ejemplo en el desarrollo de futuras aplicaciones. Mediante este “Script”, que contiene un modelo entrenado -con datos anteriores-, se pueden leer archivos que contienen datos de un determinado para poder predecir, por ejemplo, los del curso siguiente a este.

4 | Diseño de Investigación

4.1. Introducción

En un principio el proyecto comienza como parte de una necesidad que surge a la CEI. Dicha necesidad se basa en realizar explotaciones (visualizando los datos según convenga) y reportes sobre la situación actual respecto a otros años, concretamente, los últimos 10 años ¹. Véase el Anexo B.C para más información.

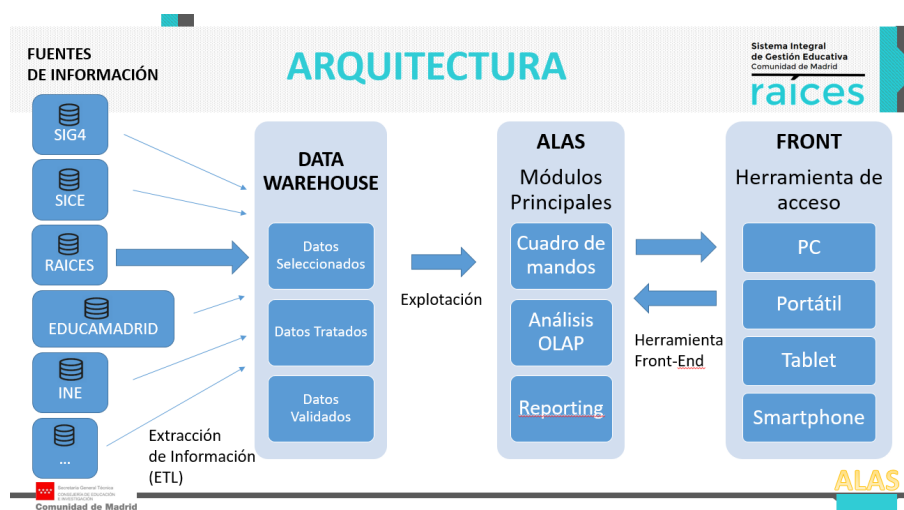
La CEI dispone de un gran número de bases de datos, por lo que obtener información sencilla a través de ellas resulta complejo. Por tanto, se requiere de un sistema de explotación que permita obtener el máximo valor de la información de dichas bases de datos.

A continuación, se expone someramente la arquitectura inicial de este sistema de explotación.

4.2. Arquitectura Inicial

Con el fin de entender conceptos posteriores, se procede a mostrar la arquitectura lógica del sistema con la que se parte inicialmente. Esta arquitectura ya está implementada y es la base del proyecto inicial y se puede observar en la figura 4.1

Figura 4.1: Arquitectura lógica del proyecto. Recuperado de la Consejería de Educación e Investigación de la Comunidad de Madrid.

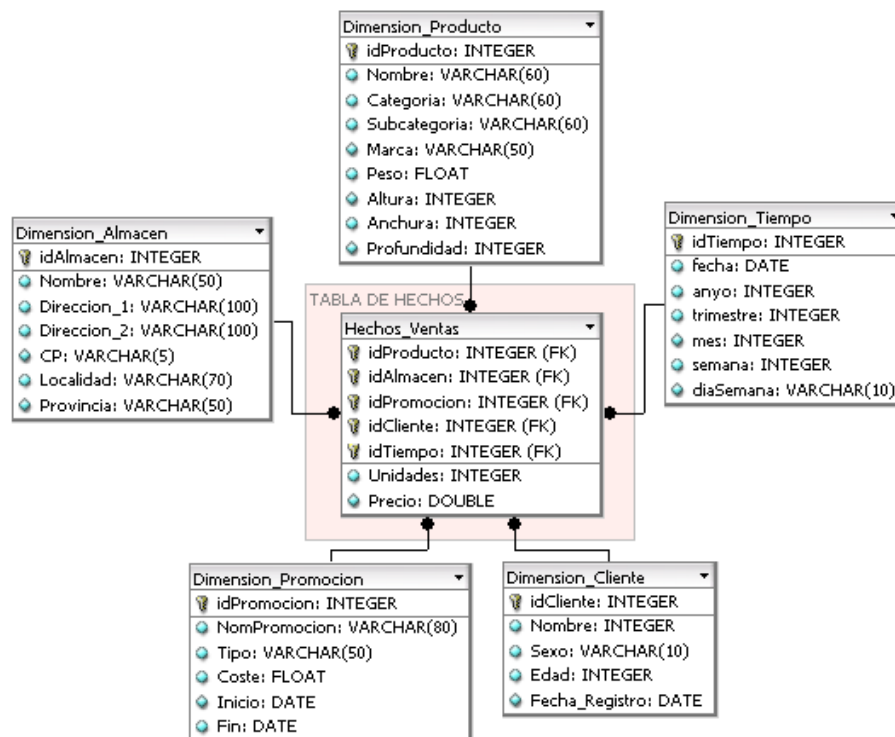


¹ Antes del curso 2007/2008 no existe un sistema de gestión centralizado y por lo tanto no se puede hacer explotaciones

En primer lugar, se tiene un gran número de bases de datos de la CEI. A partir de los datos solicitados para realizar la explotación, se tiene que realizar la búsqueda en las tablas de estas bases de datos.

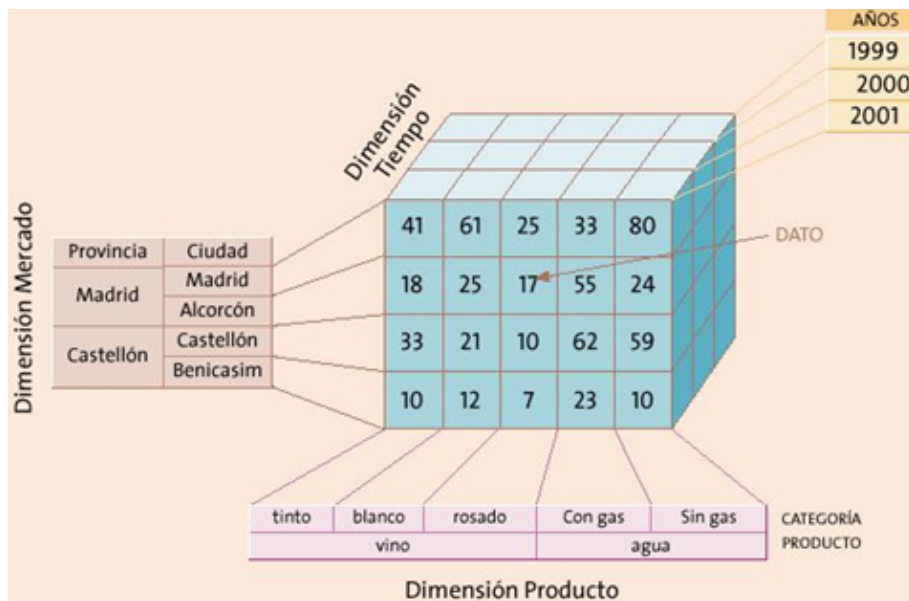
En segundo lugar, se crea un almacén de datos (Datawarehouse -DWH-), que son tablas dispuestas de una determinada forma o diseño (en este caso forma de estrella), donde se encuentran las tablas de hechos y las tablas de dimensiones. Las tablas de hechos son el corazón de un esquema en estrella, almacenan campos claves con los que se unen a las tablas de dimensiones, además incluyen métricas del negocio; contienen millones de registros. Las tablas de dimensiones son las que ofrecen más información sobre características de las tablas de hechos, normalmente contienen pocos registros.

Figura 4.2: Esquema en estrella. Recuperado de Pesquera (2019)



Cada tipo de diseño (en estrella o en copo de nieve), que permite realizar consultas multidimensionales, debe seleccionarse según las necesidades de los datos y del negocio. Esta estructura en la base de datos es la clave para realizar los esquemas de Mondrian OLAP, conocidos como cubos de Mondrian. (Hyde, 2011)

Figura 4.3: Cubo de Mondrian. Recuperado de Urquizu (2019)



Estos esquemas (que son archivos .XML) lo que hacen es “traducir” el diseño unidimensional de las tablas relacionales a un diseño multidimensional de forma que puedan ser entendidos por los servidores de Business Intelligence (BI) para la explotación de los datos.

Para crear el esquema de Mondrian, primero se tiene que decidir cuáles son las tablas de hechos y las tablas de dimensiones. Un ejemplo concreto sobre un diseño realizado en la CEI ha sido una tabla de hechos que contiene el identificador de centros, el número de alumnos, el número de grupos, el nivel escolar, la naturaleza del centro (público, privado o concertado), etc. La tabla de dimensiones contiene los propios valores de la tabla anterior, por ejemplo, para la columna naturaleza del centro de la tabla de hechos, existe una dimensión que contenga los valores público, privado o concertado.

Una vez que se tiene el diseño de tablas creadas, ya se puede realizar la fase de extracción, transformación y carga (ETL); con la que se extraen los datos de la fuente de información, se producen las transformaciones oportunas y posteriormente dichos datos se almacenan en las tablas de hechos y/o dimensiones. En esta fase es donde se limpian los datos, eliminando las filas que no contengan todos los campos o rellenando los campos con algún valor en caso que fuera necesario.

Posteriormente, con los datos limpios y cargados en las tablas de hechos y dimensiones, se tiene que subir el esquema creado al servidor y establecer una conexión con la base de datos, a partir de aquí, el trabajo es del motor OLAP del servidor, quien se encarga de traducir las opciones introducidas por el usuario en los cuadros de mandos a complejas

consultas en la base de datos. Esta información se va a mostrar en las herramientas de acceso disponibles.

En resumen, y a partir de los datos almacenados en el “Datawarehouse”(DWH), la investigación que se lleva a cabo en este TFM se centra en analizar dichos datos. Pudiéndose recurrir a las fuentes de información originales para obtener nuevos datos en caso necesario.

4.3. Diseño de la minería de datos

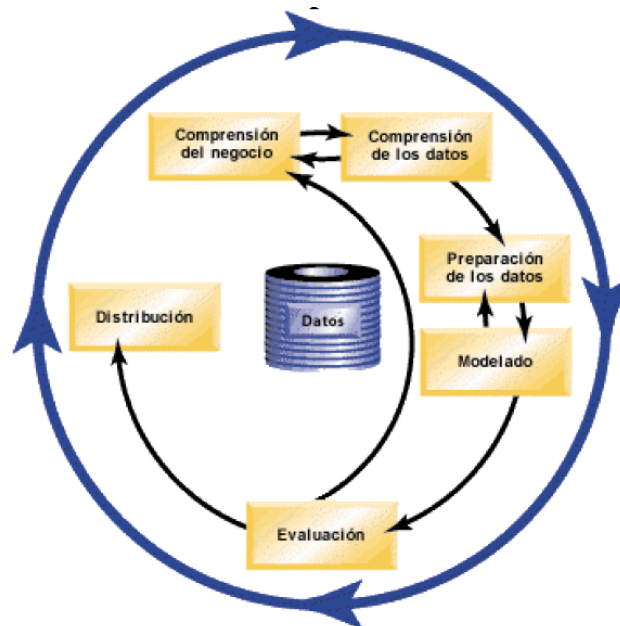
Una vez que se tiene claro la arquitectura principal del sistema, se establece el camino a seguir a la hora de realizar la investigación. Para ello, se utiliza la metodología CRISP-DM. *Manual CRISP-DM de IBM SPSS Modeler (2012)*

La metodología CRISP-DM tiene como objetivo orientar los proyectos de minería de datos.

- Como metodología: incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada una de las fases y una explicación de las relaciones entre las tareas.
- Como modelo de proceso: ofrece un resumen del ciclo vital de la minería de datos.

La metodología CRISP-DM establece un proceso genérico para satisfacer los objetivos deseados y contemplar la realización de la vigilancia e inteligencia. Este proceso se divide en distintas etapas básicas.

Figura 4.4: Fases del ciclo de vida de CRISP-DM. Recuperado de *Manual CRISP-DM de IBM SPSS Modeler* (2012).



El ciclo de vida de CRISP-DM está compuesto de seis fases. La secuencia de estas no es estricta, es más, la mayoría de proyectos avanzan y retroceden entre fases si es necesario. En la figura 4.4 se puede observar cada fase:

1. **Comprensión 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. **Comprensión 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. **Modelado.** Se debe desplegar un gran número de modelos y quedarse con aquellos que devuelvan valores óptimos para los datos utilizados.
5. **Evaluación.** Debe evaluarse y probarse los modelos. Deben compararse entre sí y comprobar que son útiles para los datos expuestos.
6. **Distribución.** Se realizan actividades usando los modelos seleccionados en el proceso de la toma de decisión.

En los próximos puntos se va a describir las actividades que se realizan en cada una de estas fases.

4.3.1. Comprensión del negocio

La primera tarea a realizar en el ciclo de CRISP-DM es obtener la máxima información posible de los objetivos de esta investigación. Desde la CEI se comunica, mediante reuniones, la información disponible y las necesidades actuales.

Una de las necesidades de la CEI es obtener la máxima información sobre la situación actual de la educación en la comunidad de Madrid. La CEI posee una gran cantidad de datos de alumnos y centros a lo largo del tiempo.

Por tanto, para sacar el mayor beneficio de los datos, se plantea el uso de herramientas y técnicas que posibiliten la obtención información no solo del momento actual, sino también de la evolución a lo largo de los últimos años.

Una de las actividades que se realizan es obtener el número de grupos y alumnos por centro, año, DAT, nivel educativo, etc.

Otra de las actividades que se realizan es obtener gráficos sobre la evolución de alumnos con necesidades especiales, alumnos de minorías étnicas e incluso porcentaje y nacionalidad de alumnos extranjeros. Véase Anexo B.C.

4.3.2. Comprensión de los datos

Esta fase implica estudiar detalladamente los datos disponibles. Es esencial para evitar problemas inesperados durante la fase siguiente.

Para realizar esta fase, debemos tener en cuenta dos consideraciones relacionadas. La primera consideración es la identificación de necesidades de información y la segunda es la identificación de fuentes internas y externas.

Identificación de necesidades de información

Para realizar la identificación de las necesidades de información se va a partir de varios factores como son:

- las demandas esperadas o manifestadas por (en este caso) una unidad de la CEI.
- el análisis, la evolución de productos, procesos, materiales y tecnologías en el ámbito de la minería de datos educativa.

Identificación de fuentes internas y externas de información

Teniendo en cuenta las principales necesidades de información, se debe identificar las fuentes de información y recursos disponibles ya sean internos o externos a la organización. En este caso, se utilizan las siguientes fuentes:

- Fundamentalmente se utilizan documentos y recursos internos de la organización como: repositorios documentales, carpetas locales, bases de datos, etc.
- Personas con conocimientos o experiencias relacionadas con la necesidad de información. En este aspecto se realizan distintas reuniones con las personas encargadas de esta unidad de la consejería de educación. Para ello se realizarán reuniones con estos responsables. A partir de estas reuniones se obtendrán las fuentes de información.
- Documentación técnica como reglamentaciones, especificaciones, propiedad industrial e intelectual o normas.
- Resultados de análisis existentes sobre las tendencias de futuro preferentemente en el ámbito educativo.

La información fundamentalmente se encuentra en bases de datos internas, no obstante, se va a acceder a bases de datos externas en caso de necesidad para cumplimentar la información.

En este aspecto, se debe recurrir a la ayuda de personas con conocimientos sobre el estado de las bases de datos. Como cualquier organización, la CEI maneja grandes volúmenes de datos, por tanto, se debe tener conocimiento sobre donde se puede encontrar la información que satisfaga las necesidades.

El desconocimiento del estado de las bases de datos conlleva la inversión de una gran cantidad de tiempo en la búsqueda de los datos relevantes, por ello, el conocimiento de la situación actual es de gran importancia.

De esta fase se espera localizar todos los datos que posteriormente se preparan y se utilizan en el modelado.

4.3.3. Preparación de los datos

Una vez que se tienen claros los datos que se utilizan, se procede a realizar la preparación para poder utilizarlos en la fase de modelado.

Algunas de las actividades que se realizan en esta fase son: la fusión de conjuntos y/o registros de datos, la selección de una muestra de un subconjunto, la agregación de registros, por contra la derivación de nuevos atributos a partir de anteriores, la eliminación o sustitución de valores en blanco o ausentes y por último la división de datos de prueba y entrenamiento.

Además, también se va a estudiar la existencia de datos perdidos y errores en estos.

Para realizar este tratamiento de datos se utilizará la técnica de ETL (extracción, transformación y carga) que consiste básicamente en obtener los datos de la fuente de

origen (bases de datos, ficheros Excel, ficheros JSON, etc.), seleccionar aquellos datos que convengan al estudio, transformarlos según las necesidades que se tenga y depurarlos (evitando así datos erróneos). (Prakash y Rangdale, 2017) (Matos, Chalmeta, y Coltell, 2006),(Gour, Sarangdevot, Tanwar, y Sharma, 2010). Para realizar este tratamiento, se utiliza Pentaho BI, que es un conjunto de programas libres para realizar entre otras muchas actividades, las técnicas de ETL. Concretamente, se utiliza la herramienta Spoon para desarrollar esta técnica. Una vez que se tienen los datos limpios y estructurados, se pueden realizar dos operaciones:

1. En primer lugar, se pueden almacenar dichos datos en una base de datos y seguir utilizando Pentaho BI para poder crear cuadros de mandos e informes o análisis OLAP.
2. En segundo lugar, se puede almacenar la información en un texto plano para poder trabajar con herramientas de análisis descriptivo y predictivo. Estos análisis se realizan a través del entorno y lenguaje de programación R, que es una referencia en el ámbito de la estadística.

Análisis Exploratorio

La primera actividad en un análisis exploratorio es estudiar el tipo de datos de cada variable a investigar, se debe clasificar las variables según sean categóricas (dicotómicas o polinómicas) o numéricas (discretos o continuos). El tipo de datos permite decidir qué tipo de análisis estadístico utilizar. Una vez que se tienen claro el tipo de datos utilizados, se utilizan los principales estadísticos como la media, la mediana, las desviaciones típicas, etc. Posteriormente se va a utilizar la matriz de varianzas y covarianzas, que indicaran la variabilidad de los datos y la información sobre las posibles relaciones lineales entre las variables.

Por otro lado, se va a estudiar la correlación de las variables mediante la matriz de correlación. Esta matriz contendrá los coeficientes de correlación.(Diazaraque, s.f.). La matriz de correlación, se utilizará fundamentalmente por pares entre las variables y la variable a predecir.

También se va a estudiar la matriz de correlaciones parciales, que estudia la correlación entre pares de variables eliminando el efecto de las restantes.(Diazaraque, s.f.)

Los datos categóricos se representan en tablas de frecuencias, gráficos de barras y gráficos de sectores. Los datos numéricos se representan mediante histogramas, boxplot y diagramas QQ-Plot o Grafico Cuantil-Cuantil. (Orellana, 2001)

Mediante el boxplot se puede observar aspectos como la posición, dispersión, asime-

tría, longitud de colas y los datos anómalos (outliers). El QQ-plot se va a utilizar para evaluar la cercanía de los datos a una distribución. (Orellana, 2001)

Por otro lado, se va a complementar el análisis descriptivo mediante el aprendizaje no supervisado, donde también se extraerán otras características de los datos.

4.3.4. Modelado

Una vez terminado el análisis descriptivo, se va a realizar un análisis predictivo. Se debe tener en cuenta, que, dentro de la ciencia de datos, existen técnicas de aprendizaje automáticas, cuyo objetivo es la construcción de un sistema que sea capaz de aprender a resolver problemas sin la intervención de un humano. (Marín, 2018).

Las técnicas de aprendizaje tienen como resultado un modelo para resolver una tarea dada. Los modelos son una representación de la realidad basado en un intento descriptivo de relacionar un conjunto de variables con otro.

Los modelos predictivos son de dos tipos: regresión, que son capaces de predecir una respuesta cuantificable; y de clasificación, que son capaces de predecir respuesta categóricas.

En este TFM se utilizan modelos de regresión, puesto que la variable que se predice es del tipo cuantitativo.

Aprendizaje automático

El **aprendizaje supervisado** consiste en la búsqueda de patrones en datos históricos relacionando todas las variables con una especial (conocida como variable objetivo). Los algoritmos que se utilizan en el aprendizaje supervisado se encarga de buscar patrones en los datos. A este proceso se conoce como entrenamiento de los datos. Una vez que se tienen los patrones, se aplican a los datos de prueba. Los datos de entrenamiento suelen ser una selección aleatoria y única de los datos históricos de un 70 % del total. Los datos de prueba son el restante 30 %. (Páez, 2017). Algunos de los algoritmos que se utilizan son:

1. Árboles de decisión

Se basa en el descubrimiento de patrones a partir de ejemplos. Un árbol de decisión está formado por un conjunto de nodos (de decisión) y de hojas (nodos-respuesta).

Los nodos están asociados a los atributos y tiene varias ramas que salen de él (dependiendo de los valores que tomen la variable asociada). Estos nodos pueden asemejarse a preguntas que, dependiendo de la respuesta que conlleve, se tomara un

flujo en las ramas salientes.

Los nodos respuesta están asociados a la clasificación que se desea proporcionar, devolviendo así la decisión del árbol con respecto al ejemplo de entrada utilizado.

Figura 4.5: Funcionamiento Árboles Decisión. Recuperado de Sayad (2019)



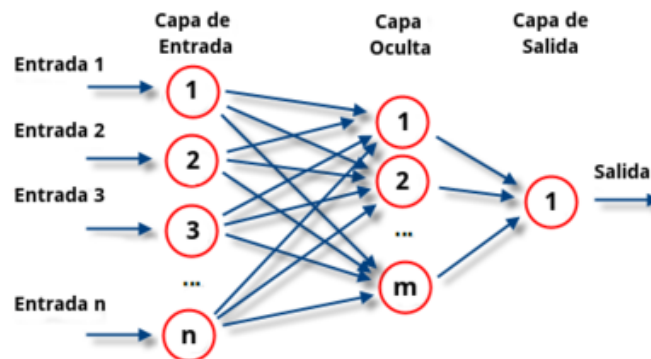
2. Regresión Lineal Simple

Es un modelo de regresión en donde una función lineal representa la relación existente entre una variable dependiente y su respectiva variable independiente. El modelo adopta la forma de $y = ax + b$, donde y es la variable dependiente, x es la variable independiente, y a , b son los coeficientes de la recta. Estos coeficientes, bajo los criterios de minimización como el de mínimos cuadrados, ofrecen el mejor ajuste de los datos de entrada. (Delgado, 2019)

3. Redes Neuronales

Las redes neuronales son un algoritmo de inteligencia artificial que se inspira en los mecanismos presentes en la naturaleza. Las neuronas envían señales eléctricas de manera fuerte o débil a otras neuronas. La combinación de todas las conexiones entre neuronas es lo que genera el conocimiento. Estas señales se envían cuando existe unos estímulos (inputs) externos a través de los sentidos. A lo largo de la vida, las neuronas aprenden que deben hacer a partir de dichos estímulos y, por lo tanto, los seres vivos aprenden a actuar ante distintas señales y situaciones. El funcionamiento de las redes neuronales en la inteligencia artificial es similar.

Figura 4.6: Red Neuronal. Recuperado de Piqueras (2017)



Como se puede observar en la figura 4.6, la primera fila (con neuronas de color rojo), se conocen como nodos de entrada y son aquellos que se encargan de recoger la información. Los nodos en la gama azul son los que se conocen como nodos de salida. Los nodos situados en el medio son aquellos que se encargan de hacer el aprendizaje, y se conocen como nodos ocultos.

En primer lugar, se obtiene la información a partir de los nodos de entrada, una vez que se tiene la información, se envía a las capas ocultas, que se activan o no dependiendo del aprendizaje previo. Los nodos ocultos se activan dependiendo de una serie del resultado de unas operaciones matemáticas. Si los nodos se activan, entonces enviarán la información a la siguiente capa.

4. Bosques aleatorios

Los bosques aleatorios son un método que se encarga de combinar los resultados de árboles de decisión independientes.

Algunas características son:

- Gran precisión.
- Eficiente para grandes bases de datos.
- Aporta estimaciones sobre la importancia de las variables en la clasificación.
- Tiene un método eficaz para la estimación de los datos faltantes y mantiene la precisión cuando falta una gran parte de los datos.

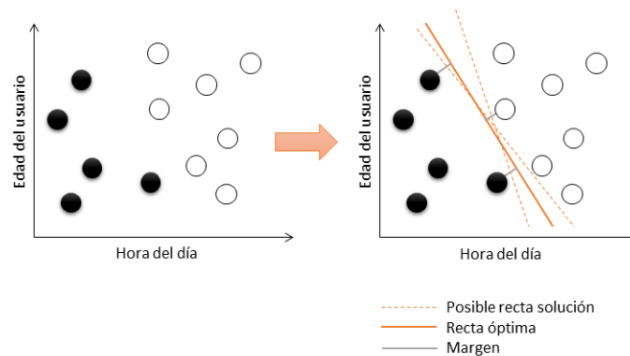
5. Maquinas de Vectores Soporte (SVM)

Constituyen un método basado en el aprendizaje para la resolución de problemas de clasificación y regresión. Para ello, recibe unas entradas y obtienen unas salidas. Busca por tanto la curva/linea que modele la tendencia de los datos.

Por ejemplo, si tenemos un anuncio en una pagina web y queremos analizar la edad

y la hora del día del usuario que hace clic o no en dicho anuncio, con SVN se obtiene la "superficie optima" que delimitara el comportamiento (clic-noclic) de un determinado usuario. (Álvarez, 2016)

Figura 4.7: Ejemplo SVM. Recuperado de Álvarez (2016)

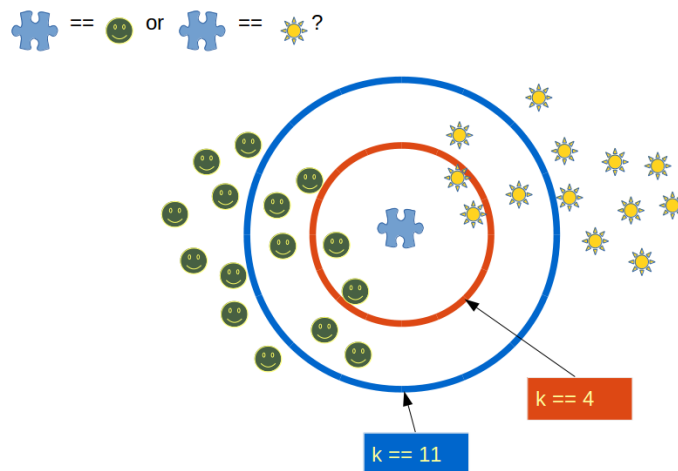


6. K-Vecinos-cercanos

K-vecinos-cercanos (conocido también como K-NN) es un algoritmo de aprendizaje supervisado en el que, a partir de unos datos iniciales, es capaz de clasificar todas las nuevas instancias.

La idea es que el algoritmo clasifica cada dato nuevo en el grupo que corresponda, según cual sea el grupo vecino (de los k grupos) más próximo. Por tanto, calcula la distancia del elemento nuevo a cada uno de los existentes e indica a que grupo debe permanecer este nuevo elemento según la menor distancia.

Figura 4.8: KNN. Recuperado de Klein (2018)



Criterio de selección

Una vez que se seleccionan las variables y los algoritmos a estudiar, es hora de realizar el propio modelado. Al realizar el modelado, debemos 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). (Martínez 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. (Martínez et al., 2009)

4.3.5. Evaluación

En esta fase de la metodología, se diferencian varias partes. La primera parte es la evaluación del propio modelo respecto a otros, por lo que se utilizarán las métricas de precisión. La segunda parte va a ser la evaluación de la propia Unidad de Secundaria la que evalúe los resultados de predicción obtenidos para un determinado curso con los existentes en la realidad para dicho curso.

Métricas de precisión El error absoluto medio (MAE) y el error cuadrático medio (RMSE) son dos de las métricas más comunes utilizadas para medir la precisión de las variables continuas en los modelos de regresión.

Error absoluto medio (MAE): mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Es el promedio sobre la muestra de prueba de la diferencia absoluta entre la predicción y la observación real.

Figura 4.9: Fórmula de MAE. Recuperado de <https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

Error cuadrático medio (RMSE): es una regla de puntuación cuadrática que mide la magnitud promedio del error. Es la raíz cuadrada del promedio de las diferencias cuadradas entre la predicción y la observación real.

Figura 4.10: Fórmula de RMSE. Recuperado de <https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Se debe destacar que cuanto menor sea el error, más se acercan los datos predichos a la realidad.

4.3.6. Distribución

La fase de distribución considera la planificación y control de la distribución de los resultados. Debe tener en cuenta la realización de un informe final.

Respecto a la fase de distribución, se utilizan los modelos generados para predecir datos de otros cursos. Concretamente se utiliza el modelo que mayor precisión obtiene con los datos aportados.

Concretamente, los datos utilizados en la predicción son aquellos del curso 2016/2017. A partir de estos datos, se obtienen varios modelos. Una vez que se selecciona el mejor modelo, ya se puede utilizar otro conjunto de datos. En este caso, el conjunto de datos a utilizar es el del curso 2017/2018.

Esta fase, por tanto, aplicada a un entorno empresarial, debería indicar que modelos deben integrarse en el sistema. En este entorno académico, simplemente se debe indicar el mejor modelo y una comparativa de todos los modelos utilizados.

4.4. Herramientas utilizadas

4.4.1. Suite de Pentaho BI

Para el desarrollo del proyecto, se tiene en cuenta la posibilidad de utilizar la herramienta de Pentaho Business Analytics, que es una suite de herramientas para la explotación de datos. Esta suite posee las herramientas que se usan y son las siguientes: Spoon, Pentaho SchemaWorkbench y Pentaho Metadata Editor.

4.4.2. Lenguaje R y RStudio

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 graficación. 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).

El Paquete *Caret*

A la hora de realizar esta investigación se utilizan distintos paquetes que proporciona R, sin embargo, el paquete fundamental y de mayor relevancia es el paquete *caret*.

El paquete **caret** es un conjunto de funciones que intenta agilizar el proceso de creación de modelos predictivos. El paquete contiene herramientas para: división de datos, pre-procesamiento, selección de características, ajuste del modelo mediante re-muestreo, estimación de importancia variable, así como otras funcionalidades. (Kuhn, 2019)

5 | ANÁLISIS DE RESULTADOS

Una vez que se establece el diseño a seguir en este TFM, se pasa a realizar el análisis de resultados.

Este análisis de resultados se va a diferenciar entre aquellos resultados obtenidos en el análisis exploratorio y los obtenidos en la aplicación de modelos predictivos.

5.1. Análisis exploratorio de datos

Antes de comenzar con el análisis, el lector puede observar en la tabla A.1 del Apéndice A.A las variables utilizadas en esta investigación. En este mismo Anexo nos encontramos también los estadísticos, que se pueden observar en la figura A.1 del Apéndice A.A.

Una vez observados los estadísticos, los vamos a representar utilizando los diagramas de caja (box-plot). Con estos diagramas vamos a observar además los datos atípicos (outliers). Un resumen de todos los diagramas de cajas se puede observar en la figura A.2 del Apéndice A.A.

Como se puede observar en la figura A.2, las variables de números de alumnos, número de unidades y ratio contienen datos anómalos. Por ejemplo, la variable "número de alumnos", como se puede apreciar en los estadísticos de la figura A.1 tiene una media de 74 alumnos. No obstante, hay niveles educativos que tiene hasta casi 700 alumnos. Esto se debe a que los centros relativos a estos alumnos tienen la modalidad de distancia para esos niveles educativos, esto implica que pueden tener un gran número de unidades debido al gran número de alumnos que se matriculan para esta modalidad. Ocurre, por tanto, exactamente lo mismo con el número de grupos y la ratio. Estos datos anómalos no se van a eliminar puesto que tienen sentido en esta investigación.

Una vez estudiado los estadísticos y los datos anómalos, se va a realizar un estudio sobre la normalidad de los datos. En este aspecto es importante destacar que después de realizar el test de Mardia se observa que los datos utilizados rechazan la hipótesis de normalidad. Este rechazo implica una mayor cota de error a la hora de predecir los datos. Además, existen determinados modelos predictivos que no asumen que sus datos provengan de determinadas distribuciones.

Uno de los aspectos más importantes en el análisis exploratorio de datos es la correlación existente entre las variables. En la figura A.10 del Anexo A.B.3 se puede observar como existe una gran correlación entre las variables de número de alumnos, número de grupos y grupos a predecir y otra gran correlación entre la variable comedor, el carácter

genérico y la naturaleza del centro. En la figura A.11 se pueden observar las mayores correlaciones entre variables ordenadas de mayor a menor.

Uno de los fines es obtener las variables que mejor correlacionan con la variable a predecir (número de grupos). En la figura A.12 se observa como las variables número de alumnos y número de grupos son las que mayor correlación tienen con la variable a predecir. No sorprende puesto que son variables determinantes en la predicción. Vemos en la figura A.13 que ambas variables tienen una relación positiva que implica que si una aumenta, la otra también.

5.2. Análisis predictivo

Una vez que se realiza el análisis descriptivo, se procede a realizar el análisis predictivo.

En primer lugar, se debe tener en cuenta las variables que se van a utilizar en los modelos, para ello, se va a utilizar Random Forest como algoritmo para obtener la importancia de las variables. En la figura A.15 podemos observar que las variables más importantes a la hora de mejorar la precisión en el modelo son: número de unidades y alumnos.

Se utiliza también otro algoritmo que es el conocido como Regresión Paso a Paso. Para este algoritmo se utilizan 3 estrategias en este TFM. En las estrategias descritas en el Apéndice A.C.2 resultando que dos de estas estrategias utilizan como máximo 5 variables (naturaleza del centro, número de curso, número de unidades, nivel de enseñanza y ratio). Estas variables se tienen en cuenta para la realización y comparación de modelos. Utilizando estas variables es cuando el modelo obtiene la mayor precisión. Debe destacarse que la mayoría de estas variables son las que mayor correlación tienen con la variable a predecir.

En esta investigación, además de realizar la comparación de los modelos, también se va a realizar una comparación teniendo en cuenta todas las variables y únicamente las 5 variables seleccionadas utilizando los algoritmos, ya comentados, de eliminación de variables.

Los resultados obtenidos, que pueden observarse en la tabla B.1 en el Apéndice B puede observarse como el algoritmo que mayor precisión ha obtenido ha sido el Árbol de decisión utilizando todas las variables. No obstante, este algoritmo requiere de un gran tiempo de entrenamiento, por lo que la regresión lineal puede ser una gran opción. Como puede observarse en la figura B.1 del Apéndice B existen varios modelos que obtienen precisiones similares (Redes neuronales, Árbol de Decisión y Regresión Lineal) por tanto es conveniente tener en cuenta los tiempos de entrenamiento que se pueden ver en la figura

B.2.

En una nueva investigación se debe tener en cuenta el tiempo de entrenamiento, sin embargo, como ya se ha realizado el entrenamiento, se va a obtener la predicción utilizando el algoritmo de Árbol de decisión (xgbTree).

Una vez seleccionado el algoritmo a usar en la predicción, se puede observar una situación que realmente ocurre, y es que pocos centros son los que aumentan o disminuyen de unidades. Concretamente, de 4436 datos con los que se trabajan, únicamente 52 han cambiado. Algunos de estos grupos se observan en la figura B.5 del Anexo B.

De estos 52 grupos que cambian: 30 de ellos son grupos que van a disminuir su número y el restante 22, son grupos que van a aumentar su número. Esto podría deberse a una posible despoblación, no obstante, debería investigarse dichos datos a fondo.

6 | CONCLUSIONES

6.1. Conclusiones

Una vez que se ha realizado el estudio y se han obtenido los resultados, es hora de realizar una pequeña valoración en forma de conclusiones.

En primer lugar, la realización de este TFM ha tenido como objetivo satisfacer las necesidades de la CEI, realizando un estudio para la obtención de modelos que sean capaces de conseguir una gran predicción.

Para realizar estas predicciones se han utilizado una serie de algoritmos como las redes neuronales, regresión lineal, bosques aleatorios, árboles de decisión, K-vecinos cercanos y soporte de máquinas vectoriales. Siendo el árbol de decisión el que mayor precisión ha aportado. Sin embargo, el “Script” que automatiza el proceso se realiza utilizando el algoritmo de regresión lineal, ya que es el que menor tiempo tarda en entrenar y aporta unos resultados similares al árbol de decisión.

Además, se ha podido observar como la eliminación de las variables en los modelos no ha alterado demasiado la precisión aunque si se ha notado diferencia en los tiempos de predicción (puesto que los algoritmos no trabajan con tantos datos). Obviamente existen variables que desde el inicio, a priori, no aportan gran información sobre el conjunto de datos, sin embargo, se decide tenerlas en cuenta por si aportaban información oculta.

Por último, mediante este TFM se ha obtenido además información intrínseca de las variables, que pueden servir para la realización de otras investigaciones.

6.2. Lineas Futuras

Existen ciertos puntos que se deben considerar en futuras investigaciones o en la ampliación de este trabajo.

En primer lugar, es necesario destacar el uso de otras variables que no se han estudiado en este TFM, por ejemplo, aspectos físicos como el acceso al centro, su comunicación con grandes vías urbanas, etc. Además, deben investigarse otras variables que no se han considerado por falta de datos, como por ejemplo la tasa de aprobados o de suspensos para un grupo de un nivel determinado.

En segundo lugar, se deben tener en cuenta otros modelos y otros parámetros para crear los modelos como por ejemplo la regresión logística, métodos robustos, combinación de modelos, etc.

Por último, se debe tener en cuenta la realización de una interfaz para sustituir el

“Script” creado por un Software en el que los usuarios simplemente tengan que cargar la plantilla Microsoft Excel en dicho Software y que este devuelva dicha plantilla rellena con las predicciones.

7 | REFERENCIAS

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 stacking 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., Mercer, 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.

Consejería de Educación e Investigación. (2018). *Instrucciones de la dirección general de educación infantil, primaria y secundaria sobre la planificación del próximo curso escolar 2018/2019 en los centros públicos que imparten eso y bachillerato, creación de nuevos centros y modificación de la red, implantación y autorización de enseñanzas y propuesta de grupos*.

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>

Delgado, R. (2019). *Introducción a los modelos de regresión en r*.

Diazaraque, J. M. M. (s.f.). *Tema 2: Estadística descriptiva multivariante*. Descargado de <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema2am.pdf>

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>

Gour, V., Sarangdevot, S., Tanwar, G. S., y Sharma, A. (2010). Improve performance of extract, transform and load (etl) in data warehouse. *International Journal on Computer Science and Engineering*, 2(3), 786–789.

Hyde, J. (2011). *Mondrian documentation*. Descargado de <https://mondrian.pentaho.com/documentation/schema.php>

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.

kassambara. (2018). *Stepwise regression essentials in r*. Descargado de <http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/>

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.

Klein, B. (2018). *k-nearest-neighbor classifier*. Descargado de https://www.python-course.eu/k_nearest_neighbor_classifier.php

Kuhn, M. (2019). *The caret package*. Descargado de <http://topepo.github.io/caret/index.html>

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).

Álvarez, J. (2016). *Machine learning y support vector machines: porque el tiempo es dinero*. Descargado de <https://www.analiticaweb.es/machine-learning-y-support-vector-machines-porque-el-tiempo-es-dinero-2/>

Manual crisp-dm de ibm spss modeler. (2012).

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)

Martinez, 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 estadísticos de rendimiento. *XX Jornadas de Paralelismo*.

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

Matos, G., Chalmeta, R., y Coltell, O. (2006). Metodología para la extracción del conocimiento empresarial a partir de los datos. *Información tecnológica*, 17(2), 81–88.

Mohamad, S. K., y Tasir, Z. (2013). Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97, 320 - 324. Descargado de <http://www.sciencedirect.com/science/article/pii/S1877042813036859> (The 9th International Conference on Cognitive Science) doi: <https://doi.org/10.1016/j.sbspro.2013.10.240>

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/>

Orellana, L. (2001). *Estadística descriptiva*. Descargado de http://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf

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>

Pesquera, C. (2019). *Qué son las tablas de hechos y de dimensión*. Descargado de <http://carlospesquera.com>

Páez, A. M. (2017). *Conceptos básicos del aprendizaje supervisado (para personas no técnicas)*. Descargado de <https://medium.com/@manguart/machine-learning-conceptos-básicos-del-aprendizaje-supervisado-para-personas-no-técnicas-142bbb222140>

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/>

- Piqueras, V. Y. (2017). *¿qué es y para qué sirve una red neuronal artificial?* Descargado de <https://victoryepes.blogs.upv.es/2017/01/07/que-es-y-para-que-sirve-una-red-neuronal-artificial/>
- Prakash, G. H., y Rangdale, P. (2017). Etl data conversion: Extraction transformation and loading data conversion. *International Journal Of Engineering And Computer Science*, 22545–22550.
- 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.
- Sayad, S. (2019). *An introduction to data science*. Descargado de https://www.saedsayad.com/data_mining_map.htm
- 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).
- Urquizu, P. (2019). *Business intelligence facil*. Descargado de <https://www.businessintelligence.info>
- 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.

ANEXOS

A | Gráficas del Análisis Exploratorio de datos

A.A. Definición de Variables

En este apartado se definen las variables y sus valores que se van a utilizar en este TFM.

Tabla A.1: Nomenclatura de las Variables

Nombre de la variable	Significado	Valores
CDGENERICO	Indica el código genérico del centro	<ul style="list-style-type: none"> ■ CIMELPR PRSEC = 16 ■ CIMPR SEC = 17 ■ CEPA = 31 ■ CREI = 39 ■ IES = 42 ■ CPR ES = 45 ■ SIES = 47 ■ CPR FPE = 58 ■ CP IFP = 68 ■ CP INF-PRI-SEC = 70 ■ CPR INF-PRI-SEC = 72 ■ CPR PRI-SEC = 73
CDNATURALEZA	Indica la naturaleza del centro. Centros públicos y privados	<ul style="list-style-type: none"> ■ Público = 1 ■ Privado = 2
CDPOSTAL	Código postal del centro	
ITCOMEDOR	Disponibilidad de comedor en el centro	<ul style="list-style-type: none"> ■ No = 1 ■ Si = 2
ITTRANSPORTE	Disponibilidad de transporte	<ul style="list-style-type: none"> ■ No = 1 ■ Si = 2
ITBILINGUE	Disponibilidad de bilingüismo	<ul style="list-style-type: none"> ■ No = 1 ■ Si = 2
CD_NIVEL	Nivel educativo del grupo	<ul style="list-style-type: none"> ■ Bachillerato = 5 ■ Educación Secundaria Obligatoria = 13 ■ Módulos Profesionales = 14 ■ Formación Profesional GM = 15 ■ Formación Profesional GS = 16
NM_CURSO	Curso del nivel educativo del grupo	<ul style="list-style-type: none"> ■ Primero = 1 ■ Segundo = 2 ■ Tercero = 3 ■ Cuarto = 4
NM_UNIDADES	Número de grupos para un determinado nivel y un número de curso	
NM_ALUMNOS	Número de alumnos para un determinado nivel y numero de curso	
RATIO	Ratio de alumnos por grupo para cada nivel y numero de curso.	
GRUPO_PREDECIR	Grupo a predecir	

Las variables existentes en las bases de datos originales son: CDGENERICO, CDNATURALEZA, CDPOSTAL, ITCOMEDOR, ITTRANSPORTE, ITBILINGUE, CD_NIVEL y NM_CURSO.

El resto de variables se obtienen a partir de las transformaciones correspondientes. Por ejemplo, las variables NM_UNIDADES NM_ALUMNOS y GRUPO_PREDECIR se obtienen de realizar consultas que agrupan por determinados aspectos. Sin embargo,

la variable **RATIO**, que indica la ratio de alumnos para un determinado grupo, se ha obtenido dividiendo la variable de **NM_ALUMNOS** entre 30 o 35 (dependiendo del nivel educativo), y este resultado se ha dividido entre el número de grupos **NM_GRUPOS** para ese nivel educativo. Estos ratios se regulan mediante el *Real Decreto 132/2010, de 12 de febrero, por el que se establecen los requisitos mínimos de los centros que impartan las enseñanzas del segundo ciclo de la educación infantil, la educación primaria y la educación secundaria*. para los niveles de enseñanza de Educación Infantil, Educación Primaria, Educación Secundaria Obligatoria y Bachillerato y el *Real Decreto 1147/2011, de 29 de julio, por el que se establece la ordenación general de la formación profesional del sistema educativo* para los niveles de Formación Profesional. Siendo las ratios los siguientes: 30 para Educación Secundaria Obligatoria y Formación profesional y 35 para Bachillerato.

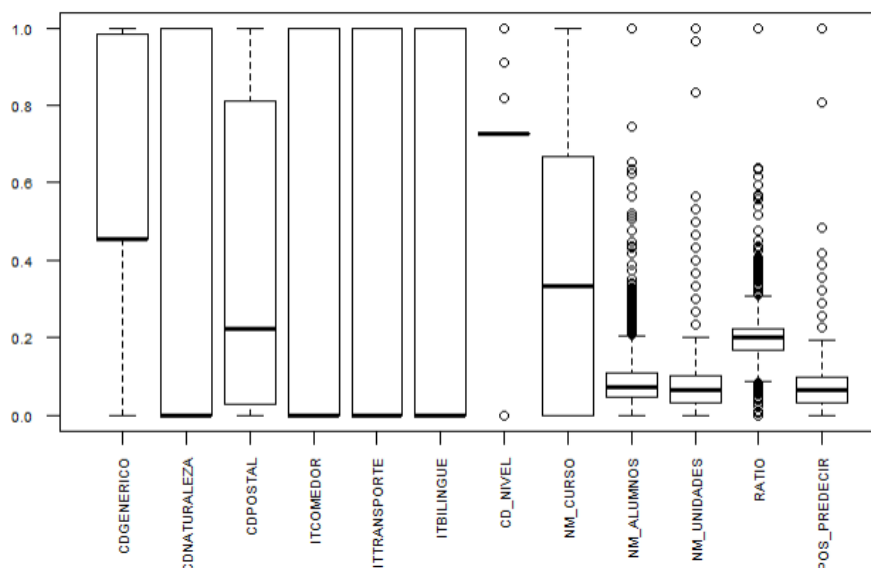
A.B. Análisis exploratorio

A.B.1. Estadísticos más relevantes

Figura A.1: Estadísticos de las variables. Elaboración propia

CDGENERICO	CDNATURALEZA	CDPOSTAL	ITCOMEDOR	ITTRANSPORTE	ITBILINGUE	CD_NIVEL
Min. :16.0	Min. :1.00	Min. :28001	Min. :1.00	Min. :1.00	Min. :1.00	Min. : 5.0
1st Qu.:42.0	1st Qu.:1.00	1st Qu.:28030	1st Qu.:1.00	1st Qu.:1.00	1st Qu.:1.00	1st Qu.:13.0
Median :42.0	Median :1.00	Median :28223	Median :1.00	Median :1.00	Median :1.00	Median :13.0
Mean :54.9	Mean :1.44	Mean :28378	Mean :1.46	Mean :1.27	Mean :1.48	Mean :12.2
3rd Qu.:72.0	3rd Qu.:2.00	3rd Qu.:28806	3rd Qu.:2.00	3rd Qu.:2.00	3rd Qu.:2.00	3rd Qu.:13.0
Max. :73.0	Max. :2.00	Max. :28991	Max. :2.00	Max. :2.00	Max. :2.00	Max. :16.0
NM_CURSO	NM_ALUMNOS	NM_UNIDADES	RATIO	GRUPOS_PREDECIR		
Min. :1.00	Min. : 0	Min. : 1.00	Min. :0.00	Min. : 1.0		
1st Qu.:1.00	1st Qu.: 47	1st Qu.: 2.00	1st Qu.:0.73	1st Qu.: 2.0		
Median :2.00	Median : 74	Median : 3.00	Median :0.88	Median : 3.0		
Mean :2.15	Mean : 85	Mean : 3.19	Mean :0.85	Mean : 3.2		
3rd Qu.:3.00	3rd Qu.: 114	3rd Qu.: 4.00	3rd Qu.:0.98	3rd Qu.: 4.0		
Max. :4.00	Max. :1035	Max. :31.00	Max. :4.35	Max. :32.0		

Figura A.2: Diagrama de cajas normalizado. Elaboración propia



A.B.2. Análisis de normalidad

En primer lugar, se pueden ver la densidad de las variables individualmente de observar a simple vista si cumplen o no con una distribución normal.

El análisis de normalidad es importante, ya que la falta de normalidad afecta a los modelos que se vayan a utilizar, por ejemplo, a su intervalo de confianza. Dicha falta de normalidad no suele afectar a las predicciones puntuales, pero si a los intervalos o “errores” en la predicción.

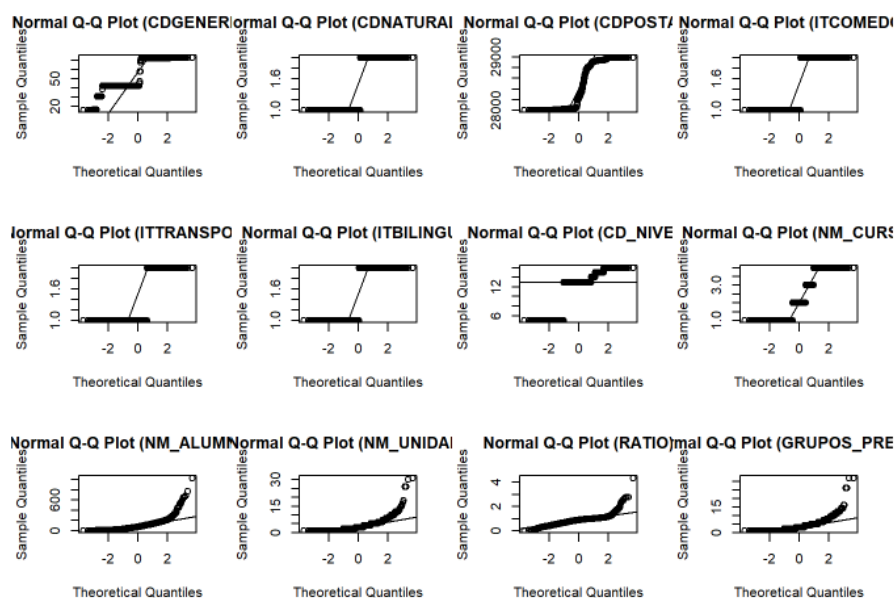
Para comprobar la normalidad en los datos se utiliza el test de Mardia que es una de las muchas pruebas que se utilizan para dicha comprobación. Existen otras pruebas como el test de Zirkler Henze, Shapiro-Wilk, etc.

Los resultados obtenidos utilizando la prueba de Mardia en el conjunto de datos son los que se pueden observar en la figura A.3

Figura A.3: Resumen de normalidad de variables. Elaboración propia

	Test	Variable	Statistic	p value	Normality
1	Shapiro-wilk	CDGENERICO	0.6763	<0.001	NO
2	Shapiro-wilk	CDNATURALEZA	0.6311	<0.001	NO
3	Shapiro-wilk	CDPOSTAL	0.7911	<0.001	NO
4	Shapiro-wilk	ITCOMEDOR	0.6339	<0.001	NO
5	Shapiro-wilk	ITTRANSPORTE	0.5518	<0.001	NO
6	Shapiro-wilk	ITBILINGUE	0.6360	<0.001	NO
7	Shapiro-wilk	CD_NIVEL	0.6230	<0.001	NO
8	Shapiro-wilk	NM_CURSO	0.8376	<0.001	NO
9	Shapiro-wilk	NM_ALUMNOS	0.8016	<0.001	NO
10	Shapiro-wilk	NM_UNIDADES	0.7886	<0.001	NO
11	Shapiro-wilk	RATIO	0.8623	<0.001	NO
12	Shapiro-wilk	GRUPOS_PREDECIR	0.7944	<0.001	NO

Figura A.4: Resumen de normalidad de variables. Elaboración propia



Como se puede observar en la figura A.4, las variables continuas (NM_ALUMNOS, NM_UNIDADES, RATIO y GRUPOS_PREDECIR) son las que más se acercan a la normalidad ya que son cuantitativas, sin embargo no se ajustan a la línea recta teórica que indica el ajuste a la normal.

Figura A.5: Distribución 1. Elaboración propia

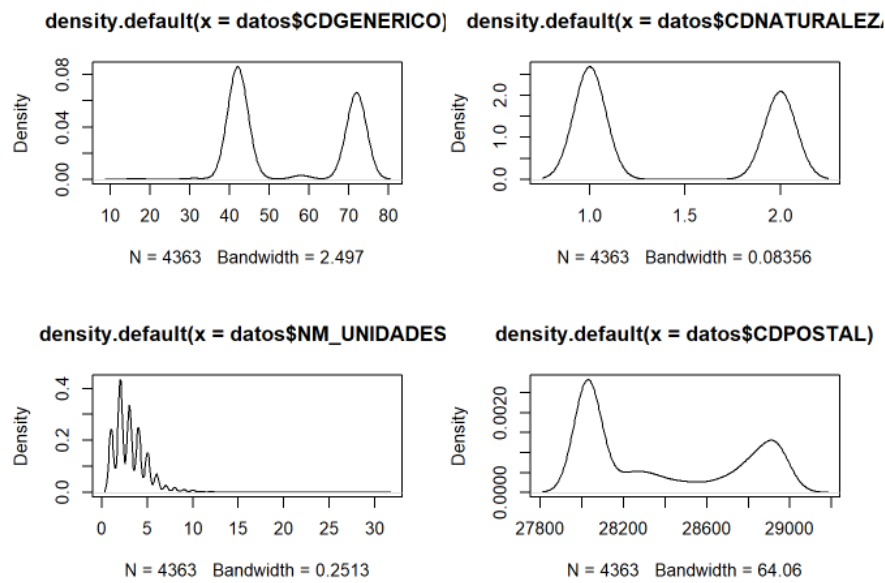
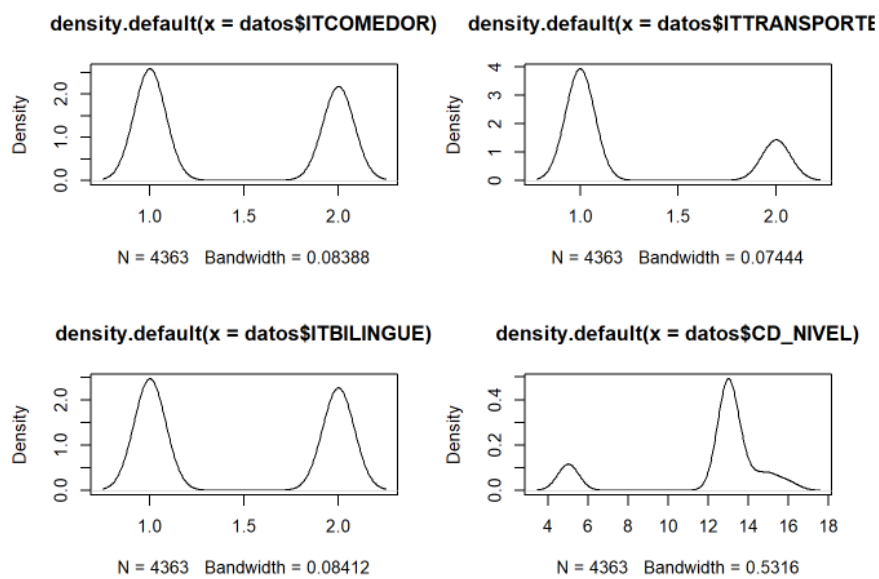


Figura A.6: Distribución 2. Elaboración propia



Las figuras A.5 y A.6, muestran de forma general los valores en los que destacan cada una de las variables. Por ejemplo, viendo la gráfica de densidad de NM_UNIDADES se puede observar que la media de los grupos por nivel de enseñanza es de 3. Por otro lado la media de alumnos por nivel educativo es de 74.

Figura A.7: Distribución 3. Elaboración propia

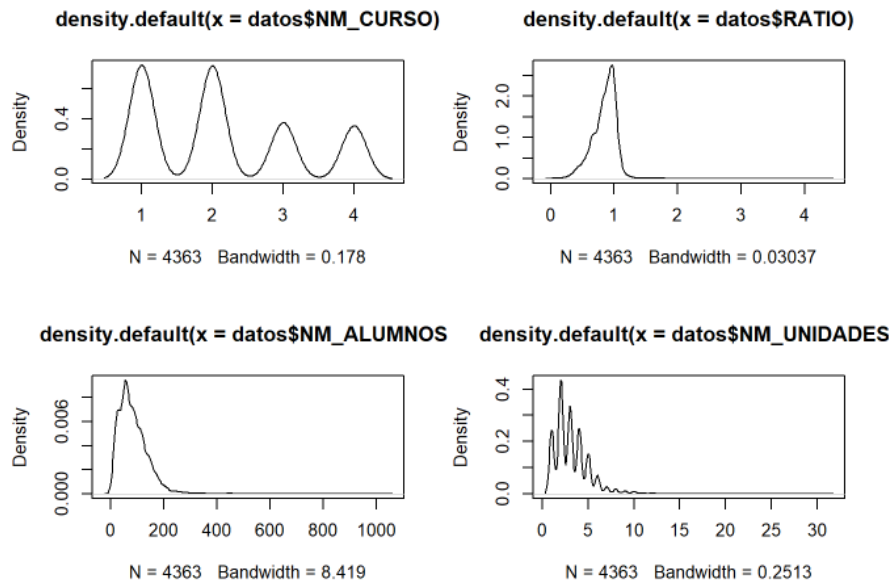


Figura A.8: Diagrama de barras 1. Elaboración propia

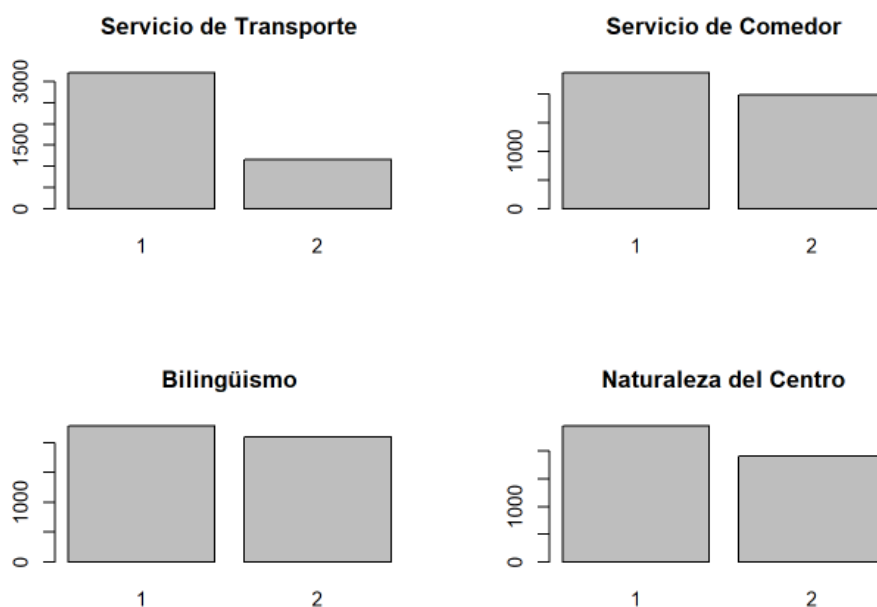
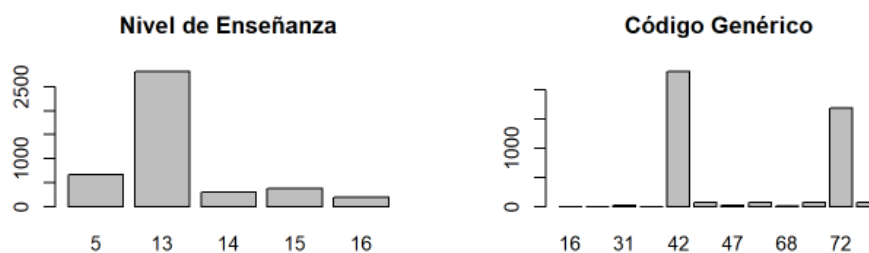


Figura A.9: Diagrama de barras 2. Elaboración propia



Tanto en la figuras A.8 y A.9 se puede hacer una comparativa sobre los servicios que ofrecen los centros. Se puede observar como la mayoría de centros no tienen transporte. Existe un ligero número de centros que tiene comedor sobre centros que no tienen, de la misma manera ocurre con el bilingüismo. La mayoría de los datos que se tiene son del nivel 13, que corresponde con la Educación Secundaria Obligatoria y la mayoría de estos datos corresponden a centros con el código genérico 42 y 72 (IES y CPR INF-PRI-SEC, respectivamente).

A.B.3. Relaciones entre variables

Para el estudio de la correlación se utiliza el Coeficiente de Correlación de Pearson (R).

Figura A.10: Matriz de correlaciones. Elaboración propia

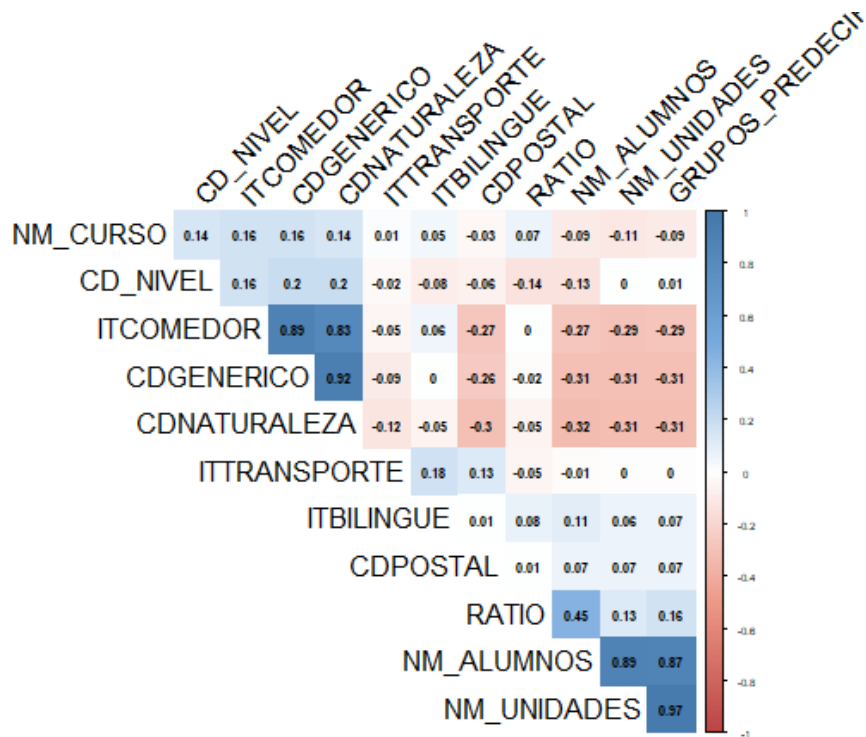


Figura A.11: Variables más correladas. Elaboración propia

First.Variable	Second.Variable	Correlation
NM_UNIDADES	GRUPOS_PREDECIR	0.9656515
CDGENERICO	CDNATURALEZA	0.9231279
CDGENERICO	ITCOMEDOR	0.8914708
NM_ALUMNOS	NM_UNIDADES	0.8898214
NM_ALUMNOS	GRUPOS_PREDECIR	0.8720157
CDNATURALEZA	ITCOMEDOR	0.8291385
NM_ALUMNOS	RATIO	0.4494119
CDNATURALEZA	NM_ALUMNOS	-0.3216512
CDNATURALEZA	GRUPOS_PREDECIR	-0.3138066
CDNATURALEZA	NM_UNIDADES	-0.3129556
CDGENERICO	NM_ALUMNOS	-0.3091014
CDGENERICO	NM_UNIDADES	-0.3090129

Figura A.12: Mayor correlación con variable a predecir. Elaboración propia

First.Variable	Second.Variable	Correlation
NM_UNIDADES	GRUPOS_PREDECIR	0.965651527
NM_ALUMNOS	GRUPOS_PREDECIR	0.872015704
CDNATURALEZA	GRUPOS_PREDECIR	-0.313806632
CDGENERICO	GRUPOS_PREDECIR	-0.307723048
ITCOMEDOR	GRUPOS_PREDECIR	-0.289707316
RATIO	GRUPOS_PREDECIR	0.160136305
NM_CURSO	GRUPOS_PREDECIR	-0.093643800
ITBILINGUE	GRUPOS_PREDECIR	0.069593706
CDPOSTAL	GRUPOS_PREDECIR	0.068786021
CD_NIVEL	GRUPOS_PREDECIR	0.005612750
ITTRANSPORTE	GRUPOS_PREDECIR	-0.003669386
GRUPOS_PREDECIR	GRUPOS_PREDECIR	0.000000000

En las figuras A.10, A.11 y A.12 se muestra prácticamente la misma información, la correlación entre variables. En la figura A.11 se muestran aquellos pares de variables que poseen la mayor correlación. Se puede apreciar en esta figura que la variable NMUNIDADES y GRUPOSPREDECIR tienen una correlación casi perfecta, lo que implica que aportarían la misma información al conjunto de datos.

En la figura A.12 se muestran aquellas variables que tienen mayor correlación con la variable a predecir. Podemos observar que las variables de NMUNIDADES y NMALUMNOS tienen una enorme correlación con la variable a predecir. Esta relación es lógica, ya que, a mayor número de alumnos o grupos, la variable a predecir aumenta. También se puede destacar que la naturaleza (pública o privada) de un centro está relacionada con el número de grupos a predecir. Al aumentar el valor de la naturaleza, disminuye el número de unidades. De esta premisa se puede deducir que los padres suelen matricular a sus hijos en centros públicos, y por lo tanto, el número de grupos tiende a aumentar. El servicio de comedor también es un factor con relación con el número de grupos, lo que implica que los padres suelen matricular a los alumnos en centros que dispongan del servicio de comedor.

Figura A.13: Correlación entre variables NM_ALUMNOS y NM_GRUPOS. Elaboración propia

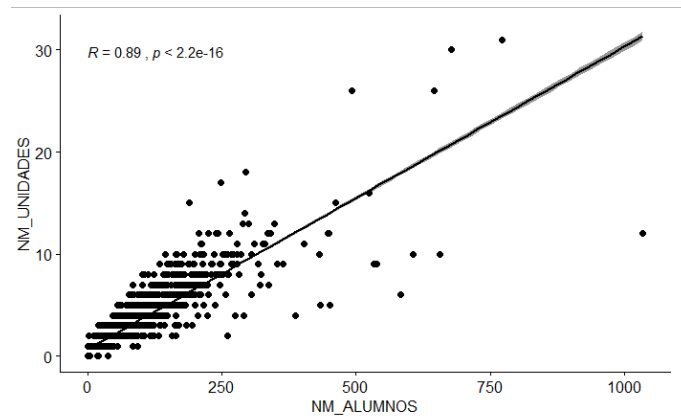
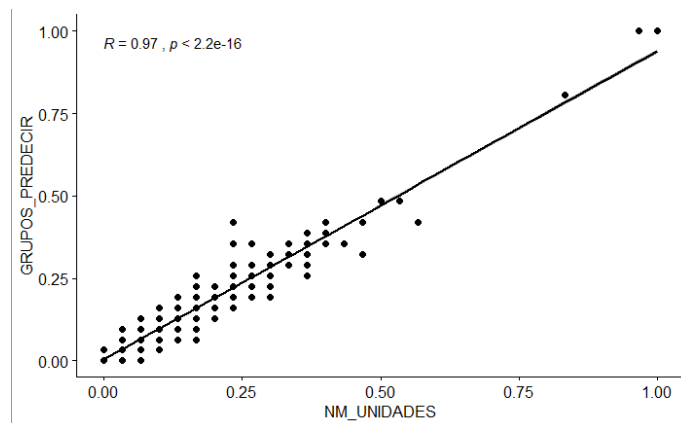


Figura A.14: Correlación entre variables NM_UNIDADES y GRUPOS_PREDECIR. Elaboración propia

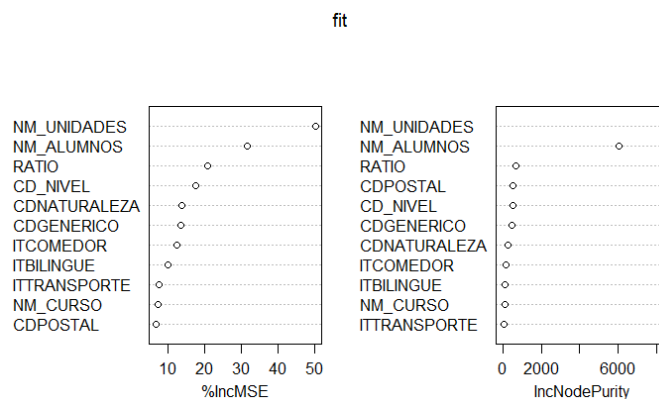


En la figura A.14 se observa la gran relación (prácticamente lineal) entre las variables NM_UNIDADES y GRUPOS_PREDECIR. Esto significa que, si aumenta el valor de una variable, linealmente aumenta el valor de la otra. En la figura A.13 existe también una relación lineal, pero no es tan pronunciada.

A.C. Selección de Variables

A.C.1. Usando Random Forest

Figura A.15: Variables más importantes usando Random Forest. Elaboración propia



La variable **IncNodePurity** se la conoce también como la media de decrecimiento de de Gini. El índice de Gini es una “medida de desorden” en este caso IncNodePurity tiene el siguiente sentido, a mayor medida, mayor importancia en los modelos creados, puesto que valores próximos a 0 implican un mayor desorden. Por tanto, si computamos la media del "decrecimiento" del índice de Gini cuanto mayor sea esta medida, mas variabilidad aporta a la variable dependiente.

Por otro lado, la variable **IncMSE** es la media de decrecimiento en la precisión, y es también un indicador sobre la importancia de las variables en el modelo.

A.C.2. Regresión Paso a Paso

La regresión por pasos (Stepwise Regression) consiste en añadir o eliminar iterativamente predicadores en el modelo predictivo, con el objetivo de encontrar el subconjunto de los datos que obtengan mayor precisión en el modelo o, dicho de otra forma, reducir el error en la predicción. (kassambara, 2018)

Existen 3 estrategias para realizar la regresión paso a paso: backward, forward y step-wise.

Usando Backward Selection

Figura A.16: Resultado Backward Selection. Elaboración propia

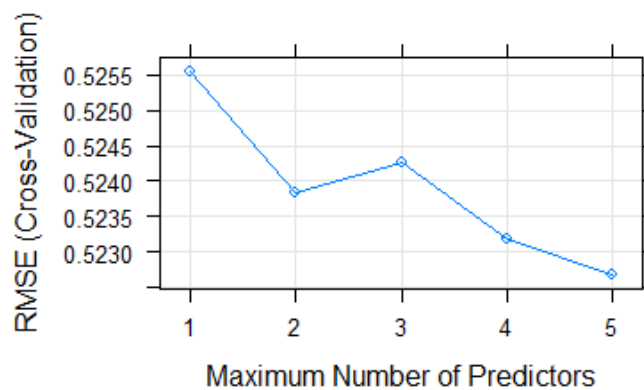
```

Selection Algorithm: backward
CDGENERICO  CDNATURALEZA  CDPOSTAL  ITCOMEDOR  ITTRANSPORTE  ITBILINGUE  CD_NIVEL  NM_CURSO  NM_ALUMNOS
1 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
2 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
3 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
4 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
5 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
NM_UNIDADES  RATIO
1 ( 1 )      " "          " "
2 ( 1 )      " "          " "
3 ( 1 )      " "          " "
4 ( 1 )      " "          " "
5 ( 1 )      " "          " "

```

Los asteriscos en los resultados indican los predictores que se deben tomar para realizar el modelo. En este caso se necesitan las variables CD_NATURALEZA, CD_NIVEL, NM_CURSO, NM_UNIDADES y RATIO.

Figura A.17: Gráfico Backward Selection. Elaboración propia



En la figura A.17 se puede observar como la mejor precision se obtiene utilizando los 5 predictores de la figura A.16.

Usando Stepwise Selection

Figura A.18: Resultado Stepwise Selection. Elaboración propia

```

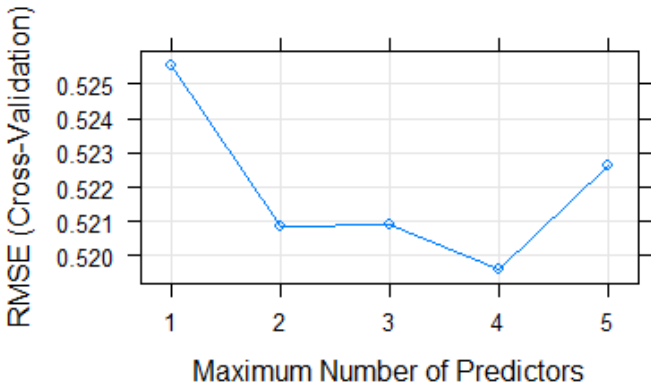
Selection Algorithm: 'sequential replacement'
CDGENERICO  CDNATURALEZA  CDPOSTAL  ITCOMEDOR  ITTRANSPORTE  ITBILINGUE  CD_NIVEL  NM_CURSO  NM_ALUMNOS
1 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
2 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
3 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
4 ( 1 )      " "          " "          " "          " "          " "          " "          " "          " "
NM_UNIDADES  RATIO
1 ( 1 )      " "          " "
2 ( 1 )      " "          " "
3 ( 1 )      " "          " "
4 ( 1 )      " "          " "

```

Nuevamente los asteriscos de la figura A.18 indican aquellos predictores que deben

utilizarse, en este caso son: CD_NATURALEZA, NM_CURSO, NM_UNIDADES y RATIO.

Figura A.19: Gráfico Stepwise Selection. Elaboración propia



En la figura A.19 se puede observar como la mayor precisión (menor valor de RMSE) se obtiene utilizando 4 variables.

Usando Forward Selection

Figura A.20: Resultado Forward Selection. Elaboración propia

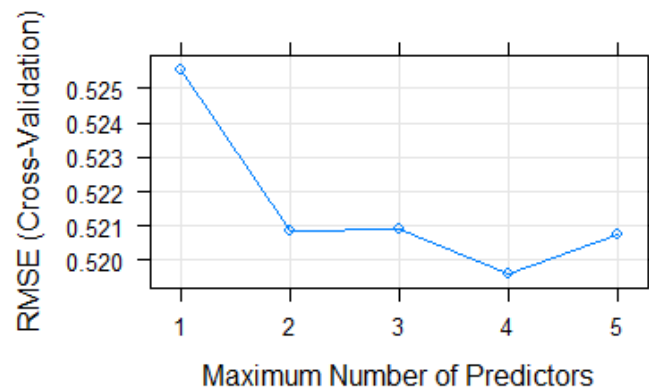
Selection Algorithm: forward

		CDGENERICO	CDNATURALEZA	CDPOSTAL	ITCOMEDOR	ITTRANSPORTE	ITBILINGUE	CD_NIVEL	NM_CURSO	NM_ALUMNOS
1	(1)	" "	" "	" "	" "	" "	" "	" "	" "	" "
2	(1)	" "	" "	" "	" "	" "	" "	" "	" "	" "
3	(1)	" "	" "	" "	" "	" "	" "	" "	" "	" "
4	(1)	" "	" "	" "	" "	" "	" "	" "	" "	" "

		NM_UNIDADES	RATIO
1	(1)	" "	" "
2	(1)	" "	" "
3	(1)	" "	" "
4	(1)	" "	" "

En los resultados de la figura A.20 se puede observar como nuevamente se tienen en cuenta las variables CD_NATURALEZA, NM_CURSO, NM_UNIDADES y RATIO.

Figura A.21: Gráfico Forward Selection. Elaboración propia



Al igual que en la estrategia de “Stepwise”, se ha obtenido el mejor resultado de RMSE utilizando 4 variables. Estas variables son las obtenidas en el resultado de la figura A.20.

Como conclusión se puede obtener que tanto usando el algoritmo de Random Forest como las varias estrategias del algoritmo de Regresión paso a paso, se han seleccionado mayoritariamente 4 variables a destacar en la utilización de un modelo y son las siguientes ordenadas por importancia: NM_UNIDADES, NM_ALUMNOS, RATIO, CD_NIVEL y CD_NATURALEZA.

B | Análisis Predictivo

Como ya se ha comentado en el diseño, a la hora de realizar un análisis predictivo utilizando técnicas de aprendizaje supervisado, es necesario dividir los datos en un 70 % para entrenar el modelo y el otro 30 % para probarlo.

Una vez que se han dividido los datos, se realiza una comparación con los modelos seleccionados. Se recuerda que los modelos utilizados a grandes rasgos han sido las redes neuronales, los k-vecinos más cercanos, los bosques aleatorios (Random Forest), la regresión logística, el soporte de vectores de máquinas y los arboles de decisión.

Estos modelos comentados a su vez tienen una serie de algoritmos distintos tanto para crear modelos de clasificación como para crear modelos de regresión. En este caso, debido a que es la predicción de datos continuos, se va a utilizar la regresión.

Los algoritmos que se han utilizado para cada modelo seleccionado son los siguientes: K-Vecinos Cercanos (kkn), Redes neuronales bayesianas regularizadas (brnn), bosques aleatorios (rf), Máquinas de soporte de vectores con núcleo lineal (svmLinear), Árbol Modelo (M5) y el Elasticnet (enet) en la Regresión Lineal.

B.A. Comparación de Modelos

Para realizar la comparación de modelos se utiliza una técnica del paquete Caret, donde se realiza un entrenamiento con los datos usando todos los modelos comentados en el apartado 4.3.4 del capítulo de Diseño de la Investigación.

Para realizar la comparación de modelos se realiza una medición no solo en la precisión de estos sino también el tiempo transcurrido en entrenar el propio modelo.

Todos los modelos se entrenan usando los mismos parámetros, de esta forma ninguno de ellos resulta favorecido. Obviamente los resultados de los modelos dependen de dichos parámetros. Realizando combinaciones de estos parámetros se obtienen una serie de resultados para cada modelo al final se selecciona el mejor resultado para cada modelo y estos resultados son los que se compararan con el resto de modelos.

Ademas de los modelos comentados, se han entrenado dichos modelos, pero teniendo en cuenta las 5 variables seleccionadas por los algoritmos de selección de variables. A continuación se muestran en la siguiente tabla B.1 los modelos, su predicción y sus tiempos.

Tabla B.1: Tabla comparación modelos (precisión y tiempo)

Modelos	Precisión	Tiempo Entrenamiento
Redes neuronales (NN)	0.5236718	52.10
Redes neuronales - 5 variables (NN2)	0.5242227	29.38
K-Vecinos Cercanos (KKNN)	0.6468044	16.69
K-Vecinos Cercanos - 5 variables (KKNN2)	0.5982330	14.86
Random Forest (RF)	0.5264411	947.03
Random Forest - 5 variables (RF2)	0.5577621	515.49
Vector de Máquinas Soporte (SVM)	0.5336122	37.93
Vector de Máquinas Soporte - 5 variables (SVM2)	0.5409171	21.22
Regresión Linear (enet)	0.5256446	2.78
Regresión Linear - 5 variables (enet2)	0.5245138	2.13
Árbol de decisión (xgbTree)	0.5146266	247.83
Árbol de decisión -5 variables (xgbTree2)	0.5311688	228.45

A continuación, se muestra gráficamente los resultados de la tabla anterior.

Figura B.1: Comparación Modelos y Precisión. Elaboración propia

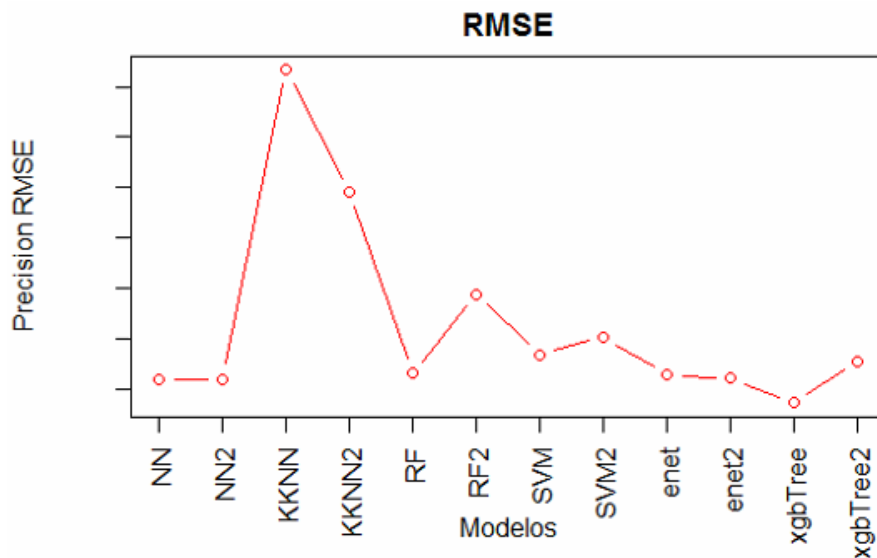
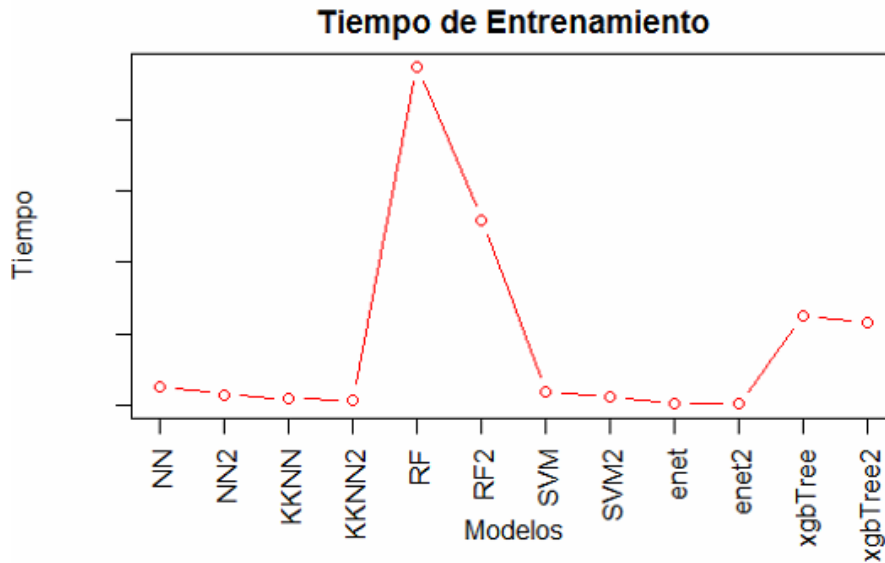


Figura B.2: Comparación Modelos y Tiempo en Entrenar. Elaboración propia



B.B. Predicción utilizando el modelo de árbol de decisión

Figura B.3: Aumento o disminución de grupos en la predicción. Elaboración propia

NM_UNIDADES	RATIO	GRUPOS_PREDECIR	GRUPOS_PRED_REDONDEADO
10	0.4233333	9.480645	9
5	1.6733333	5.953642	6
6	0.6238095	5.441341	5
6	0.3888889	4.794295	5
4	0.6083333	3.494278	3
8	0.7000000	8.803075	9
11	0.5168831	10.231400	10
13	0.4923077	10.930596	11
6	0.6380952	5.425637	5
6	0.4111111	5.148684	5
14	0.7285714	12.005297	12
16	0.4291667	13.375525	13
12	0.7666667	10.950116	11
5	2.2342857	5.594111	6
16	0.8821429	16.666368	17
14	0.9306122	13.395216	13
6	0.4222222	5.231615	5
5	0.4200000	4.387225	4
7	0.5591837	6.337043	6

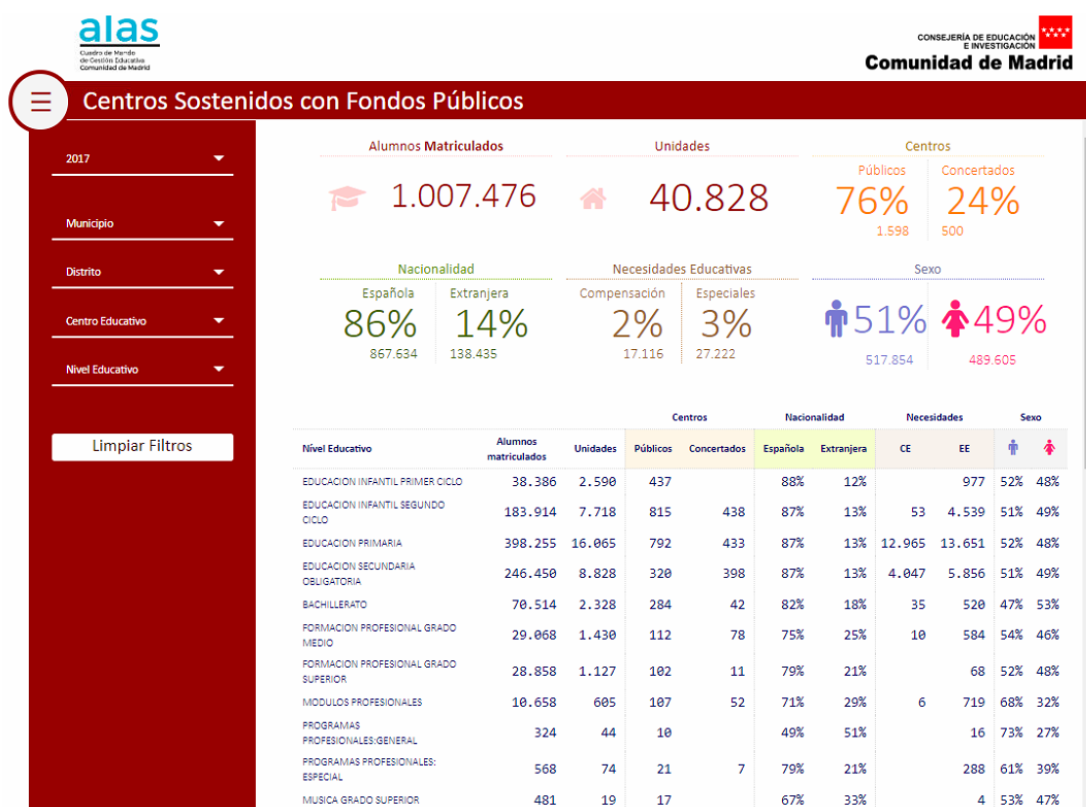
B.C. Sistema de Explotación de la Consejería de Educación e Investigación

En este apartado se van a mostrar una serie de imágenes de la aplicación de explotación de datos de la Consejería de Educación e Investigación. Concretamente se van a mostrar dos Cuadros de Mandos.

En primer lugar, se ha realizado un Cuadro de Mando relativo a los centros sostenidos con fondos públicos. En este cuadro de mando se muestran aspectos como el número de alumnos matriculados, el número de unidades, el porcentaje de centros públicos y concertados, la nacionalidad de los alumnos, el sexo, etc. Véase la figura B.4

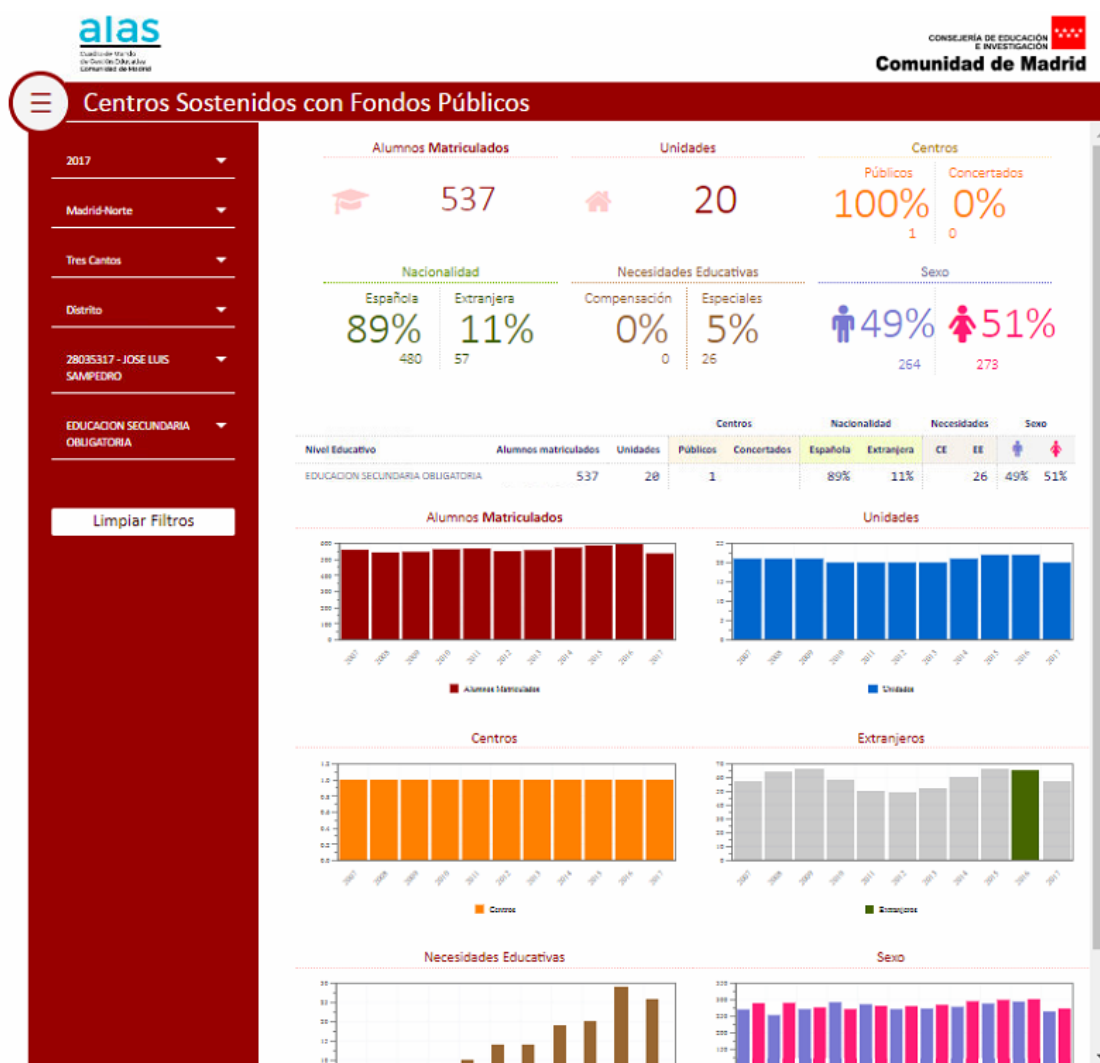
Existe un filtro, donde se puede seleccionar los parámetros de la búsqueda. Entre estos filtros se encuentran los siguientes: año, área territorial, municipio, distrito, centro educativo y nivel educativo.

Figura B.4: Cuadro de Mandos de Centros I. Elaboración propia



En la figura B.5 se ha realizado una búsqueda sobre el nivel educativo de E.S.O. en el centro educativo de “Jose Luis Sampedro”, que se encuentra en Tres Cantos (DAT-Norte).

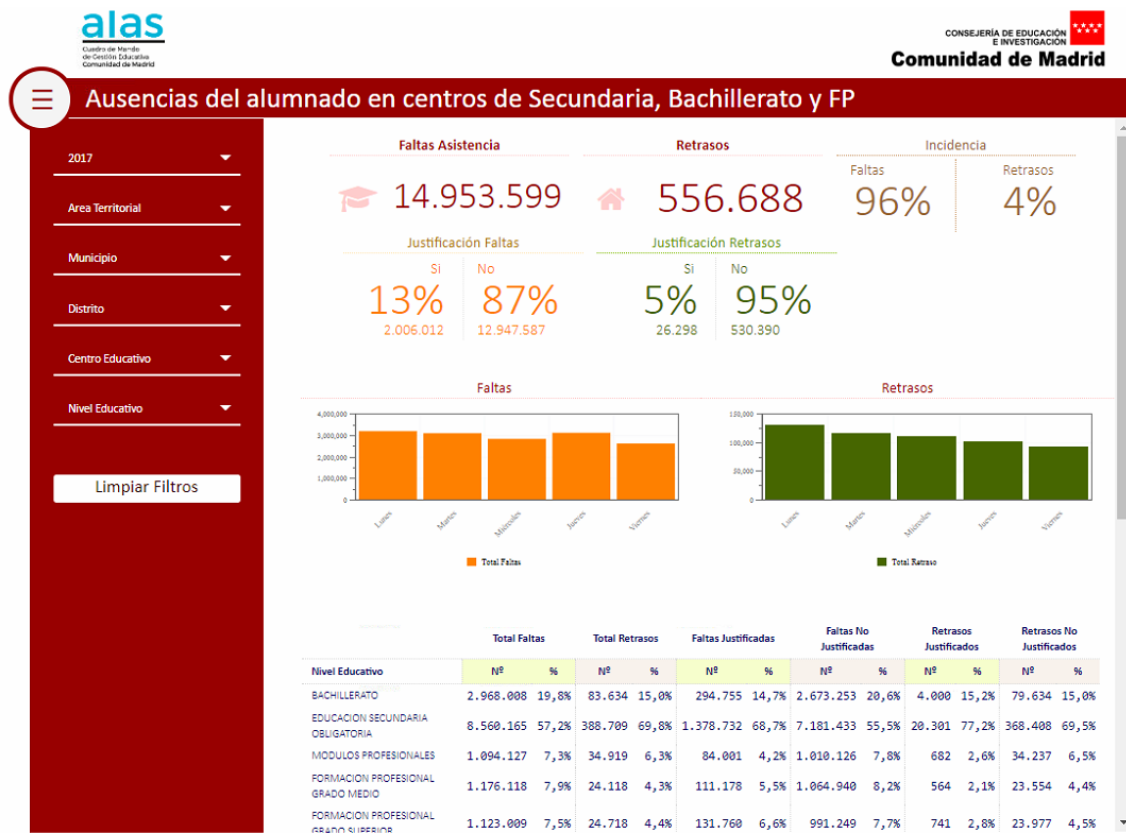
Figura B.5: Cuadro de Mandos de Centros II. Elaboración propia



Otro cuadro de mandos realizado ha sido el de ausencias de alumnados en centros de Secundaria, Bachillerato y FP.

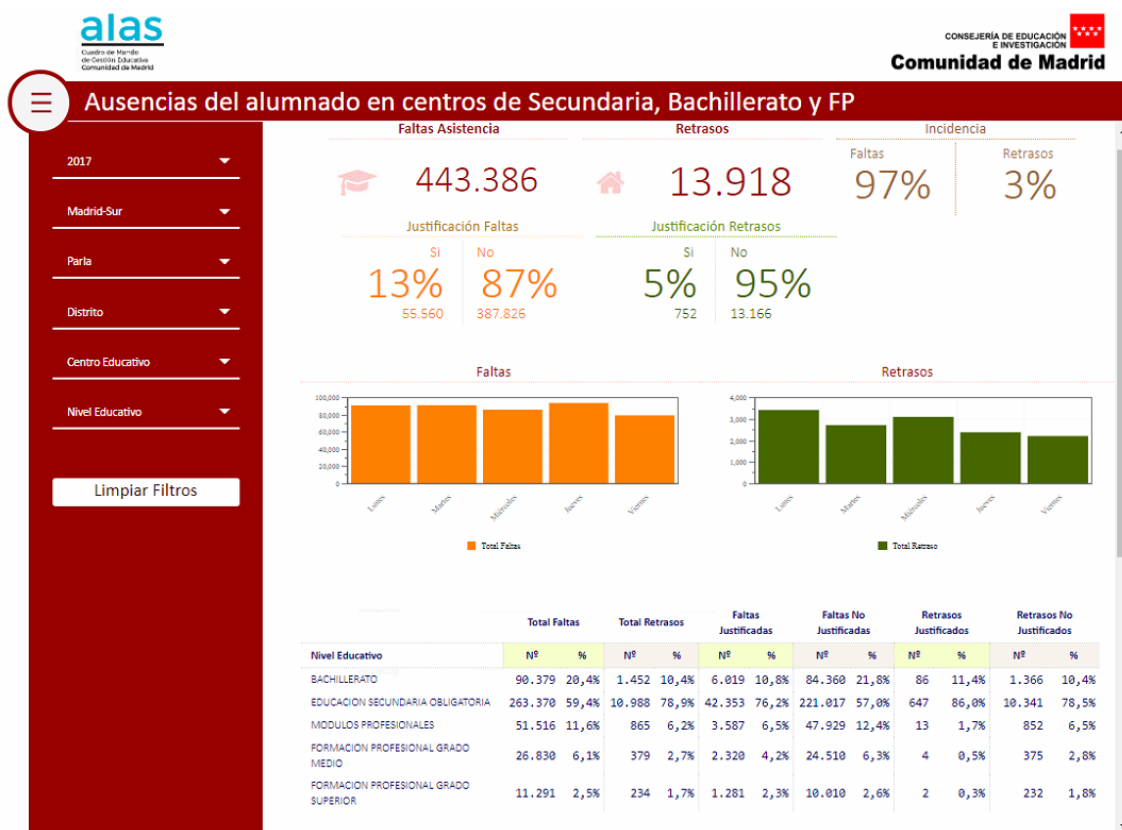
En este cuadro de mando aparecen las faltas de asistencias totales, el número de retrasos, el porcentaje sobre el total entre faltas y retrasos, el porcentaje de faltas y retrasos justificados. Se muestra un histograma también sobre el número de faltas y retrasos y los días de la semana.

Figura B.6: Cuadro de Mandos de Ausencias de Alumnado I. Elaboración propia



En la figura B.7 se filtra la búsqueda por el municipio de Parla (DAT-Sur), para el año 2017.

Figura B.7: Cuadro de Mandos de Ausencias de Alumnado II. Elaboración propia



Además, aunque no se haya creado cuadro de mandos, se ha realizado cubos OLAP de las unidades, los centros, las faltas de asistencia, las minorías étnicas, etc. En este caso se va a mostrar en la siguiente ilustración B.8 el cubo OLAP relativo a las unidades.

Figura B.8: Cubo OLAP de Unidades. Elaboración propia

STPivot

Olap MdX Gráfico Tabla

Unidades

Cancelar Ok

Columnas

Measure

Filas

Anio Naturaleza Territorio Centros Nivel Curso

Filtros

Disponibles

Objetos creados

Anio	Naturaleza	Territorio	Centros	Nivel	Curso	Medidas
						Unidades
- All Anios	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	428.529
2007	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	35.110
2008	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	36.607
2009	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	37.565
2010	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	38.578
2011	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	39.436
2012	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	39.997
2013	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	39.676
2014	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	39.693
2015	+ All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	40.300
2016	- All Naturalezas	+ Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	40.712
	CENTRO PÚBLICO	- Todos Territorios	+ All Centros	+ All Nivels	+ All Cursos	26.806
		+ Madrid-Norte	- All Centros	+ All Nivels	+ All Cursos	2.490
			FEDERICO GARCIA LORCA	- All Nivels	+ All Cursos	36
				EDUCACION INFANTIL SEGUNDO CICLO	- All Cursos	12
					3	4
					4	4
					5	4
				EDUCACION PRIMARIA	+ All Cursos	24
			ANTONIO MACHADO	+ All Nivels	+ All Cursos	18
			FRANCISCO GINER DE LOS RIOS	+ All Nivels	+ All Cursos	54
			DAOIZ Y VELARDE	+ All Nivels	+ All Cursos	20
			OBISPO MOSCOSO	+ All Nivels	+ All Cursos	18