Universidad Rey Juan Carlos

Trabajo Fin de Máster

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

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

Especialidad en Informática y Tecnología

Curso 2018-2019

AUTOR: Abel de Andrés Gómez
DIRECTOR: Aurelio Berges García

AGRADECIMIENTOS

Agrademos a...

RESUMEN

Extensión máxima de una página

SUMMARY

Extensión máxima de una página

Índice

Índice de figuras

Índice de cuadros

1. INTRODUCCIÓN

1.1. Introducción

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

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

1.2. Contexto

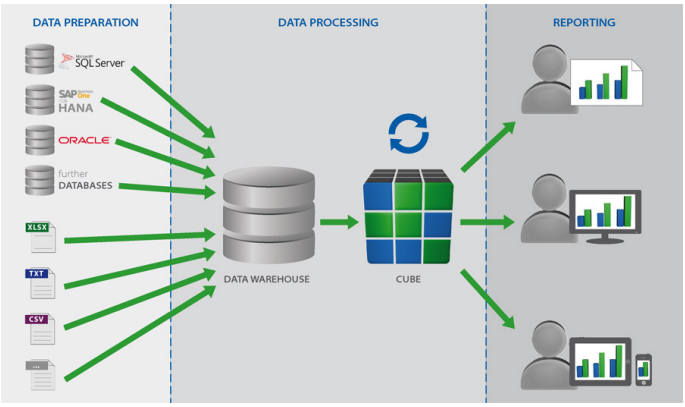
La organización educativa no ha quedado ajena a estas necesidades de una mejor comprensión de los datos. Desde la Consejería de Educación de Madrid se están realizando proyectos para conseguir sacar la máxima información del gran número de datos que se poseen.

Obviamente, debido a este gran tamaño de datos, es necesario utilizar métodos, herramientas y personas con conocimientos para obtener información concreta en un tiempo legible. Desde la Consejería de Educación se quiere tener conocimientos actuales sobre la situación educativa. Un ejemplo podría ser el número de alumnos matriculados con necesidades educativas para un determinado centro de la Dirección de Área Territorial Sur. No solo eso, también podrían obtenerse alumnos de un determinado nivel educativo o incluso grupos.

En la figura 1 se pude observar la arquitectura de un almacén de datos. Mediante esta arquitectura, los usuarios finales pueden obtener mucha información sin necesidad de realizar consultas complejas a bases de datos.

No obstante, mostrar la información actual o pasada no es suficiente. La Consejería de Educación también requiere obtener datos futuros. En este aspecto, la Consejería necesita saber cuántos alumnos podrán matricularse en el futuro, con el objetivo de destinar recursos a los centros. Por tanto, será necesario utilizar técnicas predictivas. Estas técnicas se pueden usar perfectamente en la Consejería puesto que requieren una gran cantidad de datos para realizar pronósticos ajustados, en este sentido, la Consejería tiene un gran histórico de años anteriores.

Fig. 1: Arquitectura de un almacén de datos. Recuperado de: Superior Information Technology.



1.3. Justificación y Objetivos

En este sentido (de las predicciones), una unidad de Educación Secundaria Obligatoria de la Consejería de Educación de la Comunidad de Madrid ha planteado un problema.

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

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

Estos aspectos son:

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

3. Posibilidad de creación de nuevas zonas urbanas cerca del centro.

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

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

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

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

Por ello, con este TFM se propone dar una solución al problema actual mediante el uso de herramientas y métodos flexibles que automaticen dichas tareas y proponga, además, nuevas variables o factores que puedan influir en la toma de decisión. El objetivo final consiste, por tanto, en obtener predicciones precisas para la toma de decisiones.

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

- Seleccionar variables que interesen estudiar y que aporten valor en el desarrollo de este TFM.

- Obtener modelos que se ajusten correctamente a los datos y que aporten una gran precisión.

- Realizar predicciones con datos existentes.

1.4. Metodología

Para ello, en primer lugar, se van a realizar reuniones con los responsables de la unidad de Educación Secundaria. A partir de estas reuniones se van a obtener las fuentes donde se encuentre la información relevante. Una vez que se conocen las fuentes, se van a estructura los datos existentes.

Después se seleccionarán aquellos datos que se consideran importantes para realizar las labores de predicción. Para realizar la selección, se debe estudiar los mecanismos ya existentes en esta Unidad de Secundaria para obtener aquellos en uso. Una vez que se han seleccionado los datos, se va a realizar un tratamiento de estos para poder utilizarlos en las nuevas herramientas.

Por último, se van a aplicar modelos predictivos y se van a seleccionar aquellos que mayor precisión aporten con dichos datos.

1.5. Organización del TFM

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

- **Capitulo 1. Introducción:** En el primer capítulo se van a definir las necesidades existentes que justifican el desarrollo de este trabajo. También se va a definir los objetivos que se persiguen con la realización de este. Por último, se presenta la estructura que tendrá el presente documento.
- **Capitulo 2. Justificación teórica:** En este segundo capítulo se va a realizar una investigación sobre el estado de la cuestión. Se va a realizar un estudio sobre los métodos, modelos y usos de la minería de datos en el ámbito educativo.
- **Capitulo 3. Propuesta de intervención:** En este tercer capítulo se va a plantear una solución al problema existente.
- **Capitulo 4. Diseño de la investigación:** Este capítulo va a definir los pasos que se seguirán en la realización de un proyecto de minería de datos. Se van a detallar también las tareas que se van a desempeñar en cada uno de los pasos.
- **Capitulo 5. Conclusiones:** En este capítulo se van a detallar las conclusiones obtenidas a partir de los resultados alcanzados.

2. JUSTIFICACIÓN TEÓRICA

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

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

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

Para la realización de este TFM se han analizado distintas publicaciones de la base de datos científica de ScienceDirect.

Para realizar la búsqueda se han utilizado las siguientes palabras clave: educational, data y mining. Se debe recordar que el éxito de la búsqueda depende de estas palabras claves.

También se ha realizado una búsqueda utilizando estas claves en Teseo y Google Academics, esta última búsqueda ha proporcionado otro artículo de gran interés.

De la búsqueda en ScienceDirect se han obtenido 160 artículos. Posteriormente se han tenido en cuenta aquellos de los últimos 5 años (2015, 2016, 2017, 2018 y 2019), sin olvidarse del resto. De esta forma obtenemos resultados actuales. Filtrando por fechas, hemos conseguido reducir los resultados a 73 artículos. Se ha realizado una observación sobre los artículos obtenidos y se ha comprobado la existencia de artículos que no resultaban útiles en esta investigación. Por tanto, se ha realizado otra búsqueda utilizando las claves anteriores y añadiendo la clave "prediction". Esta vez, se han obtenido 26 resultados. De todos los resultados obtenido, se han seleccionado 15 artículos que se consideran útiles y que servirán de ayuda.

Debemos destacar que la mayoría de los resultados obtenidos tratan de artículos centrados en la predicción de los resultados académicos 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.).

En este sentido es interesante realizar un análisis de dichos artículos, puesto que en primer lugar se deberá tener en cuenta cuales son las metodologías de la ciencia de datos que se están utilizando. Además, en segundo lugar, se debe tener en cuenta

los modelos que se utilizan para predecir variables de carácter educativo.

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

2.1. Usos de la minería de datos en la educación

Después de realizar un análisis sobre los artículos encontrados, se ha descubierto que la minería de datos se utiliza para resolver distintos problemas en la educación. Como ya se ha comentado, la mayoría de los artículos se centran en el rendimiento y en las calificaciones de los alumnos y, cómo teniendo en cuenta estas investigaciones, se puede mejorar la calidad educativa.

En el artículo de Fernandes et al. (2019), los datos escolares a estudiar proceden de alumnos de colegios de un Distrito Federal de Brasil durante el 2015 y el 2016. Estos datos se han obtenido a partir de la base de datos de iEducar que contiene atributos relacionados con cada alumno.

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

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

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

Por otro lado, en el artículo de Asif, Merceron, Ali, y Haider (2017), se realizan otras investigaciones relativas al rendimiento académico, donde también se utilizan variables sociales como la edad, sexo, nacionalidad, estado civil, desplazamiento (si el alumno vive fuera del distrito), necesidades especiales, tipo de admisión, situación

laboral, situación económica, etc.

Los datos utilizados en este artículo proceden de las calificaciones del cuarto año del grado de ingeniería de Tecnología Informática de una universidad de Pakistán. Se van a tomar 210 alumnos que se han matriculado en los cursos de 2007-2008 y 2008-2009. Los datos contienen variables relacionadas con las calificaciones de pre-admisión de los alumnos y de las calificaciones de estos en los siguientes 4 años del programa de grado.

El objetivo es, nuevamente, obtener información sobre el rendimiento de estudiantes para que las personas interesadas (directores y docentes) puedan mejorar el programa educativo. Los enfoques para lograr este objetivo son los siguientes:

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

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

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

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

Relacionado con el rendimiento académico, existe también un artículo en el que se realiza labores de predicción para evitar el fracaso escolar. En este artículo, (Vera, Morales, y Soto, 2012), se han seleccionado variables en el que se incluyen si el alumno fuma, bebe, si tiene alguna discapacidad física, la edad, el nivel económico entre otras muchas. Los datos de este artículo se han obtenido a partir de encuestas realizadas a alumnos, del Centro Nacional de Evaluación y del Departamento de Servicios Escolares.

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

También se ha revisado un artículo relacionado con la mejora académica de alumnos de ingeniería en los primeros 3 años de titulación. Este artículo de Adekitan y Salau (2019) utiliza datos de una universidad de Nigeria. Este articulo realiza una exploración sobre 1841 estudiantes durante sus primeros 3 años, entre los años 2002 y 2014. Inicialmente se consideraron 18 variables, sin embargo, solo se utilizaron 6 variables que son las siguientes: matriculación, genero, especialidad de los estudiantes, ciudad del estudiante, calificaciones y tipo de educación secundaria recibida previamente.

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

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

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

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

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

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

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

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

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

2.2. Metodología de Trabajo

En primer lugar, y antes de realizar cualquier trabajo, es necesario tener en cuenta una metodología valida a seguir, es decir, se debe seguir una serie de pasos

para conseguir los objetivos determinados. Existen distintas metodologías de trabajo para realizar un proyecto de minería de datos. Sin embargo, según el artículo de Moine et al. (2011), las metodologías que abarcan todas las posibles etapas de un proyecto serian CRISP-DM y Catalyst. El resto de metodologías no completan todas las fases que se debiera o simplemente establecen los pasos a seguir, pero no las tareas. La comparativa se puede mostrar en la figura 2.

Fig. 2: Comparación metodologías de Minería de Datos. Recuperado de Moine et al. (2011)

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

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

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

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

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

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

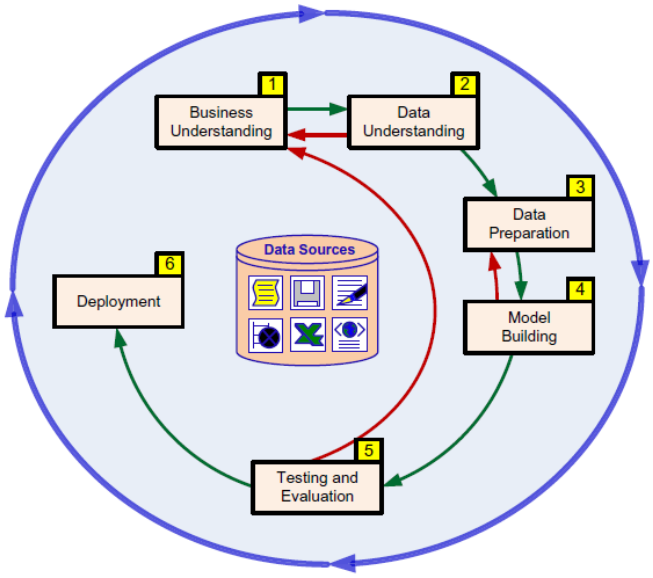
4. Construcción del modelo. Se debe desplegar un gran número de modelos y quedarse con aquellos que devuelvan valores óptimos para los datos utilizados.

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

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

En la figura 3, obtenida del artículo de Şen et al. (2012) se muestra el ciclo de CRISP.

Fig. 3: Ciclo de la metodología CRISP. Recuperado de Şen et al. (2012)



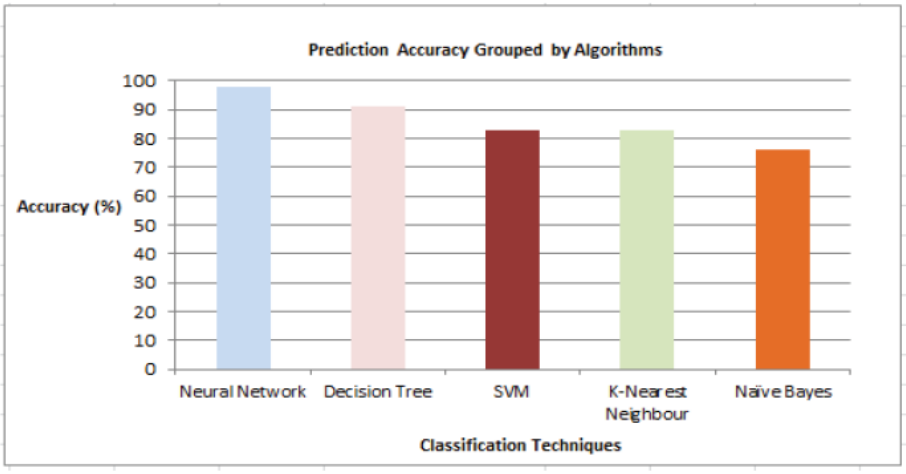
Según (Jaramillo y Arias, 2015) en su investigación se ha seleccionado Crisp-DM por las siguientes razones: "La metodología a utilizar es Crisp-DM ya que cada una de sus fases se encuentra claramente estructurada definiendo de tal forma las actividades y tareas que se requieren para lograr el objetivo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

2.3. Modelos utilizados en el ámbito educativo

En el artículo de prensa de Fernandes et al. (2019), se muestra el uso de técnicas como los métodos de clasificación y el algoritmo predictivo de GBM (Gradient Boosting Model) con el objetivo de obtener aquellas variables en el entorno del alumno, que hace que este obtenga mejores o peores resultados escolares. Este estudio, además, tiene el objetivo de aportar información útil para los representantes políticos en el ámbito educativo, el consejo escolar y los profesores con el objetivo de que estos puedan realizar políticas públicas, materiales didácticos y trabajo social para beneficiar a los estudiantes.

En el artículo de (Shahiri et al., 2015) se indica que "a priori", sin tener en cuenta la experiencia, es necesario realizar un proyecto piloto, que responda a dos preguntas en concreto. La primera pregunta que se plantea son los atributos o variables a utilizar en la investigación. La segunda pregunta planteada es sobre los métodos predictivos a utilizar. La siguiente figura 4 obtenida del artículo, muestra la precisión en la predicción de los algoritmos entre los años 2002 y 2015.

Fig. 4: Predicción en la precisión agrupada por algoritmos desde 2002 a 2015



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

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

pertenecientes a la Universidad Kashmir.

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

En cuanto al artículo de Adekitan y Salau (2019), nuevamente se han utilizado algoritmos como redes neuronales, bosques aleatorios, arboles de decisión, Naïve Bayes, combinación de árboles y regresión logística. En la 5 se puede observar la comparación entre los modelos.

Fig. 5: Comparación de los resultados obtenidos. Recuperado de Adekitan y Salau (2019)

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

En este artículo (Adekitan y Salau, 2019), se puede observar como la regresión logística obtiene la mayor precisión en los resultados. Por tanto, es capaz de clasificar correctamente los datos y, en consiguiente, obtener mejores predicciones.

2.4. Herramientas utilizadas

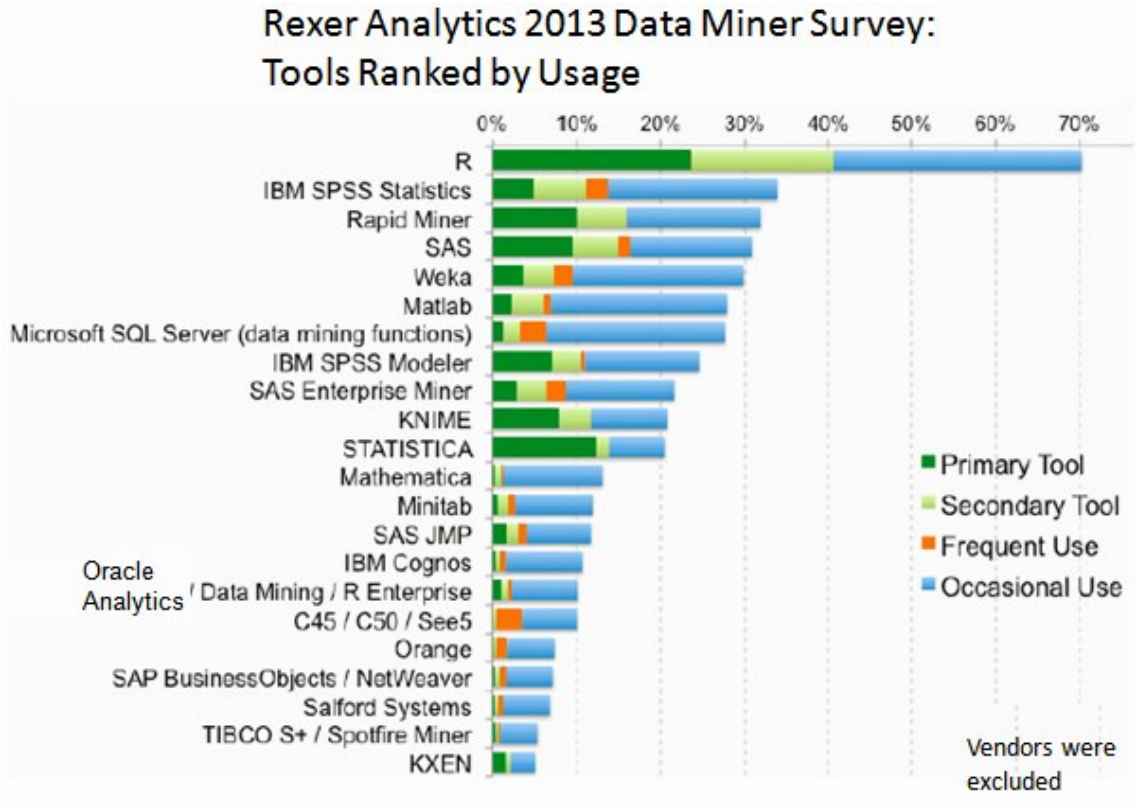
Existen diversas herramientas para realizar la minería de datos. En el artículo de (Rodríguez Suárez y Díaz Amador, 2009) se recogen algunas de las más utilizadas. Entre ellas se puede destacar SPSS Clementine, WEKA y Oracle Data Miner. Además, artículos como el de (José, 2016) han utilizado R, que es un lenguaje estadístico.

En el artículo de Jaramillo y Arias (2015), se ha realizado una breve comparación nuevamente entre las herramientas de WEKA, RapidMiner y Knime. De esta comparación, los autores han seleccionado la herramienta de RapidMiner para realizar

las investigaciones por las siguientes características: “posee una licencia libre, combinación de modelos, interfaz amiga, multiplataforma, empleo de técnicas, además permite aplicar varios algoritmos de minería de datos...” (Jaramillo y Arias, 2015)

En la figura 6 se pueden observar las 10 herramientas más utilizadas en 2013 según Rexer Analytics.

Fig. 6: Herramientas más usadas. Recuperado de (Piatetsky, 2013)



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

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

2.5. Conclusiones

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

En cuanto a las metodologías, existen muchas investigaciones que utilizan sus propias metodologías en vez de utilizar aquellas de uso frecuente.

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

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

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

3. PROPUESTA DE INTERVENCIÓN

3.1. Minería de datos

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

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

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

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

- El aprendizaje supervisado (o predictivos): se basa en algoritmos que intentan encontrar una función, que, dadas las entradas, asigne unas salidas adecuadas. Estos algoritmos se entrenan mediante datos históricos y de esta forma aprende a asignar salidas adecuadas en función de dichas entradas, dicho de otra forma, predice el valor de salida. A su vez, el aprendizaje supervisado se divide en regresión (si la salida es de tipo numérico) y clasificación (si la salida es del tipo categórico). (Recuero, 2017)

- El aprendizaje no supervisado: se utiliza en datos en los que existen variables de entrada, pero no existen variables de salida para dichas variables de entrada. Por consiguiente, solo se puede describir la estructura de los datos, para intentar conseguir algún tipo de tendencias y patrones que simplifiquen el análisis.(Recuero, 2017) (Rodríguez Suárez y Díaz Amador, 2009)

3.2. Lenguaje R y RStudio

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

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

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

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

3.3. Modelos seleccionados

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

3.3.1. Criterio de selección

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

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

al., 2009)

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

3.3.2. Métricas de precisión

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

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

Donde:

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

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

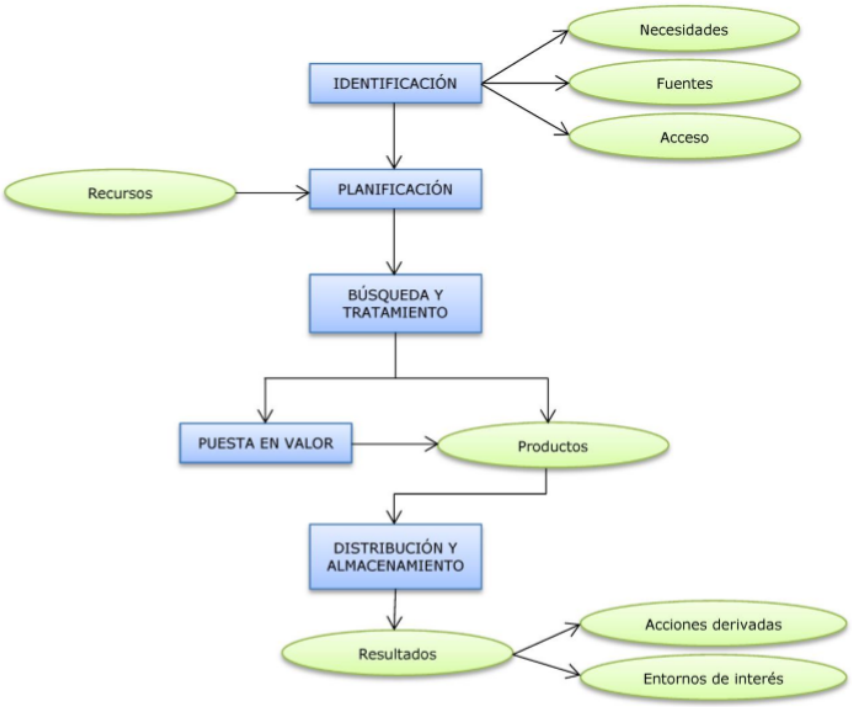
4. DISEÑO DE INVESTIGACIÓN

Uno de los pilares básicos en el diseño de una investigación es indicar el camino que se va a seguir en esta. Es importante establecer que estándar o norma se va a seguir en el desarrollo de un proyecto o una investigación. En esta investigación se va a utilizar la norma UNE 166006:2018 Gestión de la I+D+I: Sistemas de vigilancia e inteligencia. Esta norma está alineada con la norma UNE-EN ISO 9001 Sistema de Gestión de Calidad.

La norma UNE 166006 (2018) tiene como objeto facilitar la formación y estructuración del proceso de recogida, análisis y comunicación de la información sobre el entorno de una organización. No solo muestra un proceso, sino que también establece roles, responsabilidades y políticas.

La UNE 166006 (2018) 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. En la figura 7 se puede observar cada etapa.

Fig. 7: Proceso de la vigilancia e inteligencia. Recuperado de UNE 166006 (2018).



En los próximos puntos se va a describir las actividades que se van a realizar en este proceso.

4.1. Identificación de necesidades, fuentes de información y medios de acceso

4.1.1. Identificación de necesidades de información

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

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

4.1.2. 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 van a utilizar las siguientes fuentes:

- Fundamentalmente se va a utilizar documentos y recursos internos de la organización como van a ser:
  - Repositorios documentales.
  - Carpetas locales.
  - Bases de datos.
  - Etc.
- Fuentes documentales a las que tiene acceso a la organización, ya sea en soporte físico (revistas, catálogos, etc.) como en soporte electrónico. Además, se utilizarán recursos de información en Internet (portales, noticias, redes sociales, foros, etc.).
- Personas con conocimientos o experiencias relacionadas con la necesidad de información. En este aspecto se van a realizar distintas reuniones con las personas encargadas de esta unidad de la consejería de educación.
- 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.

4.2. Planificación de la realización de la vigilancia e inteligencia

Para realizar la planificación del proceso, la UNE 166006 (2018) indica que se debe realizar una búsqueda de nuevas áreas desconocidas y realizar un seguimiento sistemático de novedades en áreas previamente identificadas.

Esta etapa del proceso se ha contemplado mediante una justificación teórica, donde se ha hecho una investigación acerca de las metodologías, técnicas y herramientas.

4.3. Búsqueda y tratamiento de la información

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 consejería de educación maneja grandes volúmenes de datos, por tanto, se debe tener conocimiento sobre donde se puede encontrar la información que satisfaga con las necesidades.

El desconocimiento del estado de las bases de datos conlleva la inversión de una gran cantidad de tiempo en la búsqueda de los datos relevantes.

Una vez que se tienen los datos, muchas veces es necesario realizar un tratamiento de estos, que consiste en una limpieza y una normalización de los mismos. Muchas veces este tratamiento conlleva la conversión de datos, como por ejemplo fechas, corrección de datos, etc.

4.3.1. Proceso de Extracción, Extracción y Carga

Para realizar este tratamiento de datos se utilizará la técnica conocida como ETL (extracción, transformación y carga) que consiste básicamente en obtener los datos de la fuente de origen (bases de datos, ficheros Excel, ficheros JSON, etc.), seleccionar aquellos datos que convengan al estudio, transformarlos según las necesidades que se tenga y depurarlos (evitando así datos erróneos). (Prakash y Rangdale, 2017) (Matos, Chalmeta, y Coltell, 2006),(Gour, Sarangdevot, Tanwar, y Sharma, 2010). Para realizar este tratamiento, se ha va a utilizar Pentaho BI, que es un conjunto de programas libres para realizar entre otras muchas actividades, las técnicas de ETL. Concretamente, se ha utilizado 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.

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 van a realizar a través del entorno y lenguaje de programación R, que es una referencia en el ámbito de la estadística.

4.3.2. Análisis Exploratorio

En primer lugar, se debe 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 van a utilizar los principales estadísticos como la media, la mediana, las desviaciones típicas, etc. Posteriormente se va a utilizar la matriz de varianzas y covarianzas, que indicaran la variabilidad de los datos y la información sobre las posibles relaciones lineales entre las variables.

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

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

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

Mediante el boxplot se puede observar aspectos como la posición, dispersión, asimetría, longitud de colas y los datos anómalos (outliers). El QQ-plot se va a utilizar para evaluar la cercanía de los datos a una distribución. (Orellana, 2001)

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

4.3.3. Aprendizaje automático

Aprendizaje No Supervisado

1. Algoritmos de Clustering. El objetivo de estos algoritmos consiste en investigar si los datos pueden ser organizados en grupos o *clusters* que posean características similares. Los métodos de clustering tienen la característica común que para poder llevar a cabo las agrupaciones necesitan definir y cuantificar la similitud entre las observaciones. Por ejemplo, la distancia euclidea, la distancia de Manhattan, la correlación, el índice de Jaccard, etc. Para realizar este análisis, se va a utilizar el algoritmo de K-Means.

2. Análisis de Componentes Principales. El objetivo es transformar un conjunto de variables (originales), en un nuevo conjunto de variables denominadas componentes principales. Con este análisis, se trata de reducir la dimensión de las variables, manteniendo la suma de varianzas. (de la Fuente Fernandez, 2011)

Aprendizaje Supervisado

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

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

1. Arboles de decisión

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

   Los nodos están asociados a los atributos y tiene varias ramas que salen de él (dependiendo de los valores que tomen la variable asociada). Estos nodos

pueden asemejarse a preguntas que, dependiendo de la respuesta que conlleve, se tomara un flujo en las ramas salientes.

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

Fig. 8: Funcionamiento Árboles Decisión. Recuperado de Sayad (2019)

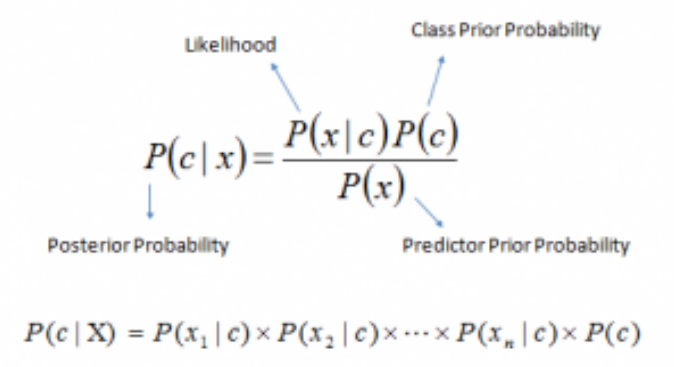
2. Clasificación de Naïve Bayes.

   Es un algoritmo que se basa en la técnica de clasificación utilizando el teorema de Bayes.

   El algoritmo es capaz de agrupar un registro mediante las características de este. Para ello aplica probabilidades condicionales de las características para determinar a qué categoría pertenece.

   Por ejemplo, una fruta puede considerarse una manzana si es de color rojo, redonda y tiene un determinado peso.

Fig. 9: Teorema de Bayes. Recuperado de (University of Cincinnati, 2018)



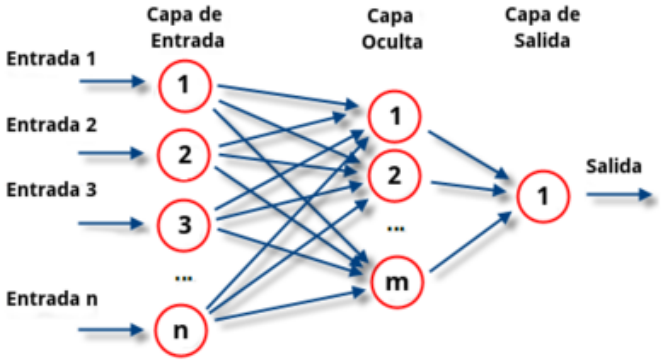
3. Regresión Logística

   Es un algoritmo de regresión que se utiliza para predecir el resultado de una variable categórica en función de las variables independientes o predictores. Para predecir el resultado, se establecen pesos en función de la puntuación dada a cada variable independiente.

4. Redes Neuronales

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

Fig. 10: Red Neuronal. Recuperado de Piqueras (2017)



Como se puede observar en la figura 10, la primera fila (con neuronas de color rojo), se conocen como nodos de entrada y son aquellos que se encargan de recoger la información. Los nodos en la gama azul son los que se conocen como nodos de salida. Los nodos situados en el medio son aquellos que se encargan de hacer el aprendizaje, y se conocen como nodos ocultos.

En primer lugar, se obtiene la información a partir de los nodos de entrada, una vez que se tiene la información, se envía a las capas ocultas, que se activan o no dependiendo del aprendizaje previo. Los nodos ocultos se activan dependiendo

de una serie del resultado de unas operaciones matemáticas. Si los nodos se activan, entonces enviaran la información a la siguiente capa.

5. Bosques aleatorios

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

   Algunas características son:

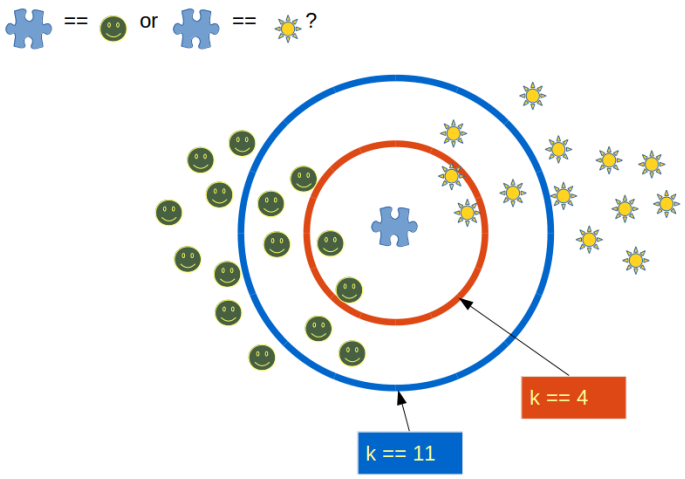
   - Gran precisión.
   - Eficiente para grandes bases de datos.
   - Aporta estimaciones sobre la importancia de las variables en la clasificación.
   - Tiene un método eficaz para la estimación de los datos faltantes y mantiene la precisión cuando falta una gran parte de los datos.

6. K-Vecinos-cercanos

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

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

Fig. 11: KNN. Recuperado de Klein (2018)



4.4. Distribución y Almacenamiento

Respecto a la distribución de la información, esta no podrá salir de la consejería de educación. Aunque se trate de datos anonimizados y agregados, se trata de datos de carácter sensible y no pueden ser distribuidos. Por tanto, dichos datos se van a almacenar en un gestor de bases de datos MySQL. Este gestor se encontrará en un servidor de la Consejería de Educación. Solo se va a poder acceder a dicho servidor desde la propia sede. Es posible que los datos también se almacenen en archivos de texto plano.

5. BIBLIOGRAFÍA

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

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

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

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

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

de la Fuente Fernandez, S. (2011). *Componentes principales.* Descargado de http://www.estadistica.net/Master-Econometria/Componentes_Principales.pdf

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

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.

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

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

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

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

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.

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

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

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

Piqueras, V. Y. (2017). *¿qué es y para qué sirve una red neuronal artificial?* Descargado de https://victoryepes.blogs.upv.es/2017/01/07/que-es-y-para-que-sirve-una-red-neuronal-artificial/

Prakash, G. H., y Rangdale, P. (2017). Etl data conversion: Extraction transformation and loading data conversion. *International Journal Of Engineering And Computer Science*, 22545–22550.

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

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

Sayad, S. (2019). *An introduction to data science.* Descargado de https://www.saedsayad.com/data_mining_map.htm

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

UNE 166006. (2018). *Gestión de la i+d+i: Sistema de vigilancia e inteligencia.*

University of Cincinnati. (2018). *Naïve bayes classifier.* Descargado de http://uc-r.github.io/naive_bayes

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*.