UNIVERSIDAD REY JUAN CARLOS

TRABAJO FIN DE MÁSTER

Uso de técnicas predictivas para la planificación de grupos en Secundaria y FP

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

TUTOR: Aurelio Berges García

DOCTOR INGENIERO DE TELECOMUNICACIÓN

AGRADECIMIENTOS

Desde estas lineas me gustaría expresar mi mas sincero agradecimiento:

A mi tutor Aurelio Berges, por ayudarme, aconsejarme y guiarme durante todo el proyecto, sobretodo en los momentos duros; y durante mi estancia en la Consejería de Educación. Gracias por todos los conocimientos aportados.

Al director de este máster, Miguel Ángel Marcos, por las indicaciones que me ha proporcionado en la redacción de este TFM.

A Felipe Retortillo, por ayudarme con toda la información técnica y de ámbito educativo 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 día 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 en la etapa de Secundaria y Formación Profesional es una tarea complicada a la que se enfrentan anualmente de forma genérica las Unidades responsables que tienen esa competencia en las diferentes Comunidades Autónomas y, en particular, la Unidad de Planificación de Centros de la Dirección General de Infantil, Primaria y Secundaria de la Consejería de Educación e Investigación (CEI) de la Comunidad de Madrid.

Gestionar los recursos de la mejor forma es vital para mejorar la calidad en el sistema educativo. Por tanto, se hace necesario el estudio de los factores educativos más importantes que permitan un mejor reparto de los recursos disponibles.

En la investigación que se realiza en este trabajo de fin de máster se utilizan datos de centros educativos como por ejemplo el número de alumnos, los números de grupos y ratio para un determinado centro; su naturaleza, etc, con el objetivo de controlar la sobrepoblación en el aula.

Con esta información se construyen varios modelos predictivos que van a ayudar a la Unidad de Planificación de Centros a predecir el número de grupos que deben planificarse para el nuevo curso y poder así repartir los recursos disponibles (profesores, suministros, etc.). De esta forma, se consigue facilitar el trabajo final de dicha Unidad de Planificación.

Ademas, en este TFM, se realiza una clasificación sobre la importancia de las variables (como por ejemplo el número de grupos, ratio, número de alumnos, naturaleza del centro, etc) existentes en el entorno educativo, además se realizan estudios estadísticos para obtener la relación que existe entre estas variables para posteriormente utilizar distintos modelos con el objeto de conseguir la mejor 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.

Estos modelos propuestos junto con las variables seleccionadas se validan mediante reuniones con la CEI, quien da el visto bueno para realizar la comparación de dichos modelos.

Palabras clave: sobrepoblación, educación, gestión, planificación, ratio, aula, minería de datos, modelos predictivos

ABSTRACT

The planning of school groups in the stage of Secondary and Vocational Education is a complicated task that is faced annually in a generic way by the responsible Units that have this competence in the different Autonomous Communities and, in particular, the Educational Center Planning Unit of the General Directorate of Children, Primary and Secondary of the Ministry of Education and Research (CEI) of the Community of Madrid.

Managing resources in the best way is vital to improve quality in the education system. Therefore, it is necessary to study the most important educational factors that allow a better distribution of available resources.

In the research carried out in this end-of-master project, data from educational centers are used, such as the number of students, the group numbers and the ratio for a given center; its nature, etc., with the objective of controlling overpopulation in the classroom. With this information, several predictive models are constructed that will help the Center Planning Unit to predict the number of groups that must be planned for the new course and thus be able to distribute the available resources (teachers, supplies, etc.). In this way, it is possible to facilitate the final work of the Planning Unit.

In addition, in this TFM, a classification is made on the importance of the variables (such as the number of groups, ratio, number of students, nature of the center, etc.) existing in the educational environment, in addition statistical studies are carried out to obtain the relationship that exists between these variables to later use different models in order to achieve the best possible prediction.

To obtain the best model, different metrics are used in order to be able to compare them according to the needs that are required.

These proposed models together with the selected variables are validated through meetings with the CEI, who gives the approval to make the comparison of these models.

Key Words: overcrowding, education, management, planning, ratio, class, datamining, predictive models

ÍNDICE

Índice de tablas

Índice de figuras

1 Introducción

1.1. Introducción

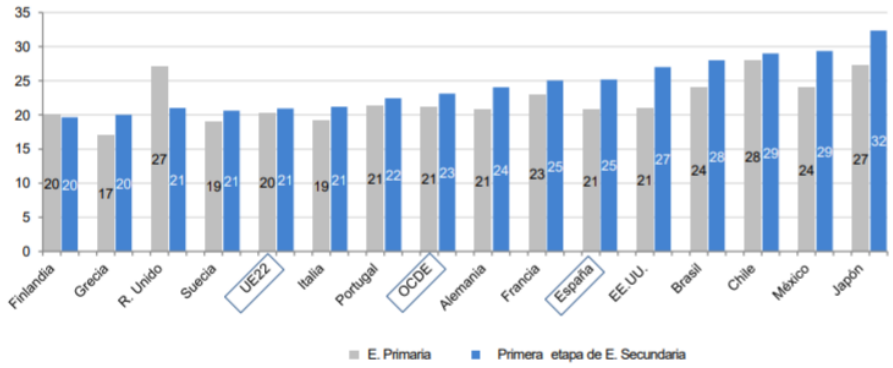
Uno de los problemas que se plantea en la actualidad en el proceso de enseñanza-aprendizaje es la cantidad de alumnos que están en una misma aula a cargo de un solo docente. Este aspecto es uno de los más debatidos en la educación. (Ministerio de Educación y Formacion Profesional, 2018)

Según Racancoj (2013), la superpoblación estudiantil es el exceso del número de estudiantes que se encuentran en un espacio determinado cuya capacidad no es adecuada para acogerlos ni cuenta con las condiciones adecuadas para el buen desenvolvimiento de los mismos.

En el caso particular de España, el tamaño de clase medio en Educación Secundaria en las instituciones públicas es superior al promedio de la OCDE y de la UE22. (Ministerio de Educación y Formacion Profesional, 2018)

Esto no implica que en todos los casos en el que la ratio sea superior exista una sobrepoblación, sin embargo, la calidad educativa se ve mermada por esta situación.

Figura 1.1: Número medio de alumnos por clase en instituciones públicas (2016). Recuperado del Ministerio de Educación y Formacion Profesional (2018)



En la figura 1.1 se puede observar como la media del número de alumnos por clase de la Unión Europea está en 21 alumnos para la etapa de Secundaria, mientras que en España la media asciende a 25 alumnos.

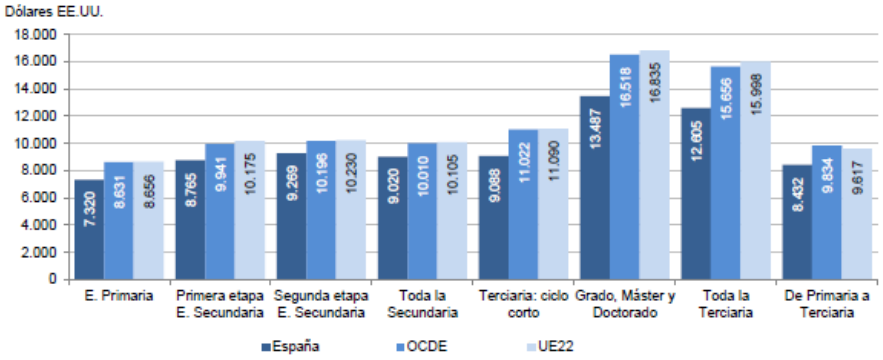
Si bien es cierto, que es bueno que se deba atender a la diversidad y a la inclusión (sobre el número de alumnos) en el aula, esto debe hacerse de forma responsable y conociendo las limitaciones que tiene el docente, que por más voluntad o capacitación que posea, tiene que atender a las necesidades personalizadas de cada alumno. (Fingermann, 2014)

Lera Rodríguez (2007) en su articulo se plantea la reducción del numero de alumnos o el aumento de asistentes en el aula con el objetivo de mejorar la calidad educativa.

La ratio, que es la relación entre el número de alumnos y profesores, es un factor importante a la hora de realizar la planificación de los recursos. En el año 2012, dicho ratio, en España aumentó un 20 %, lo que implica un ahorro en torno a 464 millones de euros. Aprobado por el Real Decreto-ley 14/2012 en el artículo 2.

Debido a que los recursos de una administración no son infinitos, es necesario tener en cuenta los recursos disponibles para la educación; concretamente, durante 2015 España presenta un gasto total por alumno inferior al promedio de los países de la OCDE y la UE22. Véase la figura 1.2. Esto implica que se dispone de menos recursos en la educación y por lo tanto se tiende a agrupar a los alumnos en grupos de mayor tamaño.

Figura 1.2: Gasto total anual (2015). Recuperado del Ministerio de Educación y Formacion Profesional (2018)



Por las razones anteriormente comentadas, es importante establecer una buena gestión sobre los recursos disponibles, realizando una óptima distribución tanto de los docentes como de los recursos asignados a estos.

1.2. Objetivos

En este trabajo fin de máster (TFM) se expone 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 de un nuevo período escolar. Para ello, se plantea el uso de herramientas y métodos flexibles que automaticen las tareas de planificación y proponga, además, nuevas variables o factores que puedan influir en la toma de decisiones.

El interés que justifica este TFM es avanzar en el conocimiento sobre los factores del entorno educativo que implican el aumento de la superpoblación de un aula. De esta

manera, se consigue que las Unidades (que tienen la competencia de gestionar la demanda de plazas escolares de un nuevo periodo escolar) tengan mayor facilidad a la hora de implantar un Sistema que sea capaz de predecir el número de grupos, a partir de los factores disponibles en cursos previos.

Las consideraciones que se han realizado en los párrafos anteriores justifican que el objetivo general de este TFM es: proponer un modelo para contribuir a la óptima planificación de los grupos escolares para los nuevos cursos, evitando así el gasto innecesario de recursos y controlando la sobrepoblación en el aula.

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

- Seleccionar variables de interés, relativas a la resolución de la necesidad anteriormente expuesta por Unidad de Planificación, que aporten valor en el desarrollo de este TFM.
- Estudiar la relación entre dichas variables con el propósito de comprender el contexto de la sobrepoblación en el aula y la planificación de grupos.
- Probar distintos modelos predictivos y seleccionar aquellos que aporten mayor precisión en la predicción de los grupos.
- Obtener y utilizar el modelo de mayor precisión para realizar predicciones con los datos existentes en el entorno educativo de la Comunidad de Madrid en los últimos 10 años y que este modelo ayude a las Unidades de la CEI a planificar grupos en los nuevos cursos.

1.3. Metodología

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

1. En primer lugar, se detecta la necesidad ya expuesta por la 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 máxima información posible acerca de las necesidades concretas y la forma de satisfacerlas. [2]

3. Teniendo en cuenta el conocimiento sobre cuáles son las necesidades, se realiza una

[1] El autor de este TFM ha colaborado con la Consejería de Educación e Investigación de la Comunidad de Madrid en el desarrollo de un proyecto (ALAS) sobre explotación de datos educativos coordinado por la Universidad Politécnica de Madrid, a lo largo del año 2018.

[2] Antes de comenzar la investigación, se debe tener un claro conocimiento sobre las necesidades existentes y se debe establecer un plan de acción.

propuesta[3] que incluya modelos, variables, y herramientas a la Unidad de Planificación para poder satisfacerlas.

4. Después de varias reuniones con la Unidad de Planificación, esta Unidad valida la propuesta.
5. Una vez validada la propuesta, se estudian distintos modelos de predicción a partir de los datos seleccionados en dicha propuesta. Se debe analizar cuál de estos modelos es el que mayor precisión obtiene.
6. Se decide, en una reunión con la Unidad de Planificación de Centros, cual es el modelo óptimo (teniendo en cuenta la precisión y el tiempo de entrenamiento de los modelos).
7. A partir del modelo acordado, se realizan predicciones con los datos de la Unidad.

1.4. 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 el desarrollo 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.
- **Capítulo 3. Propuesta de intervención:** En este capítulo se define el problema existente.
- **Capítulo 4. Diseño de la investigación:** Este capítulo establece 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 mostrarán los resultados obtenidos en este capítulo.
- **Capítulo 6. Conclusiones:** En este capítulo se detallan las conclusiones obtenidas a partir de los resultados alcanzados y los objetivos establecidos.

[3]Esta propuesta es un documento técnico donde se indica qué metodologías se van a emplear para llevar a cabo la investigación, qué modelos predictivos se van a considerar y fundamentalmente cuales son las variables que se van a usar en la predicción (teniendo en cuenta las necesidad de la Unidad de Planificación).

2 | Justificación Teórica

2.1. Introducción

La planificación de grupos no se entiende sin tener en cuenta el uso de la predicción. Para poder planificar el numero de grupos para el nuevo curso, se debe tener en cuenta los datos del curso previo; es a partir de estos datos con los que se realiza un analisis predictivo.

Uno de los intereses que se tienen a la hora de realizar cualquier proyecto de análisis de datos, "Big Data", "Business Intelligence" o análisis predictivo son las variables a utilizar, ya que son las que permiten obtener la mayoría de información.

Esta investigación se realiza con el propósito de aportar conocimiento existente sobre la importancia de determinadas variables, metodologías, modelos y la relevancia de estos últimos en la predicción, la planificación y la gestión educativa, concretamente en la previsión de la sobrepoblación en el aula.

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.

Debemos destacar que la mayoría de los resultados obtenidos tratan de artículos centrados en la predicción de los resultados académicos de los alumnos teniendo en cuenta ciertos factores internos (como las propias calificaciones a lo largo del curso) y externos (como factores etnográficos, edad, situación económica familiar, etc.). Estos factores se utilizan 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 realizan un estado del arte sobre la minería de datos en la educación, recogiendo información genérica de otros artículos. Estos artículos más representativos sirven 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, Naives Bayes, etc. En este mismo artículo, se clasifican las publicaciones en seis categorías. Podemos destacar que la categoría mayoritaria (con un 21 %) es el modelado del comportamiento del alumno seguida del rendimiento académico del alumno (con un 20 %). Romero y Ventura (2010) utiliza también categorías para clasificar las publicaciones.

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 sobre minería de datos en la 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 obtienen 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.

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 utilizan técnicas predictivas para la mejora del rendimiento académico de los alumnos. En este caso, los datos utilizados proceden de instituciones malayas. De nuevo se tienen 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 seleccionan 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 obtienen a partir de encuestas realizadas a alumnos del Centro Nacional de Evaluación y del Departamento de Servicios Escolares. Relacionado también con el rendimiento académico, esta 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, se utilizan datos de alumnos de un determinado centro público de Andalucía. Los cursos de alumnos que evalúa son 1º y 2º de Bachillerato y de ESA.

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

especialidad de los estudiantes, ciudad del estudiante, calificaciones y tipo de educación secundaria recibida previamente.

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

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

- **Facilidades Urbanas:** En este punto se incluyen las gasolineras, las tuberías de gas de alta presión y las líneas de alta tensión. Cuanto más cerca estén los centros de estas zonas, más riesgo existe para los alumnos. Se tiende por tanto a alejar los centros de estos puntos.
- **Densidad de población y áreas residenciales:** La proximidad de los colegios a zonas residenciales con una gran población de estudiantes es importante, puesto que, a menor distancia entre los estudiantes, los colegios y sus casas menor es el gasto de las familias y menor es la probabilidad de que los alumnos sean secuestrados.
- **Accesibilidad a red de carreteras urbanas:** La distancia de las calles y las autovías es otro factor importante para situar los colegios. Cuanto más cerca estén los colegios a estas vías, más facilidades tendrán los alumnos, y por lo tanto más ahorro de tiempo y costes. Sin embargo, la cercanía de los colegios a las autovías o autopistas, puede implicar mayor riesgo de accidentes. Sin embargo, si las autovías o autopistas se encuentran lejos, se reduce la accesibilidad a los colegios. Es necesario situar los centros en puntos intermedios (100-200m).
- **Servicios Urbanos:** Las distancias a los hospitales, a las estaciones de bomberos y de policía tienen mayor influencia. Sin embargo, estos deben situarse a distancias prudenciales de los centros (100-200m).
- **Centros culturales:** La proximidad de los centros culturales incrementa la salud espiritual y psicológica del alumno, incrementando así sus conocimientos. Curiosamente, si existen estos tipos de centros cercanos al colegio, entonces no es necesario que dichos colegios dispongan de estos servicios (pudiéndose ampliar las aulas, el comedor, etc)

La investigación se 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, sub-factores y los colegios.

El objetivo de este articulo 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 reducción 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 debe o simplemente establecen los pasos a seguir, pero no las tareas. La comparativa se muestra en la figura 2.1.

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

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

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

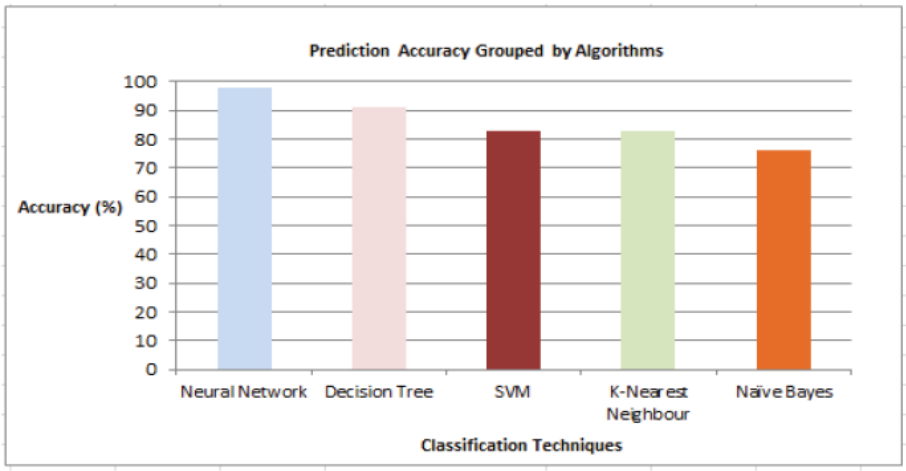
2.5. Modelos utilizados en el desarrollo de proyectos de minería de datos en el entorno educativo

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 alumnos que hacen que el modelo 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 la figura 2.2, se observa que las redes neuronales son las que obtienen mejores resultados junto con los arboles de decisiones, lo que significa que se ajustan más a los datos.

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

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

En cuanto al artículo de Adekitan y Salau (2019), nuevamente se utilizan 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. Otro artículo donde la regresión logística es la que mejor precisión ofrece 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 ám-

bito 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 amiga, 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)

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.

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

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

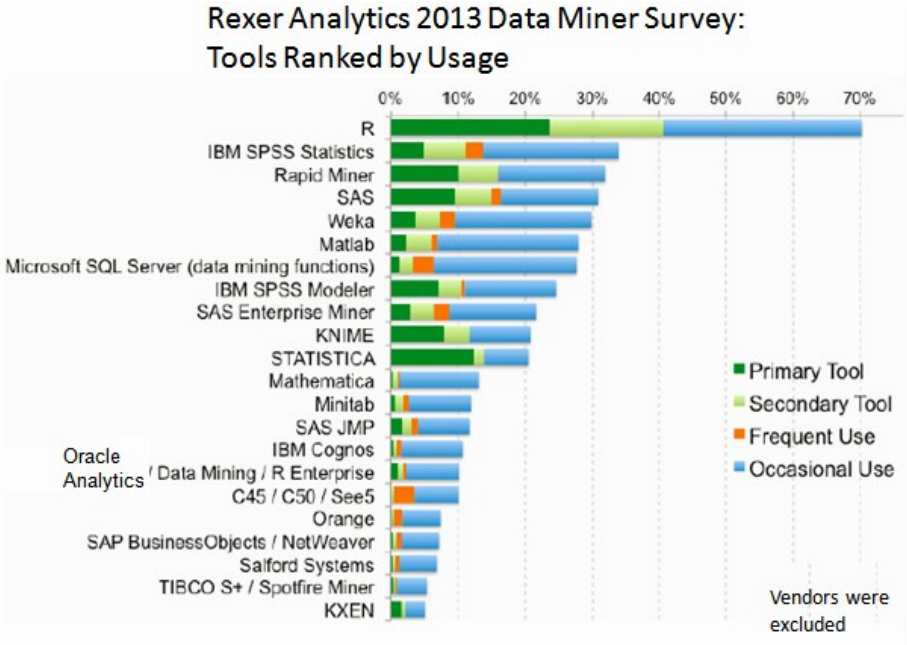


Figura 2.5: Plataformas de ciencia de datos y aprendizaje de máquina. Recuperado de Gartner (https://www.gartner.com)

2.7. Conclusiones

A partir de los artículos anteriores, se observa 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 existe 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, que 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 en ciertos artículos son más precisos que en otros. 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 regresivo.

Por último, en muchos artículos no se referencian 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 puedan 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 **demanda concreta 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 matricula.

   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 y evitando posibles sobrepoblaciones en el aula (sobrepasando los ratios). El sistema es flexible ya permite la incorporación de nuevas variables 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.

A lo largo de esta investigación, también se realizan estudios para identificar los factores que impliquen mayor demanda en un determinado centro o curso.

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 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.

Las bases de datos de este sistema de explotación se encuentran en un estado en el que es sencillo poder realizar análisis exploratorios y predictivos (ya que los datos se encuentran limpios y transformados a conveniencia), por tanto, a partir de este estado, se contempla la ampliación y la resolución de las nuevas necesidades de la CEI sobre la planificación de las unidades.

4.2. 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)

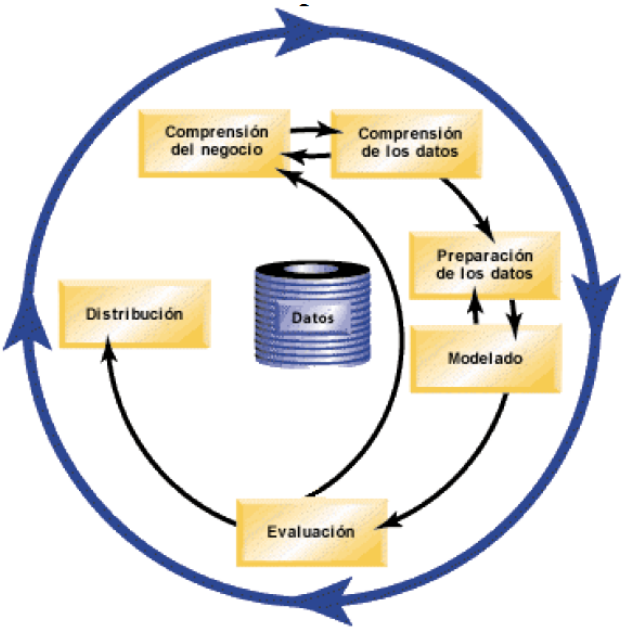
La metodología CRISP-DM tiene como objetivo orientar los proyectos de minería de datos.

- 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.

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

Figura 4.1: 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 4.1 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.

4.2.1. 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 C.

4.2.2. 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, se deben 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 parte 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:

- 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. A partir de estas reuniones se obtienen 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 a 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.

4.2.3. 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 utiliza 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 con-

vengan al estudio, transformarlos según las necesidades 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 contiene los coeficientes de correlación.(Diazaraque, s.f.). La matriz de correlación, se utiliza 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 mediante histogramas, boxplot y diagramas QQ-Plot o Grafico Cuantil-Cuantil. (Orellana, 2001)

Mediante el boxplot se observan aspectos como la posición, dispersión, asimetría, longitud de colas y los datos anómalos (outliers). El QQ-plot se utiliza 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.

4.2.4. Modelado

Una vez terminado el análisis descriptivo, se realiza 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 respuestas 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 o variable a predecir). 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: arboles de decisión, redes neuronales, bosques aleatorios, maquinas de soporte vectorial, regresión lineal y K-Vecinos mas cercanos.

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 deben desestimar. Para ello se utiliza el criterio de Akaike (AIC).

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

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

4.2.5. Evaluación

En esta fase de la metodologia, se diferencian varias partes. La primera parte es la evaluación del propio modelo respecto a otros, por lo que se utilizan 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.

A continuación se muestran las métricas utilizadas para comparar los modelos:

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 4.2: Fórmula de MAE. Recuperado de https://medium.com/human-in-a-machine-world

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

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 4.3: Fórmula de RMSE. Recuperado de https://medium.com/human-in-a-machine-world

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

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

4.2.6. Distribución

La fase de distribución considera la planificación y el 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. Para ello 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 seleccionado 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.

4.3. Herramientas utilizadas

En este apartado se van a mostrar las herramientas utilizadas para la resolución de las necesidades y la realización de este TFM.

4.3.1. 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.

4.3.2. 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 diferencia 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 5.1.

Figura 5.1: Estadísticos de las variables. Elaboración propia

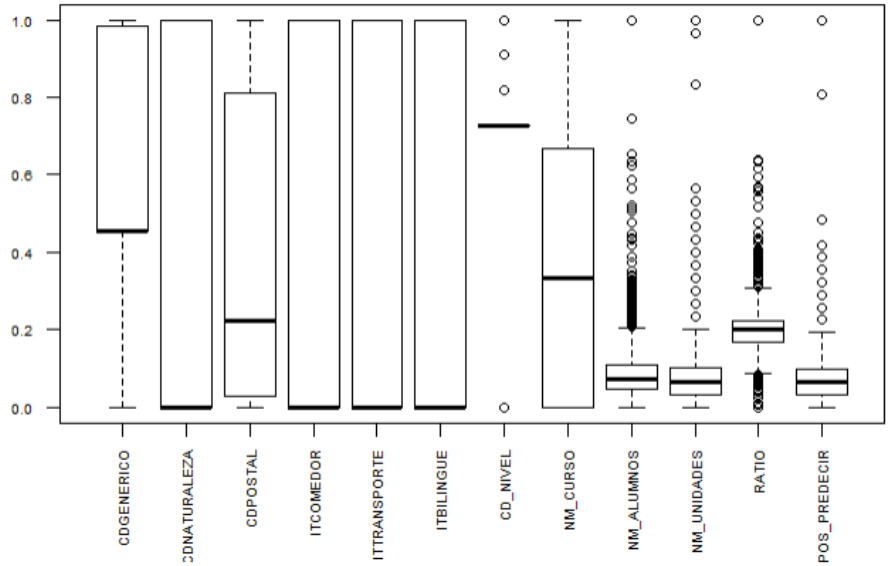
```
   CDGENERICO     CDNATURALEZA      CDPOSTAL        ITCOMEDOR      ITTRANSPORTE     ITBILINGUE       CD_NIVEL
 Min.   :16.0   Min.   :1.00   Min.   :28001   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   : 5.0
 1st Qu.:42.0   1st Qu.:1.00   1st Qu.:28030   1st Qu.:1.00   1st Qu.:1.00   1st Qu.:1.00   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
```

Una de los estadísticos más relevantes que se pueden apreciar en la figura 5.1 es la media de la variable ratio, que es de 0,88 y menor que 1. Esto implica que los centros de la Comunidad de Madrid no están sobrepoblados.

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 5.2.

Como se puede observar en la figura 5.2, las variables de número de alumnos, número de unidades y ratio contienen datos anómalos. Por ejemplo, la variable número de alumnos (nm_alumnos), como se puede apreciar en los estadísticos de la figura 5.1 tiene una media de 74 alumnos. No obstante, hay niveles educativos que tienen 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.

Figura 5.2: Diagrama de cajas normalizado. Elaboración propia



Una vez estudiados los estadísticos y los datos anómalos, se realiza 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. Véase el Anexo A.B.1

Posteriormente se realiza un estudio sobre aquellas unidades que tuvieran la ratio superior a 1 (implica que hay una sobrepoblación en el aula). En la figura 5.3 se puede observar como la DAT-Centro (5) es la que mayor sobrepoblación tiene seguida de la DAT-Sur. Esta situación quizá sea por la falta de recursos o la sobrepoblación de la zona.

Figura 5.3: Dirección de Área Territorial con aulas sobrepobladas. Elaboración propia

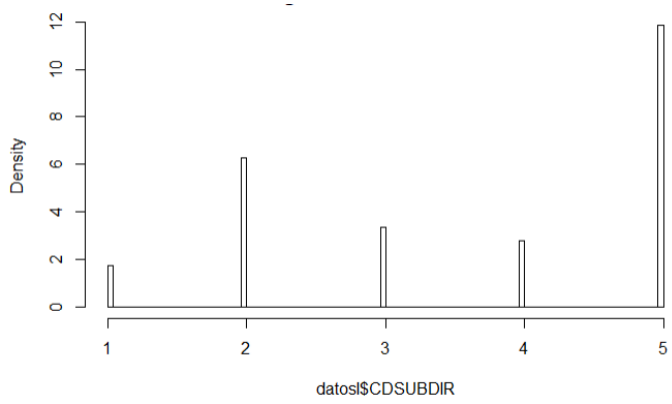
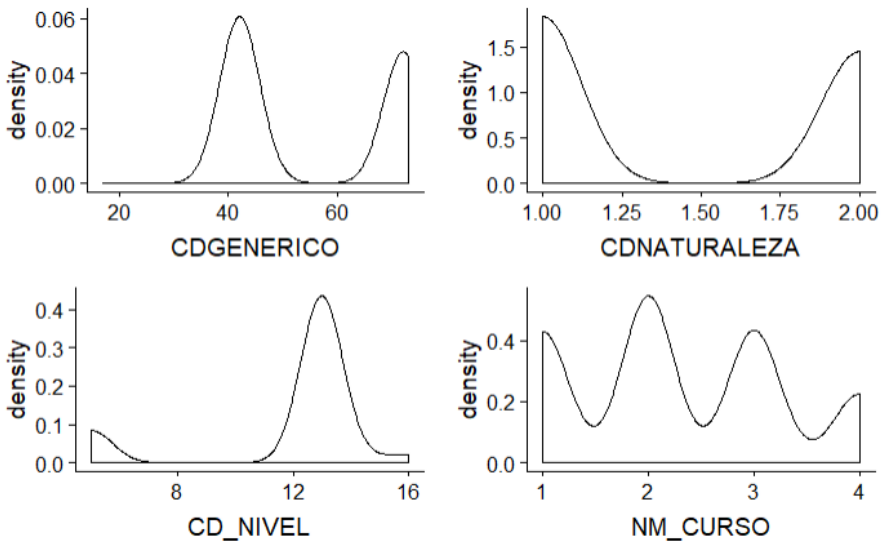


Figura 5.4: Distribución de variables cuando existe sobrepoblación. Elaboración propia



En la figura 5.4 puede observarse que cuando existe sobrepoblación en el aula, suele ocurrir en centros públicos, mas que en centros privados. Además, suele darse en los niveles de la ESO. Para cada nivel, dentro de los cursos existentes, la sobrepoblación se da para el segundo curso. Por ej: 2º ESO, 2º Bachillerato, etc. Por tanto, la sobrepoblación suele darse en el segundo curso de la ESO. Esto puede deberse al numero de repetidores.

Uno de los aspectos más importantes en el análisis exploratorio de datos es la correlación existente entre las variables. En la figura 5.5 se puede observar como existe una gran correlación entre las variables 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.

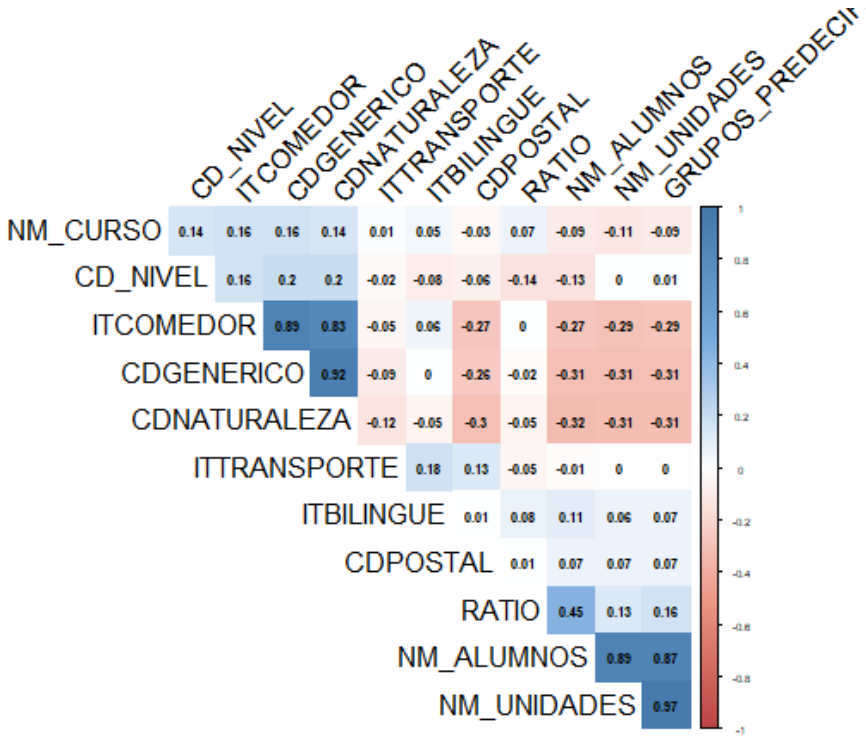
Otras relaciones observadas es que, como cabe esperar, el numero de unidades tiene una gran relación con las unidades a predecir. Véase la figura A.9 del Anexo A.B.2

En la figura 5.6 se pueden observar las mayores correlaciones entre variables ordenadas de mayor a menor.

En las figuras 5.5, 5.6 y 5.7 se muestra prácticamente la misma información, la correlación entre variables. En la figura 5.6 se muestran aquellos pares de variables que poseen la mayor correlación. Se puede apreciar en esta figura que la variable numero de unidades y grupos a predecir tienen una correlación casi perfecta, lo que implica que aportarían la misma información al conjunto de datos.

En la figura 5.7 se muestran aquellas variables que tienen mayor correlación con la variable a predecir. Podemos observar que las variables de numero de unidades y numero

Figura 5.5: Matriz de correlaciones. Elaboración propia



de alumnos 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. No sorprende puesto que son variables determinantes en la predicción. 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 tienden a matricular a sus hijos en centros públicos, y, por lo tanto, el número de grupos tiende a aumentar.

También nos interesa saber las variables que están correlacionadas con la ratio, entre estas variables nos encontramos con una correlación positiva bastante fuerte el número de alumnos. Si el número de alumnos crece, la ratio va a crecer obligatoriamente (mientras no se creen nuevos centros). También se puede destacar el servicio de comedor, que tiene una correlación negativa, lo que implica que las aulas con sobrepoblación son aquellas cuyo centro no ofrece un servicio de comedor. Este resultado se debe a que los centros de secundaria que suelen tener servicio de comedor, suelen ser centros privados, y, por lo tanto, suelen tener las ratios por debajo de 1. Lo mismo ocurre con otras variables como la naturaleza o el código genérico. Vease la figura 5.8

Figura 5.6: 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 5.7: 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

Figura 5.8: Mayores correlaciones con la Ratio. Elaboración propia

	First.Variable <fctr>	Second.Variable <fctr>	Correlation <dbl>
129	NM_ALUMNOS	RATIO	0.449411859
127	CD_NIVEL	RATIO	-0.135911631
130	NM_UNIDADES	RATIO	0.128148172
126	ITBILINGUE	RATIO	0.079067963
128	NM_CURSO	RATIO	0.066287454
122	CDNATURALEZA	RATIO	-0.053021395
125	ITTRANSPORTE	RATIO	-0.046613977
121	CDGENERICO	RATIO	-0.016884354
123	CDPOSTAL	RATIO	0.009881853
124	ITCOMEDOR	RATIO	0.002358286

Una vez obtenidos los resultados del análisis exploratorio, se muestran los resultados obtenidos en el análisis predictivo.

5.2. Análisis predictivo

En primer lugar, debe tenerse en cuenta que todas las variables comentadas no tienen la misma trascendencia en el resultado. Para detectar esta relevancia de variables se utilizan técnicas que muestren dichas variables. Utilizando Random Forest, en la figura A.10 del Anexo A.C.1 podemos observar que las más importantes a la hora de mejorar la precisión en el modelo son: el número de unidades y alumnos.

En cambio, se ha utilizado el algoritmo de Regresión Paso a Paso y se ha obtenido que se utilizan como máximo 5 variables (naturaleza del centro, numero de curso, número de unidades, nivel de enseñanza y ratio). Utilizando estas variables es cuando el modelo obtiene la mayor precisión. Son, por tanto, variables importantes en este estudio. El resto de variables aportan en la predicción, pero muy poco, por consiguiente, consideramos que no son variables que tengan influencia en la sobrepoblación del aula. Véase el Anexo A.C.2 para más información.

Una vez seleccionado el algoritmo a usar en la predicción (árbol de decisión), 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 cambian. Algunos de estos grupos se observan en la figura 5.9.

Figura 5.9: 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

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 de la zona en la que se encuentran dichos centros.

6 | CONCLUSIONES Y APORTACIONES

Una vez realizado el estudio y obtenido los resultados, es hora de realizar una valoración en forma de aportaciones y conclusiones.

6.1. Aportaciones de este TFM

En esta sección se quiere hacer especial hincapié en las contribuciones de este TFM.

En primer lugar, se debe destacar las relaciones de las variables con la ratio. Estas relaciones son útiles ya que por ejemplo a partir de variables como los servicios ofertados por un centro, se puede tener la certeza que los centros que poseen dichos servicios van a ser los más demandados y, por lo tanto, se debe tener un mayor cuidado con el número de plazas ofertadas y con la sobrepoblación. Otro claro ejemplo son los centros privados, que como se muestra, su ratio suele ser más bajo (están menos sobrepoblados).

Otra aportación que se realiza es el uso de herramientas de minería de datos para realizar la planificación de grupos en los centros educativos. Como se observa, la gran mayoría de artículos relacionados con la minería de datos en el entorno educativo, tratan sobre el rendimiento académico de forma directa. En esta investigación se ha contemplado otros aspectos del entorno educativo, como es el control de las ratios y la mejor planificación educativa.

Relacionado con la aportación anterior está el uso de modelos predictivos. En este TFM se puede observar como algunos modelos obtienen mejor predicción que otros para este conjunto de datos y en este ámbito de la educación.

Este TFM también aporta una investigación relativa a la gestión y a la planificación. Como ya se ha comentado, la mayoría de los artículos analizados se basan en la predicción del rendimiento de los alumnos y los factores que provocan dicho rendimiento (positivo o negativo), en este aspecto se abre el campo de estudio a lo referente a la planificación educativa. Destacando su importancia también en el aprendizaje y el resultado de los alumnos.

Por último, con este TFM, además de aportar un modelo predictivo que se ajusta a los datos y que ayude a la Unidad de Planificación de la CEI, se aportan variables que ayuden a controlar la ratio del aula y de esta forma, evitar que exista una sobrepoblación.

6.2. Conclusiones

Estas conclusiones se van a vincular con los objetivos propuestos en la introducción de este TFM.

La realización de este TFM tiene como objetivo principal satisfacer las necesidades de la CEI, contribuyendo a la óptima planificación de los grupos escolares para los nuevos cursos, evitando así la sobrepoblación en el aula. Para satisfacer este objetivo, se deben satisfacer unos sub objetivos establecidos. Apartado 1.2 Objetivos de la Introducción

El primer sub objetivo consiste en la selección de variables de interés, para ello se procede a utilizar algunas de las variables propuestas a partir de la instrucción de la Unidad de Planificación (Consejería de Educación e Investigación, 2018). Se tienen en cuenta inicialmente 27 variables (nombre, código y naturaleza del centro, código genérico de este, etc.), sin embargo, para el análisis predictivo se usan 11[1]. Se eliminan todas las variables del tipo descriptivo, como por ejemplo el nombre del centro (ya que no interesa, ni tampoco su código).

El segundo sub objetivo que debe cumplirse es el estudio de la relación de estas variables para comprender el contexto de la sobrepoblación en el aula. Para ello se utiliza la correlación entre las variables. A partir de estas relaciones se observa, concretamente, los factores que hacen que la ratio aumente o disminuya, y también se observa en qué grado lo hace.

Las variables más correlacionadas con la ratio son: el número de alumnos, la naturaleza, el servicio de comedor, el código genérico del centro y el servicio de bilingüismo. Estas variables son las que hacen que la ratio aumente o disminuya con más intensidad.

Una buena planificación en las aulas evita la superpoblación de estas y, por otra parte, también evita el gasto inadecuado de recursos. Este TFM no se centra en poner en duda la ratio de alumnos actual, sino más bien en optimizar los recursos según dicha ratio.

Para llevar a cabo esta planificación, en esta investigación se tienen en cuenta principalmente 10 variables: el código genérico del centro, su naturaleza, su código postal, sus servicios de transporte, comedor o bilingüismo (en caso que tengan), los niveles educativos, el curso del grupo a predecir, el número actual de unidades y la ratio de estas. Estas variables mayoritariamente son las que se utilizan actualmente en la CEI de la Comunidad de Madrid. Sin embargo, se proponen además nuevas variables como son la tasa de aprobados o suspensos de un determinado grupo, ya que, si la tasa de suspensos es alta, el número de matriculaciones para dicho grupo debe reducirse, y este factor condiciona claramente las futuras predicciones.

Finalmente, las variables utilizadas que realmente condicionan la predicción son las que se obtienen utilizando los algoritmos de eliminación de variables y son las siguientes:

[1]Las variables utilizadas son: el código genérico del centro, su naturaleza, su código postal, sus servicios de transporte, comedor o bilingüismo (en caso que tengan), los niveles educativos, el curso del grupo a predecir, el número actual de unidades, la ratio de estas y el numero de unidades reales del siguiente curso

naturaleza del centro, numero de curso, número de unidades, nivel de enseñanza y ratio. Todas estas variables supeditan la sobrepoblación del aula.

Es necesario destacar la existencia de otras variables expuestas en artículos analizados previamente. Estas variables son las que están relacionadas con el entorno del centro como por ejemplo: carreteras de acceso a este, estaciones de metro cercanas, generalmente, estructuras que puedan intervenir positiva o negativamente a la tasa de matriculaciones. Estas variables son de gran interés, puesto que pueden afectar a la planificación de grupos.

El tercer sub objetivo consiste en probar distintos modelos y seleccionar aquellos con mayor precisión, por ello, se utilizan una serie de algoritmos como las redes neuronales, regresión lineal, bosques aleatorios, arboles de decisión, K-vecinos cercanos y soporte de máquinas vectoriales. De estos modelos se obtiene que el árbol de decisión es el que mejor precisión obtiene. Sin embargo, la regresión lineal obtiene resultados parecidos y su tiempo de entrenamiento es inferior.

En la tabla B.1 del Anexo B se puede observar una comparativa sobre la precisión y el tiempo de entrenamiento de los modelos para los datos utilizados.

En el último sub objetivo se utiliza el algoritmo que mayor precisión aporta (arboles de decisión) con el fin de realizar predicciones con datos existentes (los 10 últimos años de la Comunidad de Madrid). Este modelo de árboles de decisión, por tanto, se aplica para el curso 2017/2018, obteniendo las predicciones para el curso 2018/2019. Estas predicciones no pueden ser contrastadas con las que realmente se producen para ese curso. Sin embargo, en reuniones posteriores, se admite que son pocos los centros que cambian el número de unidades de un curso a otro, coincidiendo con los resultados obtenidos.

Por último, se realiza una automatización del sistema de predicción, que ayuda a las personas de la CEI a planificar recursos respecto al número de unidades existentes. Se debe recordar que hasta ahora, la CEI, utiliza técnicas manuales que pueden tender a error en las predicciones.

Para concluir, con este TFM se han obtenido conocimientos mas allá de los adquiridos por el resto de asignaturas de este máster, con el fin también de mejorar el aprendizaje en el aula. Concretamente, este TFM se centra en nada menos que en la planificación y la gestión de la educación, parte trascendental para favorecer y mejorar la situación en el aula. Sin esta gestión y planificación, el aprendizaje dentro del aula se ve mermado. Se debe recordar que la situación actual en aula se debe gracias a la planificación y gestión de los recursos. Como docente se debe tener en cuenta el equilibrio entre el numero de grupos y los recursos, también tiene que comprenderse que, aunque la reducción del ratio es algo positivo, la mayoría de veces no se dispone de los recursos para afrontar dicha reducción.

6.3. Líneas De Trabajo Futuro

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 estudian 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 consideran por falta de datos, como por ejemplo la tasa de aprobados o de suspensos para un grupo de un nivel determinado.

Estas variables son importantes, puesto que dependiendo de esta tasa de suspensos o aprobados, se podrán aumentar o disminuir el número de matrículas para este determinado nivel. Podría incluirse, además, el número de enseñanzas del centro, sus modalidades del bachillerato, etc.

Otras variables, como ya se comenta anteriormente son las referidas a la geolocalización del centro, su posición estratégica, y como esto puede influir al aumento de la ratio y por ende a las predicciones finales.

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. Variando los parámetros de estos algoritmos se consigue una mayor precisión.

Por último, se debe tener en cuenta la realización de un Software teniendo en cuenta los modelos estudiados, que sea capaz, con una interfaz gráfica, de ayudar a los miembros de la Unidad de Planificación de Centros, sustituyendo el “Script” creado.

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

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

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

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

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

Fingermann, H. (2014). *Aulas superpobladas.* Descargado de https://educacion.laguia2000.com/general/aulas-superpobladas

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.

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.

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).

Lera Rodríguez, M. J. (2007). Calidad de la educación infantil: instrumentos de evaluación. *Revista de Educación, 343, 301-323.*.

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)

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

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.

Ministerio de Educación y Formacion Profesional. (2018). *Panorama de la educación indicadores de la ocde 2018*.

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).

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

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/

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.

Racancoj, L. V. M. (2013). *Sobrepoblación estudiantil y desempeño docente en el aula* (Thesis).

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

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

Shahiri, A. M., Husain, W., y Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414 - 422. Descargado de http://www.sciencedirect.com/science/article/

pii/S1877050915036182 (The Third Information Systems International Conference 2015) doi: https://doi.org/10.1016/j.procs.2015.12.157

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

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*.

8 | LISTA DE ACRÓNIMOS

CEI Consejería de Educación e Investigación

TFM Trabajo Fin de Máster

ESO Educación Secundaria Obligatoria

ESA Educación Secundaria para Adultos

CM Comunidad de Madrid

FP Formación Profesional

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	▪ 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	▪ Público = 1 ▪ Privado = 2
CDPOSTAL	Código postal del centro	
ITCOMEDOR	Disponibilidad de comedor en el centro	▪ No = 1 ▪ Si = 2
ITTRANSPORTE	Disponibilidad de transporte	▪ No = 1 ▪ Si = 2
ITBILINGUE	Disponibilidad de bilingüismo	▪ No = 1 ▪ Si = 2
CD_NIVEL	Nivel educativo del grupo	▪ 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	▪ 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.	
CDSUBDIR	Direccion de Area Territorial (DAT).	▪ DAT-Norte = 1 ▪ DAT-Sur = 2 ▪ DAT-Este = 3 ▪ DAT-Oeste = 4 ▪ DAT-Centro = 5
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 divido 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. 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.

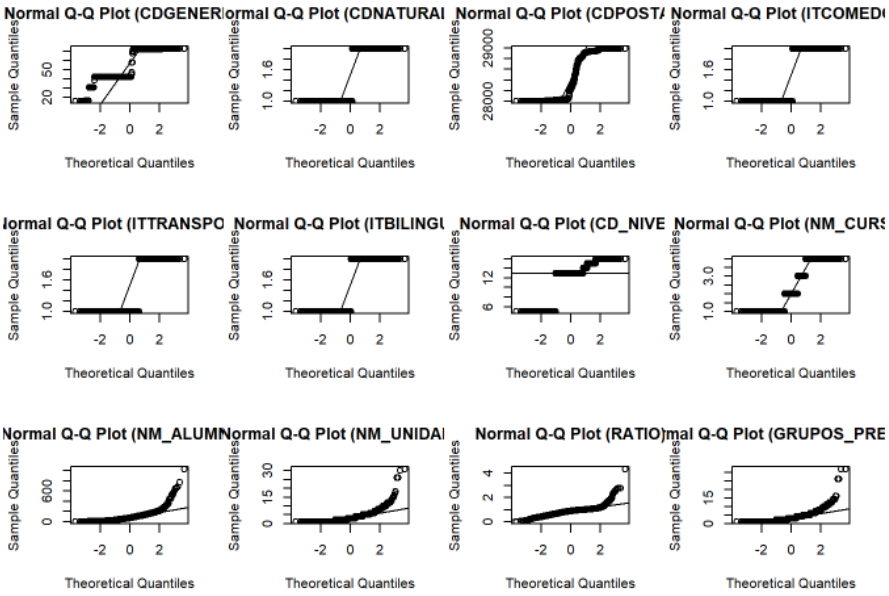
Para comprobar la normalidad en los datos se utiliza el test de Mardia que es una de las muchas pruebas que se utilizan para dicha comprobación. Existen otras pruebas como el test de Zirkler Henze, Shapiro-Wilk, etc.

Los resultados obtenidos utilizando la prueba de Mardia en el conjunto de datos son los que se pueden observar en la figura A.1

Figura A.1: 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.2: Resumen de normalidad de variables. Elaboración propia



Como se puede observar en la figura A.2, 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.3: Distribución 1. Elaboración propia

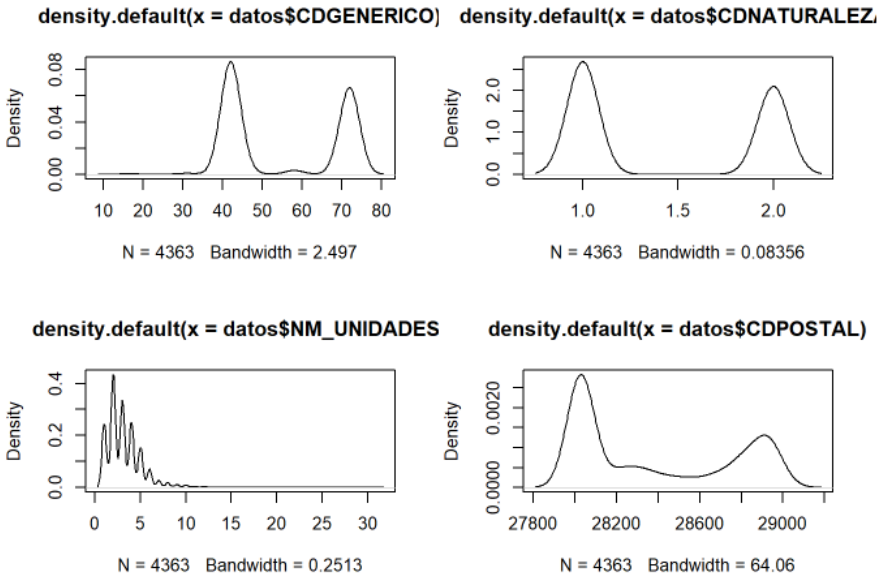
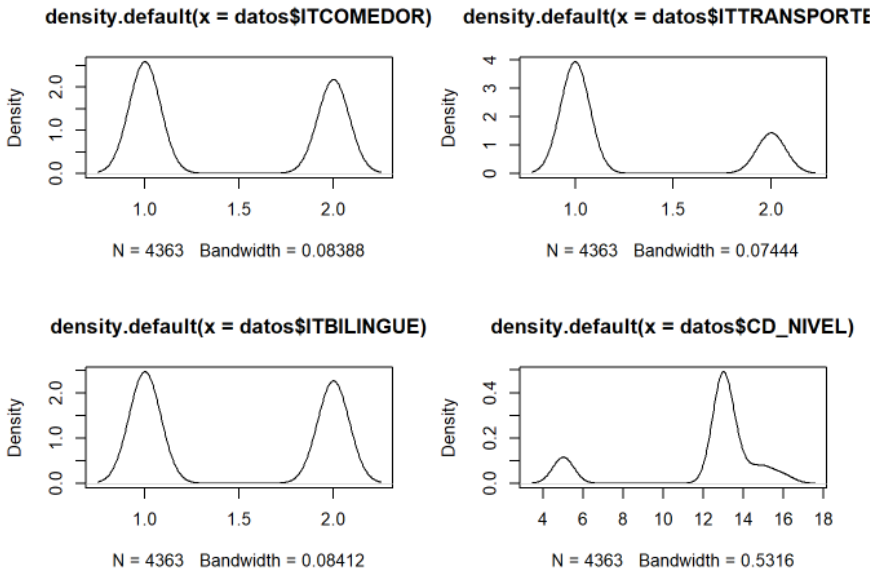


Figura A.4: Distribución 2. Elaboración propia



Las figuras A.3 y A.4, 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.5: Distribución 3. Elaboración propia

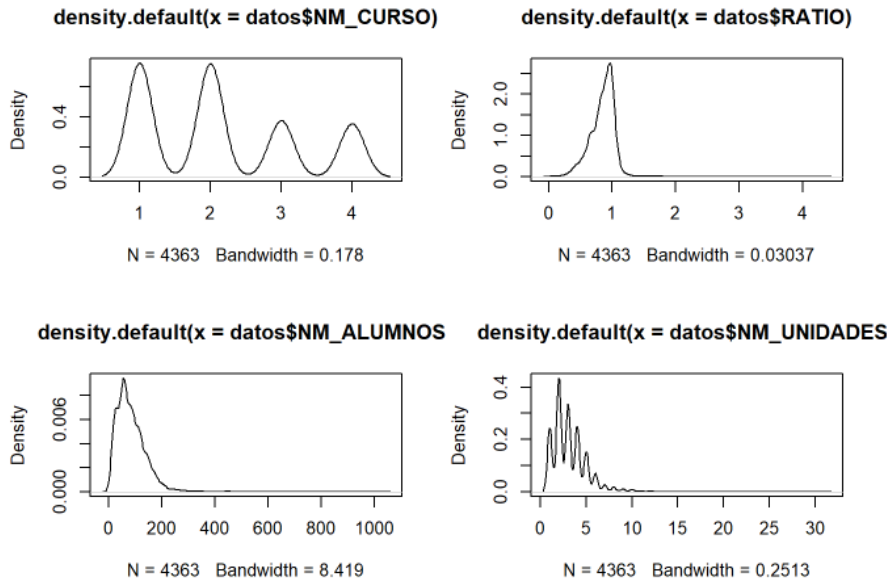


Figura A.6: Diagrama de barras 1. Elaboración propia

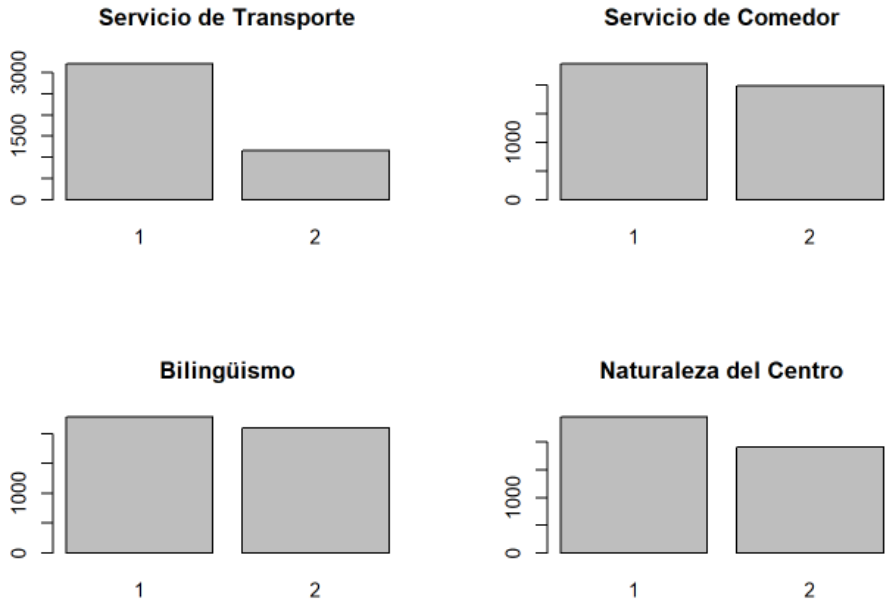
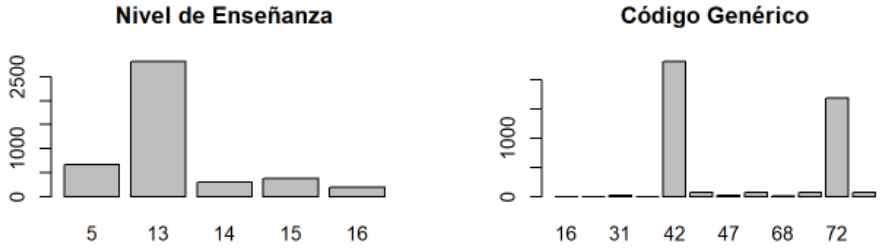


Figura A.7: Diagrama de barras 2. Elaboración propia



Tanto en la figuras A.6 y A.7 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.2. Relaciones entre variables

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

Figura A.8: Correlación entre variables NM_ALUMNOS y NM_GRUPOS. Elaboración propia

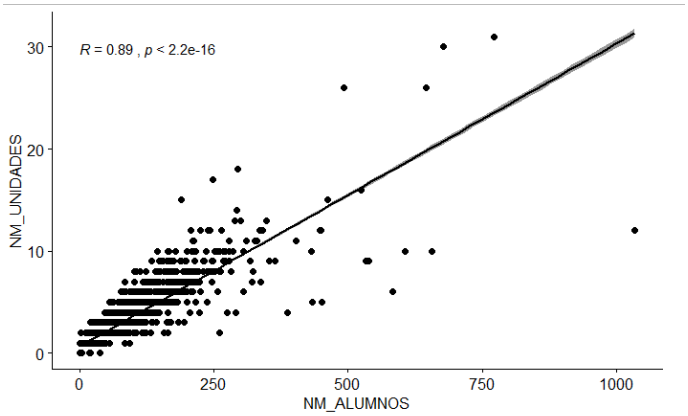
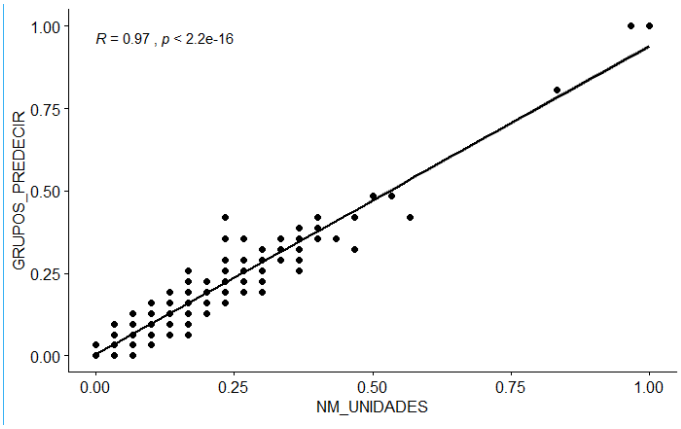


Figura A.9: Correlación entre variables NM_UNIDADES y GRUPOS_PREDECIR. Elaboración propia

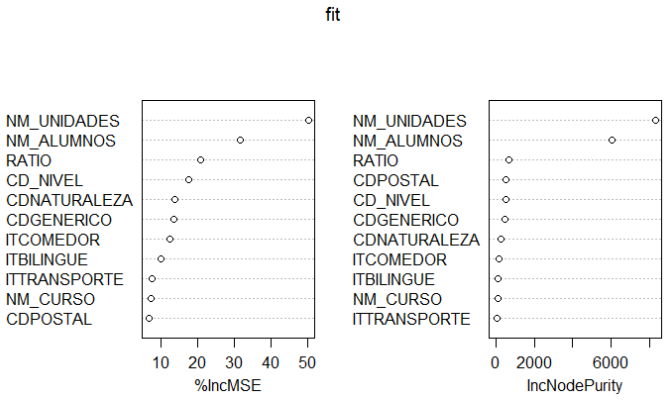


En la figura A.9 se observa la gran relación (prácticamente lineal) entre las variables NM_UNIDADES y GRUPOS_PREDECIR. Esto significa que, si aumenta el valor de una variable, linealmente aumenta el valor de la otra. En la figura A.8 existe también una relación lineal, pero no es tan pronunciada.

A.C. Selección de Variables

A.C.1. Usando Random Forest

Figura A.10: Variables más importantes usando Random Forest. Elaboración propia



La variable **IncNodePurity** se la conoce también como la media de decrecimiento de de Gini. El indice 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 indice 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 stepwise.

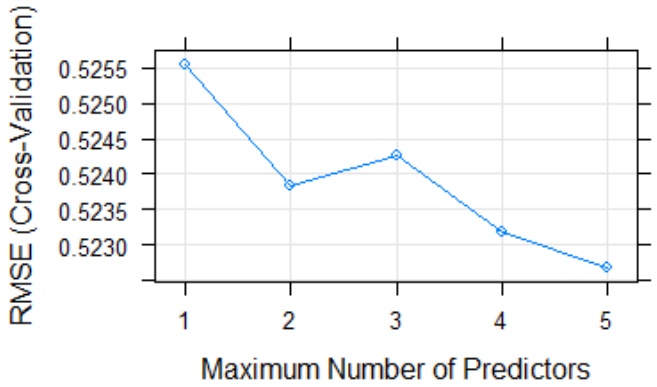
Usando Backward Selection

Figura A.11: 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.12: Gráfico Backward Selection. Elaboración propia



En la figura A.12 se puede observar como la mejor precision se obtiene utilizando los 5 predictores de la figura A.11.

Usando Stepwise Selection

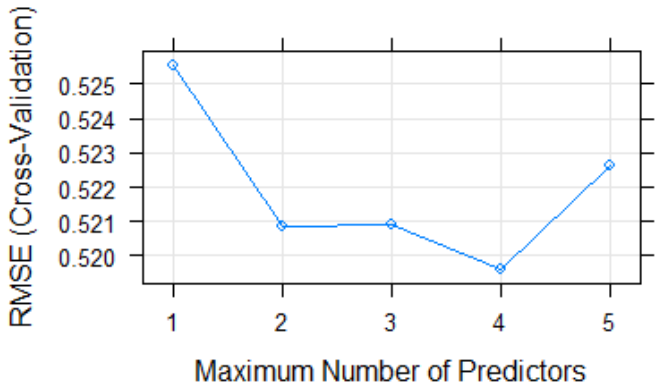
Figura A.13: Resultado Stepwise Selection. Elaboración propia

```
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.13 indican aquellos predictores que deben

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

Figura A.14: Gráfico Stepwise Selection. Elaboración propia



En la figura A.14 se puede observar como la mayor precisión (menor valor de RMSE) se obtiene utilizando 4 variables.

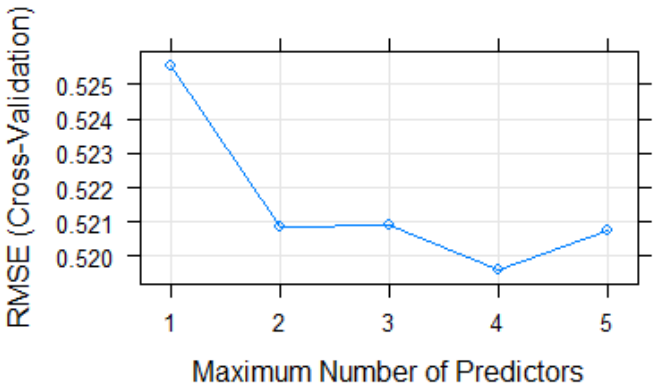
Usando Forward Selection

Figura A.15: 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.15 se puede observar como nuevamente se tienen en cuenta las variables CD_NATURALEZA, NM_CURSO, NM_UNIDADES y RATIO.

Figura A.16: 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.15.

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 (kknn), 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

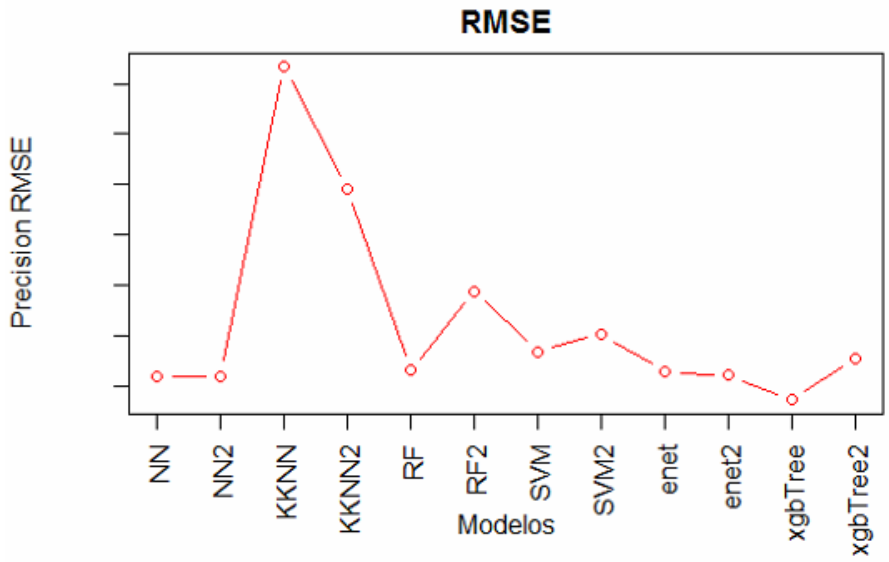
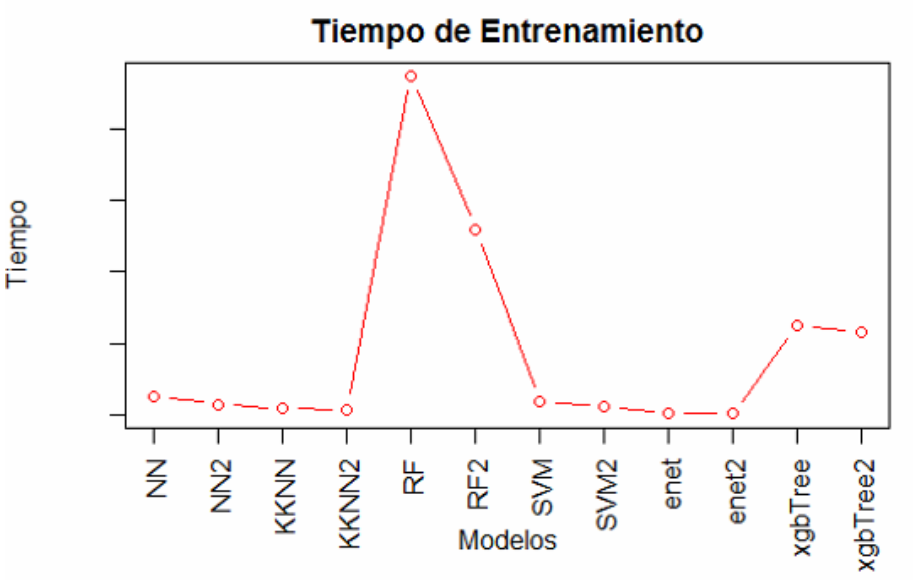


Figura B.2: Comparación Modelos y Tiempo en Entrenar. Elaboración propia



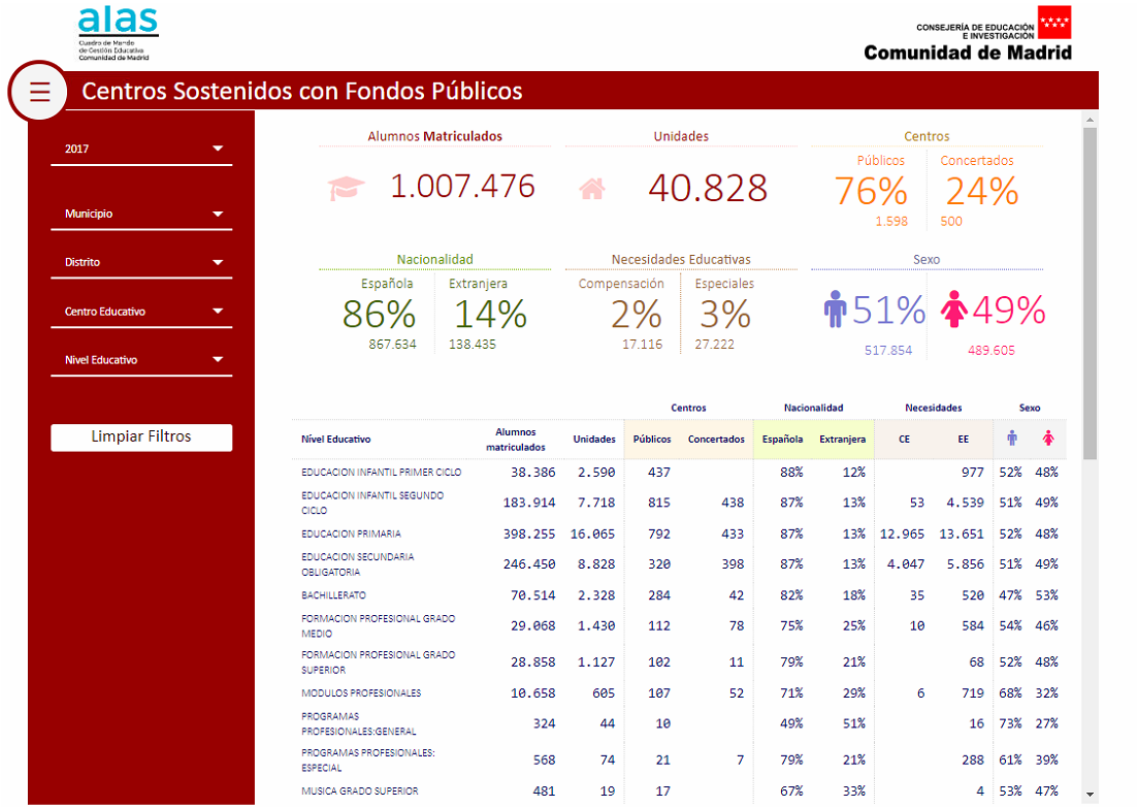
C | Sistema de Explotación de la Consejería de Educación e Investigación

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 numero de unidades, el porcentaje de centros públicos y concertados, la nacionalidad de los alumnos, el sexo, etc. Véase la figura C.1

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.

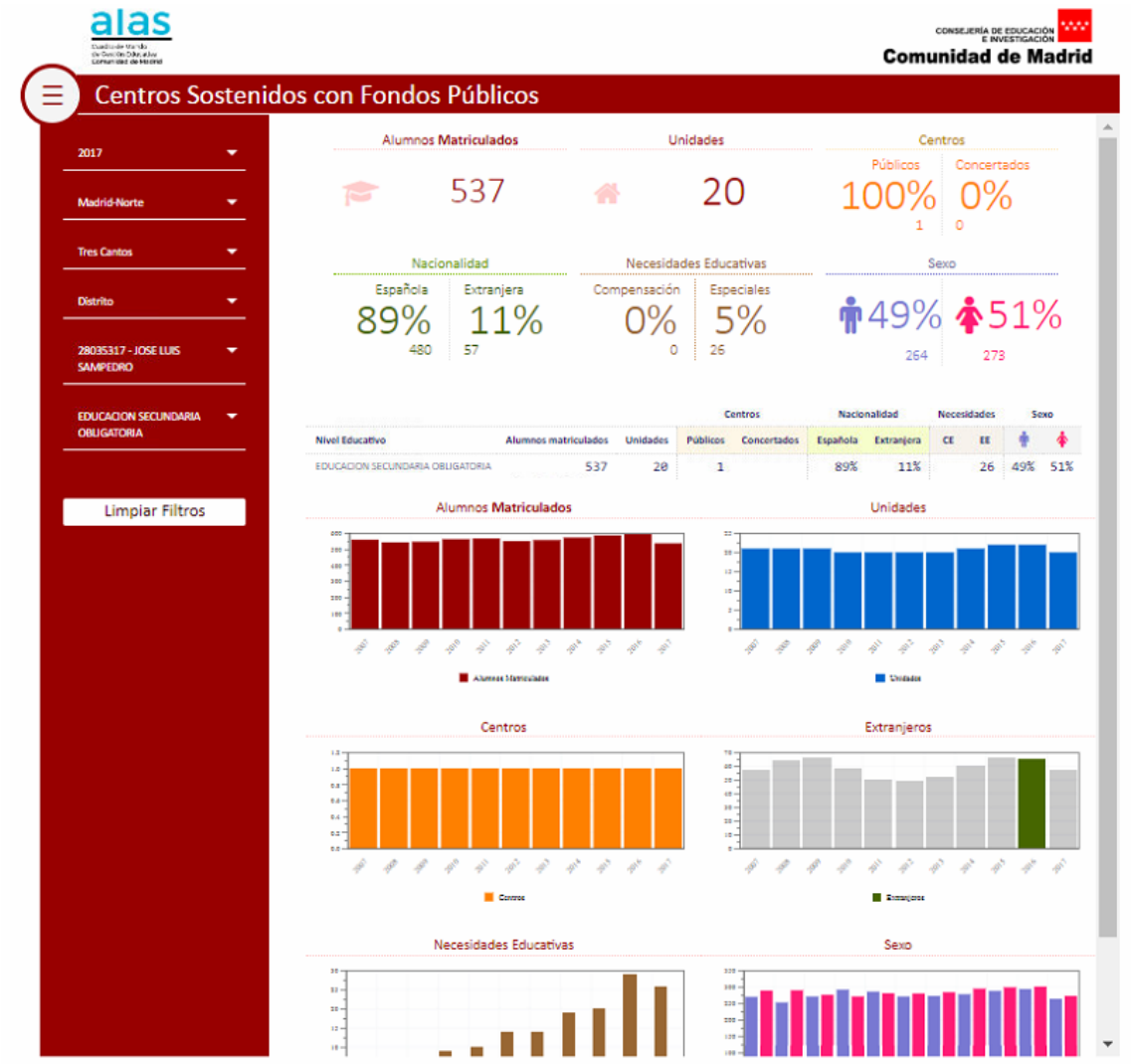
Figura C.1: Cuadro de Mandos de Centros I. Elaboración propia



Nivel Educativo	Alumnos matriculados	Unidades	Centros: Públicos	Centros: Concertados	Nacionalidad: Española	Nacionalidad: Extranjera	Necesidades: CE	Necesidades: EE	Sexo	
EDUCACION INFANTIL PRIMER CICLO	38.386	2.590	437		88%	12%		977	52%	48%
EDUCACION INFANTIL SEGUNDO CICLO	183.914	7.718	815	438	87%	13%	53	4.539	51%	49%
EDUCACION PRIMARIA	398.255	16.065	792	433	87%	13%	12.965	13.651	52%	48%
EDUCACION SECUNDARIA OBLIGATORIA	246.450	8.828	320	398	87%	13%	4.047	5.856	51%	49%
BACHILLERATO	70.514	2.328	284	42	82%	18%	35	520	47%	53%
FORMACION PROFESIONAL GRADO MEDIO	29.068	1.430	112	78	75%	25%	10	584	54%	46%
FORMACION PROFESIONAL GRADO SUPERIOR	28.858	1.127	102	11	79%	21%		68	52%	48%
MODULOS PROFESIONALES	10.658	605	107	52	71%	29%	6	719	68%	32%
PROGRAMAS PROFESIONALES:GENERAL	324	44	10		49%	51%		16	73%	27%
PROGRAMAS PROFESIONALES: ESPECIAL	568	74	21	7	79%	21%		288	61%	39%
MUSICA GRADO SUPERIOR	481	19	17		67%	33%		4	53%	47%

En la figura C.2 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).

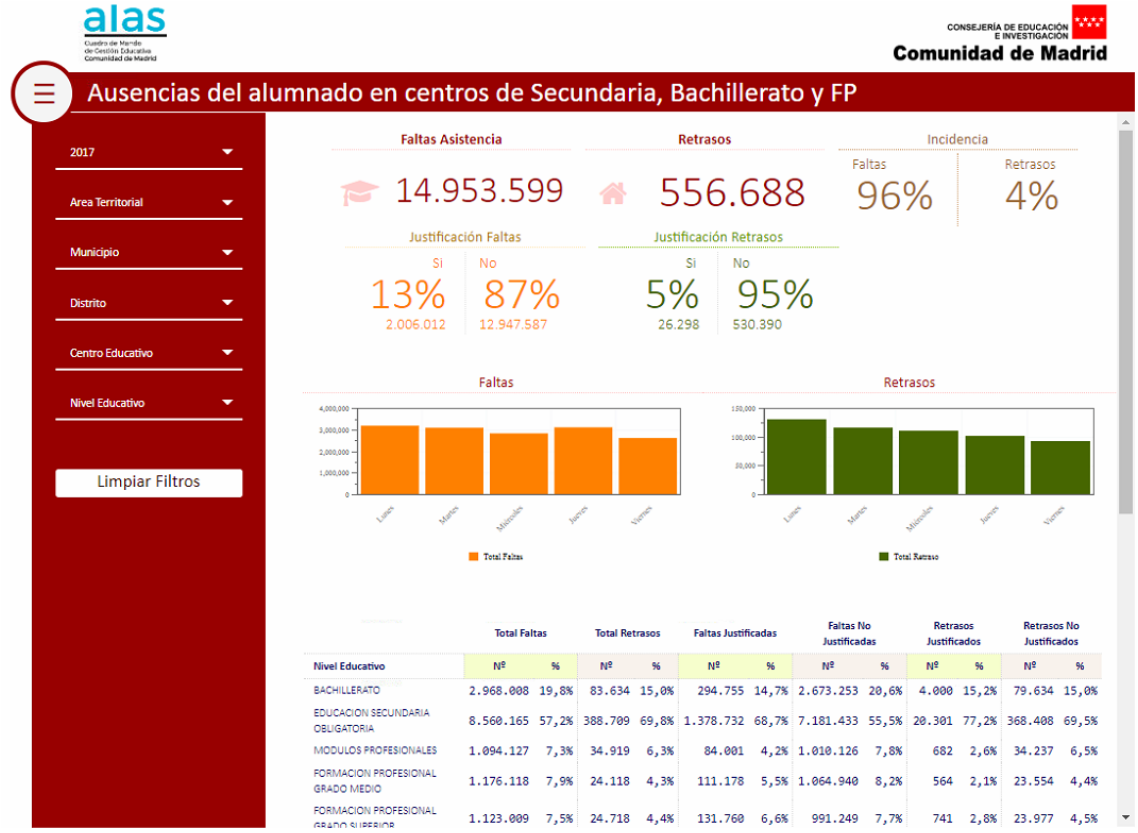
Figura C.2: Cuadro de Mandos de Centros II. Elaboración propia



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

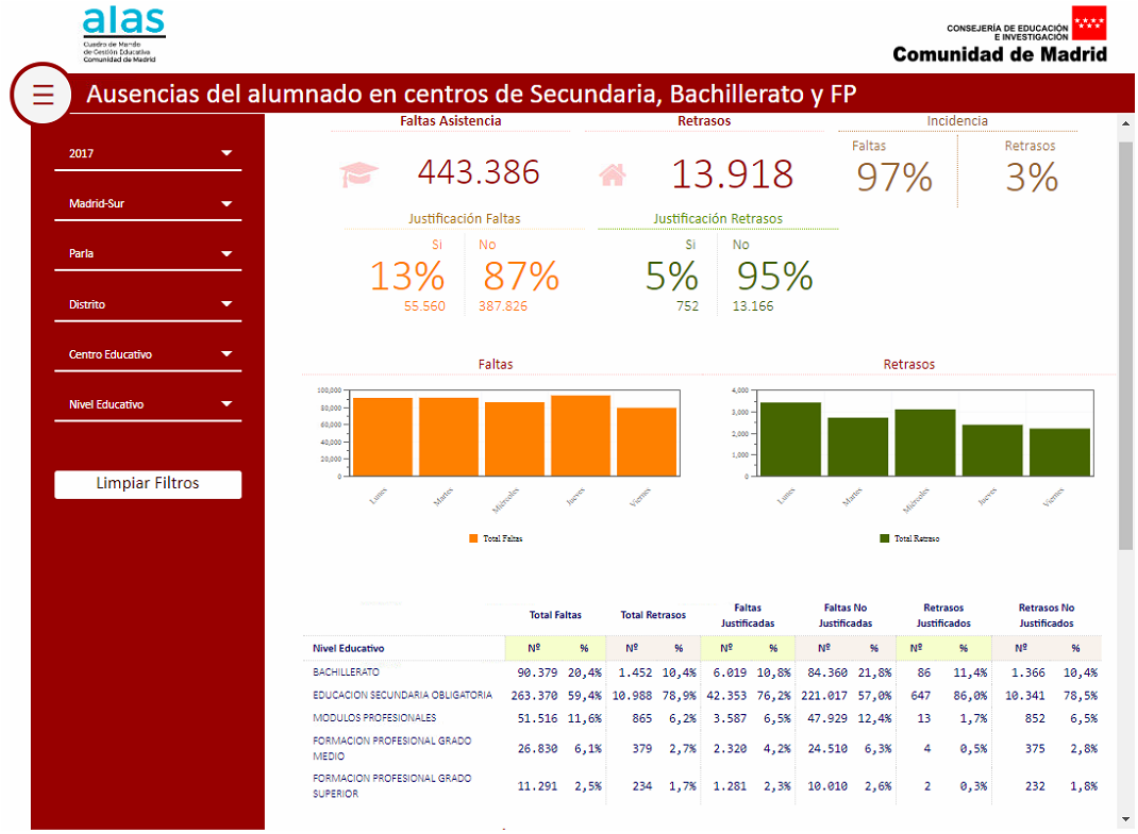
En este cuadro de mando aparecen las faltas de asistencias totales, el número de retrasos, el porcentaje sobre el total entre faltas y retrasos, el porcentaje de faltas y retrasos justificados. Se muestra un histograma también sobre el número de faltas y retrasos y los días de la semana.

Figura C.3: Cuadro de Mandos de Ausencias de Alumnado I. Elaboración propia



En la figura C.4 se filtra la búsqueda por el municipio de Parla (DAT-Sur), para el año 2017.

Figura C.4: 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 C.5 el cubo OLAP relativo a las unidades.

Figura C.5: Cubo OLAP de Unidades. Elaboración propia