# INTRODUCCIÓN

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

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

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

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

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

1. Escolaridad del curso actual.
   1. Número de alumnos y grupos de un determinado centro.
   2. El número de alumnos por aula (también conocido como ratio).
   3. Matriculación de nuevos alumnos.
      1. Principalmente alumnos que superan el nivel de 6º de primaria y pasan a 1º de ESO.
2. Bilingüismo del centro. Muchos alumnos optan por centros bilingües para su mejor formación, por lo que estos centros suelen tener más demanda de alumnos.
3. Posibilidad de creación de nuevas zonas urbanas cerca del centro.
4. Posibilidad de apertura o cierre de centros educativos. El cierre de por ejemplo, de un centro privado, provocara una mayor tasa de matriculación de los centros contiguos.
5. Porcentaje de aprobados. Los alumnos que están ya matriculados tienen prioridad sobre los nuevos alumnos, por lo tanto, si existe una alta tasa de suspensos, quedan pocas plazas de admisión de nueva matricula.
6. El número y la aparición de nuevas enseñanzas. La oferta de nuevas enseñanzas atraerá a nuevos alumnos al centro, incrementando así el número de matriculaciones.

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

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

# JUSTIFICACIÓN TEÓRICA

<https://www.thetechedvocate.org/8-ways-machine-learning-will-improve-education/>

En primer lugar, y, antes de comenzar la investigación, se ha acudido a los datos del INE (Instituto Nacional de Estadística) y a los del ministerio de educación, ciencia y deportes (MECD).

Como la investigación va dirigida a los alumnos de la ESO, se ha buscado información respectiva a este nivel educativo.

# PROPUESTA DE INTERVENCIÓN

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. (<http://datos.gob.es/es/noticia/ciencia-de-datos-machine-learning-y-deep-learning>). 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. Se puede observar también la correlacion entre las variables, sobretodo

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

* El aprendizaje supervisado 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. ( <https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html>). 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).
* 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 estructura u organización que simplifique el análisis. (<https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html>)

Estos procesos, aplicados a los datos que se tienen sobre la educación, nos proporcionara nuevos factores que se tengan en cuenta en la actualidad. Además, mediante el uso de técnicas predictivas, se va a buscar la automatización de las labores de la unidad de Educación Secundaria Obligatoria.

# DISEÑO DE INVESTIGACIÓN

<https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques>

En este apartado, se va a presentar la forma en la que se va a realizar la investigación. En primer lugar, se va a realizar un proceso ETL, posteriormente se va a realizar un análisis descriptivo mediante sus técnicas que se explicaran posteriormente, además se va a incluir técnicas de aprendizaje no supervisada en este análisis.

Una vez que se ha realizado el análisis descriptivo, se va a realizar un análisis predictivo. En este análisis se va a utilizar técnicas de aprendizaje supervisadas.

## Proceso de Extracción, Transformación y Carga

En primer lugar, se va a realizar un tratamiento de datos, para ello, 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, transfórmalos según las necesidades que se tenga y depurarlos (evitando así datos erróneos). (Prakash, 2017) (Guillermo Matos, 2006) (Sharma, 2014)

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:

* 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.
* 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.

## Análisis Descriptivo

El análisis predictivo (también conocido como estadísticas predictivas) se encarga de resumir los datos en bruto para que puedan ser interpretados. Estos análisis son útiles ya que permiten aprender sobre comportamientos o patrones pasados e entender cómo pueden influir en los resultados futuros. En este tipo de análisis se van a utilizar tanto métodos graficos como medidas resumen

En primer lugar, se debe estudiar el tipo de datos de cada variable a estudiar, 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. (<http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema2am.pdf>). 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. ( http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema2am.pdf)

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. (<http://cms.dm.uba.ar/academico/materias/verano2015/estadisticaQ/descriptiva.pdf>)

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. (<http://cms.dm.uba.ar/academico/materias/verano2015/estadisticaQ/descriptiva.pdf>)

<https://www.sergas.es/gal/documentacionTecnica/docs/SaudePublica/Apli/Epidat4/Ayuda/Ayuda_Epidat_4_Analisis_descriptivo_Octubre2014.pdf>

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

### Aprendizaje No supervisado

Algunos de los algoritmos que se utilizaran son:

1. Algoritmos de clustering
   1. K-mean
2. Análisis de componentes principales
3. Descomposición en valores singulares
4. Análisis de componentes independientes
5. Stepwise Regression

<https://www.fisterra.com/mbe/investiga/10descriptiva/10descriptiva.asp#top>

<http://www.uco.es/zootecniaygestion/img/pictorex/27_12_49_7.pdf>

<https://machinelearningmastery.com/descriptive-statistics-examples-with-r/>

<http://cms.dm.uba.ar/academico/materias/verano2015/estadisticaQ/descriptiva.pdf>

## Análisis Predictivo

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. (<http://datos.gob.es/es/noticia/ciencia-de-datos-machine-learning-y-deep-learning>).

### Aprendizaje Supervisado

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%. (https://medium.com/@manguart/machine-learning-conceptos-básicos-del-aprendizaje-supervisado-para-personas-no-técnicas-142bbb222140)

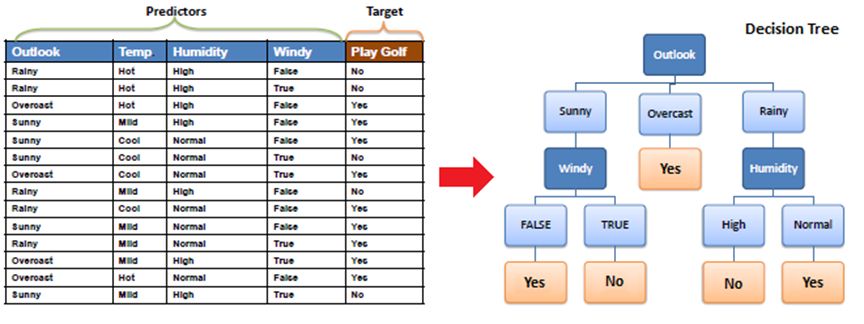
Algunos de los algoritmos que se van a utilizar son:

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.



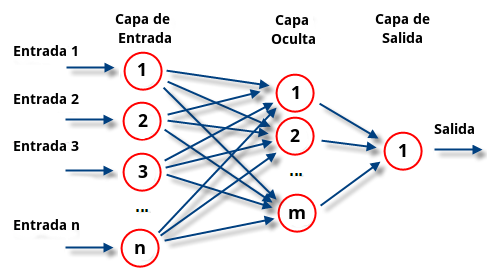
(http://www.cs.us.es/~fsancho/?e=104)

1. Redes Neuronales

Son sistemas de aprendizaje basados en el comportamiento de las redes neuronales biológicas presentes en el cerebro de los animales.

Estas redes se construyen, al igual que en los animales, a partir de nodos llamados neuronas que se organizan en forma de capas. Existe una capa de entrada, una capa de salida y, al menos, una capa intermedia (llamada capa oculta).

Las capas ocultas se activan o se desactivan dependiendo del valor obtenido mediante funciones de activación (suele ser 0 o 1). Estas funciones de activación tienen como entrada los valores de las neuronas de la capa de entrada.



1. Redes Bayesianas
2. Bosques Aleatorios (random Forest)
3. Boosting
4. Gradient Boosting
5. Clasificación de Naïves Bayes
6. Regresión por mínimos cuadrados
7. Regresión Logística

En esta investigación se va a utilizar técnicas de regresión en el aprendizaje supervisado debido a que la función de salida, que en este caso es el número de alumnos matriculados, es de tipo numérico. En consecuencia, se debe utilizar algoritmos de regresión.

http://www.diegocalvo.es/aprendizaje-supervisado-y-no-supervisado/

## Modelo “Ensemble”

# ANÁLISIS DE RESULTADOS

# CONCLUSIONES

# LINEAS FUTURAS

# REFERENCIAS

<http://www.educacionyfp.gob.es/servicios-al-ciudadano-mecd/dms/mecd/servicios-al-ciudadano-mecd/estadisticas/educacion/indicadores-publicaciones-sintesis/datos-cifras/Datosycifras1617esp.pdf>

<http://www.comunidad.madrid/sites/default/files/doc/educacion/sgea_datosycifras_presentacion_17_18.pdf?language=es>

<http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117memòria.pdf>

# Bibliografía Básica

166006., U. (Abril de 2018). Gestión de la I+D+I: Sistema de vigilancia e inteligencia.

1666006:2018, N. U. (s.f.).

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *QÜESTIIÓ*.

Archanco, R. (12 de 12 de 2018). *Listado de técnicas de análisis de datos útiles para tu empresa*. Obtenido de https://papelesdeinteligencia.com/tecnicas-de-analisis-de-datos/

Eduardo Fernandes, M. H. (7 de Febrero de 2018). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *ELSEVIER*.

*Educational Data Mining*. (2018). Obtenido de http://educationaldatamining.org

Guillermo Matos, R. C. (2006). Metodología para la Extracción del Conocimiento Empresarial a partir de los Datos. *Información Tecnológica*, 81-88.

*MathWorks*. (13 de 12 de 2018). Obtenido de https://es.mathworks.com/discovery/predictive-analytics.html

MECD. (s.f.). *Estadísticas Nacionales*. Obtenido de http://www.mecd.gob.es/servicios-al-ciudadano-mecd/estadisticas/educacion.html

Nielsen, J. (6 de Enero de 2018). *NN/g Nielsen Norman Group.* Obtenido de https://www.nngroup.com/articles/law-of-bandwidth/

OCDE. (s.f.). *Estadística OCDE*. Obtenido de https://stats.oecd.org

Prakash, G. H. (2017). ETL Data conversion: Extraction, transformation and loading Data Conversion. *International Journal of Engineering And Computer Science*.

Sharma, S. (2014). Modeling ETL Process in Data warehouse: An Exploratory Study. *2014 Fourth International Conference on Advanced Computing & Communication Technologies*.